ACCIDENT PREDICTION MODELS FOR UNSIGNALIZED INTERSECTIONS

by

LUIS FELIPE RODRIGUEZ

B.Sc. (Civil Engineering), Universidad de los Andes, Bogotá, Colombia, 1990
M.Sc. (Civil Engineering), Universidad de los Andes, Bogotá, Colombia, 1992

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

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF CIVIL ENGINEERING

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

APRIL, 1998

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that the permission for the extensive copying of this thesis for scholarly purpose may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Civil Engineering
The University of British Columbia
2324 Main Mall
Vancouver, B.C.
Canada, V6T 1Z4

Date: April 15th 1998

ABSTRACT

The main objective of this thesis is to develop Accident Prediction Models (APM) for estimating the safety potential of urban unsignalized (T and 4-leg) intersections in the Greater Vancouver Regional District (GVRD) and Vancouver Island on the basis of their traffic characteristics. The models are developed using the generalized linear regression modeling (GLIM) approach, which addresses and overcomes the shortcomings associated with the conventional linear regression approach. The safety predictions obtained from GLIM models can be refined using the Empirical Bayes' approach to provide, more accurate, site-specific safety estimates. The use of the complementary Empirical Bayes approach can significantly reduce the regression to the mean bias that is inherent in observed accident counts.

The thesis made use of sample accident and traffic volume data corresponding to unsignalized (both T and 4-leg) intersections located in urban areas of the Greater Vancouver Regional District (GVRD) and Vancouver Island. The data included a total of 427 intersections located in the cities of Victoria, Surrey, Nanaimo, Coquitlam, Burnaby and Vancouver. The information available for each intersection included the total number of accidents in the 1993-1995 period, traffic volumes for both major and minor roads given in Average Annual Daily Traffic (AADT) and type of intersection (T or 4-leg). Four categories of models were developed in this study: (1) models for the total number of accidents; (2) separate models for T and 4-leg intersections; (3) separate models for different regions (Vancouver Island, the Lower Mainland and Surrey); and (4) a model for Surrey including intersection control.

Five applications of APM were used in this thesis. Four of them relate to the use of the Empirical Bayes refinement: identification of accident-prone locations, developing critical accident frequency curves, ranking the identified accident-prone locations and before and after safety evaluation. The fifth application provides a safety-planning example, comparing the safety of a 4-leg intersection to two staggered T-intersections. These applications show the importance of implementing APM as a tool to assess in a reliable fashion traffic safety, and design different safety strategies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This thesis represents not only my last step toward an M.A.Sc. Degree, but also represents the fulfillment of a dream that I have had for many years, to obtain a graduate degree from a university outside my country. Many people helped me achieve this goal, which has been the most important achievement in my professional career. This has been a real teamwork comprised by my family, friends and the faculty staff, who in different ways supported me during these last two years.

First of all I would like to thank my supervisor, Dr. Tarek Sayed, for all the time he spent in explaining to me all the ideas and theory behind the phenomenon of traffic accident occurrence, and for encouraging me to keep the enthusiasm during the different stages of this research. I am also very grateful to him for having given me the opportunity to work as a research assistant in two of his projects.

I want to thank the Insurance Corporation of British Columbia (ICBC) for providing the financial support to this research and Delcan Corporation for providing the accident data used to develop the accident prediction models. Without them completing this thesis would have been impossible. I am also grateful to the Instituto Colombiano de Fomento para Estudios Técnicos en el Exterior, ICETEX, for its financial support during my first year in Canada. I also thank my cousin Felipe Hernández for the whole week that he spent in correcting my English and even sometimes my Spanish.

Last but not least, I want to dedicate this work and milestone to the persons that have given me the greatest support in my life, my father Luis and my mother Cecilia. Thank you guys for showing me the way to get where I am now. I have learned from you that despite all difficulties and sometimes frustrations, there are no impossible goals to reach. I feel very happy to inform you that "lo logré!".

CHAPTER I

INTRODUCTION

1.1 Background

Since the dawn of the automobile age about a century ago, traffic safety problems have been a serious concern: an enormous economic and human toll has been exacted as a result of the public's ongoing love affair with the motor vehicle. It is commonly accepted that there are many costs associated with vehicular mobility such as air pollution, noise, and accidents. However, the economic and social costs associated with road accidents greatly exceed other mobility costs due to the loss of property, injury, pain, grief and deaths attributed to road accidents.

In British Columbia, 500 people are killed and 50,000 injured as a result of road accidents. The annual direct claim costs for the Insurance Corporation of British Columbia (ICBC) due to road accidents are estimated to exceed $2 billion (ICBC, 1996 Annual Report), themselves far exceeded by their related social costs. Consequently, the importance of reducing the social and economic costs of road accidents can not be overstated.

Recognizing the traffic safety problem and the importance of reducing the frequency and severity of road accidents, the majority of road authorities have established Road Safety Improvement Programs (RSIPs). The objective of these programs is to identify accident-prone locations, determine possible causes and countermeasures, and to implement the most effective countermeasures in order to alleviate the problems at these locations. The success of these RSIPs can be enhanced by developing statistically reliable accident prediction models, which provide accurate estimates for the traffic safety at road sections and intersections. These safety estimates can be used in identifying accident prone locations and evaluating the effectiveness of remedial measures.

The main objective of this thesis is to develop accident prediction models for estimating the safety potential of urban unsignalized intersections as functions of traffic volumes on both major and minor roads, and type of intersection (T and 4-leg). The data used for this thesis included accident records and traffic volume data for intersections located in the GVRD and urban areas of Vancouver Island. The methodology used to derive these models is based on the Generalized Linear Regression Models (GLIM) approach. The GLIM approach addresses and overcomes the problems associated with conventional linear regression. Several researchers have shown that conventional linear regression lack the distributional property to describe the occurrence of accidents. Some of the potential applications of accident prediction models include: Identifying and ranking accident prone locations, before and after safety evaluation, and safety planning.

The work reported in this thesis is part of the ongoing research at the Civil Engineering Department of the University of British Columbia on accident prediction models. Models have been developed for urban signalized intersections (Feng and Sayed, 1997). Currently, models are being developed for rural signalized intersections, urban and rural corridors.

1.2 Thesis Structure

This thesis is divided into six chapters. Chapter One provides an overview of the thesis and its structure. Chapter Two summarizes previous work on accident prediction models, the theoretical background of the GLIM approach, and its applications to accident prediction models. Chapter Three describes with the accident and traffic volume data used and the models developed. Chapter Four discusses several statistical issues related to the GLIM approach. Chapter Five discusses several applications of the models. The applications include: identification of accident prone locations; developing critical frequency curves, ranking of accident prone locations; before-and-after safety evaluation; and the use of the models in safety planning. Chapter Six provides suggestions for follow up work and the summary and conclusion of the thesis.

CHAPTER II

LITERATURE REVIEW

2.0 Introduction

The relationship between traffic accidents and traffic volumes has been the subject of numerous studies. Most of the earlier studies used the conventional linear regression approach to develop models relating accidents to traffic volumes. However, the past decade has seen a significant development and advances in accident data analysis and modeling. Accident prediction models are no longer limited to conventional linear regression approach, as more accurate and less restrictive nonlinear models are considered. In addition, the use of Empirical Bayes' approach for refining the estimates obtained from accident prediction models has also been an important development. This chapter describes the statistical theory behind the accident prediction models, as well as previous research and developments.

2.1 Shortcomings Associated with Conventional Linear Regression Models

The conventional linear regression model is defined as follows:

$$Y_i = a_0 + \sum_{j=1}^{k} a_j x_{ij} + \varepsilon_i \qquad (2.1)$$

where,

Y_i = estimated or dependent variable

a_o, a_j = estimated coefficients

x_{ij} = independent variables

ε_i = estimated error, assumed to be normally distributed

Several researchers (Jovanis and Chang, 1986, Saccomanno and Buyco, 1988, Miaou and Lum, 1993) have shown that conventional linear regression models lack the distributional property to adequately describe random, discrete, non-negative, and typically sporadic events which are all characteristics of traffic accidents.

Jovanis and Chang (1986) identified three shortcomings associated with the assumption of a normal distribution error structure. The first shortcoming is found in the relationship between the mean and the variance of accident frequency. Jovanis and Chang (1986) demonstrated that as volume of traffic increases, so does the variance of accident frequency. Under a normal distribution assumption, the variance remains constant. The second shortcoming is associated with the non-negativity of accident occurrence. Predicted negative values under conventional linear models might occur when there exists low accident frequencies in the data set. A way to avoid this problem is by using non-linear models, which are linearized in a logarithm fashion in order to estimate their parameters. The third problem is related with the non-normality of the error distribution, due to the characteristics of non-negativity and small value of discrete dependent variable. Jovanis and Chang (1986) found that the best way to overcome these problems is to assume a Poisson distribution error structure. Their results were demonstrated by modeling accidents at highway sections in Indiana.

Miaou and Lum (1993) identified the same shortcomings when performing a comparison between four accident prediction models applied to trucks on highways. Two of these models were developed under the assumption of normal distribution error structure, while the others were assumed to be Poisson distributed. It was found that predicted values, from models using the Poisson distribution assumption were much closer to the observed values and its estimated coefficients had higher t-statistics, which denote higher significance. For the normally distributed models, it was found that some of the estimated coefficients had signs contrary to the expectation. These results confirmed all the shortcomings associated with the conventional linear regression technique and its applicability for developing accident prediction models.

2.2 Generalized Linear Models (GLIM)

As seen in the previous section, GLIM has the advantage of overcoming all the shortcomings associated with the conventional linear regression approaches. As well, GLIM has the flexibility of assuming different error distributions and link functions that allow the conversion of non-linear models into linear models. Recognizing the advantages of the GLIM approach, it will be utilized in this thesis.

The GLIM approach used herein is based on the work of Kulmala (1995) and Hauer *et-al*, (1988). Assuming that Y is a random variable that describes the number of accidents at an intersection in a specific time period, and y is the observation of this variable during a period of time. The mean of Y is Λ which can also be regarded as a random variable. Then for $\Lambda=\lambda$, Y is Poisson distributed with parameter λ:

$$P(Y = y|\Lambda = \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}; \; E(Y|\Lambda = \lambda); \; Var(Y|\Lambda = \lambda) = \lambda \qquad (2.2)$$

Since each site has its own regional characteristics with a unique mean accident frequency Λ, Hauer *et-al*, (1988) have shown that for an imaginary group of sites with similar characteristics, Λ follows a gamma distribution (with parameters κ and κ/μ), where κ is the shape parameter of the distribution. That is:

$$f_{\Lambda}(\lambda) = \frac{(\kappa/\mu)^{\kappa}\lambda^{\kappa-1}e^{-(\kappa/\mu)\lambda}}{\Gamma(\kappa)} \tag{2.3}$$

with a mean and variance of:

$$E(\Lambda) = \mu;\ Var(\Lambda) = \frac{\mu^2}{\kappa} \tag{2.4}$$

Kulmala (1995) has also shown that the point probability function of Y based on equations (2.3) and (2.4) is given by the negative binomial distribution:

$$P(Y = y) = \frac{\Gamma(\kappa + y)}{\Gamma(\kappa)y!}\left(\frac{\kappa}{\kappa + \mu}\right)^{\kappa}\left(\frac{\mu}{\lambda + \mu}\right)^{y} \tag{2.5}$$

with an expected value and variance of:

$$E(Y) = \mu;\ Var(Y) = \mu + \frac{\mu^2}{\kappa} \tag{2.6}$$

As shown in equation (2.6), the variance of observed accidents for the entire sample has two sources: the second term (μ^2/κ) from the variance of the predicted number of accidents, and the first term (μ) from the variation of the number of accidents (Kulmala, 1995). Notice that when

$\kappa \rightarrow \infty$, the variance of equation (2.6) equals the mean, which is identical to the Poisson distribution.

As described earlier, for the GLIM approach, the error structure that best fits the accident occurrence is usually assumed to be Poisson or negative binomial. The main advantage of the Poisson error structure is the simplicity of the calculations, because the mean and variance are equal and its method for calculation is readily included in the GLIM software package (NAG, 1994). However, this advantage is also a limitation. It has been shown (Kulmala and Roine, 1988, and Kulmala, 1995) that most accident data is likely to be overdispersed (the variance is greater than the mean) which indicate that the negative binomial distribution is the more realistic assumption.

Miaou and Lum (1993) identified three possible sources of overdispersion in accident data. The first is related to omitted variables that explain accident occurrence. Traffic accidents depend on numerous variables including geometric characteristics, weather, time of day, and human factors. Many of these variables are not discernible from accident records. The second possible source of overdispersion is related to uncertainties in vehicle exposure data, derived from error during collection of data. The third source comes from non-homogeneous roadway environments, which can explain why accident rates are different during daylight and night times or during rainy versus sunny days.

The main difficulty associated with using the negative binomial distribution error structure is the determination of the shape parameter κ. Kulmala (1995) proposed an iterative approach using the method of moments. The GLIM software package (V 4.0) includes a macro library in which the parameter κ is calculated by three different iterative methods: the maximum likelihood, the mean deviance estimate, and the mean χ^2 estimate (NAG, 1996). A comparison between the four methods will be provided in Chapter 4. The method of the maximum likelihood is used in this thesis.

Bonneson and McCoy (1993) proposed a methodology to decide whether to use a Poisson or negative binomial error structure. First, the model parameters are estimated based on a Poisson distribution error structure. Secondly, a dispersion parameter (σ_d) is calculated. The dispersion parameter is defined as:

$$\sigma_d = \frac{Pearson\ \chi^2}{n - p} \qquad (2.7)$$

where n is the number of observations and p is the number of model parameters (The *Pearson* χ^2 test will be described in detail in next section).

If σ_d is greater than 1.0, then the data have greater dispersion than is explained by the Poisson distribution, and a further analysis using a negative binomial distribution is required. If σ_d is near 1.0, then the assumed error structure approximately fits the Poisson distribution. This method has the advantage of testing the model under the Poisson distribution first, which is easier to estimate than the negative binomial distribution.

2.3 Testing the Models Significance

The significance of GLIM models is usually assessed using the Scaled Deviance (SD) and the *Pearson* χ^2 test. The SD is defined as the likelihood test ratios measuring the difference between the log likelihood of the studied model, and the saturated model (Kulmala, 1995). The general equation for SD is defined as follows:

$$SD = 2\log f(\mathbf{y},\mathbf{y}) - 2\log f(\mathbf{E}(\Lambda),\mathbf{y}) \tag{2.8}$$

where $\log f(\mathbf{E}(\Lambda),\mathbf{y})$ is the natural logarithm for the probability density function.

Mc Cullagh and Nelder (1983) have shown that for the Poisson the SD is defined as:

$$SD = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{E(\Lambda_i)}\right) \tag{2.9}$$

and for the negative binomial distribution the SD is defined as:

$$SD = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{E(\Lambda_i)}\right) - (y_i+\kappa)\ln\left(\frac{y_i+\kappa}{E(\Lambda_i)+\kappa}\right)\right] \tag{2.10}$$

The scaled deviance is asymptotically χ^2 distributed with *n-p-1* degrees of freedom. Therefore, for a well-fitted model with appropriate link function, error distribution and functional form, the expected value of SD will approximately equal the number of degrees of freedom (Maycock and Hall, 1984)

Another measure to assess the significance of the GLIM models is the *Pearson* χ^2 statistic defined as (Bonneson and McCoy, 1993):

$$Pearson\ \chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - E(\Lambda_i)\right]^2}{Var(y_i)} \tag{2.11}$$

where y_i is the observed number of accidents at intersection i, $E(\Lambda_i)$ is the predicted number of accidents obtained from the accident prediction model, and $Var(y_i)$ is the variance of the observed accidents defined in equation (2.2) and (2.6) for Poisson and negative binomial distributions, respectively. The *Pearson* χ^2 statistic follows the χ^2 distribution with n-p-1 degrees of freedom, where n is the number of observations, and p is the number of model parameters.

In addition, useful subjective measures of the model goodness of fit are graphical methods. One of them is to plot the predicted accident frequency versus the observed accident frequency. A well fitted model should have all points in the graph clustered symmetrically around the 45° line. A second graphical method is to plot the average of squared residuals versus the predicted accident frequency. For a well fitted model, all points should be around the variance function line as defined in equation (2.6) for the negative binomial distribution.

Another graphical method is to calculate the Prediction Ratio (*PR*) and plot it against the predicted values. *PR* is defined as the normalized residual, which is the difference between the predicted and observed accidents, divided by the standard deviation (Bonneson and McCoy, 1997). *PR* can be calculated according to the following equation:

$$PR_i = \frac{E(\Lambda_i) - y_i}{\sqrt{Var(y_i)}} \tag{2.12}$$

For a well fitted model PR_i should be clustered around the zero axis in a Predicted Accidents vs. *PR* graph.

Finally, the T-ratio test is used to measure the statistical significance of the variable coefficients. The t-ratio test is defined as the ratio between the estimated GLIM parameter and its standard error. For a significant variable at 95% level of confidence, the t-ratio should be greater than 1.96.

All six tests described in this section were used to access the significance of the models developed for this thesis.

2.4 Model Structure

Intersection accident prediction models can be generally classified into two types. The first type relates accidents to the sum of traffic flows entering the intersection, while the second relates accidents to the product of traffic flows entering the intersection. The latter type has been shown to be more suitable to represent the relationships between accidents and traffic flows at intersections (Hauer *et-al*, 1988). In this kind of structure, accident frequency is a function of the product of traffic flows raised to a specific power (usually less than one). This approach has been used in this thesis. That is:

$$E(\Lambda) = a_0 \times V_1^{a_1} \times V_2^{a_2} \tag{2.13}$$

where,

$E(\Lambda)$	= predicted accident frequency
V_1	= major road traffic volume
V_2	= minor road traffic volume
a_0, a_1, a_2	= model parameters

As mentioned earlier, accident occurrence is not a function of traffic flows only, but also other variables (e.g. weather, intersection type, geometric features, etc.). Kulmala (1995) and Maher and Summersgill (1996) proposed to model these additional variables along with traffic flows as follows:

$$E(\Lambda) = a_0 \times V_1^{a_1} \times V_2^{a_2} \times e^{\sum_{j=1}^{m} b_j x_j} \tag{2.14}$$

where x_j represents any of the m additional variables.

2.5 Location Specific Prediction: The Empirical Bayes Refinement

There are two types of clues to the safety of a location: its traffic and road geometric design characteristics, and its historical accident data (Hauer, 1992, Brüde and Larsson, 1988). The Empirical Bayes (EB) approach makes use of both clues. The EB approach is used to refine the estimate of the expected number of accidents at a location by combining the observed number of accidents at the location with the predicted number of accidents obtained from the GLIM model, to yield more accurate, location-specific safety estimate.

