

Collision Prediction Models For British Columbia

Prepared For:

Engineering Branch
BC Ministry of Transportation & Infrastructure

Prepared by:

Tarek Sayed, PhD, P.Eng.,
Professor of Civil Engineering
University of British Columbia

Paul de Leur, PhD, P.Eng.,
Highway Safety Engineer
de Leur Consulting Limited

December 2008

TABLE of CONTENTS

List of Tables	2
List of Figures	2
1.0 Introduction	3
1.1 Background	3
1.2 Road Safety Estimation Techniques.....	5
1.3 What Is A CPM?	6
1.4 Why Should the MOT Use CPMs?.....	6
1.5 Developing CPMs for BC - Project Objectives.....	7
1.6 Report Structure.....	8
2.0 Data for Developing Collision Prediction Models	9
2.1 Highway Segments	9
2.2 Highway Intersections	10
3.0 Collision Prediction Models Developed.....	13
3.1 Background.....	13
3.1.2 Highway Segment CPMs:.....	13
3.1.3 Highway Intersection CPMs:	14
3.2 Summary of Collision Prediction Models Developed.....	14
4.0 Applications for Collision Prediction Models	17
4.1 Location Specific Prediction and the Empirical Bayes Refinement	17
4.2 Identification of Collision Prone Locations	18
4.3 Ranking of Collision Prone Locations	20
4.4 Before and After Studies.....	21
Appendix A: Theoretical Background of Collision Prediction Models and Development	
Methodology	23
A.1 Model Form, Regression Technique, and Error Structure.....	23
A.2 Assessment of Model Goodness of Fit	24
A.3 Outlier Analysis	24
Appendix B: Identification Of Collision Prediction Models	26
Appendix C: All Models Developed.....	27
Appendix D: References	29

List of Tables

Table 3.1: Segment Models15
Table 3.2: Signalized Intersection Models..... 15
Table 3.3: Un-Signalized Intersection Models.....16

List of Figures

Figure 4.1 Identification of Collision Prone Locations.....19
Figure 4.2 Critical Curves for Identifying Collision Prone Location..... 20

1.0 Introduction

1.1 Background

Traffic collisions represent a major problem on a worldwide basis. Every year, around one million people die in traffic collisions across the world and a further 20 million are injured or disabled. Based on the 20 years of records that are available from the Ministry's Highway Accident System (HAS), there have been over 338,000 police reported collisions that occurred on BC Highways from 1987 through 2006. On average over this time period, there have been 238 fatal collisions, 6627 injury producing collisions and 10,049 property damage only (PDO) incidents per year. Using the current Ministry of Transportation (MOT) values for collisions costs (\$5.6M per fatal, \$100K per injury and \$7350 per PDO), the annual cost of collision on BC highways is in excess of \$2 billion. Given the magnitude of the traffic safety problem, The BC MOT is working hard to improve highway safety in an effort to reduce the economic and societal costs associated with collisions that occur on BC Highway.

It is widely accepted that traffic collisions can be mainly attributed to one or a combination of the following highway system components: the road user (driver and pedestrian), the vehicle, and the road. Collision reduction can be achieved by targeting these components and developing measures to improve their safety performance. Transportation engineering is primarily concerned with the safety performance of the road component of the highway system.

There are two main engineering approaches for dealing with traffic safety problems: the reactive approach and the proactive approach. The reactive approach, or retrofit approach, consists of making the necessary improvements to existing hazardous sites in order to reduce collision frequency and severity at these sites. The proactive approach, on the other hand, is a collision prevention approach that tries to prevent unsafe road conditions from occurring in the first place.

The primary objective of the proactive approach is to ensure that road safety is an explicit priority in transportation planning and design. This stated objective of the proactive approach might seem redundant at first. After all, the most common definition of transportation engineering is that it is the application of science and technology in order to provide for the *safe* and efficient movement of people and

goods. This definition gives the impression that existing transportation planning and design policies and standards lead to the construction of safe roads. However, the level of safety that is built into the road system by following these policies and standards is largely unknown. There is no doubt that these standards and codes were written with safety in mind, but the level of safety that they introduce into the road system is largely unknown.

The reactive approach to safety management focuses on identifying and remedying safety problems in existing road networks. Most road authorities have established road safety improvement programs, otherwise known as black spot programs, whose main goal is to identify hazardous locations in road networks and to establish countermeasures for correcting the problems at these locations. In broad terms, road safety improvement involves the following three stages:

- *Detection* of black spots, otherwise known as hazardous or collision-prone locations.
- *Diagnosis* of the problems that cause the detected locations to be hazardous.
- *Remedy* of the diagnosed problems by establishing and implementing countermeasures that are effective in alleviating them.

The success of safety management (both proactive and reactive) in reducing collision occurrence hinges upon the existence of methods that give reliable estimates of road safety. Under proactive safety management, the quantitative safety evaluation of different transportation plans and designs is synonymous with estimating the level of safety associated with each. Under reactive safety management, the successful detection of black spots hinges upon the ability to accurately estimate the true safety level of any road location. Additionally, the successful remedy of detected black spots requires the ability to estimate the change in safety level brought about by changes in various road characteristics.

The next section gives a brief history describing the early safety estimation techniques, the reasons they are becoming less popular, and the techniques that are replacing them (Collision Prediction Models).

1.2 Road Safety Estimation Techniques

For many years, the prevailing measure of safety was the collision rate¹. Two reasons were most likely behind this practice. First, the collision rate reflects exposure, which researchers have always believed to be the major road-related factor affecting collision occurrence. Second, the collision frequency is not a stand-alone measure of safety. The highly random nature of traffic collisions alerted safety researchers that a location's observed collision rate does not reflect its true safety level. As a result, they fitted probability distributions to the collision rates observed on similar road locations and took the means of these distributions as estimates of the locations' true collision rates. Later, they used Bayesian techniques to further refine these estimates.

But still, there was a need for quantitative relationships that enable the prediction of the safety of road entities based on their various characteristics. Consequently, safety researchers resorted to statistical modeling in order to capture systematic relationships between collisions, traffic volumes, and road geometry. Initially, they employed ordinary or normal linear regression for this purpose. However, they gradually realized that normal linear regression is inappropriate for collision modeling. Several of them demonstrated that traffic collision data do not meet the standard conditions that make this type of regression an appropriate modeling technique.

