An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression

Feng Wei, Gordon Lovegrove*

School of Engineering, Faculty of Applied Science, University of British Columbia, Kelowna, BC Canada

**Abstract**

Today, North American governments are more willing to consider compact neighborhoods with increased use of sustainable transportation modes. Bicycling, one of the most effective modes for short trips with distances less than 5 km is being encouraged. However, as vulnerable road users (VRUs), cyclists are more likely to be injured when involved 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 two types of statistical associations between collisions and each explanatory variable: (1) An increase in bicycle–auto collisions is associated with an increase in total lane kilometers (TLKM), bicycle lane kilometers (BLKM), bus stops (BS), traffic signals (SIG), intersection density (INTD), and arterial–local intersection percentage (IALP). (2) A decrease in bicycle collisions was found to be associated with an increase in the number of drive commuters (DRIVE), and in the percentage of drive commuters (DRP).

These results support our hypothesis that in North America, with its current low levels of bicycle use (<4%), we can initially expect to see an increase in bicycle collisions as cycle mode share increases. However, as bicycle mode share increases beyond some unknown ‘critical’ level, our hypothesis also predicts a net safety improvement. To test this hypothesis and to further explore the statistical relationships between bicycle mode split and overall road safety, future research needs to pursue further development and application of community-based, macro-level CPMs.

© 2012 Published by Elsevier Ltd.

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 our society, economy and environment than do automobile-dominated transportation systems. The benefits of cycling are generally understood to include: relatively low costs, emissions, and energy use, together with improved health, and convenient parking. Bicycles traveling at an average speed of 15 km/h, are well suited to the short and medium-distance trips. Bicycling is typically the fastest mode for trips less 5 km (Dejister and Schollaert’s, 1999). However, as vulnerable road users (VRUs), cyclists are more likely than auto drivers to be injured in traffic collisions. Research on VRU safety at road intersections has addressed many of the impacts of intersection traffic volume and geometric design on VRUs safety (e.g. Ekman, 1996; Leden et al., 2000; Grey et al., 2010). Although these reactive road safety measures have been effective in improving VRU safety on existing facilities, there is still a lack of planning-level assessment empirical tools that forecast safety effects of VRU mode split on the whole traffic stream (e.g. total collisions over all modes). For example, in many North American (NA) communities, the public perceive that bicycling is dangerous, leading to low bicycle mode share, while in many European Union communities, no such public perception exists and cycling mode share is over 30% with some of the lowest total collisions worldwide. Our hypothesis is that increased VRU mode share will lead to decreased total collisions. However, in order to test this hypothesis, we must address the empirical gap, and develop reliable tools that allow our community planners and engineers to proactively forecast the level of road safety of increased VRU use. As a start toward quantifying the road safety benefits of increased bicycle use, this

* Corresponding author. Tel.: +1 250 807 8717; fax: +1 250 807 9850. E-mail address: gord.lovegrove@ubc.ca (G. Lovegrove).

0001-4575/$ – see front matter © 2012 Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.aap.2012.05.018
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.

The three objectives of the research presented in this paper were to:

1. Conduct a literature review on micro-level and macro-level bicycle collision prediction models (CPMs).
2. Use negative binomial (NB) regression to develop community-based, macro-level bicycle CPMs using data from the Central Okanagan Regional District (CORD), in BC, Canada.
3. Recommend statistical and data issues in this model development for future research.