The EB estimated number of accidents for any intersection can be calculated by using the following equation (Hauer *et-al*, 1992):

$$EB_{safety\ estimate} = \alpha \times E(\Lambda) + (1 - \alpha) \times count \tag{2.15}$$

where,

$$\alpha = \frac{1}{1 + \frac{Var(E(\Lambda))}{E(\Lambda)}} \tag{2.16}$$

count = observed number of accidents

E(Λ) = predicted number of accidents as estimated from the GLIM model

Var(E(Λ)) = variance of the GLIM estimates

Using the variance of the predicted accidents, *Var(E(Λ))*, defined in equation (2.4), equation (2.15) can be rearranged to yield:

$$EB_{safety\ estimate} = \left(\frac{\kappa}{\kappa + E(\Lambda)}\right) \times E(\Lambda) + \left(\frac{E(\Lambda)}{\kappa + E(\Lambda)}\right) \times count \tag{2.17}$$

In addition, the variance of the EB refined estimate can be calculated using the following equation (Kulmala, 1995):

$$Var(EB_{safety\ estimate}) = \left(\frac{E(\Lambda)}{\kappa + E(\Lambda)}\right)^2 \times \kappa + \left(\frac{E(\Lambda)}{\kappa + E(\Lambda)}\right)^2 \times count \tag{2.18}$$

Equation (2.17) shows that the EB refined estimate lies between the observed and the predicted number of accidents, combining both the individual accident history of the location and the GLIM model prediction (Figure 2.1).

The κ parameter also plays an important role in the calculation of the EB estimate. Kulmala (1995) showed that for high values of κ, the variance of the predicted accidents is low (equation 2.4), and therefore, there is a small uncertainty and the EB estimate is closer to the GLIM estimate. Conversely, when κ is low, the variance of the predicted value is high as is the uncertainty of the GLIM model. Therefore, the EB estimate is closer to the observed value. Figure 2.1 shows how the κ value affects the EB estimate.

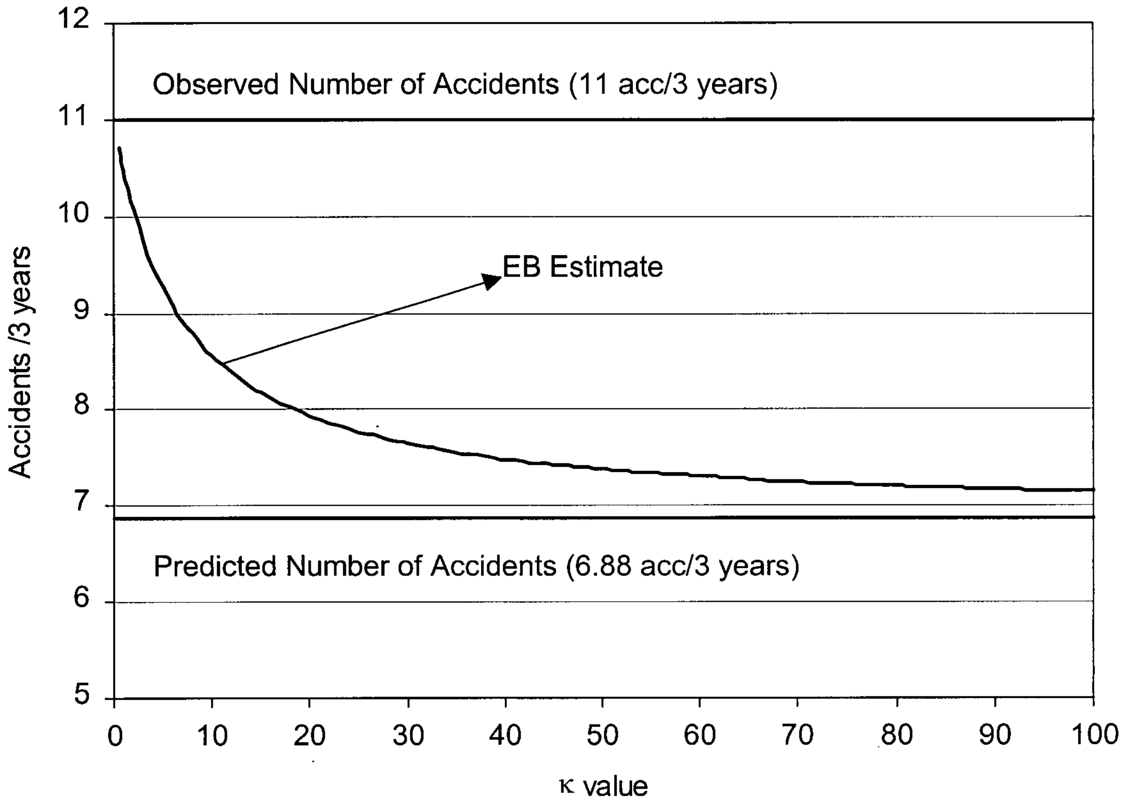


Figure 2.1 Empirical Bayes' Estimate for Different κ Values

In addition to combining the two types of safety clues and providing site-specific safety estimates, it has also been shown that the EB procedure significantly reduces the regression to the mean effects that are inherent in observed accidents count (Brüde and Larsson, 1988). The regression to the mean is a statistical phenomenon by which a randomly large number of accidents for a certain entity during a before period, is normally followed by a reduced number of accidents during a similar after period, even if no measures have been implemented (while the opposite applies in the case of a randomly small number of accidents).

The EB refinement is important for various applications of GLIM models, such as identification and ranking of accident-prone locations, and assessment of effectiveness of safety measures. The

EB estimate combined with reliable GLIM models, has the advantage of overcoming the difficulties associated with defining reference groups to perform before and after studies (Mountain and Fawaz, 1996).

2.6 Previous work

There are few studies dealing with accident prediction models at junctions, even though most of accidents occurs at these kind of locations.

Satterthwaite (1981), made an extensive review of over 80 studies dealing with the relationship between traffic accidents and traffic volumes. Most of models reported in this study consider accidents at road sections, and only 14 of the references reported deal with accident models at intersections.

For accident at intersections Satterthwaite (1981) found some non-linear relationships between accidents and traffic volumes at T-intersections located in rural areas. The proposed models are desegregated in accidents of vehicles turning left and right from the minor road (non-through road) to the major road (through road). The relationships found are similar to equation (2.13) but in one study it was found that the a_1 and a_2 coefficients are approximated to 0.5, while in a subsequent study a_1 is approximated to 1 while a_2 is again approximated to 0.5. These models were developed during the 50's and 60's and were estimated by using the conventional regression analysis.

Also reported were similar studies conducted at other intersections where traffic control and layout variables were taken into account, however there is no report related to accident prediction models at urban unsignalized intersections. At the conclusion of the study, it was found that results concerning accidents at intersections were not consistent and it was suggested that more research should be done.

Bonneson and McCoy (1993), using data from 125 two-way stop controlled intersections in Minnesota, developed the following model:

$$Accidents\,/\,year = 0.692\left(\frac{AADT_{major\ road}}{1000}\right)^{0.256}\left(\frac{AADT_{minor\ road}}{1000}\right)^{0.831} \tag{2.19}$$

Using a similar approach, Bélanger (1994) developed several models using data from 149 4-leg unsignalized intersections in western Quebec. The models included the "total-accidents model" for different ranges of speed; "accident-type models" such as right angle, rear end etc.; and models including other variables such as the existence of flashing beacons, sight distance and turning lanes. For instance, the total-accidents model for all speeds developed by Bélanger is as follows:

$$Accidents\,/\,year = 0.00193\left(AADT_{major\ road}\right)^{0.42}\left(AADT_{minor\ road}\right)^{0.51} \tag{2.20}$$

Both Bonneson and McCoy and Bélanger models were developed for intersections in rural areas assuming a negative binomial distribution error.

In a more recent study, Maher and Summersgill (1996), using selected data recorded all over the UK, developed the following model for T-intersections on urban single carriageways based on the negative binomial distribution:

$$Accidents\,/\,year = 0.049\left(\frac{AADT_{major\ road}}{1000}\right)^{0.80}\left(\frac{AADT_{minor\ road}}{1000}\right)^{0.36} \qquad (2.21)$$

In addition, Mountain and Fawaz (1996), using the same approach (negative binomial), derived a model for 390 unsignalized intersections located in 12 UK counties. Out of the 390 intersections, 338 were T-intersections and approximately 35% were located in urban areas. The model developed is as follows:

$$Accidents\,/\,year = 0.141\left(\frac{AADT_{major\ road}}{1000}\right)^{0.64}\left(\frac{AADT_{minor\ road}}{1000}\right)^{0.24} \qquad (2.22)$$

Since this thesis deals only with intersections located in urban areas, only the models in equations (2.21) and (2.22) will be compared with the models developed in this thesis. This comparison is shown in next chapter.

2.7 Conclusion

Developing accident prediction models has been a concern for the last four decades. During the 50's 60's, 70's, the models were limited by the use of the conventional linear regression analysis, leading to inconsistencies and misinterpretation in describing traffic accidents occurrence. The advancements in computer and software technology during the last two decades, and the development of more sophisticated statistical tools, has resulted in the development and release of software packages such as GLIM and SAS, which are capable of solving non-linear regression models by specifying any type of error structure consistent with the data.

Several researchers have found that accident occurrences follow the negative binomial distribution, rather than the Poisson distribution, because it has been shown to be the most appropriate way to model overdispersion.

With respect to the use of GLIM to develop safety models for unsignalized intersections, only a few studies were found. Most of these studies deal with the rural environment. Most of GLIM accident prediction models have been developed during the last 10 years and are focused on signalized intersections, rural areas, and road sections. More work is needed in developing accident prediction models for urban unsignalized intersections.

CHAPTER III

DATA COLLECTION AND MODEL DEVELOPMENT

3.0 Introduction

This chapter is divided into three sections. The first section contains a detailed description of the data used to develop the accident prediction models. It also includes a procedure to identify outliers which may affect the quality of the models. The second section describes the models developed and their goodness of fit. Finally, the third section shows a comparison between the developed models and similar models found in the literature.

3.1 Data Collection

This thesis made use of sample accident and traffic volume data corresponding to unsignalized (both T and 4-leg) intersections located in urban areas of the Greater Vancouver area and the Vancouver Island.

3.1.1 Accident and Traffic Volume Data

Three years of accident data was available for analysis on each intersection (1993-1995). The source of the accident data is the MV 104 accident reporting form, British Columbia's accident police report. The data set contained 427 intersections from the cities of Surrey, Victoria, Coquitlam, Vancouver, Burnaby and Nanaimo. The information available for each intersection includes the total number of accidents that occurred during the 1993-1995 period. The explanatory variables of accident occurrence included the traffic volumes on the both the major and minor roads given in Average Annual Daily Traffic (AADT), and the type of intersection (T or 4-leg).

Another explanatory variable taken into account for this thesis, is the type of intersection control, which was only available for Surrey intersections. Traffic control types included 2-way Stop, 4-way Stop, and one-way Stop at T-intersections. Tables 3.1 and 3.2 provide a statistical summary of the data.

City	Number of Intersections			Number of Accidents		Average AADT	
	Total	T	4-leg	Acc/year	Acc/yr/Int.	Major Road	Minor Road
Surrey	56	18	38	285	5.08	17,937	3,075
Victoria	340	162	178	360	1.06	12,355	1,494
Nanaimo	10	0	10	23	2.25	7,172	3,242
Coquitlam	8	2	6	34	4.25	10,004	1,514
Burnaby	9	3	6	36	4.04	12,984	2,837
Vancouver	4	1	3	17	4.33	22,191	1,408
Total	**427**	**186**	**241**	**755**	**1.77**	**13,186**	**1,770**

Table 3.1 Summary of Accident, Intersection Control and Traffic Volume Data

City	Number of Accidents*			AADT minor Road			AADT Major Road		
	Max.	Min.	Std. Dev.	Max.	Min.	Std. Dev.	Max	Min	Std Dev
Surrey	11.0	1.7	2.3	9,300	500	2,060	42,600	2,100	10,385
Victoria	8.3	0.0	1.2	11,000	100	1,483	47,800	500	9,397
Nanaimo	4.8	0.6	1.5	6,025	1,968	1,307	15,739	2,771	4,132
Coquitlam	8.7	1.0	2.7	2,360	730	542	32,310	730	10,642
Burnaby	10.3	0.3	3.0	7,415	365	2,252	29,020	5,715	7,492
Vancouver	8.3	0.3	3.5	2,550	860	775	37,295	7,835	12,070
Total	**11.0**	**0.0**	**2.1**	**11,000**	**100**	**1,673**	**47,800**	**500**	**9,673**

* Indicates average annual accidents per intersection

Table 3.2 Statistical Summary of Accidents

As shown in Table 3.1, the average number of accidents per intersection for the cities located in the Lower Mainland (Surrey, Coquitlam, Burnaby and Vancouver) is much higher than the average of number of accidents per intersection for the cities located in Vancouver Island (Victoria and Nanaimo). About 44% of the intersection are T-intersections, while the rest 56%

are 4-leg intersection. This indicates that there is not an absolute predominance of either one of the intersection types in the database, unlike the studies made by Mountain and Fawaz (1996) and Maher and Summersgill (1996), where their data set included mainly T-intersections. This condition is desirable when developing a total model of accidents as the model will not be biased in favor of one of the intersection types.

As previously mentioned, intersection control type data is available only for Surrey intersections. Of the 56 intersections, 32 are two-way Stop controlled, 8 are 4-way Stop controlled, and the remainders 16 intersection are classified as one-way Stop-T intersections.

3.1.2 Outlier Analysis

Outliers are defined as data points that split off or are very different from the rest of the data (Stevens, 1986). Outliers can be caused by irregularities or errors occurred during the data recording or observation process or when the data is genuinely different from the rest. These points deserve further investigation in order to decide whether or not to remove them.

Kulmala (1995) proposed a procedure to identify outliers based on the calculation of the leverage statistic. The leverage of a point is a measure of how far the x-value of the point is away from the average of the rest of the x-values (NAG, 1994). The leverage values are the diagonal elements of the hat matrix, which is the matrix that multiplies the observed vector in order to yield the predicted vector. One of the properties of the leverage values, h_i, is that the sum over the n-values, yields the number of parameters, p, in the model. According to this statement the average value of the leverage is *p/n*, and many authors (NAG, 1994, Stevens, 1988) consider that a high leverage is one that exceeds 2*p/n*, and should be subject to further examination.

However, it has been shown (NAG, 1994) that the leverage alone is not a good indication of whether the parameters estimate is being affected by specific observations. A measure which does this is the Cook's distance (NAG, 1994). The Cook's distance measures the influence of observations on the model. The higher the Cook's distance value for a given observation, the stronger its influence on the model. The Cook's distance is calculated as follows:

$$c_i = \frac{h_i}{p(1-h_i)}\left(r_i^{PS'}\right)^2 \tag{3.1}$$

where,

h_i	= leverage value
p	= number of parameters
$r_i^{PS'}$	= standardized residual

The main disadvantage of using the Cook's distance is that there is no clear rule for what constitute a high c_i. NAG (1994) proposes to sort the data according to the Cook's distance values, and in a stepwise procedure, remove the points with the highest values, and for every point removed, assess the change in the scaled deviance.

Maycock and Hall (1984) have found that the difference in scaled deviance in two models with degrees of freedom df_1 and df_2, is χ^2 distributed with parameters (df_1 - df_2). This means that if only one point with a high Cook's distance is removed, then the difference in the scaled deviance must be greater than 3.8 (the χ^2 value for 95% level of confidence and 1 degree of freedom).

GLIM has the capacity of extracting both leverage and Cook distance values, from each model. The procedure to identify outliers in the models developed in this thesis is to visually examine the relationship between the observed number of accidents for each intersection and the Cook's distance. Intersections with exceptionally large values of c_i are then removed and the change in scaled deviance is determined. If this change is significant the intersections are removed.

The previous analysis was performed to all models of this thesis. After the analysis none of the critical points were classified as outliers that should be removed. Figure 3.1 and Table 3.3 show the results of this procedure for the total accident model. From visual examination of Figure 3.1 it was determined to select five intersections for removal (Cook's Distance greater than 0.02 and the intersections are tagged 1 through 5 in the figure). As shown in Table 3.3, the cumulative drop in scaled deviance is always below the χ^2 statistics. This indicates that removing these intersections from the data set is not warranted. The analysis summarizing the results for the remaining models is shown in Appendix I.

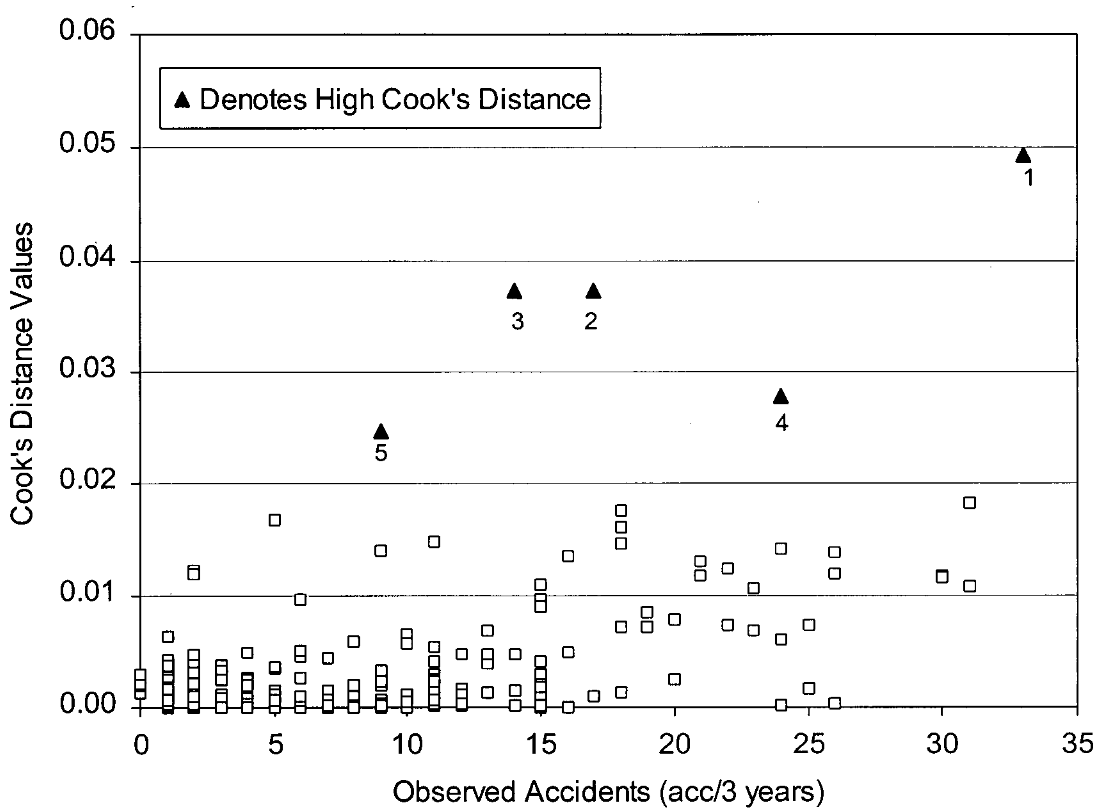


Figure 3.1 Identification of the Highest Cook's Distance Values for Total Model

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	14	426	397.4	1.5	1.5	3.8
2	22	425	395.3	2.1	3.5	6.0
3	157	424	393.2	2.1	5.7	7.8
4	5	423	391.0	2.2	7.8	9.5
5	33	422	389.0	2.0	9.8	11.1

Table 3.3 Identification of Outliers for Total Model

3.2 Model Development

The main task of this research is to develop multivariate models to estimate the predicted number of accidents. Four categories of models were developed in this thesis: (1) models for the total number of accidents; models for T and 4-leg intersections; (3) separate models for every region

(Vancouver Island, the Lower Mainland, and Surrey); and (4) a model for Surrey including intersection control type.