It has become recognized that the technique of generalized linear regression modeling (GLM) offers the most appropriate and sound approach for developing collision prediction models. Recently, GLM has been used almost exclusively for this purpose. A great number of models developed by this technique reveal that the relationship between collision frequency and exposure is frequently nonlinear, which means that the collision rate is not an appropriate representative of safety. This finding has led a growing number of safety researchers to abandon the use of collision rate as a measure of road safety. Today, collision prediction models constitute the primary tools for estimating road safety. This project has developed collision prediction models for the BC MOT, thereby allowing the Ministry to use this technique in the analysis of highway safety.

¹ Collision rate is the number of collisions per unit of exposure. Exposure for intersections is commonly expressed in million entering vehicles while that for road sections is commonly expressed in million vehicle-kilometers.

1.3 What Is A CPM?

A collision prediction model (CPM) is a regression model that produces an estimate of the collision frequency for a location based on the site-specific characteristics of the location. For a highway segment, the traffic volume and the road length are inputs to the CPM, which produces a collision frequency estimate that represents “normal” safety performance based on the type of highway. For intersections, the CPM utilizes the number of vehicles entering the intersection from the major and minor legs of the intersection.

Historically, many researchers developed CPMs using conventional linear regression, but more recently it has been demonstrated that conventional linear regression models lack the distributional property to adequately describe collisions. This inadequacy is due to the random, discrete, non-negative, and typically sporadic nature that characterize the occurrence of a traffic collision. To overcome these limitations, a generalized linear regression modelling approach (GLM) is used as it assumes a non-normal distribution error structure (usually Poisson or negative binomial). A GLM modelling exercise to produce CPMs requires advanced statistical knowledge and modelling techniques and tests to ensure robust and accurate models are developed.

1.4 Why Should the MOT Use CPMs?

As mentioned earlier, several approaches are available to estimate safety performance, ranging from simple collision rates to collision prediction models. Currently, the MOT uses collision rates as the basis for most safety analysis, however there are several reasons why the MOT should consider using CPMs for safety analysis, as described below.

Collision prediction models have the ability to overcome some of the limitations and difficulties with traditional highway safety measurement. The MOT currently uses collision rates as the foundation for safety measurement and evaluation. Although collision rates do consider exposure (i.e., the amount of traffic on highway segment (MVKm) or entering an intersection (MEV)), the use of collision rates to gauge the safety of a location can be misleading. Several researchers have shown that the relationship between collision frequency and exposure is frequently non-linear, which indicates that collision rates are not appropriate representatives of safety. This finding has led most safety researchers to discard the use of collision rates as a measure of road safety and currently, collision prediction models constitute the primary tool for estimating road safety.

CPMs can facilitate the accurate and consistent quantification of safety performance at all stages of highway planning and design. Traditional safety performance measures of collision rates and collision severities provide very little information concerning the safety of a location and cannot accurately reflect the future potential safety benefits of road improvements. Conversely, CPMs can be used either alone or in combination with collision modification factors (CMFs) to accurately assess and predict the safety of planned highway improvements.

The use of CPMs allows for an improvement in the accuracy of safety measurement, and can facilitate the establishment of acceptable safety performance benchmarks and thresholds. Currently, the MOT has a series of provincial average collision rate tables for different highway classifications and volume ranges. A quick review of these rate tables indicates that there is considerable instability in the rates over time and between volume ranges. This can produce significant problems in the assessment of safety, particularly when the time periods of assessment do not match (i.e., the provincial average collision rate in the table to the collision rate calculated for a project). The use of CPMs will overcome this problem and allow for meaningful safety performance thresholds to be established.

The use of CPMs for safety analysis is consistent with evolving safety evaluation techniques. In fact, CPMs will soon serve as the foundation for the standard practice for road authorities in evaluating highway safety. For example, the Federal Highway Administration's (FHWA) Interactive Highway Safety Design Model (IHSDM) uses CPMs as the basis for the model. Furthermore, the FHWA's Highway Safety Manual (HSM) will also dictate the use of CPMs for safety analysis and evaluation.

Finally, there is a great opportunity to introduce CPMs into the Ministry Highway Accident System (or any system that would replace the HAS). The CPMs can be used for network screening and identifying locations that have a high potential for improvement.

1.5 Developing CPMs for BC - Project Objectives

The objective of this project has been to use multivariate statistical modeling techniques to develop a set of accurate and reliable collision prediction models for intersections and segments for the types of facilities that exist on BC highways. The

goal of the project was to develop a set of “foundation” CPMs that will reflect all major types of provincial highway segments and intersections in British Columbia.

A foundation model for segments is a CPM where the dependant variable is the expected collision frequency (per year or number of years) and the independent variables are the length of the segment (L in km) and the traffic volume (V given in AADT). A foundation model for intersections is similar, except the independent variables are the traffic volumes entering the intersection from the major and minor roadways (V given in AADT). A common functional form of the foundation models for intersections and roadway segment are shown in the equations below.

$$\text{Segment Model Form: } E(\Lambda) = a_0 V_1^{a_1} L_1^{a_2}$$

$$\text{Intersection Model Form: } E(\Lambda) = a_0 V_1^{a_1} V_2^{a_2}$$

1.6 Report Structure

Chapter 1 of this report provided a short introduction, introducing some general background information and listing the objectives of the report. Chapter 2 provides a discussion of the data elements used in the model development. Chapter 3 lists all the models developed. Chapter 4 discusses several model applications. A comprehensive list of references as well as several appendices are provided at the end of this report.

2.0 Data for Developing Collision Prediction Models

The successful development of collision prediction models is highly dependant on the data that is available for the modeling exercise. A significant effort was devoted to the collection, validation and compilation of the data that is necessary to develop the models. This chapter describes the data that was used for the project.

2.1 Highway Segments

For highway segments, there are three data elements that were required for the model development. First was the classification of the highway segments, which was based on the MOT Highway Classification system, which included the following:

- Land use: Urban or Rural
- Road Class: Arterial, Expressway, or Freeway
- Median: Divided or Un-Divided
- Lanes 2 Lanes or 4+ Lanes

This combination of classifications and the types of highways facilities that exist in BC lead to a total of 9 classifications as listed below.