For some time, it has been generally understood and demonstrated by evidence from previous studies suggests that decreasing auto use, accompanied by increased sustainable transport mode splits (e.g., bicycling, walking, or public transit), is at least statistically (if not causally) associated with lower traffic collisions and road fatalities (Osberg and Stiles, 1998; Newman and Kenworthy, 1999; Marshall and Garrick, 2011). Osberg and Stiles (1998) compared bicycle use and road safety in Boston, Paris and Amsterdam, which have similar social-economics, but very different levels of bicycle use: lowest use in Boston, slightly higher cycle use in Paris; and highest use in Amsterdam. They found that Amsterdam had the lowest total road fatalities rate, at 5.8 fatalities/100,000 population (POP), despite experiencing the highest cyclist fatality rate of these three cities with 1.8 fatalities/100,000 POP. Compare these rates to 0.6 cyclist fatalities/100,000 POP, and 0.3 cyclist fatalities/100,000 POP, and 10 total fatalities/100,000 POP and 8.8 total fatalities/100,000 POP for Paris and Boston, respectively. Newman and Kenworthy (1999) reported that Amsterdam and Copenhagen are considered high bicycle use cities among developed countries, yet they suffer only half the road traffic fatality rate of US cities (5.8 fatalities/100,000 POP for Amsterdam, 7.5/100,000 for Copenhagen and 14.6 fatalities/100,000 POP for US cities on average). Marshall and Garrick (2011) examined the road safety data from 24 California cities and found that cities with higher bicycle use generally showed much lower risks of fatalities for all road users. It is reasonable to expect that, similar automobile transport, as cycling use grows, there will be increased investment in engineering, education and enforcement, in support of mobility, safety, and efficiency cycling needs. Moreover, as those investments occur, cycling-related collisions, injuries, and fatalities would reduce as well. Consequently, it is not unreasonable to expect that as the environmental ‘friendliness’ toward cycling improves, the NA public perception toward cycling and other vulnerable road users (VRU) would also be expected to improve.

Our hypothesis then follows from these empirical evidences and subsequent reasoned assumptions. The red line in Fig. 1 depicts our hypothesis regarding a causal relationship between overall level of road safety and bicycle mode share that increased bicycle use will lead to a significant reduction in total traffic collisions. The blue line in Fig. 1 depicts an associated hypothesis for NA bicycling levels, that initial increases in bicycle mode share above our currently low levels to some unknown medium level of cycling use will likely cause increased bicycle-related collisions, but as bicycle use increased beyond that unknown medium level there would be a decrease in bicycle-related collisions. Our research seeks to test these hypotheses by development and application of reliable empirical tools. Community-based, macro-level collision prediction models (CPMs) have been shown to provide reliable empirical evidence, and was used in this paper to test our second (blue line in Fig. 1) hypothesis.

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, negative binomial regression, and Poisson lognormal regression.

2.2. Poisson and zero-inflated poisson 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; Lord et al., 2005). 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 (Qin et al., 2004; Kumara and Chin, 2003; Lee and Mannering, 2002).

2.3. Negative binomial 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 (Miaou and Lord, 2003; Lord, 2006; Sawalha and Sayed, 2006; Hadyayeghi et al., 2003; Lovegrove and Sayed, 2006; Ladrón de Guerra et al., 2004). Model results from these studies demonstrated that an NB model was superior to a
Poisson model. The formulations for NB regression are presented as follow:

\[ \frac{Y_i}{E_i} \sim \text{Poisson}(E_i) \]  
\[ E_i = \mu_i \]  
\[ \ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} \]

where \( Y_i \) = the number of collisions at location \( i \); \( E_i, \lambda, \kappa_i = \) the distribution parameters, \( X_{ij} = \) explanatory covariates. So in Eq. (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 \( \lambda_i \) and \( \kappa_i \) (see Eq. (4)). Under the NB model, the mean and the variance are presented in Eq. (5) as:

\[ E(Y_i) = \mu_i \text{ and } \text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\kappa_i} \]

### 2.3.1. Poisson lognormal regression

Poisson lognormal regression (PLN) model also can be reflective of the extra-Poisson variations (Kim et al., 2002; El-Basyouny and 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 Eqs. (6)–(9):

\[ \frac{Y_i}{E_i} \sim \text{Poisson}(E_i) \]  
\[ E_i = \mu_i \exp(u_i) \]  
\[ \ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} \]  
\[ E_i \sim \text{Lognormal}(\mu_i, \sigma^2) \exp(\exp(u_i) \sim (0, \sigma^2) \text{or } u_i \sim N(0, \sigma^2) \]