Since the average number of accidents per year per intersections is relatively small (especially for intersections in Victoria and Nanaimo), it was decided to use the number of accidents in a three year period.

The models developed are assumed to follow the negative binomial distribution, which is included in the GLIM software package, through a macro designed by NAG (1996). Out of the six goodness of fit tests described in section 2.3, the graphs describing the predicted accidents vs. the Prediction Ratio for every model are shown in Appendix II, while the rest of tests are shown with the description of each model. In general, the Prediction Ratio graphs show similar dispersions as the ones obtained for the observed vs. predicted accident graphs. Appendix III shows the GLIM output of all models, which in addition to the models' parameters, includes the scaled deviance, the κ value (represented by THETA in the GLIM output), and the standard error of the parameters. Note that the model under Poisson distribution assumption is developed first.

3.2.1 Model for the Total Number of Accidents

A model relating the total number of accidents to the traffic volumes for minor and major roads was developed. The whole data set is used for this analysis and Table 3.4 shows the parameter estimates of the model and its goodness of fit.

Model Form	t-ratio		SD (dof)	κ	Pearson χ^2 (χ^2 test)*
$Acc/3\,yrs = 1.4929 \times \left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.3839} \times \left(\frac{AADT_{min\,rd}}{1000}\right)^{0.7044}$	a_o a_1 a_2	3.2 7.8 12.4	399 (424)	1.97	459 (472)

* Denotes significance at a 95-percent confidence level

Table 3.4 Model for the Total Number of Accidents

The Pearson χ^2 indicates significance at the 5% confidence level. The t-ratios are significant for all the variables included in the model, and the scaled deviance value is smaller than the number of degrees of freedom. Figure 3.2 shows the relationship between the observed and predicted number of accidents for the model. The results are symmetrically clustered around the 45° line to a reasonable extent, which is desirable. In addition, Figure 3.3 shows the fit of the variance of the observed accidents (assuming a negative binomial distribution) to the average squared residuals. Each point represents the average of predicted accident frequency for a sequenced group of intersections (e.g. the first twenty intersections sorted by predicted accident frequency). The figure shows a reasonably good fit.

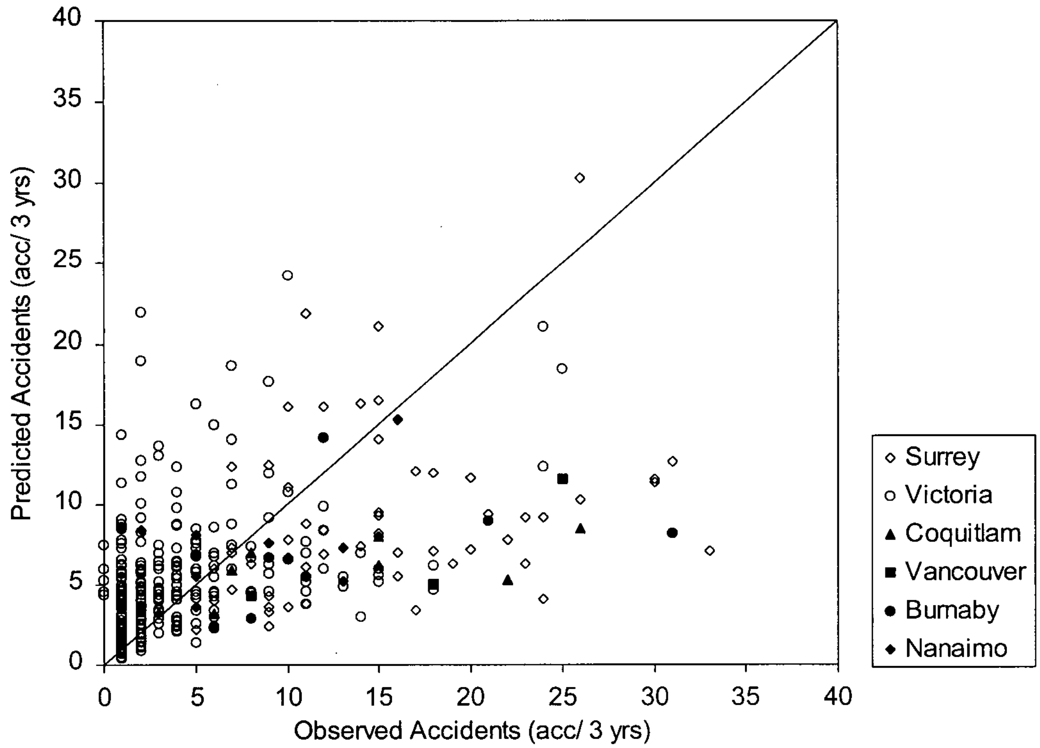


Figure 3.2 Total Model: Observed vs. Predicted Number of Accidents

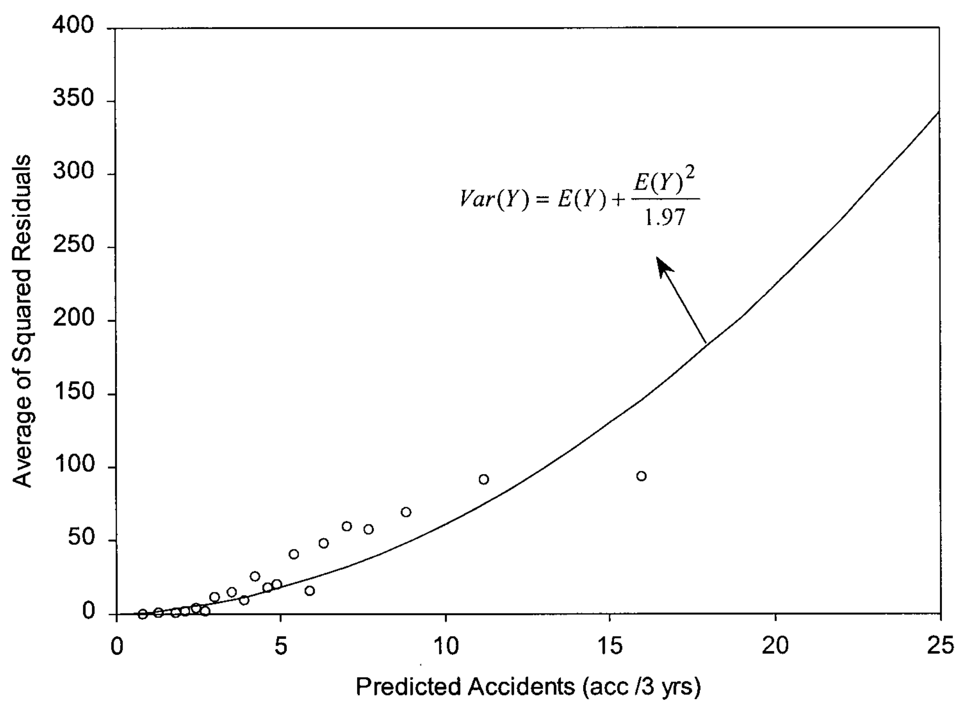


Figure 3.3 Total Model: Predicted Accidents vs Estimated Variance

Figure 3.2 also indicates that intersections in the Lower Mainland are generally different from those in Vancouver Island. Therefore, separate models for the Lower Mainland and Vancouver Island intersections should be developed.

3.2.2 Models for T and 4-leg Intersections

There are two ways to approach to these kinds of models. The first is to develop separate models for T and 4-leg intersections. Alternatively, one model can be developed using the entire sample size as the total model with the intersection type variable (T or 4-leg intersection) included within the model.

Using the first approach, the sample size for T intersections is 186, and for 4-leg intersections is 241. Table 3.5 shows the parameter estimates for each model, as well as the different goodness of fit test. Both models have a relatively good fit with respect to the scaled deviance, and the χ^2 values are significant at the 95% confidence level. The t-test ratios for all the independent variables are significant, which indicates that the models are more dependent on the explanatory variables rather than a constant coefficient, which is also desirable.

Model Form	t-ratio		SD (dof)	κ	Pearson χ^2 (χ^2 test)*
T-intersection model $Acc/3\,yrs = 0.9333 \times \left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.4531} \times \left(\frac{AADT_{min\,rd}}{1000}\right)^{0.5806}$	a_o a_1 a_2	-0.3 5.5 7.4	164 (183)	2.34	205 (214)
4-leg intersection model $Acc/3\,yrs = 1.6947 \times \left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.4099} \times \left(\frac{AADT_{min\,rd}}{1000}\right)^{0.7065}$	a_o a_1 a_2	3.6 6.8 9.1	230 (238)	2.17	251 (274)

* Denotes significance at a 95-percent confidence level

Table 3.5 Models for T and 4-leg Intersections

Figures 3.4 through 3.7 show the relationships between the observed and the predicted number of accidents, and the fit of the variance of the observed accidents to the average squared residuals, for both models. The results for both models are symmetrically clustered around the 45° line and the average squared residuals fits the variance equation well.

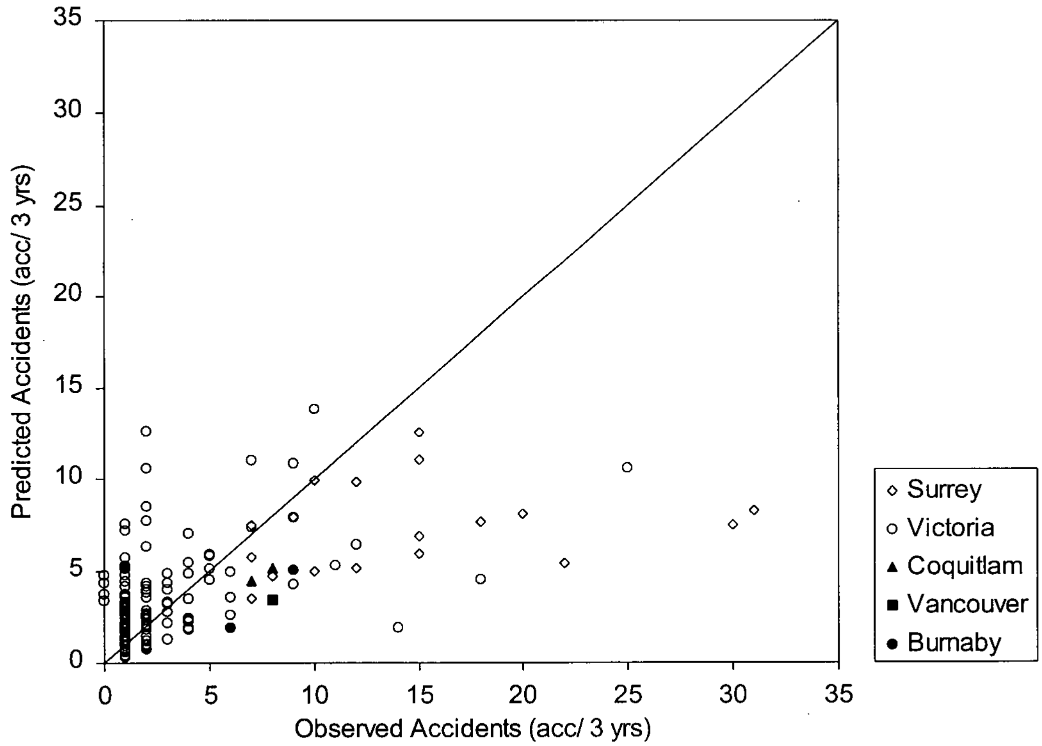


Figure 3.4 T-Intersection Model: Observed vs. Predicted Number of Accidents

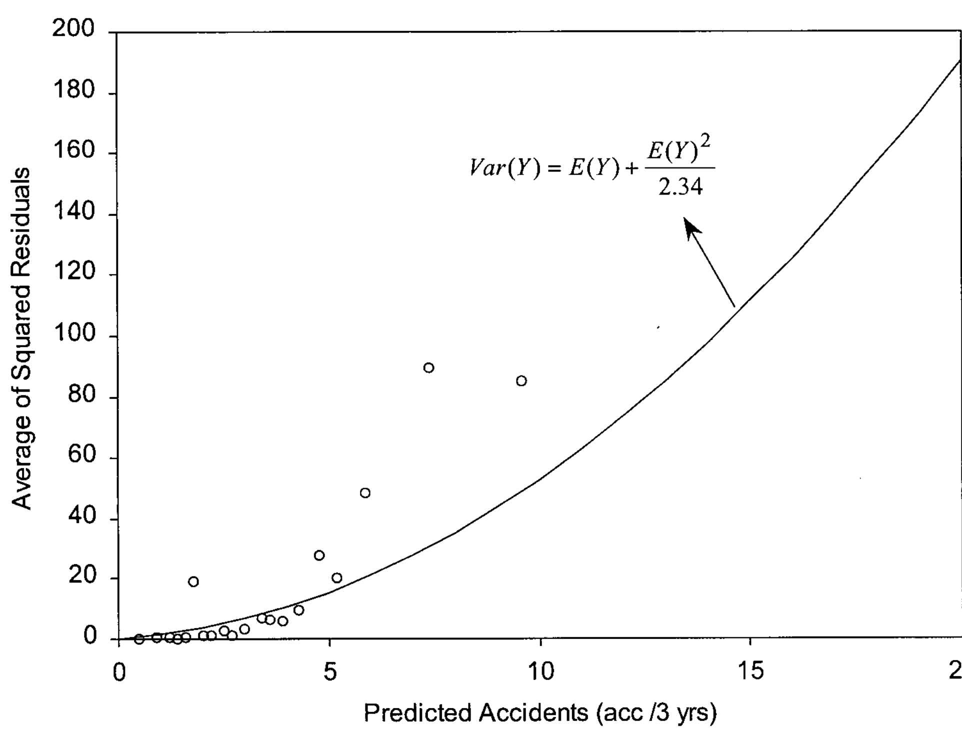


Figure 3.5 T-Intersection Model: Predicted Accidents vs Estimated Variance

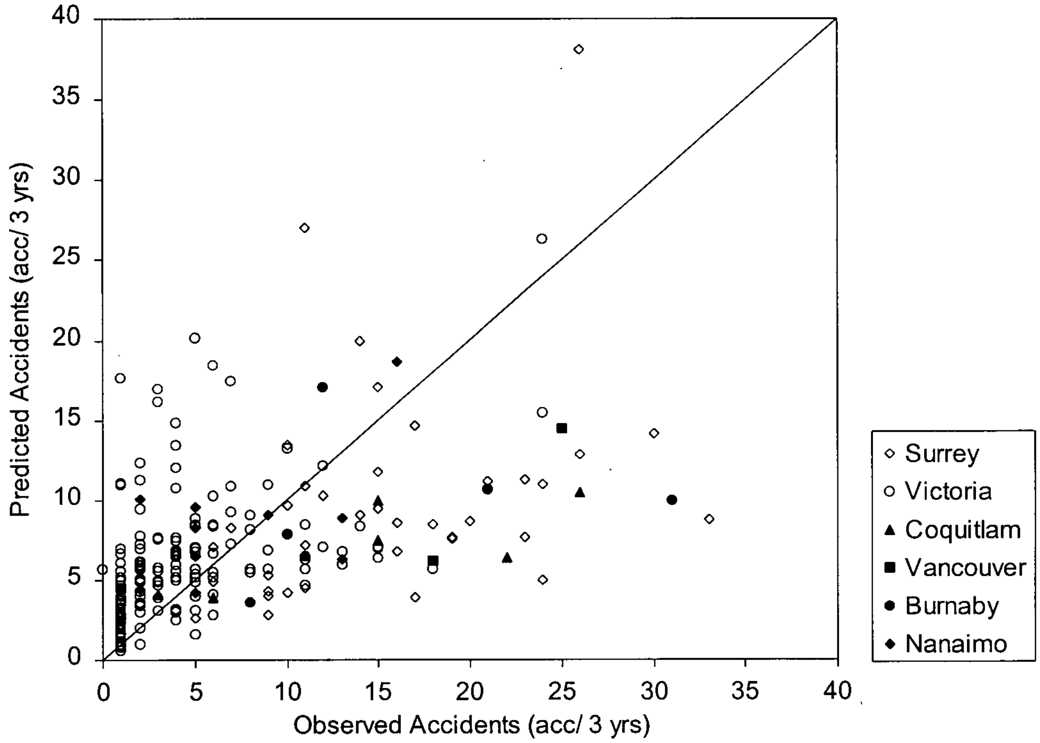


Figure 3.6 4-leg Intersection Model: Observed vs. Predicted Number of Accidents

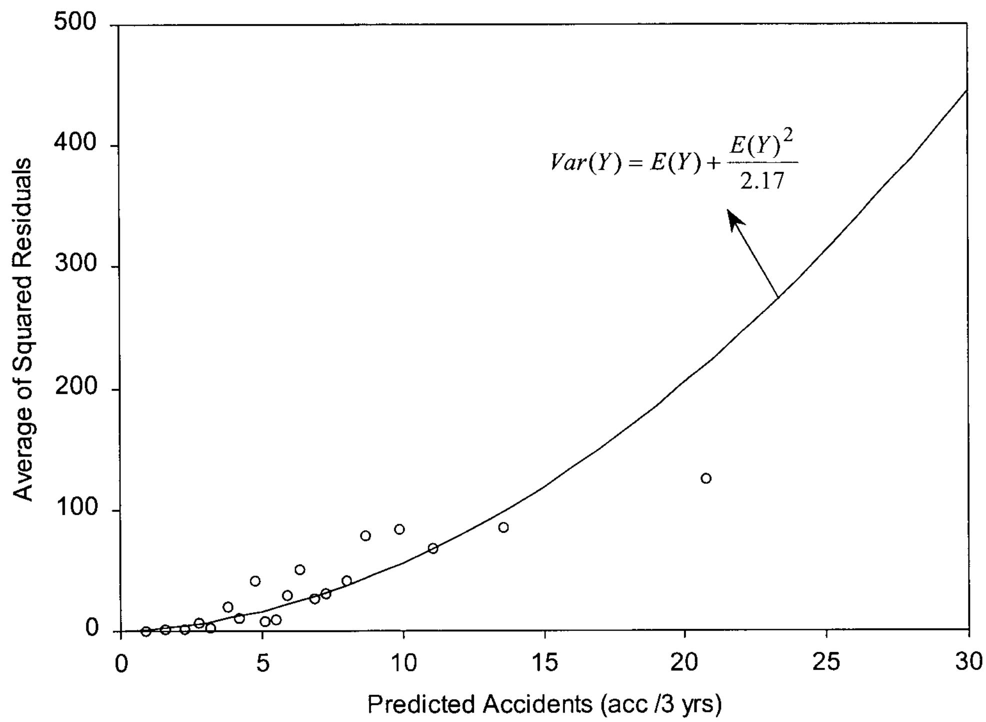


Figure 3.7 4-leg Intersection Model: Predicted Accidents vs Estimated Variance

Using the second approach a total model with the effect of intersection type model was developed by using only one equation that includes the effect of both T and 4-leg intersections. This model follows the same structure described in equation (2.14). Table 3.6 shows the estimates results of this model as well the goodness of fit test. The variable *Type*, which indicates the intersection type, has two values: 1 for T-intersections and 2 for 4-leg intersections. According to the results described in Table 3.6, all variables are significant in the model, the scaled deviance is also closed to the degrees of freedom, and the Pearson χ^2 test indicates a significance at the 95% confidence level.