- 1) Rural Arterial Un-Divided 2-Lane Highway
- 2) Rural Arterial Un-Divided 4-Lane Highway
- 3) Rural Arterial Divided 4-Lane Highway
- 4) Rural Freeway Divided 4-Lane Highway
- 5) Urban Arterial Un-Divided 2-Lane Highway
- 6) Urban Arterial Un-Divided 4-Lane Highway
- 7) Urban Arterial Divided 4-Lane Highway
- 8) Urban Expressway Divided 4-Lane Highway
- 9) Urban Freeway Divided 2-Lane Highway

With the highway network disaggregated into these classifications, the next step was to obtain the collision data. The collision data was obtained from the Ministry's Highway Accident System (HAS). Collision data from 2001 to 2005 inclusive, was extracted from the HAS and was used to develop the CPMs. For the segments, the records for all collisions were obtained for the entire highway network and were broken into discrete segments, based on the LKI segment and the traffic volumes defined by the HAS. It is noted that collisions that occurred at intersection collisions were excluded from the data set.

Each of the nine different segment classifications was populated with corresponding highway segments and the collision history. Some of the highway classifications had a

very large sample of segments, such as the rural 2-lane un-divided arterial, since this road type represents the majority of the highway that exist on the BC highway network. In contrast, some classifications had a relatively small sample, because of the limited number of kilometers of the corresponding highway type.

The HAS collisions were broken down into collision severity categories, including fatal collisions, injury collisions and PDO collisions. This was completed so that collision prediction models for different severities could be developed. It is noted however, that in the modeling exercise, the fatal and injury collisions were combined to have a category known as “severe” collisions.

The next step was to verify the traffic volumes for each highway segment, noting that for the collision prediction model, the average annual daily traffic (AADT) is used. The Highway Accident System does have traffic volumes assigned to each segment and these were extracted at the same time as the collision data. In general, this traffic volume data was used for the development of the CPMs, however, there were some locations where the HAS reported traffic volume was suspect, based on the ratio of collisions to traffic volume. For these locations, the Ministry’s website that contains the traffic volumes was accessed to verify traffic volumes.

Both the collision frequency and the traffic volumes were extracted on an annual basis from 2001 to 2005 inclusive (5 years), since the traffic volumes can change over the years. The volume and collision data over the five years were then combined for the model development.

2.2 Highway Intersections

The first step in preparing the data for the development of collision prediction models for highway intersections was to categorize the types of intersections that should be developed. Based on input from staff at the MOT, the following factors were considered important in determining the intersection models to be developed:

- Traffic Control: Signal or Stop-Controlled
- Intersection Type: 4-Leg or 3-Leg (T-type intersection)

Based on these characteristics, a total of six CPMs were developed for intersections on BC Highways, as listed below. It is noted that minor access points and driveways that connect to a highway were not considered as an intersection and not included in the list below.

- 1) Signalized Intersections (ALL intersections)
- 2) Signalized Intersections (Four-Leg intersections)
- 3) Signalized Intersections (Three-Leg intersections)
- 4) Stop-Controlled Intersections (ALL intersections)
- 5) Stop-Controlled Intersections (Four-Leg intersections)
- 6) Stop-Controlled Intersections (Three-Leg intersections)

It is noted that in actuality, there are only 4 different intersection types since category 1 and 4 capture both 4-leg and 3-leg intersections.

With the intersection types defined, the next step was to identify the intersections that should be included in each category. For the signalized intersections, the electrical drawings of all signalized intersections in the province were received and these drawings were used to determine which intersections were 3-leg intersections and which were 4-leg intersections. For the stop-controlled intersections, this information was provided by staff at HQ and was supplemented by a review of other data sources, such as strip maps for different corridors, corridor management plans, safety studies that had been completed over the years, and so on.

It was known that it would be impossible to collect data for all of the stop-controlled intersections on BC Highways. It was also known, that this would not be required, but rather only a sample of locations would be required that would allow for the successful development of the CPM. With this in mind, a target of approximately 100 locations was set for stop-controlled intersections in each category (3-leg and 4-leg intersections).

Another complication for the data for the intersections was the traffic volume, since the Highway Accident System does not include the minor leg traffic volume, but rather only includes the highway (major leg) volumes. If the minor leg volumes were not available, then the site could not be used in the development of the CPM.

Once an adequate sample of intersections was determined for each intersection classification, the HAS was used to extract the collision data at each site. Similar to the data for the highway segments, the collision data for the intersections was extracted over a 5-year period from 2001 to 2005 inclusive, using the LKI location information that was available. The data was extracted by the different severity categories (Fatal, Injury and PDO), which were later collapsed into 'severe; and

'PDO' collision categories.

Perhaps the most difficult step in the entire data preparation exercise was to obtain the traffic volume data for intersections. Staff from the MOT was very helpful in assembling some of the traffic volume data, but unfortunately, to obtain an adequate sample for the model development exercise, it was necessary to look to other information sources for traffic volume, such as the MOT traffic volume website or other studies where the minor road traffic volume was captured. If reliable minor road traffic volume could not be found for a site, then the site would have to be dropped and new sites found.

Several iterations of data extraction followed by a modeling attempt were made before an adequate set of CPMs could be developed. This is primarily due to the limitations of the traffic volumes, and in particular, the lack of the minor road traffic volumes on roads that intersect highways.

In the end, there was an adequate sample of both signalized intersections and stop-controlled intersection to allow for the successful development of the requisite CPMs. However, it is duly noted that more and better minor road traffic volumes for intersections on BC Highways is critical for reliable evaluation of intersections. This is a recognized weakness of the Ministry's Highway Accident System, which only uses the major road volumes in calculating an accident rate for an intersection. It is known that the accident rate calculation using on the mainline volumes cannot effectively capture the safety of an intersection. Ministry staff should examine this issue and if possible, attempts should be made to improve the traffic volume information that is used by the HAS for intersections.