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:

\[ E(Y_i) = \mu_i \exp(0.5\sigma^2) \text{ and } \text{Var}(Y_i) = E(Y_i) + [E(Y_i)]^2\exp(\exp(\sigma^2 - 1)) \]

### 2.4. Micro- and macro-level models for bicycle collisions

Regression models for collision prediction can be divided into two types: (1) Micro-level CPMs for single-facility locations (e.g. intersections or road segments) and (2) Macro-level CPMs for area-wide or macro-level regions (e.g. neighborhoods, cities, or districts). Micro-level CPMs have been well researched and applied successfully in reactive road safety improvement programs (RSIPs), to improve the safety performance of existing road facilities. Macro-level CPMs can be used in reactive RSIPs in existing communities, but can also be applied proactively to evaluate the road safety level of planned facilities, with potential for significant reductions in collision frequencies below that achieved to date using reactive techniques (De Leur and Sayed, 2003; Lovegrove and Sayed, 2006; Lovegrove, 2007). De Leur and 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.

#### 2.4.1. 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 annual average 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 and 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.

#### 2.4.2. Macro-level bicycle CPMs

From the macro-level aspect, Jacobsen (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 (20.4 = 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 and Sayed, 2006; Lovegrove, 2007), 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 and 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) above, where \( k \) is a positive constant known as the dispersion parameter. In this study, bicycle CPMs with NB regression were developed using GenStat, a generally available statistics software package. Lovegrove and Sayed (2006) proposed a generalized linear model form for macro-level CPMs, which is presented in Eq. (11) as:

\[
E = a_0 Z^a e^{\Sigma h x_i}
\]

where \( E \) is the predicted collision frequency (over 3 years for motor collisions); \( a_0, a_1, b, h \) = model parameters; \( Z \) = leading exposure variables (i.e. VKT – vehicle kilometers for modeled data or TLK=total lane kilometers for measured data); and, \( x_i \) = other explanatory variables. This form not only takes account of the influence of different independent variables, but also fairly represents the non-negative, nonlinear, and non-normal nature of collisions. Based on this form, four possible model forms predicting bicycle collisions were tested using the same datasets, as follows:

- (Model form 1)
  \[
  E_B = a_0 Z^a e^{\Sigma x_i}
  \]
- (Model form 2)
  \[
  E_B = a_0 Z^a e^{\Sigma x_i}
  \]
- (Model form 3)
  \[
  E_B = a_0 [(B+1)^a] e^{\Sigma x_i}
  \]
- (Model form 4)
  \[
  E_B = a_0 (B+1)^a Z^a e^{\Sigma x_i}
  \]

where \( E_B \) is the predicted bicycle collision frequency over 5-year period; \( B \) is the leading exposure variables of bicycle use. In these model forms, only bicycle lane kilometers meters (BLKM) was used as the leading exposure variable of bicycle use. Other bicycle exposure variables, such as bicycle kilometers traveled (BKT), could not be used due to lack of available BKT data, and is left for future research. All of the model forms support the “product-of-exposure-to-power” relationship, which has been demonstrated by previous macro-level CPM studies (Jacobsen, 2003; Hadayeghi et al., 2003; Ladrón 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 log error (i.e. zero ‘B’ leads to zero ‘EB’) as bicycle collisions could happen in the locations without bicycle lanes. In the Model form 1–3, 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 variables were TLKM and BLKM. The decision to keep a variable in the model was based on it meeting four criteria (Sawalha and Sayed, 2006; Lovegrove and Sayed, 2006). First, the logic (i.e. \( \pm \)) of the estimated parameter was intuitively associated with collisions, relying on previous research evidence as well as contextual community traits in specific study areas. Second, the \( t \)-statistic for each parameter needed to be significant at the 95% confidence level (i.e. \( t > 1.96 \)). Third, each added variable had to have minimal correlation (i.e. \( <0.3 \)) with other independent variables in the same model. Fourth, the added variable had to produce a significant drop in the scaled deviance (SD) statistic at a 95% confidence level (i.e. drop in SD>3.84) and this step is to do an analysis of deviance procedure for comparing two nested models. Other researchers have suggested using an AIC model goodness-of-fit method to fit models built using data with low mean values. However, we did not use this one as the SD drop was used to test whether a given model is qualified in model fitting, instead of choosing the best fitting model with data in many different options. So this variable selection procedure more recently recommended by Sawalha and Sayed (2006) was used as the main methodology. This procedure carries out a likelihood ratio test to determine whether the additional variable can significantly increase the modeling likelihood of observed collision data. As the scaled deviance is asymptotically \( \chi^2 \) distributed with \( n-p \) degree of freedom, if adding one variable can cause a drop in scaled deviance exceeding \( \chi^2_{0.05,1} \) (3.84) then the likelihood is considered to increase significantly (Sawalha and Sayed, 2006). In this case, any variable meeting the first three criteria but not the last one could be excluded to avoid the model over-fitting problem.