Model Form	t-ratio		SD (dof)	κ	Pearson χ^2 (χ^2 test)*
Total Model with Intersection Type $Acc/3\,yrs = 0.5776 \times \left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.4221} \times \left(\frac{AADT_{min\,rd}}{1000}\right)^{0.6480} \times e^{0.5379 \times Type}$	a_o a_1 a_2 b_1	-2.8 8.7 11.7 6.3	394 (423)	2.23	449 (471)

* Denotes significance at a 95-percent confidence level

Table 3.6 Total Model Including the Effect of Intersection Type

A brief comparison between this model and the total model developed in section 3.2.1 shows a smaller scaled deviance and Pearson χ^2 for the first model. Note that decreasing the degrees of freedom by 1, lead to a drop in scaled deviance of 5, which is greater than 3.8 (the 95-percent value of the χ^2 square distribution with 1 degree of freedom). This indicates the importance of including in the model as many explanatory variables as possible in order to get a better fit.

Figures 3.8 and 3.9 show the relationships between the observed and the predicted number of accidents, and the fit of the observed accidents variance to the average squared residuals, respectively. The results in Figure 3.8 show that the points are closer to the 45° line than the results displayed in Figure 3.3 (Total Model). Figure 3.9 also shows the tendency of the model's average squared residual to follow the variance equation.

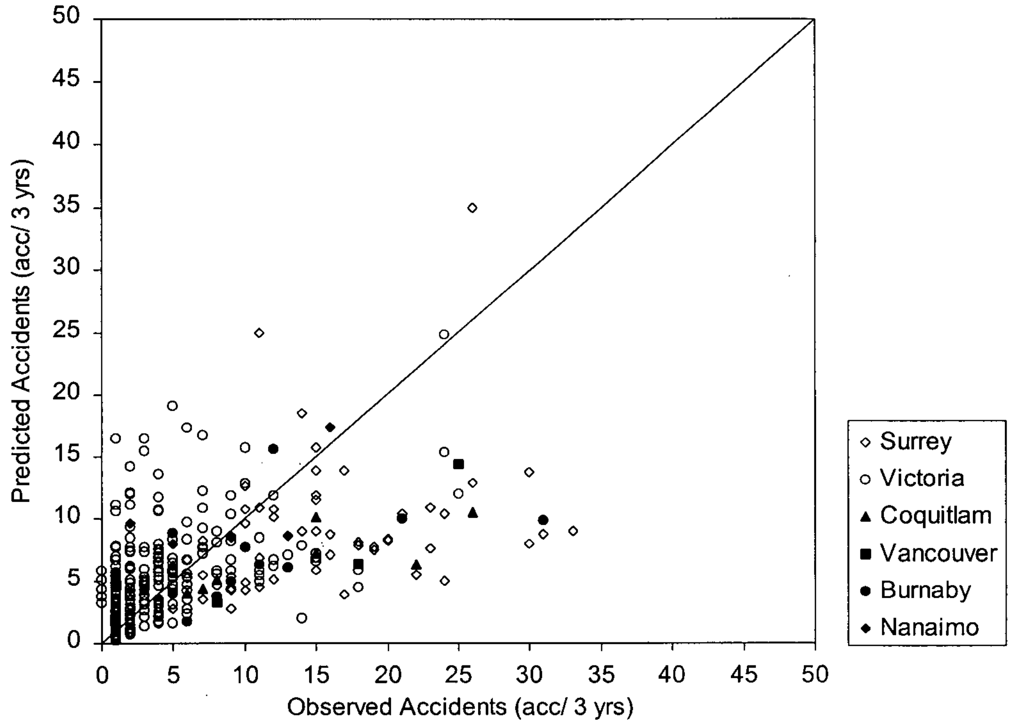


Figure 3.8 Total Model with the Effect of Intersection Type: Observed vs. Predicted Number of Accidents

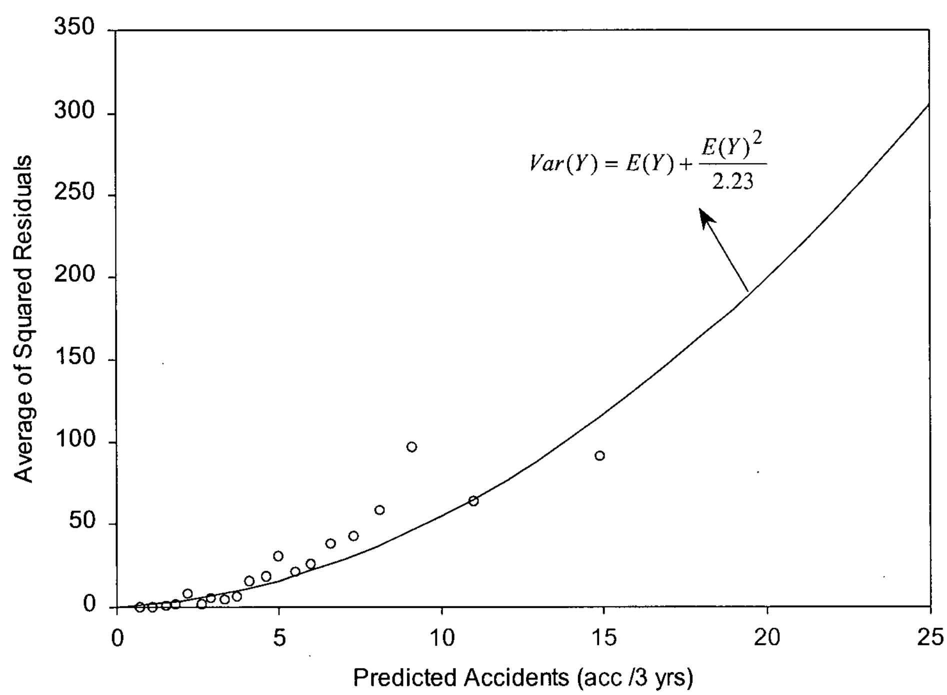


Figure 3.9 Total Model with the Effect of Intersection Type: Predicted Accidents vs Estimated Variance

A comparative analysis between the two approaches developed in this section for intersection type model is shown in Table 3.7. The analysis made use of the entire sample of this study (427 intersections), and the number of predicted accidents was calculated by using the separate models and the total model including the intersection type. The results obtained in this analysis shows that using the single model provides a slightly lower Pearson χ^2 test, and sum-of-squared error. This indicates that the total model with intersection type variables performs slightly better than using the two separate models. However the difference between both approaches is not significant, since the single model fits better only for 51% of the data.

Parameters	Separate Models	Single Model
Pearson χ^2	456	449
χ^2 test$_{0.05,\ n\text{-}p\text{-}1}$	472	471
Sum of Error^2	12,117	11,882
Closer Estimates	208 (49%)	219 (51%)

Table 3.7 Intersection Type Model: 2 Separate Models vs Single Model

Figure 3.10 shows the predicted accidents as a function of major road traffic volume, for both approaches. This figure shows that for 4-leg intersections, the separate model curve is slightly above the single model curve. For T-intersections, at low major road traffic volumes (AADT<12,000 veh/day), both curves are practically the same, but at high traffic volume the separate model curve is also slightly above the single model curve.

The results of the comparison show that using a single model seems to be slightly accurate than using separate models. The differences between these two approaches are relatively small.

Therefore, using either single or separate models will yield practically the same results, and both approaches are valid.

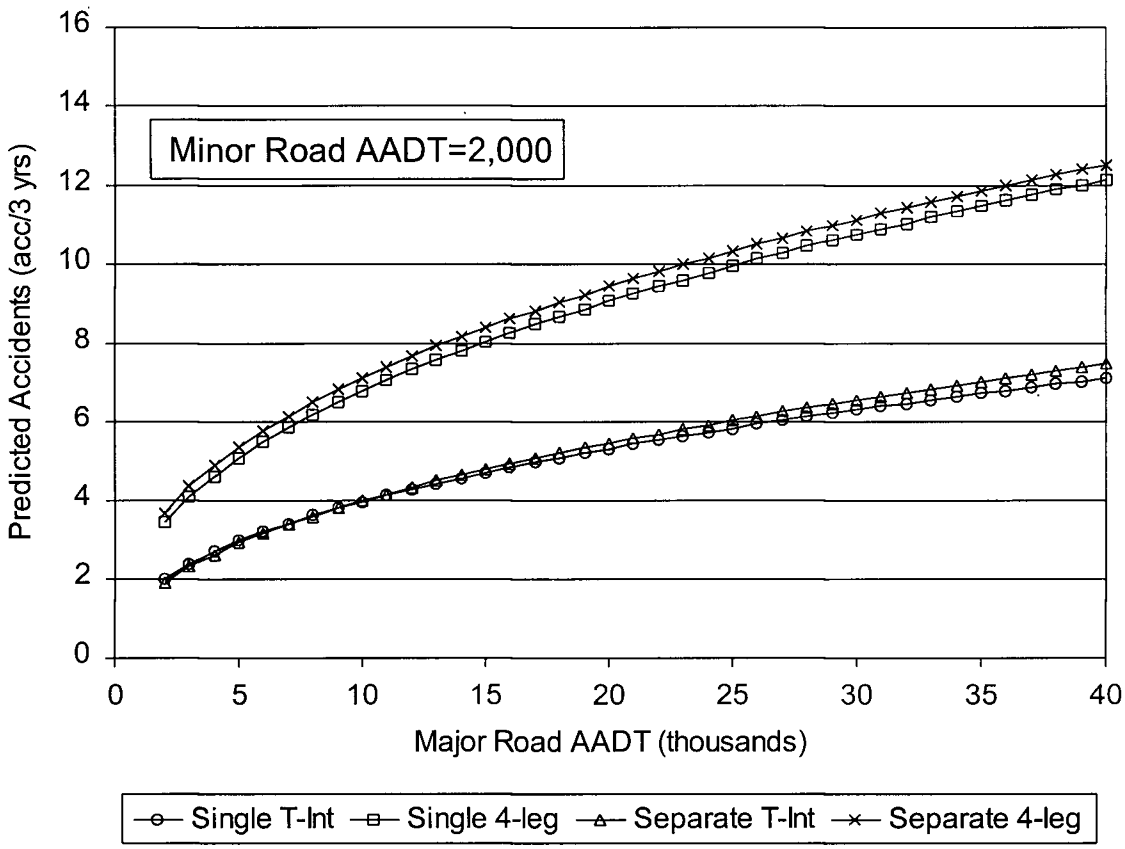


Figure 3.10 Intersection Type Model: Separate vs. Single Models

In addition, Figure 3.10 shows that T-intersections are approximately 50% safer than 4-leg intersections. A more detailed analysis regarding this point will be introduced in Chapter 5.

3.2.3 Regional Models

As described earlier, intersections in the Lower Mainland are generally different from those in Vancouver Island. Therefore, three regional models were developed: (1) a model for the Lower Mainland which comprises intersections located in the cities of Surrey, Coquitlam, Vancouver and Burnaby; (2) a model for the Vancouver Island which comprises intersections located in Victoria and Nanaimo; and (3) a model for Surrey. It was decided to develop a model for Surrey because it has the highest average number of accidents per intersection.

The sample size are 77, 350, and 56 for the Lower Mainland, Vancouver Island and Surrey models respectively. Table 3.8 shows the results of each model. For all models, the Pearson χ^2 values indicate significance at the 95% confidence level. The scaled deviance is also smaller than the degrees of freedom. The t-ratios are significant at 95% confidence level for all parameters, except for the major road traffic volume for the Surrey model, which is significant only at the 90% confidence level. Therefore, it is suggested that a larger sample size be used for the Surrey model or more explanatory variables should be added in order to obtain a more reliable model.

Model Form	t-ratio		SD (dof)	κ	Pearson χ^2 (χ^2 test)*
Vancouver Island model $Acc/3\,yrs = 1.3807\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.3042}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.5488}$	a_o a_1 a_2	2.6 6.0 9.0	302 (347)	2.92	383 (390)
Lower Mainland model $Acc/3\,yrs = 6.5929\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.2011}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.2864}$	a_o a_1 a_2	7.6 2.4 3.7	81 (74)	6.27	76 (94)
Surrey model $Acc/3\,yrs = 8.441\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.1516}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.1907}$	a_o a_1 a_2	8.7 1.8 2.3	56 (53)	8.89	58 (70)

* Denotes significance at a 95-percent confidence level

Table 3.8 Regional Models

Figures 3.11 through 3.16 show the relationships between the observed and the predicted number of accidents, and the fit of the variance of the observed accidents to the average squared residuals. The results for Vancouver Island and Mainland models (Figures 3.11 through 3.14) are symmetrically clustered around the 45° line and the average squared residuals follow the variance to a satisfactory extent. For Surrey model, the results shown in Figures 3.15 and 3.16, indicates a larger dispersion, which confirms the need either use a larger sample or add more explanatory variables to the model.