3.0 Collision Prediction Models Developed

3.1 Background

This section presents the results of the collision prediction model development. As previously mentioned models were developed for both highway segments and intersections. After examining available data and sample size requirements, the models included the following categories:

3.1.2 Highway Segment CPMs:

1. Rural Arterial Un-Divided Two-Lane Highways (RAU2)
 - a. PDO Collision Model
 - b. Severe Collision Model
2. Rural Arterial Un-Divided Four-Lane Highways (RAU4)
 - a. PDO Collision Model
 - b. Severe Collision Model
3. Rural Arterial Divided Four-Lane Highways (RAD4)
 - a. PDO Collision Model
 - b. Severe Collision Model
4. Rural Freeway Divided Highways (RFD4)
 - a. PDO Collision Model
 - b. Severe Collision Model
5. Urban Arterial Un-Divided Two-Lane Highways (UAU2)
 - a. PDO Collision Model
 - b. Severe Collision Model
6. Urban Arterial Un-Divided Four-Lane Highways (UAU4)
 - a. PDO Collision Model
 - b. Severe Collision Model
7. Urban Arterial Divided Four-Lane Highways (UAD4)
 - a. PDO Collision Model
 - b. Severe Collision Model
8. Urban Expressway Divided Four-Lane Highways (UED4)
 - a. PDO Collision Model
 - b. Severe Collision Model
9. Urban Freeway Divided Highways (UFD4)
 - a. PDO Collision Model
 - b. Severe Collision Model

3.1.3 Highway Intersection CPMs:

1. ALL Signalized Intersections
 - a. PDO Collision Model
 - b. Severe Collision Model
2. Four-Leg Signalized Intersections
 - a. PDO Collision Model
 - b. Severe Collision Model
3. Three-Leg Signalized Intersections
 - a. PDO Collision Model
 - b. Severe Collision Model
4. ALL Stop-Controlled Intersections
 - a. TOTAL Collision Model
5. Four-Leg Stop-Controlled Intersections
 - a. TOTAL Collision Model
6. Three-Leg Stop-Controlled Intersections
 - a. TOTAL Collision Model

The methodology used in developing these models and the theoretical background is described in Appendix A.

3.2 Summary of Collision Prediction Models Developed

As described earlier the model functional forms for intersections and roadway segment are:

$$\text{Segment Model Form: } E(\Lambda) = a_0 V_1^{a_1} L_1^{a_2}$$

$$\text{Intersection Model Form: } E(\Lambda) = a_0 V_1^{a_1} V_2^{a_2}$$

where,

- $E(\Lambda)$ = collision frequency (collisions/5yrs in this case)
- L = section length
- V = section annual average daily traffic (AADT)
- a_0, a_1, a_2 = model parameters

Table 3.1 provides the parameters of the highway segment models, while Tables 3.2 and 3.3 show the signalized and un-signalized intersection models.

Table 3.1: Segment Models

CPM Project Results						
Segments			Volume	Length	Dispersion	
			a0	a1	a2	k
1	RAU2	PDO	0.005706	0.7523	0.9222	2.90
		Severe	0.005242	0.7279	0.9403	5.02
2	RAU4	PDO	0.000737	0.9394	0.9832	1.92
		Severe	0.008982	0.6875	0.9676	2.39
3	RAD4	PDO	0.000562	0.9862	0.8044	2.51
		Severe	0.000412	1.0000	0.7859	2.78
4	RFD4	PDO	0.155700	0.4140	0.7593	4.26
		Severe	0.097700	0.3977	0.8663	5.33
5	UAU2	PDO	0.091600	0.4689	1.0000	1.86
		Severe	0.057100	0.4763	1.0000	2.75
6	UAU4	PDO	2.470000	0.1358	1.0000	1.87
		Severe	0.458900	0.2923	1.0000	1.90
7	UAD4	PDO	0.030320	0.6291	0.2514	2.12
		Severe	0.003546	0.8170	0.2149	1.89
8	UED4	PDO	0.001677	0.9158	0.8005	1.78
		Severe	0.000388	1.0000	0.8531	3.41
9	UFD4	PDO	0.140800	0.4594	0.8983	3.48
		Severe	0.075800	0.4692	0.9343	5.89

Table 3.2: Signalized Intersection Models

CPM Project Results						
Signalized Intersections			Major Volume	Minor Volume		
			a0	a1	a2	k
10	ALL	PDO	0.005196	0.5077	0.2802	2.138
		Severe	0.000872	0.7582	0.1860	2.167
11	4-Leg	PDO	0.003114	0.5946	0.2418	2.178
		Severe	0.001401	0.7642	0.1334	2.282
12	3-Leg	PDO	0.042330	0.2193	0.3217	2.401
		Severe	0.002395	0.4196	0.3163	2.365
13	4-Leg Rural	PDO	0.000247	0.6839	0.4431	2.307
		Severe	0.001368	0.7013	0.2264	1.95
14	4-Leg Urban	PDO	0.005862	0.4820	0.2852	2.706
		Severe	0.000892	0.7290	0.2082	3.198
15	4-Leg 2-Lane	PDO	0.007119	0.4616	0.3073	2.124
		Severe	0.000217	0.8749	0.2249	2.439
16	4-Leg Multi-Lane	PDO	0.000038	1.0260	0.2410	2.905
		Severe	0.002184	0.7379	0.1150	2.205

Table 3.3: Un-Signalized Intersection Models

CPM Project Results			Major	Minor	Dispersion	
Unsignalized Intersections			Volume	Volume		
			a0	a1	a2	k
17	All	Total	0.000326	0.5442	0.6431	74.100
18	4-Leg	Total	0.002074	0.3471	0.6719	263.200
19	3-Leg	Total	0.000023	0.8141	0.6348	55.500

Only models for the total number of collisions are provided for un-signalized intersections because of the limited sample size and the low collision frequency at these types of locations. To estimate the frequency of PDO and Severe collisions at a 4-leg stop controlled intersection, the predicted collision frequency from the Total model should be multiplied by 0.60 to obtain the frequency of PDO collisions and by 0.40 to obtain the frequency of severe collisions. For a 3-leg stop controlled intersection, the predicted collision frequency from the Total model should be multiplied by 0.55 to obtain the frequency of PDO collisions and by 0.45 to obtain the frequency of severe collisions.

Two statistical measures were used to assess the significance and goodness of fit of the prediction models, including the *Pearson* χ^2 statistic and the scaled deviance (SD), the details of which can be provided in Appendix A. All the models developed showed good fit to the data according to the two statistical measures.

4.0 Applications for Collision Prediction Models

Four applications of collision prediction models are described: location specific prediction, identification of collision-prone locations, ranking the identified collision-prone locations, and before and after safety evaluation.