3.3. Model goodness-of-fit tests

As each candidate variable was added, the model fit was re-assessed. Scaled Deviance (SD) and Pearson \( \chi^2 \) are common goodness of fit measures for Poisson or NB regression (McCullagh and Nelder, 1989), defined in Eqs. (12) and (13) as follows:

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

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

where \( y_i \) and \( E(A_i) \) are observed and predicted collision frequency in location \( i \), respectively; \( Var(y_i) \) is the variance for location \( i \); and \( \kappa \) is the over dispersion parameter for the model (an output of the GLM regression analysis). Both SD and Pearson \( \chi^2 \) approximate a \( \chi^2 \) distribution 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 \( \chi^2 \) values of any model were smaller than \( \chi^2_{0.05, n-p} \), this model was seen as successfully fit. However, some studies have suggested that SD performed better in large samples than small samples, and did not work well when there are many extreme observations (such as zeroes) (Maycock and Hall, 1984; Maher and Summersgill, 1996; Wood, 2002; Agawal and Lord, 2006). Indeed in this case study, many observations (bicycle collisions in
different zones) were zeroes, suggesting that the use of an SD test may not be wise. Wood (2002) proposed one more complex method, which is called the grouped SD that included five steps for solving this problem. His grouped SD method is based on the knowledge that through grouping, the terms of $\chi^2$ will better fit a normal distribution. As recommended by Wood (2002), the grouped SD values were also developed for each model, and compared with the corresponding new critical $\chi^2$ value.

4. Data description and extraction

Data used for this study came from the Central Okanagan Regional District (CORD) in the Province of British Columbia, Canada. This mainly rural area is roughly 44,000 hectares and comprised of 4 member municipalities (i.e. Kelowna, West Kelowna, Lake Country and Peachland). There were approximately 160,000 residents, 66,000 households, and 29,000 total lane kilometers of road in the CORD (Statistic Canada, 2006a,b). The major highway corridor through this region is Highway 97, linking the commercial areas of all four municipalities. Fig. 2 shows the location of this area in Canada (in the indent map), four municipalities, traffic analysis zone (TAZ) boundaries, the land use of TAZs (i.e. rural and urban), and all bicycle lanes and paths. All bicycle lanes are in the City of Kelowna and grouped into five categories, including: on-road bicycle lanes (one/both side(s)), off-road paths and trails, and shared roadways (i.e. without bicycle lane marks, but with very low traffic volume for bicycling).

For development of community-based, macro-level CPMs, collision data and all independent variable data needed to be aggregated into areal units. The aggregation units in the CORD were 500 TAZs derived from the 2005 regional 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 (≥3–5 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 were used for model stratification based on previous macro-level CPM research (Lovegrove and Sayed, 2006; Lovegrove’s 2007), including: four variable themes (i.e. exposure, socio-demographics, transportation demand management, or network), two land use types (i.e. rural or urban), and two exposure data derivations (i.e. derived from modeling or direct GIS measurement). 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. Although the earlier research listed 16 model group stratifications, modeled exposure data for the CORD was not available for this study, which eliminated use of vehicle kilometers traveled (VKT), and volume/congestion (VC) variables. Moreover, as only models in urban areas were developed due to extremely low observed bicycle collisions in rural areas, only 4 of 16 groups of models were planned to develop (i.e. urban, measured models in exposure, socio-demographic, transportation demand management, and road network themes). Model variable definitions and their statistical summary are presented in Table 1 according to different variable themes.