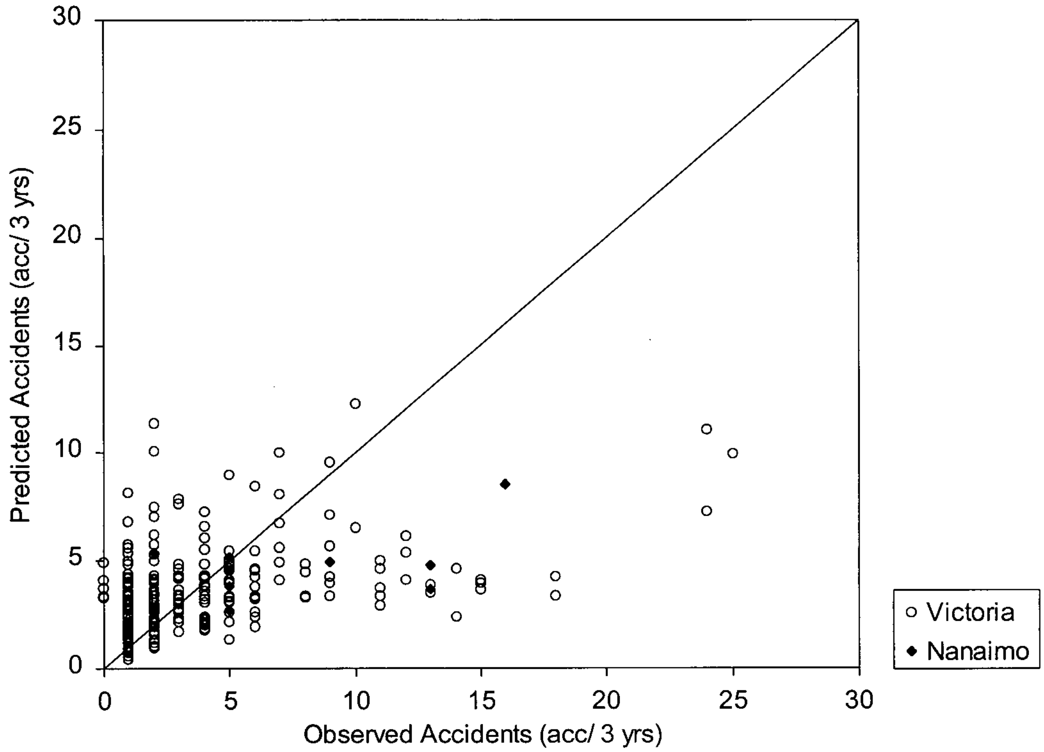


Figure 3.11 Vancouver Island Model: Observed vs Predicted Number of Accidents

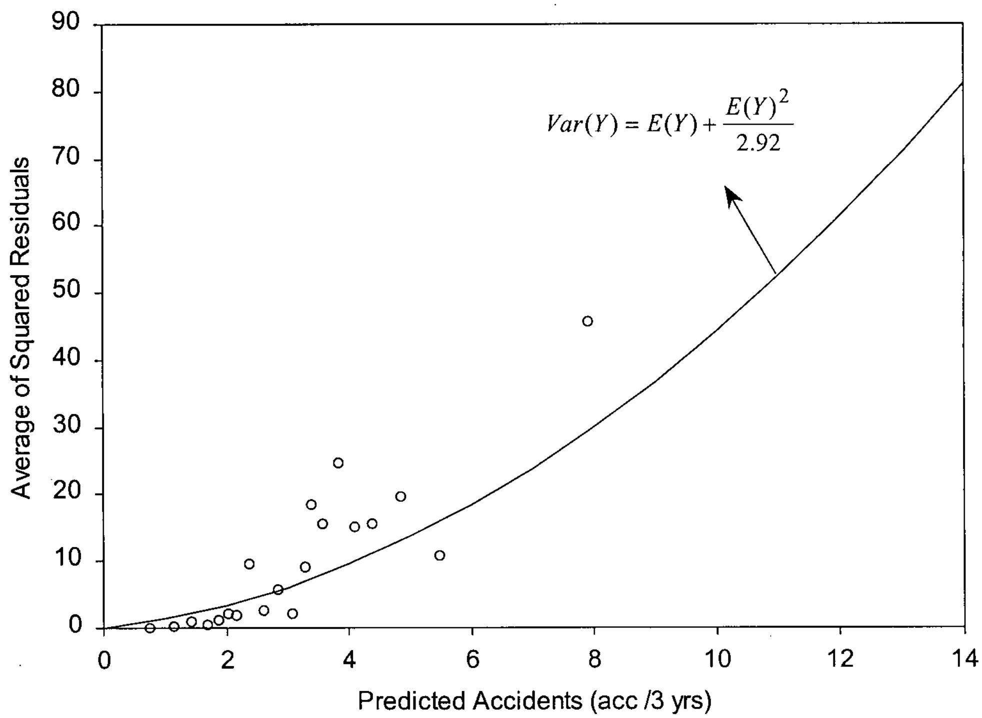


Figure 3.12 Vancouver Island Model: Predicted Accidents vs Estimated Variance

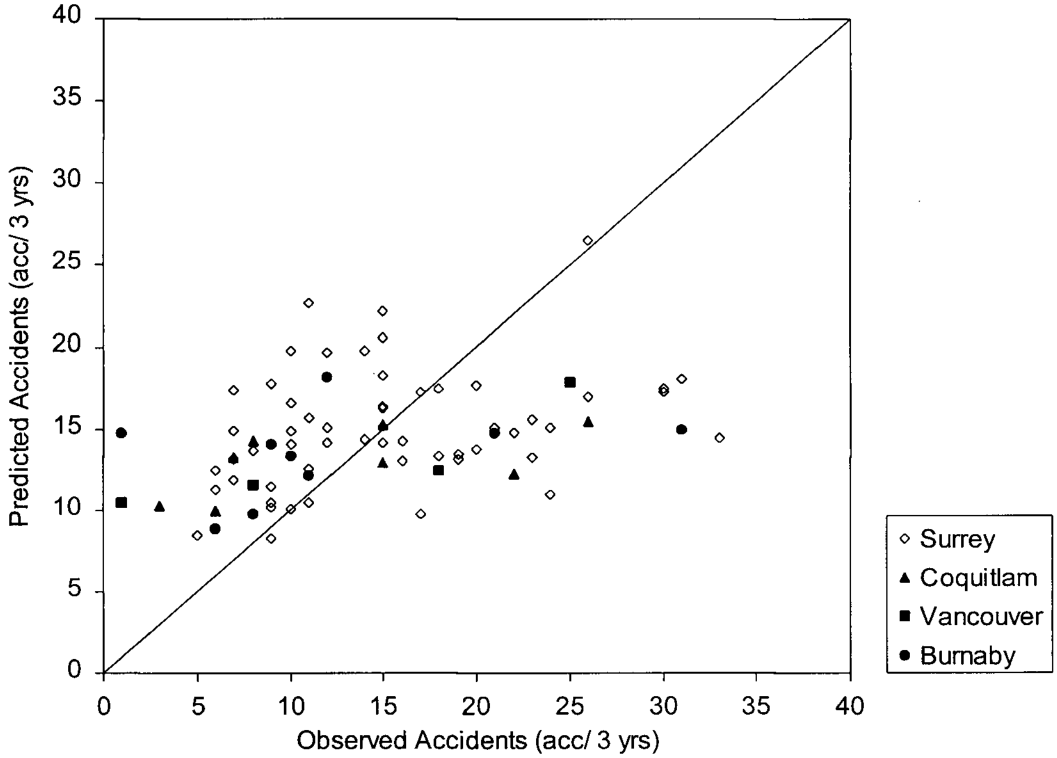


Figure 3.13 Lower Mainland Model: Observed vs Predicted Number of Accidents

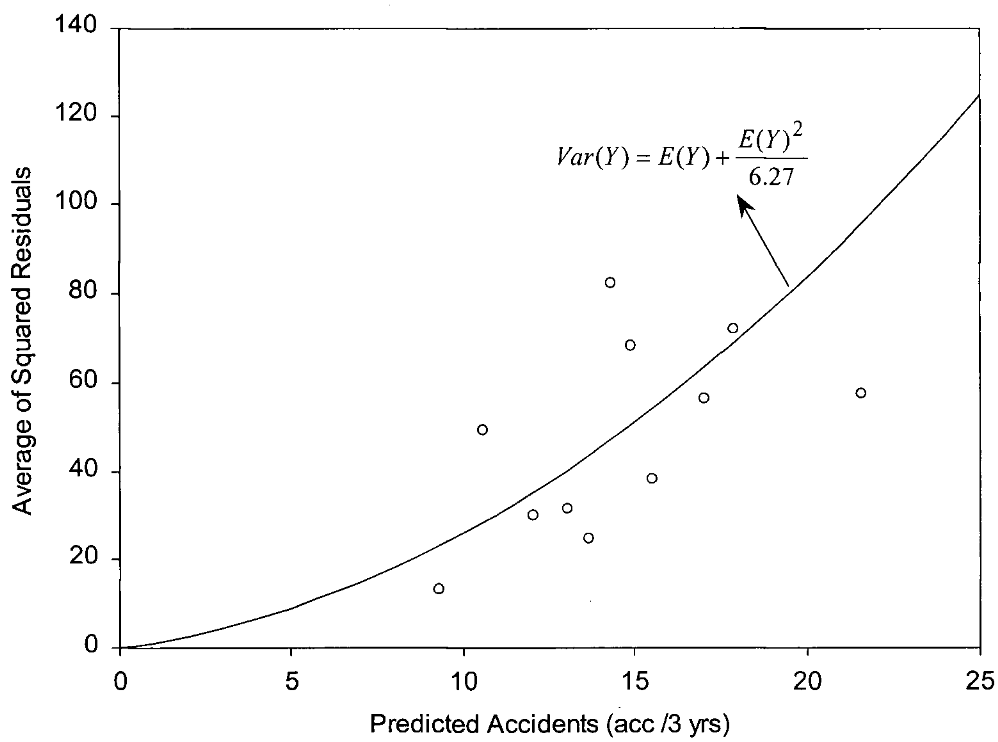


Figure 3.14 Lower Mainland Model: Predicted Accidents vs Estimated Variance

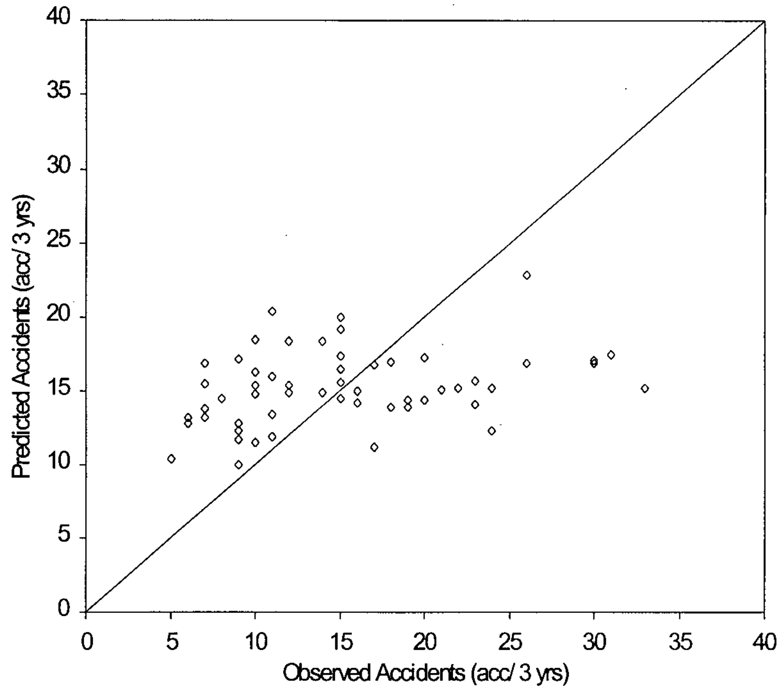


Figure 3.15 Surrey Total Model: Observed vs Predicted Number of Accidents

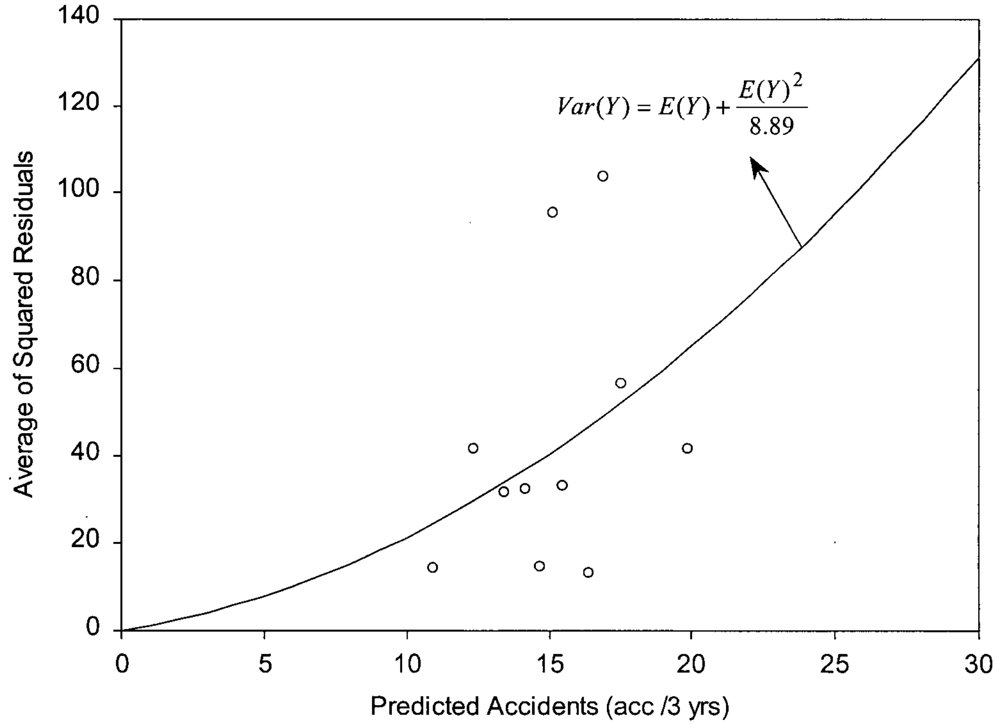


Figure 3.16 Surrey Total Model: Predicted Accidents vs Estimated Variance

Figure 3.17 shows a comparison of the total model estimated in section 3.2.1 with the three regional models. It should be noted that the total model lies between the Vancouver Island and the Lower Mainland models, which is expected. The total model curve is closer to the Vancouver Island model because more than 80% of the data comes from the cities of Victoria and Nanaimo.

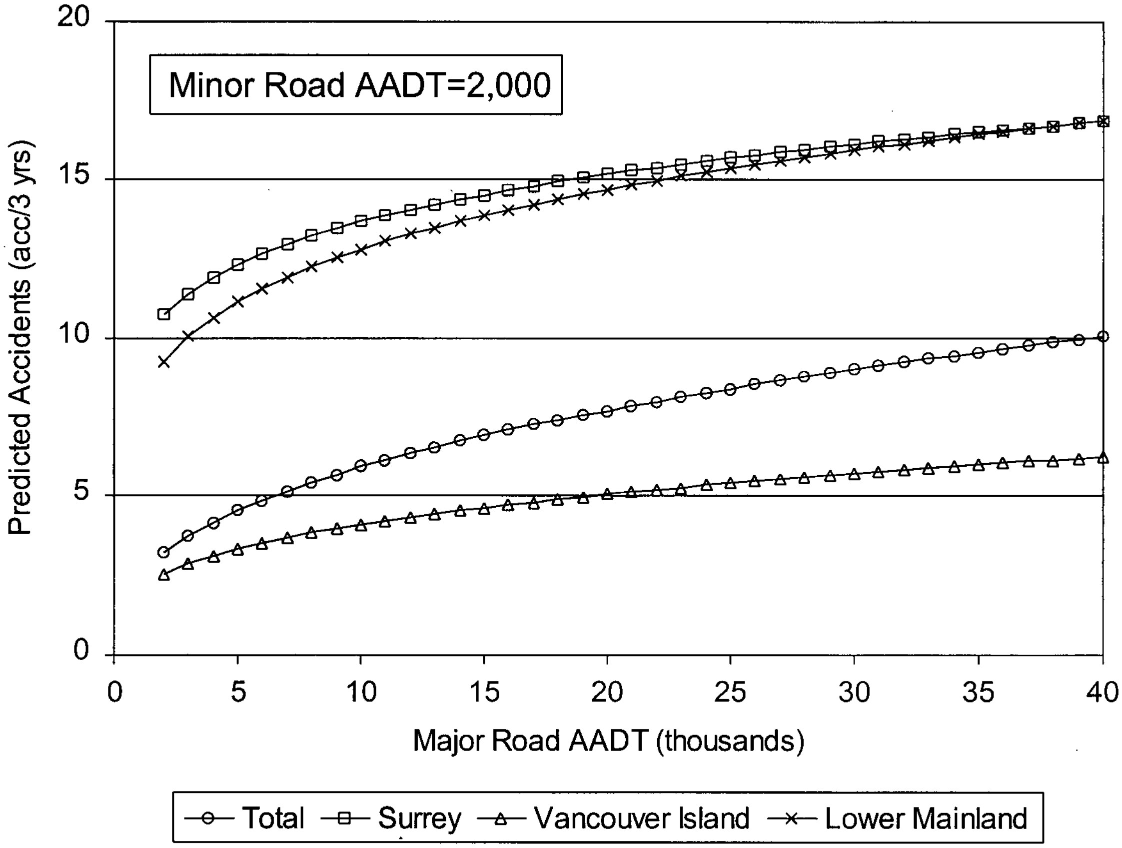


Figure 3.17 Comparison of Total Model with Regional Models

3.2.4 Effect of Intersection Control Type

Since data on intersection control type was only available for Surrey intersections, a Surrey total model including control type was estimated. As mentioned earlier, of the 56 intersections for Surrey data, 32 are classified as 2-way controlled, 8 as 4-way controlled and the remainders 16 as one-way stop controlled T intersections. The type variable in the equation is denoted by 1, 2 and 3 respectively for each control type.

Table 3.9 shows the results of this model. It can be noted that the a_0 parameter is much more significant than the three variables included in the model. This parameter also has a relatively high value. This is considered a deficiency since it indicates that the number of accidents is less dependent on traffic volumes and the control type. The t-ratio for the control type is not significant. As in the previous Surrey total model, it is therefore suggested that a larger size be used to develop this model.

Model Form	t-ratio		SD (dof)	κ	Pearson χ^2 (χ^2 test)*
$Acc/3\,yrs = 8.8906 \times \left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.1645} \times \left(\frac{AADT_{min\,rd}}{1000}\right)^{0.2256} \times e^{-0.06994 \times CONType}$	a_o	8.8	55 (52)	9.01	58 (69)
	a_1	2.0			
	a_2	2.6			
	b_1	-1.0			

* Denotes significance at a 95-percent confidence level

Table 3.9 Surrey Total Model with Control Type

Figures 3.18 and 3.19 show the relationship between the observed and the predicted number of accidents, and the fit of the variance of the observed accidents to the average squared residuals. In both figures the points are dispersed around the lines, which indicates a relatively poor fit.

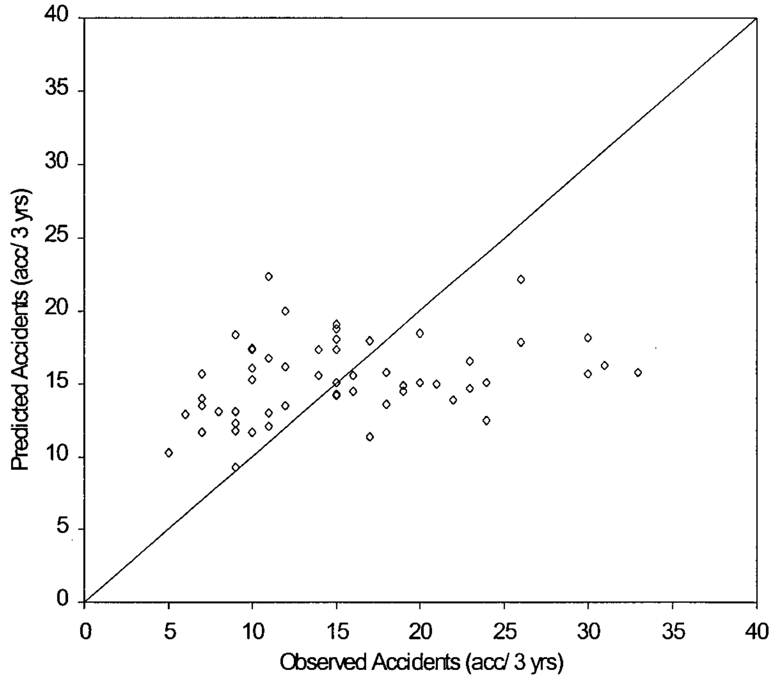


Figure 3.18 Surrey Total Model with Control Type: Observed vs Predicted Number of Accidents

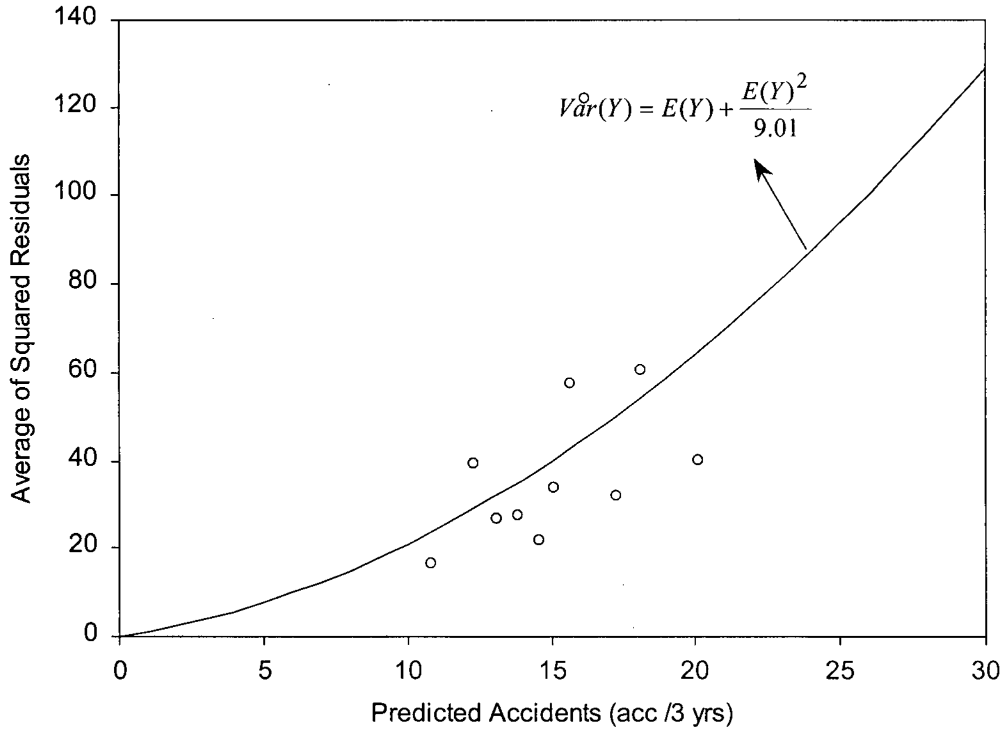


Figure 3.19 Surrey Total Model with Control Type: Predicted Accidents vs Estimated Variance

3.3 Comparison with Previous Results

As mentioned in Chapter 2, there are few studies which developed accident prediction models for urban unsignalized intersections. Therefore, this section will only compare the models developed herein to those developed by Maher and Summersgill (1996), and Mountain and Fawaz (1996). Since the dada base used to obtain these models comprised mainly of T-intersections, then the comparison was performed on the separate T-intersection model described in section 3.2.2.

Figure 3.20 shows the results of these three models for a constant minor traffic volume of 2,000 vehicles per day. The T-intersection model developed in this thesis has higher frequencies than the other two models. The difference in results may be attributed to the fact that Maher and Summmersgill' model included only T-intersections on urban single carriageways while Mountain and Fawaz's model include both urban and rural intersections. As well, there are differences in regional characteristics and the accident reporting practice between the UK and British Columbia (different reporting limit, police attendance, etc.)