4.1 Location Specific Prediction and the Empirical Bayes Refinement

There are two types of clues to the safety performance of a location: its traffic and road characteristics (represented by collision prediction models), and its historical collision data. In the absence of data on historical collision record for a location, only the first clue is used. For example, if safety prediction is required for a segment of a new highway design, the appropriate collision prediction model can be used to estimate the safety of this specific segment.

Example 4.1

A rural arterial un-Divided two-Lane highway (RAU2) segment has a length of 1.1 km and average annual daily traffic (AADT) of 12000 vehicles per day. Estimate the predicted collision frequency for the segment.

Solution:

The predicted collision frequency per 5 years can be estimated for PDO, and Injury collisions as:

$$PDO \text{ Collisions} / 5 \text{ yrs} = 0.005706 \times (12000)^{0.7523} \times (1.1)^{0.9222} = 7.3 \text{ Collisions}$$

$$Injury \text{ Collisions} / 5 \text{ yrs} = 0.005242 \times (12000)^{0.7279} \times (1.1)^{0.9403} = 5.34 \text{ Collisions}$$

$$Total \text{ Collisions} / 5 \text{ yrs} = 7.3 + 5.34 = 12.64 \text{ Collisions}$$

EB approach is used to refine the estimate of the expected number of collisions at a location by combining the observed number of collisions at the location with the predicted number of collisions obtained from collision prediction model to yield a more accurate, location-specific safety estimate.

The EB estimate of the expected number of collisions at any location can be calculated by using the following equation:

$$EB_{\text{safety estimate}} = \alpha \times prediction + (1 - \alpha) \times observed$$

where

$$\alpha = \frac{k}{k + \text{prediction}} \quad (2.14)$$

observed = observed number of collisions/5yrs

prediction = predicted number of collisions as estimated from the collision prediction model (collisions/5yrs)

k = Model dispersion parameter as given in Tables 3.1 to 3.3

Example 4.2

For the highway segment of Example 4.1, observed collision frequency is given as 8 and 6 collisions/5yrs for PDO and Injury collisions respectively.

Calculate the refined safety estimate for the location.

Solution:

$$\text{For PDO collisions } \alpha = \frac{2.90}{2.90 + 7.30} = 0.28$$

$$\text{PDO Collisions / 5yrs} = 0.28(7.3) + (1 - 0.28)(8) = 7.8 \text{ Collisions}$$

$$\text{For Injury collisions } \alpha = \frac{5.02}{5.02 + 5.34} = 0.48$$

$$\text{Injury Collisions / 5yrs} = 0.48(5.34) + (1 - 0.48)(6) = 5.68 \text{ Collisions}$$

$$\text{Total Collisions / 5yrs} = 7.8 + 5.68 = 13.48 \text{ Collisions}$$

4.2 Identification of Collision Prone Locations

Collision-prone locations (CPLs) are defined as locations that exhibit a significant number of collisions compared to a specific norm. Because of the randomness inherent in collision occurrence, statistical techniques that account for this randomness should be used when identifying CPLs. The EB refinement method can be used to identify CPLs as demonstrated graphically in Figure 4.1. In the Figure, the prior distribution is obtained from the predicted number of collisions and its variance using the appropriate collision prediction model. The posterior distribution is obtained from the EB safety estimate and its variance. The process is described in detail in Appendix B.

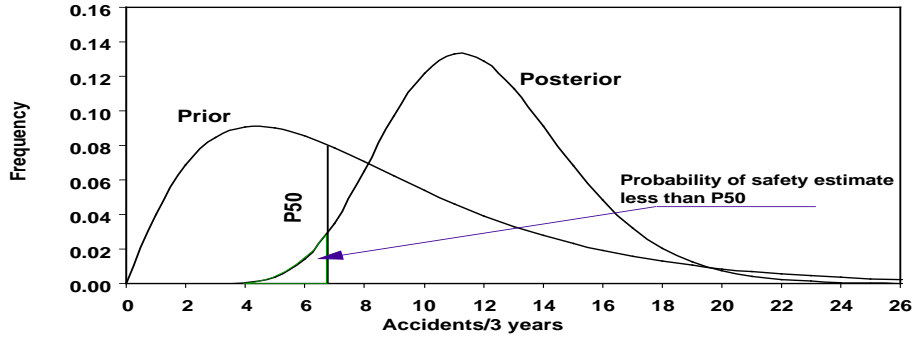


Figure 4.1 Identification of Collision Prone Locations

The process described graphically in Figure 4.1 to identify collision prone locations is cumbersome and requires considerable computational effort. To facilitate this process, critical collision frequency curves can be developed for each GLIM model. A critical curve is one that indicates the number of collisions that should be observed in order to identify a location as collision-prone for a given collision prediction model and a confidence level.

Figure 4.2 shows these curves. The curves are shown representing the 95% confidence level. To illustrate the use of these curves, consider the same example as before. For the given traffic volume and segment length, 7.3 PDO collisions/5 years are predicted. For this number of collisions and for 95% confidence level, and for a model with a k value of 2.9, at least 11.9 collisions/5 years need to be observed to consider this highway segment as collision-prone.