5. Model results and discussion

5.1. Data observations

In the process of model development, when reviewing the available data, several observations were made that should be
Table 1
Variable definitions and data summary (n = 500 TAZs, urban = 242, rural = 258).*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Symbol</th>
<th>Source</th>
<th>Years</th>
<th>Zonal min</th>
<th>Zonal max</th>
<th>Zonal Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicycle/vehicle collisions</td>
<td>B5</td>
<td>ICBC</td>
<td>02–06</td>
<td>0</td>
<td>6</td>
<td>0.37</td>
</tr>
<tr>
<td>Total lane km</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>63.25</td>
<td>5.79</td>
<td></td>
</tr>
<tr>
<td>Total bicycle lane km</td>
<td>TLKM</td>
<td>CORD</td>
<td>2006</td>
<td>0.00</td>
<td>7.12</td>
<td>0.70</td>
</tr>
<tr>
<td>Total bicycle lane km – on road</td>
<td>BLKMO</td>
<td>CORD</td>
<td>2006</td>
<td>0.00</td>
<td>3.61</td>
<td>0.11</td>
</tr>
<tr>
<td>Total bicycle lane km – off road</td>
<td>BLKMF</td>
<td>CORD</td>
<td>2006</td>
<td>0.00</td>
<td>6.39</td>
<td>0.58</td>
</tr>
<tr>
<td>Zonal area (hectares)</td>
<td>AR</td>
<td>CORD</td>
<td>2005</td>
<td>1.33</td>
<td>163.12</td>
<td>88.72</td>
</tr>
<tr>
<td>Urban zones</td>
<td>URB</td>
<td>CORD</td>
<td>2005</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Rural zones</td>
<td>RUR</td>
<td>CORD</td>
<td>2005</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Population</td>
<td>POP</td>
<td>Census</td>
<td>2006</td>
<td>0</td>
<td>2858</td>
<td>320</td>
</tr>
<tr>
<td>Population density (POP/AR)</td>
<td>POPD</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>85.35</td>
<td>11.26</td>
</tr>
<tr>
<td>Population aged ≤ 30 (POP%)</td>
<td>POPSO</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>53.57</td>
<td>31.39</td>
</tr>
<tr>
<td>Male/female ratio</td>
<td>MJF</td>
<td>Census</td>
<td>2006</td>
<td>0.50</td>
<td>1.67</td>
<td>1.04</td>
</tr>
<tr>
<td>Home</td>
<td>NH</td>
<td>Census</td>
<td>2006</td>
<td>0</td>
<td>1012</td>
<td>132</td>
</tr>
<tr>
<td>Home density</td>
<td>NHD</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>56.45</td>
<td>5.32</td>
</tr>
<tr>
<td>Participation in labor force (EMP/POP) (%)</td>
<td>PARTP</td>
<td>Census</td>
<td>2006</td>
<td>12.34</td>
<td>84.76</td>
<td>62.13</td>
</tr>
<tr>
<td>Employed residents</td>
<td>EMP</td>
<td>Census</td>
<td>2006</td>
<td>1</td>
<td>1539</td>
<td>162</td>
</tr>
<tr>
<td>Employed percentage (EMP/POP%)</td>
<td>EMPP</td>
<td>Census</td>
<td>2006</td>
<td>12.27</td>
<td>82.72</td>
<td>58.98</td>
</tr>
<tr>
<td>Unemployed residents</td>
<td>UNEMP</td>
<td>Census</td>
<td>2006</td>
<td>0.01</td>
<td>58.59</td>
<td>5.50</td>
</tr>
<tr>
<td>Unemployed percentage (UNEMP/POP) (%)</td>
<td>UNENMP</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>19.15</td>
<td>5.04</td>
</tr>
<tr>
<td>Average income ($)</td>
<td>INCA</td>
<td>Census</td>
<td>2006</td>
<td>6100</td>
<td>69600</td>
<td>32000</td>
</tr>
<tr>
<td>Commuter density (TCM/AR)</td>
<td>TCM</td>
<td>Census</td>
<td>2006</td>