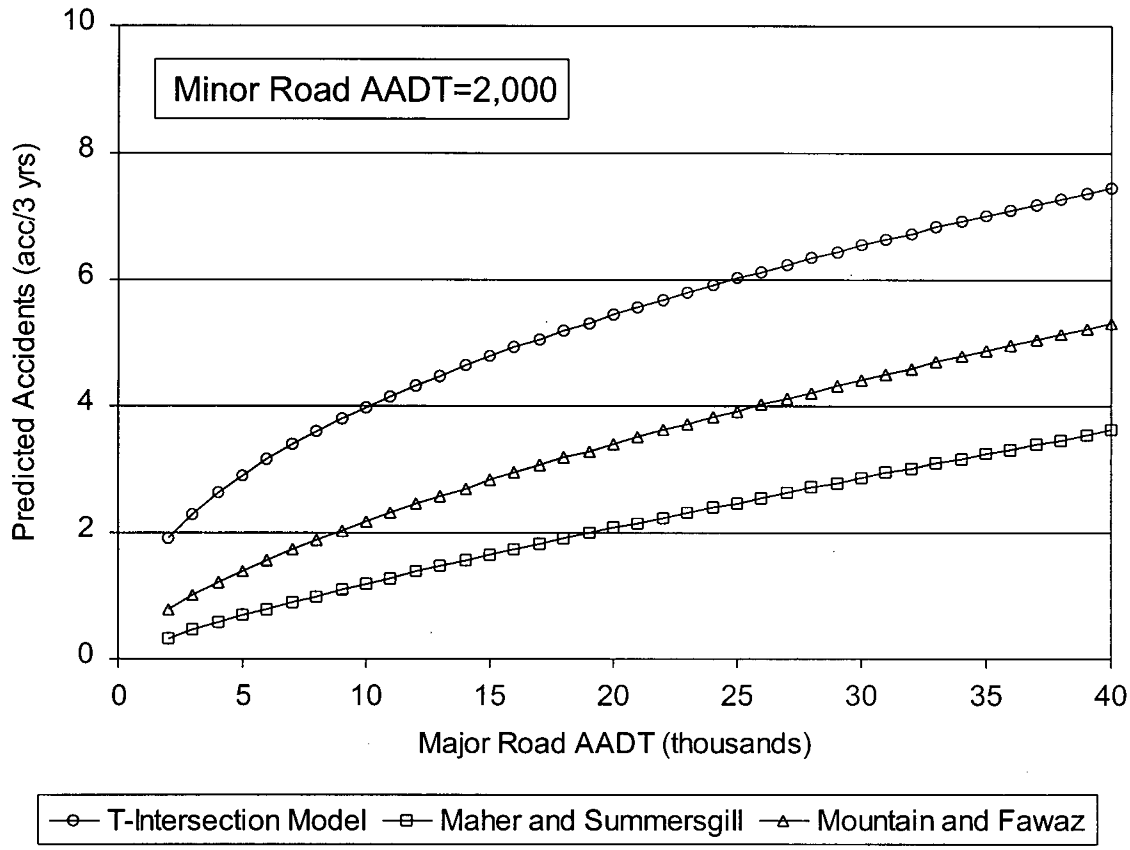


Figure 3.20 Comparison of T-Intersection model with Previous Studies

3.4 Conclusion

Using the negative binomial distribution approach eight different accident prediction models were developed. The first model developed included the entire data set and related accident frequency with traffic volumes for the major and minor roads. The rest of the models were classified according to certain characteristics such as intersection type (T and 4-leg intersections), regional characteristics, and intersection control type.

According to the various quality tests performed in this chapter, six out of the eight models showed a good statistical fit. The two models that showed poor fit, were characterized by having a lower sample size. It was suggested to increase the sample size or to include more explanatory variables into the models.

For the intersection type model, two different approaches were utilized: (1) by developing two separate models for each intersection type and; (2) by developing a single model that includes the intersection type as one of the variables. The differences between these two models were relatively small, and the effect of intersection type can be measured by using either approach.

Finally, a procedure to identify outliers in the data set was performed according to the Cook's distance values. The procedure indicated that there were no outliers in the data.

CHAPTER IV

STATISTICAL CONSIDERATIONS

4.0 Introduction

In this chapter several statistical issues will be discussed. The first issue relates to the error structure distribution. As described earlier, for the GLIM approach, the error structure is usually assumed to be Poisson or negative binomial. A comparison will be made between the two error structure distributions. The second issue relates to the method of calculating the parameter κ of the negative binomial distribution. A comparison of several approaches to calculate κ will be presented.

4.1 Poisson vs. Negative Binomial Distribution Error Structure

As mentioned in Chapter Two, dispersion parameters (σ_d, defined in equation 2.7) can be used to decide whether to use the Poisson or the negative binomial distribution error structure. If the dispersion parameter in the Poisson distribution model is greater than one, then going for the negative binomial distribution is recommended.

The Poisson distribution was used as a first step to develop all eight models discussed in Chapter 3. Appendix 2 shows the GLIM session results of the Poisson distribution. Table 4.1 shows a comparative analysis between these two approaches. Note that the dispersion parameter for the Poisson distribution is considerable high for all models, ranging from 2.57 for the Vancouver Island model, to 4.38 for the 4-leg intersection model. These high values are explained, by the lack of significance of the *Pearson* χ^2 tests. This indicates that for all models the data has greater dispersion than can be explained by the Poisson distribution, and it is necessary to assume a negative binomial distribution error structure. Under the latter distribution, σ_d ranges from 1.02 for the Lower Mainland model, to 1.12 for T-intersection and Surrey control type models. This indicates that the data dispersion is satisfactorily explained by the negative binomial distribution.

PARAMETERS	Poisson	Neg bin	Poisson	Neg bin	Poisson	Neg bin	Poisson	Neg bin
	Total Model		Total Intersection Type		T-Intersection		4-Leg Intersection	
a_o	1.4833	1.4929	0.5906	0.5776	0.6717	0.9333	1.9007	1.6947
a_1	0.4067	0.3839	0.4336	0.4221	0.5809	0.4531	0.3884	0.4099
a_2	0.6086	0.7044	0.5937	0.6480	0.5902	0.5856	0.5944	0.7065
b_1			0.5268	0.5379				
Dispersion Parameter, σ_d	4.30	1.08	3.86	1.06	3.26	1.12	4.38	1.06
κ		1.97		2.2		2.35		2.17
Scaled Deviance	1577	399	1436	394	485	164	942	230
Deg. of Freedom	424	424	423	423	183	183	238	238
Pearson χ^2	1823	459	1634	449	597	205	1042	251
χ^2(95%)	472	472	471	471	214	214	274	274
Error^2	12494	12996	11687	11883	3267	3347	8272	8770
Closer Estimates	44%	56%	44%	56%	52%	48%	43%	57%
	Vancouver Island		Lower Mainland		Surrey Total		Surrey Control Type	
a_o	1.3327	1.3807	6.7666	6.5929	8.5677	8.4401	8.9442	8.8906
a_1	0.3231	0.3042	0.2036	0.2011	0.1529	0.1516	0.1647	0.1645
a_2	0.5240	0.5488	0.2474	0.2864	0.1720	0.1907	0.1958	0.2256
b_1							-0.0570	-0.0699
Dispersion Parameter, σ_d	2.57	1.10	3.35	1.02	2.97	1.10	3.00	1.12
κ		2.92		6.27		8.89		9.10
Scaled Deviance	734	302	250	81	152	56	150	55
Deg. of Freedom	347	347	74	74	53	53	52	52
Pearson χ^2	893	383	248	76	158	58	156	58
χ^2(95%)	390	390	94	94	70	70	69	69
Error^2	3879	3892	3633	3681	2431	2440	2417	2438
Closer Estimates	42%	58%	44%	56%	41%	59%	52%	48%

Table 4.1 Comparison between Poisson and Negative Binomial Distribution

In addition to the dispersion parameter, Table 4.1 also shows other parameters to compare both model approaches such as the scaled deviance, *Pearson* χ^2, error squared and the share of the predicted accidents closer to the observed accidents.

The scaled deviance in the Poisson distribution is considerably greater for all models and exceeds the number of degrees of freedom from 112% for the Vancouver Island model, to 296% for 4-leg model. In contrast, for the negative binomial distribution models, the scaled deviance is relatively close to the degrees of freedom, which indicates a reasonably good fit.

Regarding the other comparative tests such as the sum of error squared and the closer predicted values, Table 4.1 shows that the sum of error squared is slightly smaller in the Poisson models than in the negative binomial models. However, for this latter assumption, there are more predicted values closer to the observed data. This indicates that, while most of the data fits the negative binomial distribution model better, the estimates that fit the Poisson distribution better have higher differences with the observed values when using negative binomial distribution models.

4.2 Approaches for Estimating the Negative Binomial Distribution Parameter κ

There are several approaches to estimate the parameter κ of the negative binomial distribution error (Famoye, 1997). The macro library of the GLIM software package contains three methods: maximum likelihood and two methods of moments called mean χ^2 and mean deviance. In addition Kulmala (1995), following Maycock and Hall (1984), proposed a method of moments, in which the parameter κ is initially calculated from the estimates obtained from the Poisson distribution model. All these methods are iterative.

The method of maximum likelihood has been the most widely used (Hauer et-al, 1988, Bonneson and McCoy, 1993, Maher and Summersgill, 1996). According to Lawless (1987) this method is based on the log-likelihood function, which is the natural logarithm of the joint probability function of the negative binomial distribution (equation 2.5). This is a function of μ and κ, where μ is also a function of the parameter estimates, a_j. The iterative process is aimed at maximizing the log-likelihood function with respect to the parameter estimates, a_j, for selected values of κ. The iterative process continues until the maximum value of κ has been reached.

The mean χ^2 method consists of fitting the Pearson χ^2 value to the number of degrees of freedom. As a first iteration, κ is solved from the Pearson χ^2 equation, and the initial estimates are calculated by using the Poisson distribution. Having an initial value of κ, the new parameters are estimated. Then the process is repeated until convergence.

The mean deviance method is similar to the mean χ^2 with the main difference being that the scaled deviance is forced to equal the number of degrees of freedom.

The method of moments proposed by Kulmala (1995) and Maycock and Hall (1984) consists of estimating a first value of κ, based on the following equation:

$$\kappa \approx \frac{\sum_{i=1}^{n} E(\Lambda)_i^2}{\sum_{i=1}^{n}(error_i^2 - E(\Lambda)_i)} \tag{4.1}$$

where the predicted values $E(\Lambda)_i$ are initially estimated based on the Poisson distribution model. Then, the κ value is the run to a GLIM macro to estimate the parameters under the negative binomial distribution. As the previous methods of moments, the process is repeated until convergence. According to Kulmala (1995) the estimates obtained in this method deviate less than 5% from those produced by the maximum likelihood method.

The previous four methods were used to estimate the parameter κ. The results obtained in the GLIM sessions are shown in Appendix IV. Table 4.2 summarizes the results obtained for each method. The table shows that the parameters' values are equal up to the first two decimal points for all methods, and the t-ratios show that for all cases the variables are significant. This shows a relative similarity between the four methods.

PARAMETER	MAXIMUM LIKELIHOOD		MEAN χ^2		MEAN DEVIANCE		MOMENTS (KULMALA)	
	Value	t-ratio	Value	t-ratio	Value	t-ratio	Value	t-ratio
a_o	1.4929	3.2	1.4963	3.1	1.4905	3.3	1.4947	3.4
a_1	0.3839	7.8	0.3827	7.5	0.3850	8.0	0.3832	8.0
a_2	0.7044	12.4	0.7058	12.0	0.7023	12.8	0.7054	12.8
κ	1.97		1.76		2.15		1.85	
Scaled Deviance	399		370		424		382	
Deg. of Freedom	424		424		424		424	
Pearson χ^2	459		424		489		439	
χ^2(95%)	472		472		472		472	
Error^2	12996		13006		12981		13003	

Table 4.2 Results of Different Negative Binomial Methods in the Total Model

With Regard to the parameter κ, there are more differences than the model's parameters. The highest κ value is obtained through the method of mean deviance, followed by the method of maximum likelihood. According to the criteria of maximizing κ, which reduces the variance, the best method would be the mean deviance, while the worst would be the mean χ^2 method. However, by analyzing the Pearson χ^2 statistic, the method with the highest κ is not significant at the 95% confidence level. Therefore, the best method would be the maximum likelihood, which has the second highest κ parameter.

Table 4.2 also shows that, the scaled deviance value for all methods is significant compared with the degrees of freedom. The Pearson χ^2 statistic is significant at 95% of confidence level for all models, except the mean deviance model, and the sum of error squared show that the estimates are quite similar, which is a result of the similarity in the model's parameters.

Figure 4.1 shows the predicted accidents of the total model as a function of major road traffic volume for the different methods. This Figure shows all curves turning into one curve, which confirms the similarities, found in Table 4.1.

This analysis shows that with exception of the mean deviance method, which was not significant according to the Pearson χ^2 statistic, the other three methods yield approximately the same results. However, out of the three significant methods, the maximum likelihood yields the highest κ value, and for this reason it is regarded as the most appropriate method.

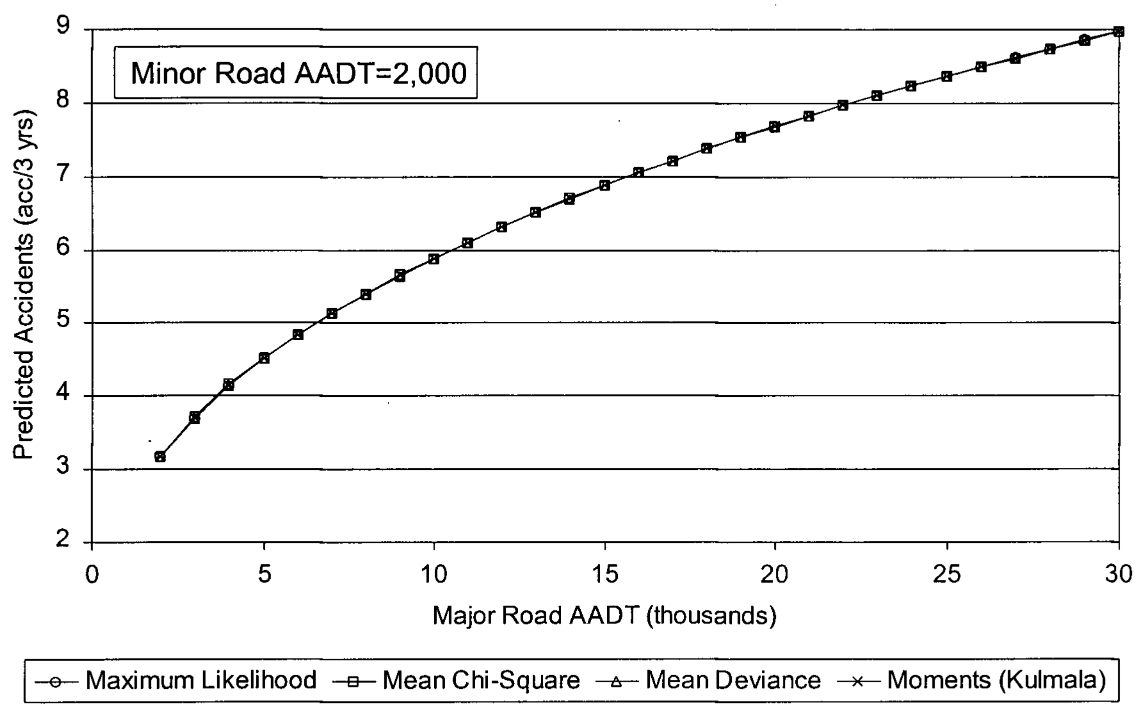


Figure 4.1 Predicted Accidents using Different Methods to Obtain κ

4.3 Conclusion

This chapter was intended to demonstrate the advantages of the methodology used in chapter three to derive the accident prediction models. First, it was demonstrated in a comparative fashion that the accident prediction models for urban unsignalized intersections follow the negative binomial distribution rather than the Poisson distribution.

Next, it was also demonstrated that the maximum likelihood method is the most appropriate to calculate the negative binomial model's parameter because it yields the maximum value of κ for significant models. However, it was found that the methods of the mean χ^2 and the method of moments proposed by Kulmala, yielded significant results which were similar to the maximum likelihood method.

CHAPTER V

APPLICATIONS

5.0 Introduction

As described earlier, there are several applications of accident prediction models. This chapter describes five different applications. The first four applications relate to the use of the Empirical Bayes refinement: identification of accident-prone locations, developing critical accident frequency curves, ranking the identified accident-prone locations and before and after safety evaluation. The fifth application provides a safety-planning example, comparing the safety performance of a 4-leg intersection and two staggered T-intersections for the same traffic volume.

Empirical Bayes refinement applications are demonstrated using the model relating the total number of accidents to traffic flows (Table 3.4) because it is the most general model. The Vancouver Island and the Lower Mainland models are also used in the identification and ranking of accident-prone locations.

5.1 Empirical Bayes Refinement

As mentioned in Section 2.5, the main goal of using the Empirical Bayes refinement is to yield more accurate, location-specific safety estimate by combining the observed number of accidents at the location, with the predicted number of accidents obtained from the GLIM model.

To illustrate this process, assume that an unsignalized intersection has the following data:

Major road ADT = 15,000 veh/day

Minor road ADT = 2,000 veh/day

Observed accidents = 11 acc/3 years

Using the model from Table 3.4, the safety of this intersection is:

$$pred = 1.4929 \times \left(\frac{15{,}000}{1{,}000}\right)^{0.3839} \times \left(\frac{2{,}000}{1{,}000}\right)^{0.7044} = 6.88\, acc/3\ years$$

Using equations 2.17 and 2.18 the empirical safety estimate and its variance respectively, can be calculated as:

$$EB_{safety\ estimate} = \left(\frac{1.97}{1.97 + 6.88}\right) \times 6.88 + \left(\frac{6.88}{1.97 + 6.88}\right) \times 11 = 10.08\, acc/3\ years$$

$$Var(EB_{safety\ estimate}) = \left(\frac{6.88}{6.88 + 1.97}\right)^2 \times 1.97 + \left(\frac{6.88}{6.88 + 1.97}\right)^2 \times 11 = 7.84\, (acc/3\ years)^2$$

In this example the expected number of accidents is reduced from 11 to 10.08 which corresponds to about eight percent regression to the mean correction.

Figure 5.1 illustrates the Empirical Bayes refinement estimation versus the values predicted from the GLIM model. Notice that the EB estimates are much closer to the 45° line, indicating an average regression to the mean correction of 35% although for some extreme cases, the corresponding correction is up to 150%.