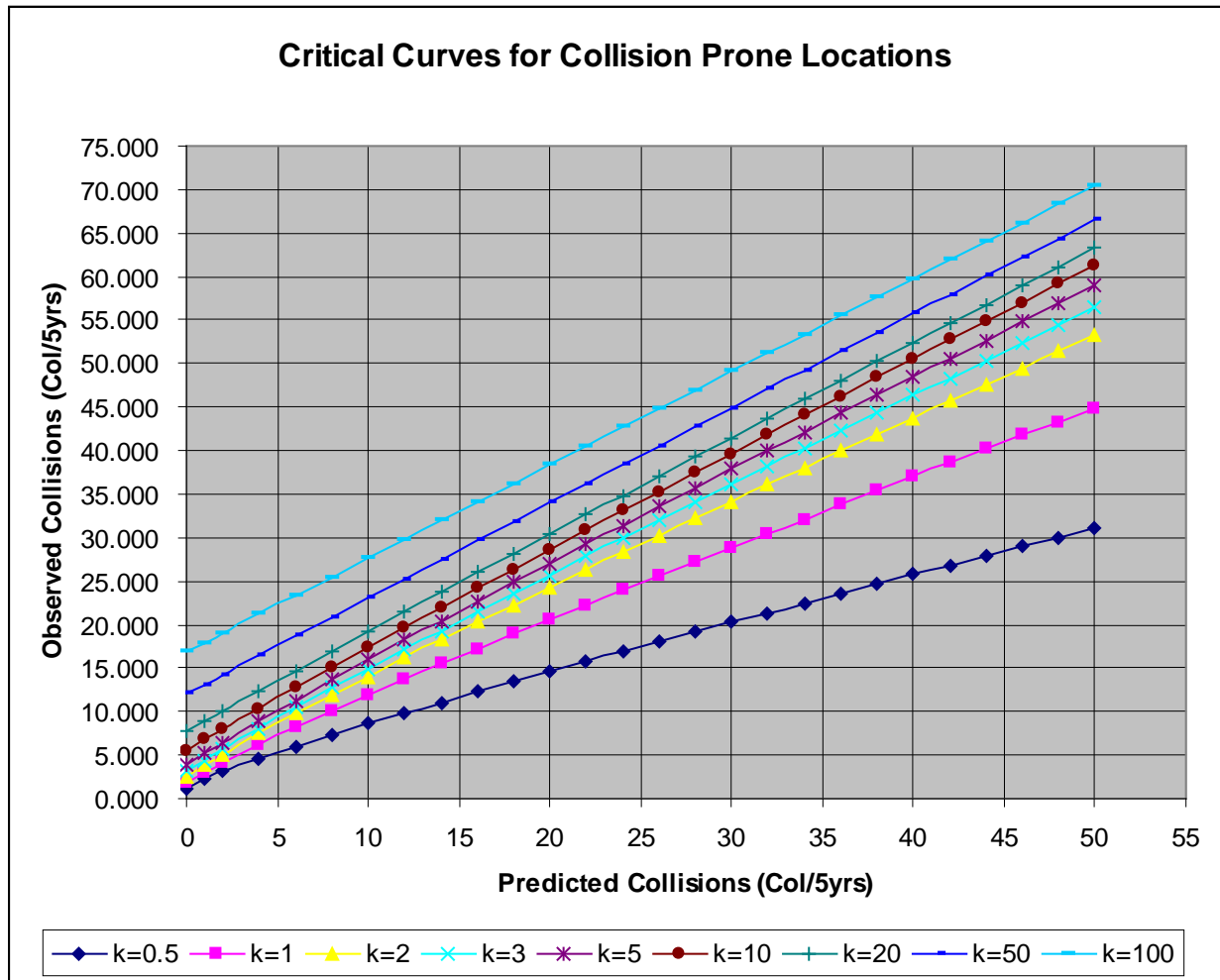


Figure 4.2 Critical Curves For Identifying Collision Prone Location

4.3 Ranking of Collision Prone Locations

The ranking of collision-prone locations is important when road agencies can afford to examine and treat only a limited number of collision-prone locations. In this case they have to select from their list of collision-prone locations those ones whose treatment is likely to achieve the highest rate of return on their investment. In addition to their obligation to utilize their funding most efficiently, road authorities have a moral obligation to the public to ensure that all locations within a certain jurisdiction have approximately equal levels of risk of collision involvement, regardless of the collision frequency observed.

To achieve the goal of high cost-effectiveness, the most logical ranking criterion is the difference between the EB safety estimate and the predicted collision frequency for the collision-prone location. This difference is a good indication of the expected

safety benefits of treatment. To fulfill the moral obligation towards the public, the most logical ranking criterion is the ratio of the EB safety estimate to predicted collision frequency for the collision-prone location. This ratio is a good indication of the risk of collision involvement. The higher this ratio, the higher the risk of collision involvement.

Road authorities can combine the two ranking criteria. The list of collision-prone locations can be sorted according to the sum of the two ranks in ascending order. In this case the two ranking criteria are given equal weight. Alternatively, different weights can be used. For example, a higher weight can be given to the first criterion to achieve a higher cost-effectiveness.

4.4 Before and After Studies

In addition to providing site-specific safety estimates, it has also been shown that the EB procedure significantly reduces the regression to the mean bias that is inherent in collision occurrence. The regression to the mean is a statistical phenomenon by which a randomly large number of collisions for a certain entity during a before period, is normally followed by a reduced number of collisions during a similar after period, even if no measures have been implemented. The effect of a safety measure is often studied by comparing the number of collisions observed after the implementation of the measure, to the expected number of collisions had the measure not been implemented. In simple before and after studies, the observed number of collisions in the period before the implementation is used to estimate the latter value. However, because of the random variations in collision occurrence (e.g. the regression to the mean effect), the observed number of collisions before the implementation may not be a good estimate of what would have happened had no measure been implemented. An alternative and more accurate approach is to use the EB refinement process.

Example 4.3

Using the same example as before, assume that a specific safety measure to reduce the number of collisions at the highway segment was implemented. The observed number of collisions in the next five years following the implementation is given as 6 and 5 collisions/5yrs for PDO and Injury collisions respectively.. Therefore, the effectiveness of the measure can be calculated as:

$$\text{Measure of Effectiveness for PDO Collisions} = 1 - \frac{6}{7.8} = 0.23$$

$$\text{Measure of Effectiveness for Injury Collisions} = 1 - \frac{5}{5.68} = 0.12$$

which indicates reductions by 23% and 12% in PDO and Injury collisions because of the treatment.

Appendix A: Theoretical Background of Collision Prediction Models and Development Methodology

Model development involves the choice of model form and regression technique, determination of the appropriate error structure to use, assessment of model goodness of fit, and identification and removal of model outliers.