<td>0</td>
<td>1315</td>
<td>145</td>
</tr>
<tr>
<td>Core area (hectares)</td>
<td>CORE</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>293.67</td>
<td>17.88</td>
</tr>
<tr>
<td>Car passenger commuter percentage (%)</td>
<td>CRP</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>100.00</td>
<td>42.34</td>
</tr>
<tr>
<td>Transit commuter percentage (%)</td>
<td>BUS</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>18.11</td>
<td>7.07</td>
</tr>
<tr>
<td>Biking commuter percentage (%)</td>
<td>BIKE</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>14.05</td>
<td>1.88</td>
</tr>
<tr>
<td>Pedestrian percentage (%)</td>
<td>WALK</td>
<td>Census</td>
<td>2006</td>
<td>0.00</td>
<td>31.44</td>
<td>5.20</td>
</tr>
<tr>
<td>No. of driving commuters</td>
<td>DRiver</td>
<td>Census</td>
<td>2006</td>
<td>0</td>
<td>1179</td>
<td>118</td>
</tr>
<tr>
<td>Driving commuter percentage (%)</td>
<td>DRB</td>
<td>Census</td>
<td>2006</td>
<td>47.17</td>
<td>100.00</td>
<td>78.85</td>
</tr>
<tr>
<td>Bus stops</td>
<td>BS</td>
<td>BC Transit</td>
<td>2006</td>
<td>0</td>
<td>14</td>
<td>1.60</td>
</tr>
<tr>
<td>Bus stop density</td>
<td>BSD</td>
<td>BC Transit</td>
<td>2006</td>
<td>0.00</td>
<td>2.82</td>
<td>0.07</td>
</tr>
<tr>
<td>Road network, urban, measured (group 4)</td>
<td>SIG</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>Signal density</td>
<td>SIGD</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0.00</td>
<td>0.89</td>
<td>0.02</td>
</tr>
<tr>
<td>No. of intersections</td>
<td>INT</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0</td>
<td>50</td>
<td>6.16</td>
</tr>
<tr>
<td>Intersection density</td>
<td>INTD</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0.00</td>
<td>1.50</td>
<td>0.15</td>
</tr>
<tr>
<td>No. of intersections/TLKM</td>
<td>INTK</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0.00</td>
<td>7.67</td>
<td>1.09</td>
</tr>
<tr>
<td>No. of 3 way intersections/INT (%)</td>
<td>I3WP</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0.00</td>
<td>100.00</td>
<td>66.05</td>
</tr>
<tr>
<td>No. of arterial-local intersections/INT (%)</td>
<td>IAALP</td>
<td>CORD/GE</td>
<td>2006</td>
<td>0.00</td>
<td>100.00</td>
<td>15.02</td>
</tr>
<tr>
<td>No. of collector lane-km</td>
<td>ALK</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>19.43</td>
<td>0.84</td>
</tr>
<tr>
<td>No. of collector lane-km/TLKM</td>
<td>CLK</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>32.57</td>
<td>0.73</td>
</tr>
<tr>
<td>No. of local lane-km</td>
<td>LKLK</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>37.65</td>
<td>4.16</td>
</tr>
<tr>
<td>No. of arterial lane-km/TLKM (%)</td>
<td>ALKP</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>100.00</td>
<td>13.75</td>
</tr>
<tr>
<td>No. of collector lane-km/TLKM (%)</td>
<td>CLKP</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>100.00</td>
<td>11.89</td>
</tr>
<tr>
<td>No. of local lane-km/TLKM (%)</td>
<td>LLKP</td>
<td>CanMap®</td>
<td>2006</td>
<td>0.00</td>
<td>10.00</td>
<td>68.13</td>
</tr>
</tbody>
</table>