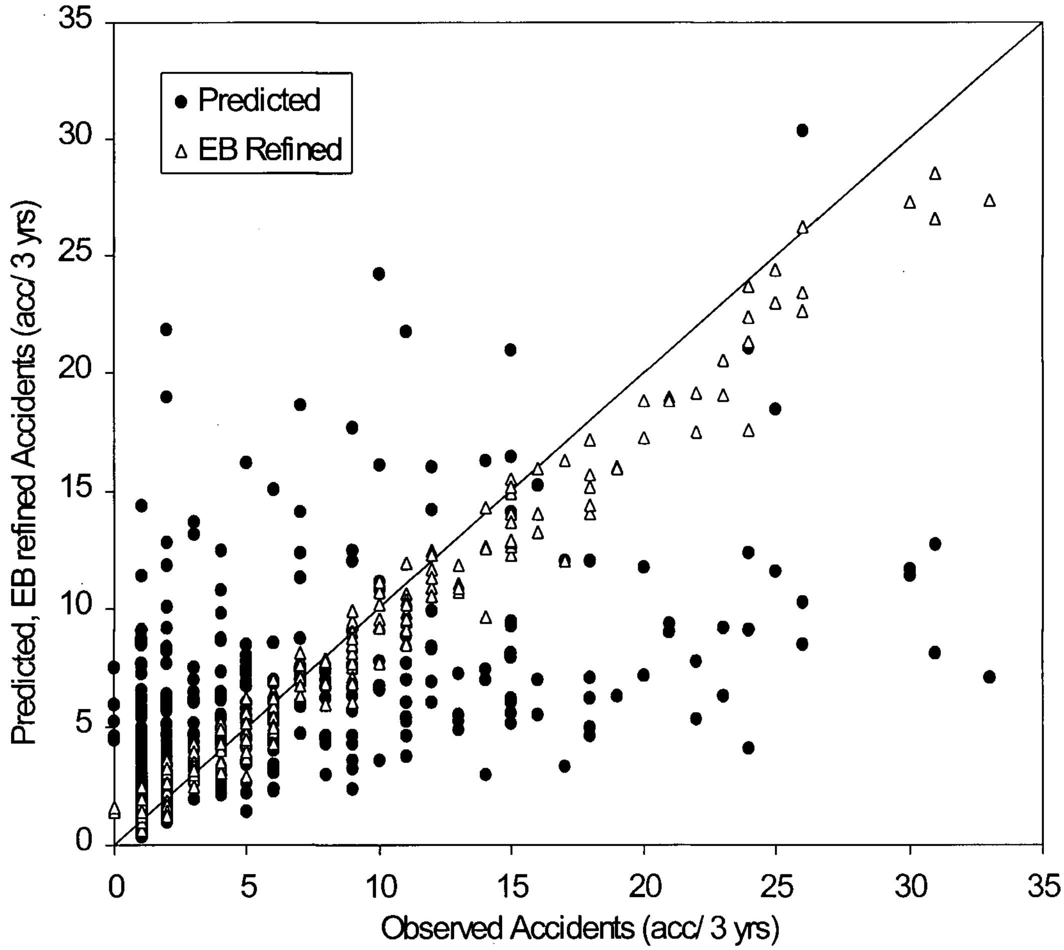


Figure 5.1 Predicted vs. EB Refined Number of Accidents for Total Model

5.2 Identification of Accident Prone Locations

Accident prone locations (APLs) are defined as the locations that exhibit a significant number of accidents compared to a specific norm. Because of the randomness inherent in accident occurrence, statistical techniques that account for this randomness should be used when identifying APLs. The EB refinement method can be used to identify APLs according to the following process (Bélanger, 1994):

1. Estimate the predicted number of accidents and its variance for the intersection, using the appropriate GLIM model. This follows a gamma distribution (the prior distribution) with parameters α_1 and β_1, where:

$$\beta_1 = \frac{E(\Lambda)}{Var(\Lambda)} = \frac{\kappa}{E(\Lambda)} \quad and \quad \alpha_1 = \beta_1 \cdot E(\Lambda) = \kappa \tag{5.1}$$

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

$$\int_0^{P_{50}} \frac{(\kappa / E(\Lambda))^{\kappa} \cdot \lambda^{\kappa-1} \cdot e^{-(\kappa / E(\Lambda))\lambda}}{\Gamma(\kappa)} d\lambda = 0.5 \tag{5.2}$$

3. Calculate the EB safety estimate and its variance from equations (2.17) and (2.18) respectively. This is also a gamma distribution (posterior distribution) with parameters α_2 and β_2:

$$\beta_2 = \frac{EB}{Var(EB)} = \frac{\kappa}{E(\Lambda)} + 1 \text{ and } \alpha_2 = \beta_2 \cdot EB = \kappa + count \tag{5.3}$$

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

$$f_{EB}(\lambda) = \frac{(\kappa / E(\Lambda) + 1)^{(\kappa + count)} \lambda^{\kappa + count - 1} e^{-(\kappa / E(\Lambda) + 1)\lambda}}{\Gamma(\kappa + count)} \tag{5.4}$$

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

$$\left[1 - \int_0^{P_{50}} \frac{(\kappa / E(\Lambda) + 1)^{(\kappa + count)} \lambda^{\kappa + count - 1} e^{-(\kappa / E(\Lambda) + 1)\lambda}}{\Gamma(\kappa + count)} d\lambda \right] \geq \delta \tag{5.5}$$

where δ represents the confidence level desired (usually 0.95)

For the example given in the previous section, the predicted number of accidents and its variance is 6.88 acc/3yr and 24.67 (acc/3yr)^2 respectively. Then using equation 5.2 to obtain the P_{50} value:

$$\int_{0}^{P_{50}} \frac{(1.97/6.88)^{1.97} \cdot \lambda^{1.97-1} \cdot e^{-(1.97/6.88)\lambda}}{\Gamma(1.97)} d\lambda = 0.5$$

solving the integral for 0.5, the P_{50} value is 5.75 acc/3yr.

From the pervious section, the EB estimate and its variance is 10.08 acc/3yr and 7.84 $(acc/3yr)^2$ respectively. Using equation 5.5 the left-hand side of the equation is:

$$1 - \int_{0}^{5.75} \frac{(1.97/6.88+1)^{(1.97+11)} \lambda^{1.97+11-1} e^{-(1.97/6.88+1)\lambda}}{\Gamma(1.97+11)} d\lambda = 0.96$$

This indicates that there is a significant probability (96%) of exceeding the P_{50} value and the intersection can be considered accident-prone. Figure 5.2 shows a graphical representation of this example:

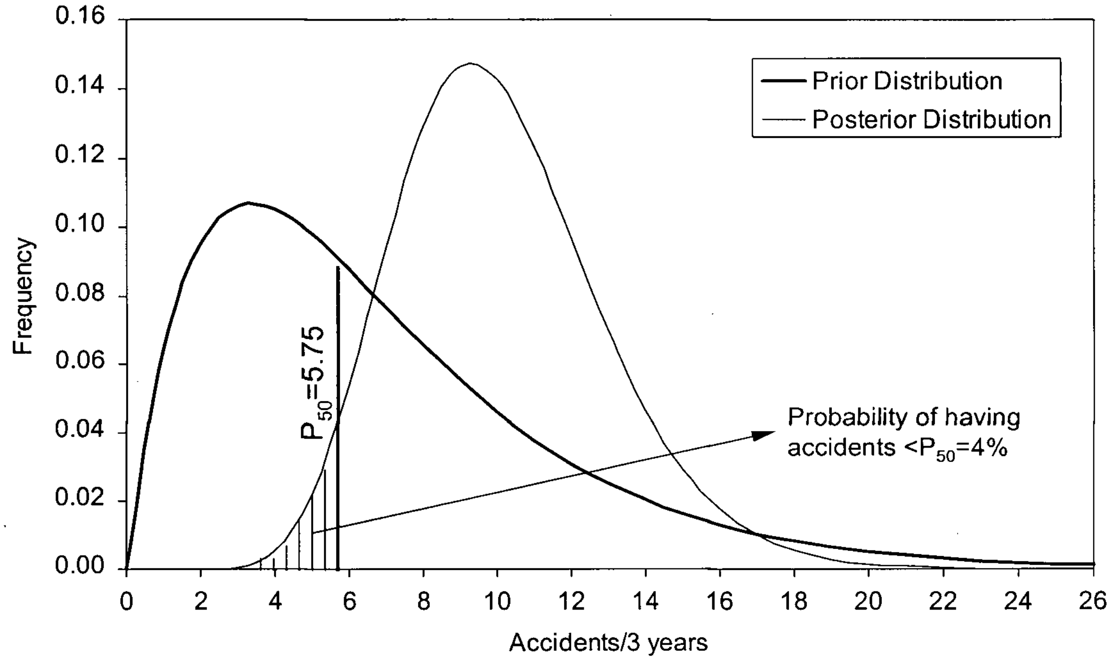


Figure 5.2 Identification of Accident Prone Locations

5.3 Critical Accident Frequency Curves

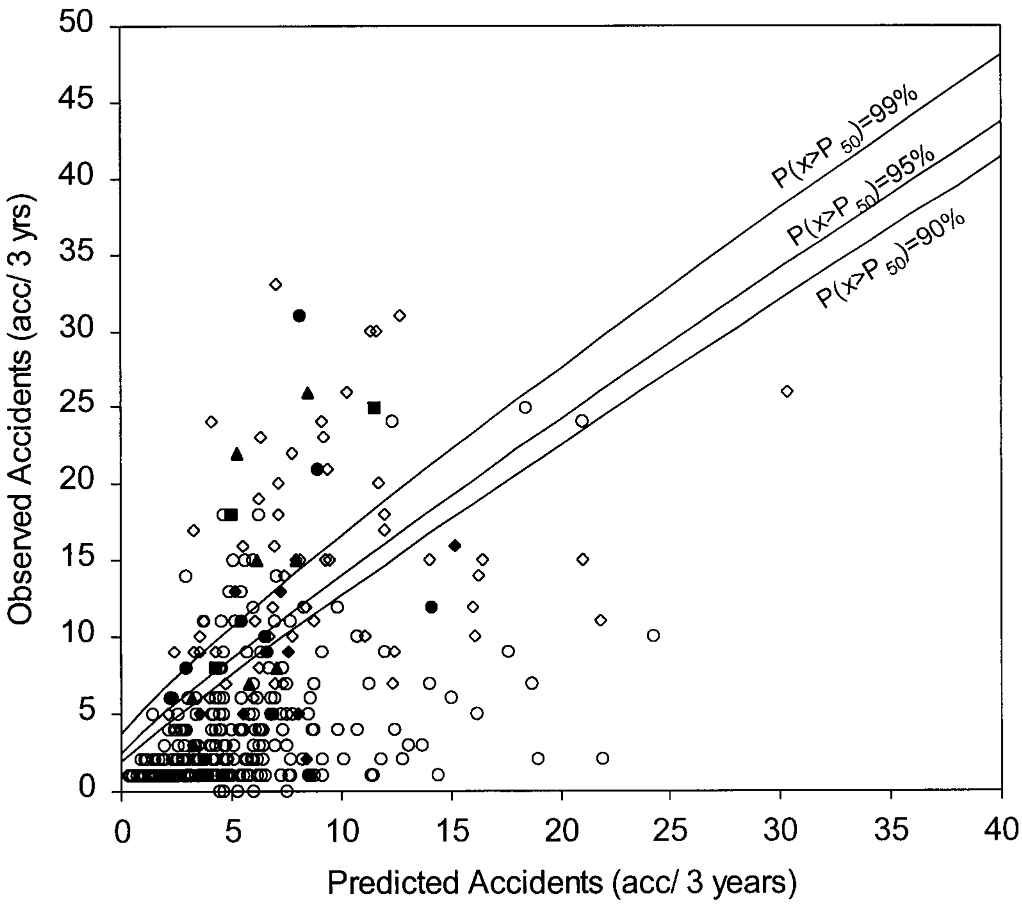
The process of identifying accident-prone locations, as described in the previous section, involves considerable computational effort. To facilitate this process, critical accident frequency curves can be developed for each GLIM model. A critical curve is one that indicates the number of observed accidents that must be exceeded in order to classify the location as accident-prone for a given GLIM model and a confidence level.

The procedure to obtain these critical curves is iterative and makes use of equations (5.2) and (5.5). The initial data is the number of predicted accidents based on a GLIM model with its κ parameter. For every predicted accident, the P_{50} value is calculated by using equation (5.2). This value is used in equation (5.5), where for a given level of confidence, the equation is solved in an iterative fashion, in order to find the observed number of accidents (variable *count* in equation (5.5)) that fits the given level of confidence. The critical curve is obtained by joining all the critical points in a Predicted versus Observed Accidents chart.

As an example, Figures 5.3, 5.4 and 5.5 show these curves for the total model (Table 3.4), Vancouver Island and Lower Mainland models (Table 3.8). Three curves are shown in each figure, representing the 90%, 95%, and 99% confidence levels. To illustrate the use of these curves, consider the example described in Section 5.1. Using the total model and the given traffic volumes, 6.88 accidents/3 years, are estimated. For this number of accidents and for 99% confidence level, at least 13 accidents/3 years need to be observed to consider this intersection as accident-prone (Figure 5.3). Table 5.1 shows the number of APLs identified by the three models for different significance levels.

MODEL	LEVEL OF CONFIDENCE		
	90%	95%	99%
Total Model	82	67	51
Vancouver Island Model	38	30	21
Lower Mainland Model	21	14	6

Table 5.1 Number of Accident Prone Locations



◇ Surrey ○ Victoria ▲ Coquitlam ■ Vancouver ● Burnaby ◆ Nanaimo

Figure 5.3 Critical Curves for Total Model

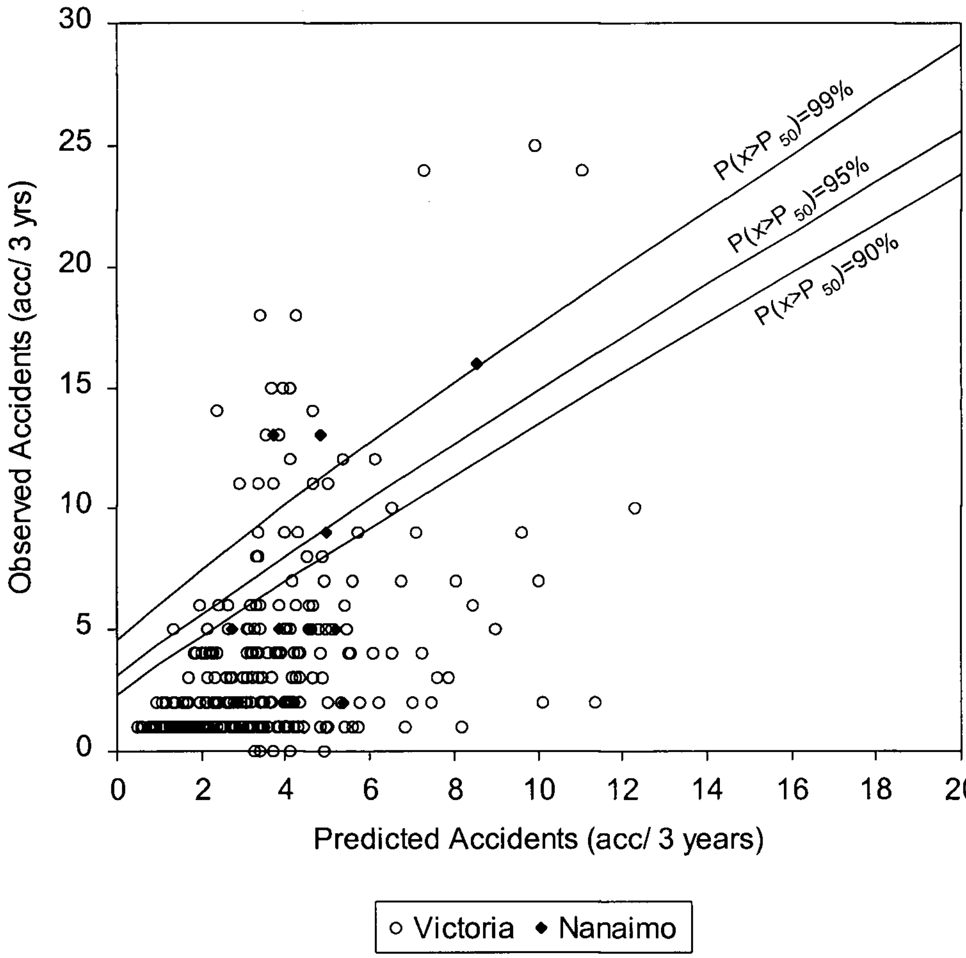


Figure 5.4 Critical Curves for Vancouver Island Model

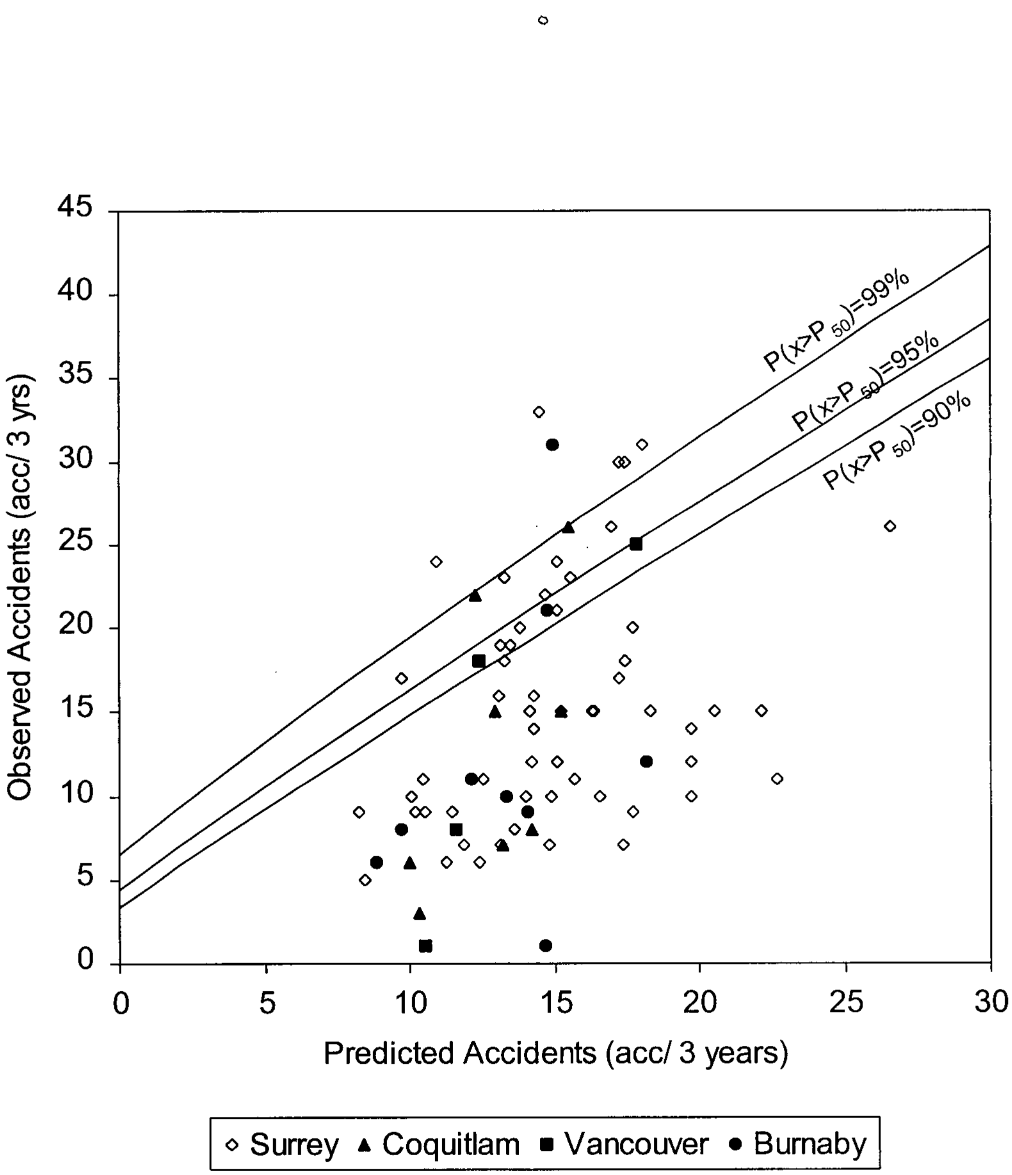


Figure 5.5 Critical Curves for Lower Mainland Model

An extension of the critical curves can be also applied for different κ values. Figure 5.6 shows the critical curves for eight different values of κ and a confidence level of 95%. The advantage of this kind of curves is that they can be used for any negative binomial model. The data required to use this curve is a negative binomial model from which the predicted number of accidents is calculated and according to the model's κ value, the critical number of accidents is estimated by using the curve for the corresponding κ. The disadvantage of this method is that the results are not as accurate as the previous ones, because in most cases there is not a curve for the specific κ value (i.e. κ=2.17) and the critical value is estimated by approximating the κ value to the closest curve.