A.1 Model Form, Regression Technique, and Error Structure

The following model forms, the merits of which have been discussed extensively in the literature, are adopted:

$$\begin{aligned} \text{Segment Model Form:} \quad & E(\Lambda) = a_0 V_1^{a_1} L_1^{a_2} \\ \text{Intersection Model Form:} \quad & E(\Lambda) = a_0 V_1^{a_1} V_2^{a_2} \end{aligned} \quad \text{A.1}$$

where,

$$\begin{aligned} E(\Lambda) &= \text{collision frequency (collisions/5yrs in this case)} \\ L &= \text{section length} \\ V &= \text{section annual average daily traffic (AADT)} \\ a_0, a_1, a_2 &= \text{model parameters} \end{aligned}$$

The inappropriateness of conventional linear regression for modeling traffic collision occurrence was demonstrated in Chapter 2. Therefore, the estimation of model parameters is carried out using the GLM modeling approach available through the GLIM 4 statistical software package (Numerical Algorithms Group, 1994). The GLM approach to modeling traffic collision occurrence assumes an error structure that is Poisson or negative binomial. The decision on whether to use a Poisson or negative binomial error structure is based on the following methodology. First, the model parameters are estimated based on a Poisson error structure. Then, the dispersion parameter (σ_d) is calculated as:

$$\sigma_d = \frac{\text{Pearson}\chi^2}{n - p} \quad \text{A.2}$$

where n is the number of observations, p is the number of model parameters, and $\text{Pearson}\chi^2$ is defined as:

$$\text{Pearson}\chi^2 = \sum_{i=1}^n \frac{[y_i - E(\Lambda_i)]^2}{\text{Var}(Y_i)} \quad \text{A.3}$$

where y_i is the observed number of collisions on section i , $E(\Lambda_i)$ is the predicted number of collisions for section i , and $\text{Var}(Y_i)$ is the variance of the collision

frequency for section i . The dispersion parameter, σ_a , is noted by McCullagh and Nelder (1989) to be a useful statistic for assessing the amount of variation in the observed data. If σ_a turns out to be significantly greater than 1.0, then the data have greater dispersion than is explained by the Poisson distribution, and a negative binomial regression model is fitted to the data.

A.2 Assessment of Model Goodness of Fit

Two statistical measures are used to assess the goodness of fit of the GLM collision prediction models. The two statistical measures used are those cited by McCullagh and Nelder (1989) for assessing model goodness of fit. These are 1) the *Pearson* χ^2 statistic, defined in equation 3.3, and 2) the scaled deviance. The scaled deviance is the likelihood ratio test statistic measuring twice the difference between the maximized log likelihoods of the studied model and the full or saturated model. The full model has as many parameters as there are observations so that the model fits the data perfectly. Therefore, the full model, which possesses the maximum log likelihood achievable under the given data, provides a baseline for assessing the goodness of fit of an intermediate model with p parameters.

McCullagh and Nelder (1989) have shown that if the error structure is Poisson distributed, then the scaled deviance is as follows:

$$SD = 2 \sum_{i=1}^n y_i \ln \left(\frac{y_i}{E(\Lambda_i)} \right) \quad A.4$$

while if the error structure follows the negative binomial distribution, the scaled deviance is:

$$SD = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{E(\Lambda_i)} \right) - (y_i + \kappa) \ln \left(\frac{y_i + \kappa}{E(\Lambda_i) + \kappa} \right) \right] \quad A.5$$

Both the scaled deviance and the *Pearson* χ^2 have exact χ^2 distributions for Normal theory linear models, but are asymptotically χ^2 distributed with $n - p$ degrees of freedom for other distributions of the exponential family (Aitkin et al., 1989).

A.3 Outlier Analysis

Most data sets contain a few unusual or extreme observations that are not typical of the rest of the data. Such data points are termed “outliers”, and they occur in a set of data either because they are genuinely different from the rest of the data or because errors took place during data collection and recording. In general, these data points deserve special investigation in order to determine whether they were recorded erroneously. However, in the case of traffic collision data, this investigation is not possible due to the fact that such data cannot be replicated unlike the situation with laboratory data for example.

Outliers pose a problem if they are influential points. Influential points strongly influence the values of the regression model parameter estimates. Model users in

general are not in favor of allowing a relatively small percentage of the data to have a significant role in determining the model parameters since this would render the model misrepresentative of the rest of the data. An appropriate measure of influence is Cook's distance, and observation points with high Cook's distance are influential points. Unfortunately, there is no clear rule for what constitutes a high Cook's distance.

The following procedure is used in this project for identifying outliers that are influential:

1. The data are sorted in descending order according to Cook's distance.
2. Starting with the data point having the largest Cook's distance, data points are removed one by one, and the drop in scaled deviance is observed after the removal of each point.
3. The points whose removal causes a significant drop in the scaled deviance at the 95 percent confidence level are considered influential outliers and removed from the database.

Since the scaled deviance is asymptotically χ^2 distributed with $n - p$ degrees of freedom, the removal of one data point with a high Cook's distance is considered to cause a significant drop in the scaled deviance at the 95 percent confidence level only if the resulting drop is equal to or greater than $\chi^2_{0.05,1}$. As mentioned before, the scaled deviances of two negative binomial regression models, with different values of κ , cannot be directly compared. Performing an outlier analysis for a negative binomial regression model therefore requires that the value of κ for the model with n data points be imposed upon the model with $n - 1$ data points; then the difference in scaled deviance can be compared to $\chi^2_{0.05,1}$ in order to assess whether the removed data point is an influential outlier.

Appendix B: Identification Of Collision Prediction Models

Collision prone locations (CPLs) are defined as the locations that exhibit a significant number of collisions compared to a specific norm. Because of the randomness inherent in collision occurrence, statistical techniques that account for this randomness should be used when identifying CPLs. The EB refinement method can be used to identify CPLs according to the following process (Sayed and Rodriguez, 1999; Hagle and Witkowski, 1988; Bélanger, 1994):

1. Estimate the predicted number of collisions and its variance for the intersection, using the appropriate GLM model. This follows a gamma distribution (the prior distribution) with parameters α and β , where:

$$\beta = \frac{E(\Lambda)}{Var(\Lambda)} = \frac{\kappa}{E(\Lambda)} \quad \text{and} \quad \alpha = \beta \cdot E(\Lambda) = \kappa \quad (\text{B.1})$$

2. Determine the appropriate point of comparison based on the mean and variance values obtained in step (1). Usually the 50th percentile (P_{50}) is used as a point of comparison. P_{50} is calculated such that:

$$\int_0^{P_{50}} \frac{(\kappa/E(\Lambda))^\kappa \cdot \lambda^{\kappa-1} \cdot e^{-(\kappa/E(\Lambda))\lambda}}{\Gamma(\kappa)} d\lambda = 0.5 \quad (\text{B.2})$$