* Abbreviations: GE: google earth; POP15: population aged 15 and over in 2006; ICBC: Insurance Corporation of British Columbia.

noted. First, all of the models developed in this study predict only total collisions, and more specifically, only total bicycle–vehicle collisions. 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. Further stratifications were left as a topic for future research, pending improved bicycle collision data. The data that was found suggested that 95% of bicycle–vehicle collisions in the CORD are severe (i.e. involving injury and/or fatalities). Second, approximately 90% of bicycle–vehicle collisions occurred at intersections, while only 10% of them happened at mid-blocks. This suggests that moving motor vehicles usually crash with bicycles at intersections and how to improve VRU safety at intersection should also be a future of research. Bicycle collision data that did not involve motor vehicles (e.g., bicycle–fixed objective, bicycle–pedestrian, or bicycle–bicycle collisions) was not available. Third, 55% of bicycle–vehicle collisions in the CORD occurred on roads with on-road bicycle lanes, versus only 45% on roads without bicycle lanes. The higher proportion of collisions on roads with bicycle lanes is more probably due to high bicycle and vehicular flows on these roads. In accordance with bikeway design guidelines, bicycle lanes are typically only required where high bicycle and vehicle volumes warrant for safety. Therefore, evaluating the safety of on-road bicycle lanes based on the percentage split alone (55% versus 45%) would be a biased interpretation, and future research should be conducted to provide an unbiased evaluation, including development and use of macro-level CPMs. Last, while 16% of CORD bicycle paths consist of off-road bicycle paths (versus on-road), no bicycle–vehicle collision data was available for them, because non-vehicular incidents (i.e. that would not be covered by an ICBC vehicle insurance claim record) that occur on separated bicycle paths and tracks were not included in the ICBC collision database. Hence, the safety benefits/costs evaluations for both on-road bicycle lanes and off-road...
bicycle paths should be a future research topic once this off-road, non-vehicular data becomes geo-spatially coded to incident location and severity in hospital records.

5.2. Model formulation testing

Successfully fit bicycle CPMs in different model forms and groups, together with their statistic summary and goodness of fit test results are presented in Table 2. Four model forms were tested with the same CORD datasets, with models of three of four forms successfully developed. The fourth model form included both TLKM and BLKM as leading exposure variables, and in each regression analysis, the BLKM variable t-statistic did not meet the 95% level test and had to be removed. This result suggests that the TLKM exposure variable has much more influence on bicycle–vehicle collisions than the BLKM exposure variable, and may be a better leading predictor in modeling bicycle–vehicle collisions than BLKM. Also, in model form 2, in which the TLKM variable was set as the leading variable and BLKM was tested as one of the secondary exponential function variables, the t-statistics of BLKM parameter was not significant. So in all form 2 models derived, no BLKM variables were successfully included. These two results indicate that BLKM, as a lone exposure variable, may not be a good predictor for bicycle collisions. Moreover, this BLKM exclusion would be a serious deficiency in development of CPMs that are meant to predict bicycle collisions, unless some other measure(s) of bicycle flow or bicycle kilometers traveled could be found, pending availability of such data in future research.

As the data sample had very low mean, the SD values of these models (i.e. around 270) were less than the critical \( \chi^2 \) with corresponding degree of freedom (i.e. around 270), suggesting a fit too good to be true. For reference, the grouped SD value for each model was derived and compared with the corresponding new critical \( \chi^2 \) value, which is shown in the last column of Table 2. It was found that all but one model considered successfully fit using methods and goodness-of-fit tests described in Lovegrove and Sayed (2006) also comply with the grouped SD method described in Wood (2002). Model Form 3 and Group 1 shows a grouped SD value of 127.0 that is larger than its new critical \( \chi^2 \) value of 74.5 at the 95% confidence level test.

It was also noted that macro-level CPMs for bicycle–vehicle collisions in the Socio-Demographic group (i.e. group 2) were not developed, because none of the socio-demographic variables were significant according to t-statistics. This result suggests possible interpretations: (1) that socio demographic factors may not play a significant role in bicycle–vehicle collisions, and/or (2) that more research is needed to identify those socio-demographic factors that do play a significant role in bicycle–vehicle collisions.

Model results revealed intuitive relationships between bicycle–vehicle collisions and explanatory variables. The associations between bicycle–vehicle collisions and traffic exposure (i.e. total lane kilometers, and bicycle lane kilometers), confirmed intuitive expectations that more vehicle and/or bicycle exposure contributes to more bicycle–vehicle collisions. The association between increased collisions and increased bus stops (BS) is consistent with Kim’s research (2010). The association between increased collisions and increased arterial–local intersection percentage (IALP) suggests that arterial–local intersections are high risk locations for cyclists, likely due to their typically high traffic volumes, high speeds, and many conflicting un-signalized turning movements. In addition, increased collisions were also found to be associated with increased signals (SIG) and intersection density (INTD). Inverse associations between bicycle–vehicle collisions and...
drive commuters (DRIVE)/drive commuter percentage (DRP) were observed. 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 Fig. 1.

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.

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 Fig. 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 authors’ hypothesis about the relationship between safety and bicycle use in Fig. 1 actually exists.

6. Conclusions

In their pursuit of more sustainable communities, governments across North America (NA) are planning and building infrastructure and built forms that will promote increased use of sustainable transportation modes, including increased levels of bicycling. To address possible concerns about the safety ramifications of increased use of cycling in NA, empirical tools are needed to evaluate planned communities and maximize road safety levels.

This paper has hypothesized a global model on the relationship between bicycle use and road safety levels (see Fig. 1), that suggests bicycle collisions in North America will get worse before they get better, beyond some levels seen in European Union countries (e.g. above 30%). As a means to start testing this hypothesis, it then reviewed generalized linear regression methods (GLM) for collision prediction models (CPMs) and summarized previous studies in bicycle collision prediction at the micro and macro-levels. Four formulations of community-based, macro-level CPMs were proposed. Based on negative binomial GLM, several bicycle CPMs were successfully fit using urban data from the Central Okanagan Regional District (CORD) in Canada. These urban models revealed that bicycle–auto collisions were directly associated with total lane kilometers (TLKM), bicycle lane kilometers (BLKM), bus stops (BS), signals (SIG), intersection density (INTD), and arteriole–local intersections (IALP); but were inversely associated with drive commuters (DRIVE) and drive commuter percentage (DRP). With the models developed in this research, several conclusions can be drawn regarding current levels of cycling in North America, including:

- An increase in roadway infrastructure in general – roads, signals, intersections – can be expected to be associated with an increase in bicycle–auto collisions.
- An increase in bicycle lanes can be expected to be associated with an increase in bicycle–auto collisions.
- An increase in cycling can be expected to be associated with an increase in bicycle–auto collisions.
- These results appear to support at least in part the proposed hypothesis at low levels of cycling.

Data issues in the research process need to be addressed in future research in order for additional models to be developed, and for additional findings to be made regarding the hypothesis and the relationship between road safety and bicycle use in North America. In view of ongoing planning and construction to promote increased cycling in NA now underway, this research would seem urgently needed, especially regarding off-road facilities and bicycle-only collisions.

Acknowledgements

The Insurance Corporation of BC, and the City of Kelowna for their data used in this study. Ms Vicky Wei, research assistant, for her tireless work on this research. The Canadian Natural Sciences and Engineering Research Council for their continuing support of this important sustainable road safety research on saving lives in our communities.

References

Agawal, R., 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 1950, 35–43.


City of Kelowna, 2009a. Geo-referenced Intersections of Kelowna, BC (4.3.11).

City of Kelowna, 2009b. Geo-referenced Bicycle Lanes of Kelowna, BC (6.3.11).