3. Calculate the EB safety estimate and its variance from equations 17 and 18. This is also a gamma distribution (posterior distribution) with parameters α_1 and β_1 :

$$\beta_1 = \frac{EB}{Var(EB)} = \frac{\kappa}{E(\Lambda)} + 1 \quad \text{and} \quad \alpha_1 = \beta_1 \cdot EB = \kappa + count \quad (\text{B.3})$$

Then, the probability density function of the posterior distribution is:

$$f_{EB}(\lambda) = \frac{(\kappa/E(\Lambda) + 1)^{(\kappa+count)} \lambda^{\kappa+count-1} e^{-(\kappa/E(\Lambda)+1)\lambda}}{\Gamma(\kappa + count)} \quad (\text{B.4})$$

4. Identify the location as collision-prone if there is a significant probability that the intersection's safety estimate exceeds the P_{50} value. Thus, the location is identified as collision prone if:

$$\left[1 - \int_0^{P_{50}} \frac{(\kappa/E(\Lambda) + 1)^{(\kappa+count)} \lambda^{\kappa+count-1} e^{-(\kappa/E(\Lambda)+1)\lambda}}{\Gamma(\kappa + count)} d\lambda \right] \geq \delta \quad (\text{B.5})$$

where δ represents the confidence level desired (usually 0.95)

Appendix C: All Models Developed

CPM Project Results

Segments				Volume	Length	Dispersion	
				a0	a1	a2	k
1	1A	RAU2	PDO	0.005706	0.7523	0.9222	2.90
	1B		Severe	0.005242	0.7279	0.9403	5.02
	1C		Total	0.020530	0.6877	0.8641	3.39
2	2a	RAU4	PDO	0.000737	0.9394	0.9832	1.92
	2b		Severe	0.008982	0.6875	0.9676	2.39
	2c		Total	0.005560	0.7882	1.0000	2.69
3	3a	RAD4	PDO	0.000562	0.9862	0.8044	2.51
	3b		Severe	0.000412	1.0000	0.7859	2.78
	3c		Total	0.000660	1.0000	0.8022	2.70
4	4a	RFD4	PDO	0.155700	0.4140	0.7593	4.26
	4b		Severe	0.097700	0.3977	0.8663	5.33
	4c		Total	0.241800	0.4145	0.7969	4.41
5	5a	UAU2	PDO	0.091600	0.4689	1.0000	1.86
	5b		Severe	0.057100	0.4763	1.0000	2.75
	5c		Total	0.128500	0.4926	1.0000	1.72
6	6a	UAU4	PDO	2.470000	0.1358	1.0000	1.87
	6b		Severe	0.458900	0.2923	1.0000	1.90
	6c		Total	3.595000	0.1590	1.0000	1.72
7	7a	UAD4	PDO	0.030320	0.6291	0.2514	2.12
	7b		Severe	0.003546	0.8170	0.2149	1.89
	7c		Total	0.027110	0.6963	0.3076	1.97
8	8a	UED4	PDO	0.001677	0.9158	0.8005	1.78
	8b		Severe	0.000388	1.0000	0.8531	3.41
	8c		Total	0.001222	0.9882	0.8324	2.24
9	9a	UFD4	PDO	0.140800	0.4594	0.8983	3.48
	9b		Severe	0.075800	0.4692	0.9343	5.89
	9c		Total	0.5936	0.366	0.9111	2.65

Segment Models

**CPM Project Results
Signalized Intersections**

				Major Volume	Minor Volume	Dispersion	
			a0	a1	a2	k	
10	10a	ALL	PDO	0.005196	0.5077	0.2802	2.138
	10b		Severe	0.000872	0.7582	0.1860	2.167
	10c		Total	0.007140	0.5229	0.2930	2.648
11	11a	4-Leg	PDO	0.003114	0.5946	0.2418	2.178
	11b		Severe	0.001401	0.7642	0.1334	2.282
	11c		Total	0.007830	0.6129	0.1916	2.252
12	12a	3-Leg	PDO	0.042330	0.2193	0.3217	2.401
	12b		Severe	0.002395	0.4196	0.3163	2.365
	12c		Total	0.022760	0.3367	0.3219	8.130

Signalized Intersection Models

**CPM Project Results
Unsignalized Intersections**

				Major Volume	Minor Volume	Dispersion
			a0	a1	a2	k
13	All	Total	0.000326	0.5442	0.6431	74.100
14	4-Leg	Total	0.002074	0.3471	0.6719	263.200
15	3-Leg	Total	0.000023	0.8141	0.6348	55.500

Unsignalized Intersection Models²

² Only models for the total number of collisions are provided for unsignalized intersections because of the sample size and the low collision frequency at these locations.

Appendix D: References

Aitkin, M., Anderson, D., Francis, B., Hinde, J., 1989. *Statistical Modelling in GLIM*. Oxford University Press, New York.

Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of safety at signalized intersections. *Transportation Research Record* 1185, 48-61.

Hauer, E., 1997. *Observational before-after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety*. Pergamon Press, Elsevier Science Ltd., Oxford, U.K.

Hinde, J., 1996. Macros for fitting overdispersion models. *Glim Newsletter* 26, 10-26.

Hinde, J., Demetrio, C.G.B., 1998. Overdispersion: models and estimation. *Computational Statistics & Data Analysis* 27, 151-170.

Hoaglin, D.C., Welsch, R.E., 1978. The hat matrix in regression and ANOVA. *Am. Statist.* 32 (1), 17-22.

Jorgensen, B., 1993. *The Theory of Linear Models*. Chapman and Hall, New York.

Jovanis, P.P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. *Transportation Research Record* 1068, 42-51.

Kulmala, R., 1995. *Safety at rural three and four-arm junctions: Development and application of accident prediction models*. VTT Publications, Espoo, Finland.

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

Miaou, S., Lum, H., 1993. Modeling vehicle accident and highway geometric design relationships. *Accident Analysis & Prevention* 25 (6), 689-709.

Numerical Algorithms Group, 1994. *The GLIM System Release 4 Manual*, The Royal Statistical Society, Oxford, UK.

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

Sawalha, Z., Sayed, T., 2001. Evaluating the Safety of Urban Arterial Roadways. *Journal of Transportation Engineering, ASCE*, Vol. 127(2), pp. 151-158.

Sawalha, Z. and Sayed, T., 2006. Traffic Accidents Modeling: Some Statistical Issues. *Canadian Journal of Civil Engineering*, Vol. 33(9), pp. 1115-1124.