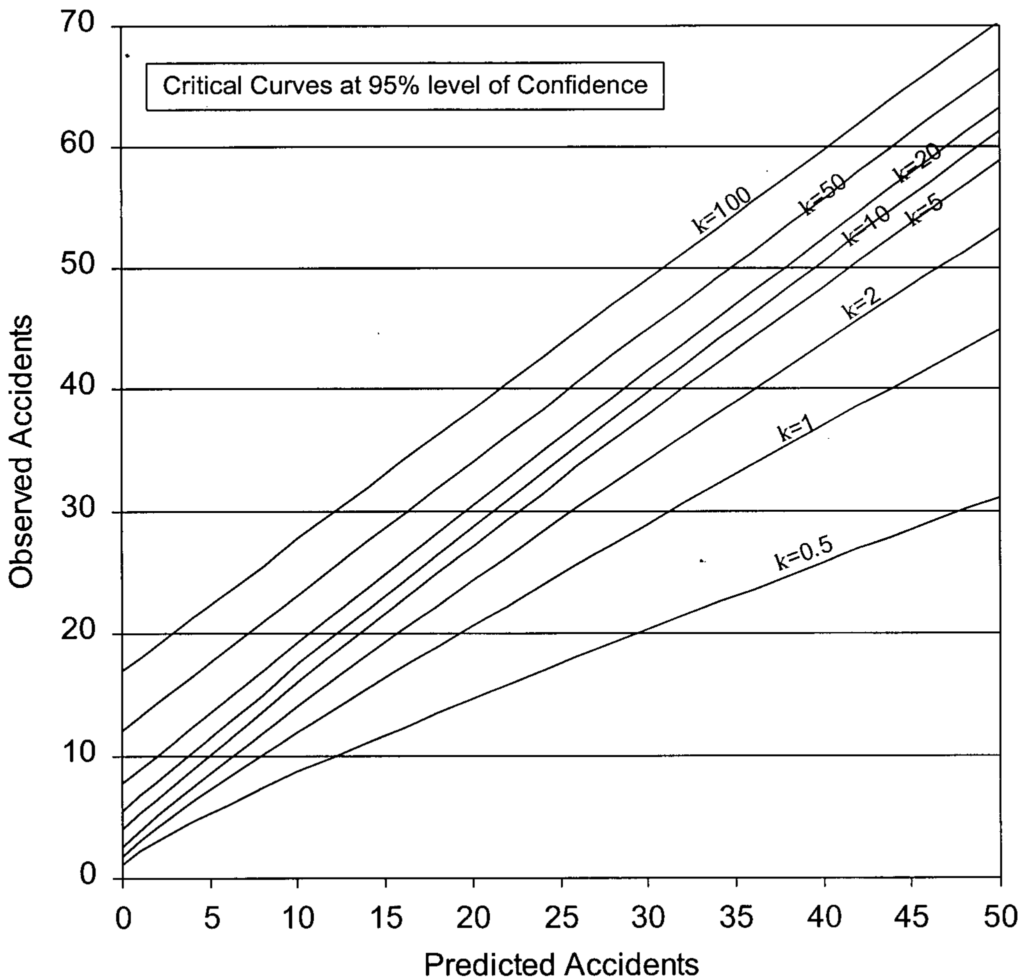


Figure 5.6 Critical Curves for Different Values of κ

Note that the higher the κ value, the higher the critical number of accidents. The rationale for this is illustrated in Figure 5.7.

Figure 5.7-a shows the same example as in Section 5.1, but in this case it is assumed to have a κ value of 1.0 (low κ) and an observed number of accidents of 9.05 acc/3 years (this is the critical number of accidents at 95% of confidence level). Under these conditions the predicted number of accidents is the same 6.88 acc/3 years, but due to the change in the κ value and the observed number of accidents, the P_{50} is 4.77 acc/3 years and the EB estimate is 8.77 acc/3 years. The probability of having accidents greater than P_{50} value is 95%, a critical condition.

Figure 5.7-a shows that at low values of κ, the prior distribution is skewed left, and the EB estimate is close to the observed number of accidents. The reason of this is that low κ values increase the variance leading to more uncertainty about the predicted value. Therefore, the EB estimate is closer to the observed value rather than the predicted one.

Figure 5.7-b, shows the same model but the κ value is considerably higher (κ=20). The observed number of accidents is the same as in Figure 5.7-a, but in this case due to the increase in κ, this value is no longer critical. The EB estimate is now closer to the predicted number of accidents instead of the observed one, because the variance has decreased leading to more reliability about the GLIM model estimate. The prior distribution is less skewed and closer to the posterior distribution.

In order to find the critical number of accidents for the conditions in this case (Figure 5.7-b), it is necessary to raise considerably the observed number of accidents. Figure 5.7-c shows that the critical value is 15.65 accidents/3 years, which represents an increase of 6.5 accidents/3 years compared with the previous conditions, while the EB estimate has also increased but only by 1.6 acc/3 years. This latter value remains closer to the predicted number of accidents.

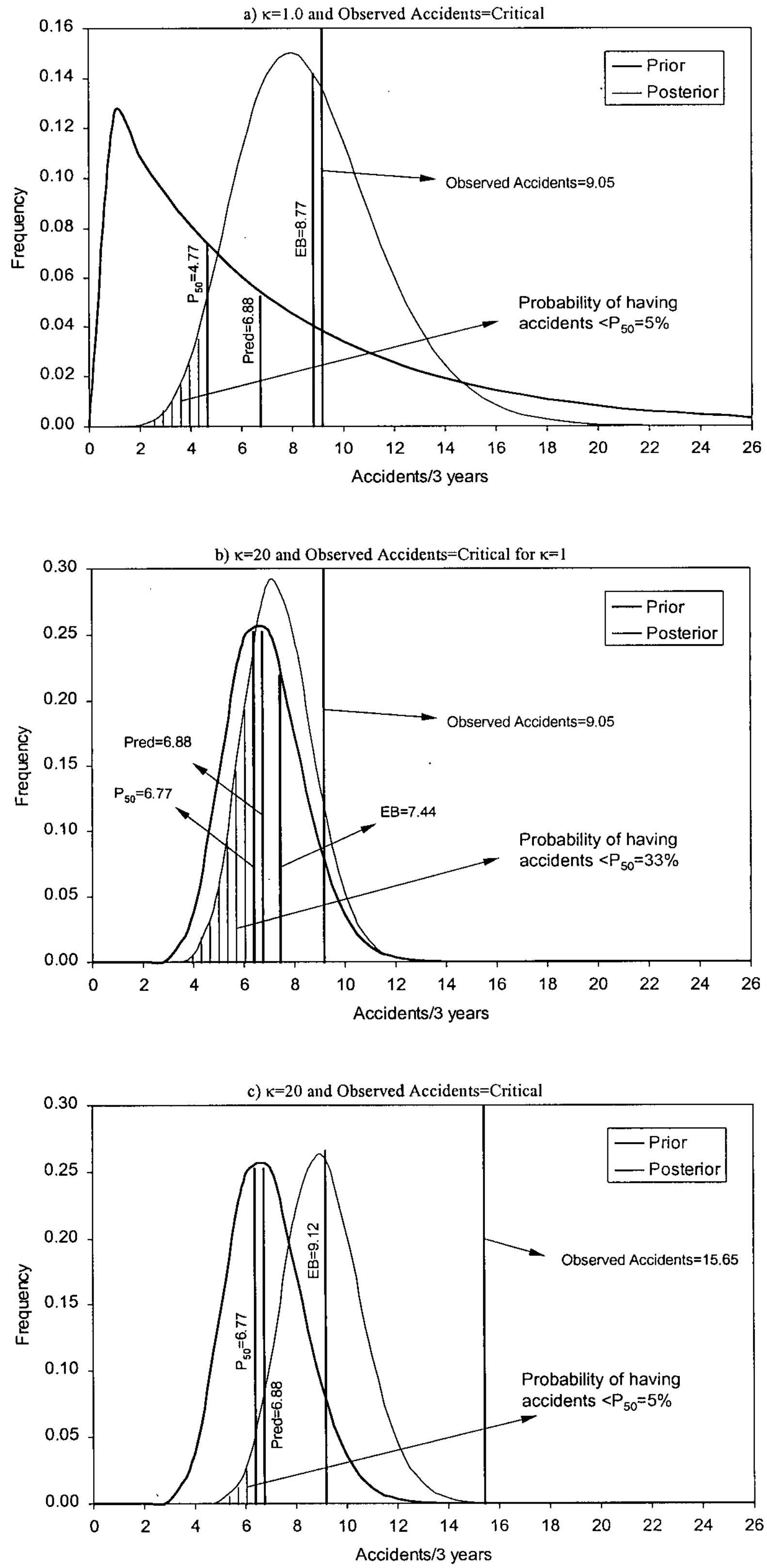


Figure 5.7 Comparison of Critical Accidents for Different κ Values

5.4 Ranking of Accident Prone Locations

The methods used to identify accident-prone locations explained in the previous two sections can be also useful in ranking these locations. Two ranking criteria can be used. The first is to calculate the ratio between the EB estimate and the predicted frequency (as obtained from the GLIM model) for the accident prone locations identified in the previous section. This ratio represents the deviation of the intersection from the "norm". The higher this ratio the more accident prone the intersection is. The justification for using this ranking criterion is to ensure that the safety level at each criterion is comparable to other intersections with similar characteristics.

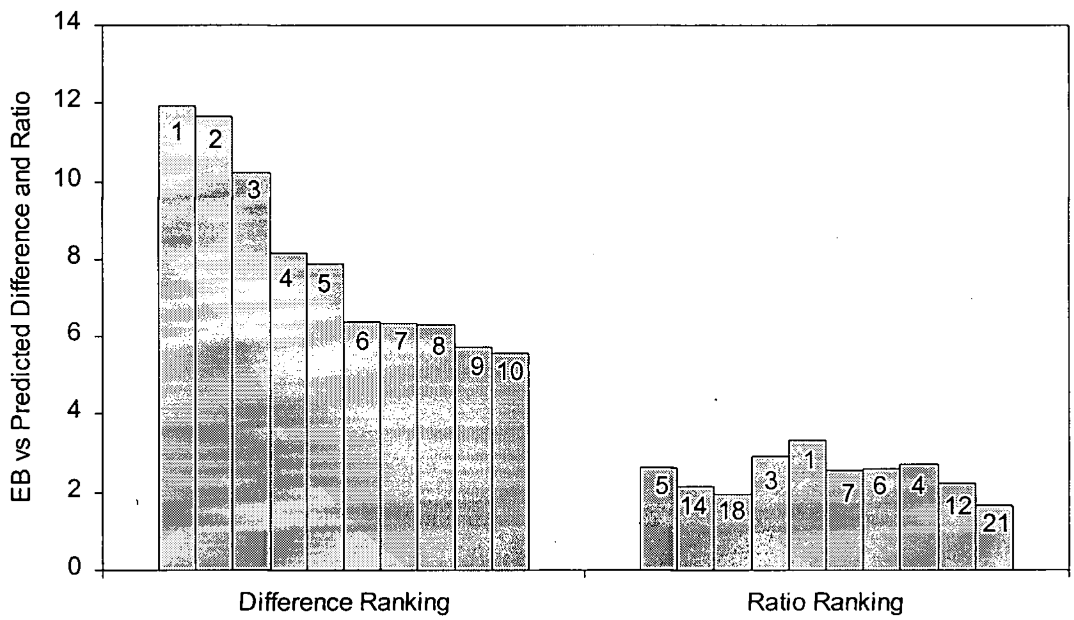
Another criterion is to calculate the difference between the EB estimate and the predicted frequency for the accident prone locations. This difference is a good indication of the expected safety benefits and is useful for carrying out the estimation of the pre-implementation safety benefits of countermeasures. Unlike the previous criterion, this one is useful to quantify economical benefits.

A comparison of the two ranking criteria is shown in Table 5.2 for the Vancouver Island Model. Twenty-one accident prone locations (APLs) were identified at the 99% confidence level. The table shows the values of both the difference (EB - Predicted) and ratio (EB/Predicted) for all APLs. As shown in Table 5.2, the difference in rank between the two criteria ranges between 1 and 15 with an average value of 4.9. The difference in rank seems to be higher for the top ranked intersections. The reason for this difference can be explained by the different goals of the two criteria. The first criterion favors intersections with high accident frequency which are usually more cost-effective to treat. The second criterion considers the deviation from the expected values and its variance regardless of the number of accidents observed. This criterion can be considered by road authorities to ensure that the safety of different locations is within acceptable levels.

Int. No.	Intersection	Observed Frequency	Predicted Accidents	EB Refined	EB-Pred	EB/Pred	Rank EB-Pred	Rank EB/Pred	Diff Rank
1	Blanshard-Topaz	24	7.3	19.2	11.9	2.6	1	5	4
2	Cook-Kiwanis	25	9.9	21.6	11.7	2.2	2	14	12
3	Douglas-Tolmie	24	11.1	21.3	10.2	1.9	3	18	15
4	Finlayson-Nanaimo	18	4.3	12.4	8.2	2.9	4	3	1
5	Government-Discovery	18	3.4	11.3	7.9	3.3	5	1	4
6	Vancouver-Balmoral	15	4.1	10.5	6.4	2.5	6	7	1
7	Douglas-Princess	15	3.9	10.3	6.4	2.6	7	6	1
8	Cook-View	15	3.7	10.0	6.3	2.7	8	4	4
9	Southgate-Vancouver	14	4.7	10.4	5.7	2.2	9	12	3
10	Bowen-Pine-Access	16	8.6	14.1	5.5	1.6	10	21	11
11	Dallas-Douglas	14	2.4	7.6	5.2	3.2	11	2	9
12	Douglas-Discovery	13	3.9	9.1	5.2	2.3	12	11	1
13	Albert-Fourth-Pine-Park	13	3.7	8.9	5.2	2.4	13	9	4
14	Quadra-Topaz	13	3.6	8.7	5.2	2.5	14	8	6
15	Wakesiah-Fourth	13	4.8	9.9	5.1	2.1	15	17	2
16	Fairfield-Foulbay	12	4.1	8.7	4.6	2.1	16	15	1
17	Douglas-Spruce	12	5.4	9.7	4.3	1.8	17	20	3
18	Shelbourne-Pearl	11	3.4	7.5	4.1	2.2	18	13	5
19	Quadra-Burdett	11	3.7	7.8	4.1	2.1	19	16	3
20	Hillside-Graham	11	2.9	6.9	4.0	2.4	20	10	10
21	Quadra-Pembroke	11	4.6	8.6	3.9	1.8	21	19	2

Table 5.2 Ranking of APLs for The Vancouver Island Model

Figure 5.8 shows the values of the two ranking criteria for the top 10 APLs. The figure shows that intersections 1, 3, 4, 5, 6, and 7 are among the top 10 intersections for both methods with an average ranking difference of 5.6. Intersections 2, 8, 9, and 10 are included in the top 10 using the "difference" ranking, but are not included in the "ratio" ranking. The degree of proneness of these intersections, despite of showing high expected benefits, is not among the top 10 intersections. The same applies to intersections 12, 14, 18 and 21 which show a high degree of proneness, but its indication of expected benefits is not among the top 10 intersections.



Number inside bar denotes intersection number according to Table 5.2

Figure 5.8 Ranking of Top10 APL for Island Model

There are other ranking criteria relating both a ratio and a difference. These other methods involves parameters of accident prediction models such as the predicted vs. observed number of accidents, observed vs. critical curve value, observed vs. EB estimates, etc. These criteria can be implemented by using the same methodology of this section. There is little research concerning the ranking criteria when using accident prediction models. This is an area that surely needs further research.

5.5 Before and After Studies

The effect of a safety measure is often studied by comparing the number of accidents observed after the implementation of the measure, to the expected number of accidents had the measure not been implemented. In simple before and after studies, the observed number of accidents in the period before the implementation is used to estimate the latter value. However, because of the random variations in accident occurrence (e.g. the regression to the mean effect), the observed number of accidents 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.

Using the same example as before, assume that a specific safety measure to reduce the number of accidents at the intersections was implemented. The observed number of accidents in the next three years following the implementation is 8. Therefore, the effectiveness of the measure can be calculated as:

$$\textit{Measure of Effectiveness} = 1 - \frac{8}{10.08} = 0.21$$

which indicates a reduction by 21% in total accidents because of the treatment.

The importance of using accident prediction models in before and after studies is highlighted by the difficulty associated in developing this analysis in a traditional fashion, via a reference group of comparison. This group should be of sufficient size and homogeneity to carry out an accurate analysis. The difficulty lies in defining a group with these features. Accident prediction models overcome this difficulty, since they represent of local conditions and replace the role of the traditional reference group.

5.6 Safety Comparison of Staggered T and 4-leg Intersections

Several researches have compared the safety performance of 4-leg intersections and staggered T-intersections. Kulmala (1995) found that, in general, the staggering of 4-leg intersections into two staggered T-intersections reduces the number of injury accidents if the percentage of traffic entering the junction form the minor road is greater than 5% of the total traffic. He also found that if 50% of the total traffic enter the junction from the minor road, the staggering would reduce the number of injury accidents by 23%. Kulmala (1995) found his results consistent by comparing them with some Nordic studies, where the staggering was found to decrease the number of injury accidents by 0% to 20%.

In order to confirm these results, a safety comparison of 4-leg and staggered T-intersections was carried out using the models developed in Table 3.5 (the separate T and 4-leg intersection models). According to the analysis made in Section 3.2.2, it is also valid to use the total model with intersection type (Table 3.6) which yields approximately the same results. The following assumptions were made:

1. The traffic volumes on the major and minor roads for the 4-leg intersections are V_1 and V_2, respectively (expressed in AADT).

2. For the two staggered T-intersections, the traffic volume on the major road is V_1, while the minor approaches have traffic volumes of $V_2/2$ (expressed in AADT). This assumption ensures that the traffic volume in both scenarios is the same.

3. The two staggered intersections will not affect each other (isolated intersections). This assumption depends, of course, on the distance between the two intersections.

The results of the comparison are shown in Figure 5.9 for three different minor road traffic volumes. The results indicate that the staggering is effective in reducing the predicted number of accidents. This reduction increases as the traffic volume on the major or minor road increases. It should be noted that the degree of reduction would vary with different ratios of traffic volumes on the major and minor roads.

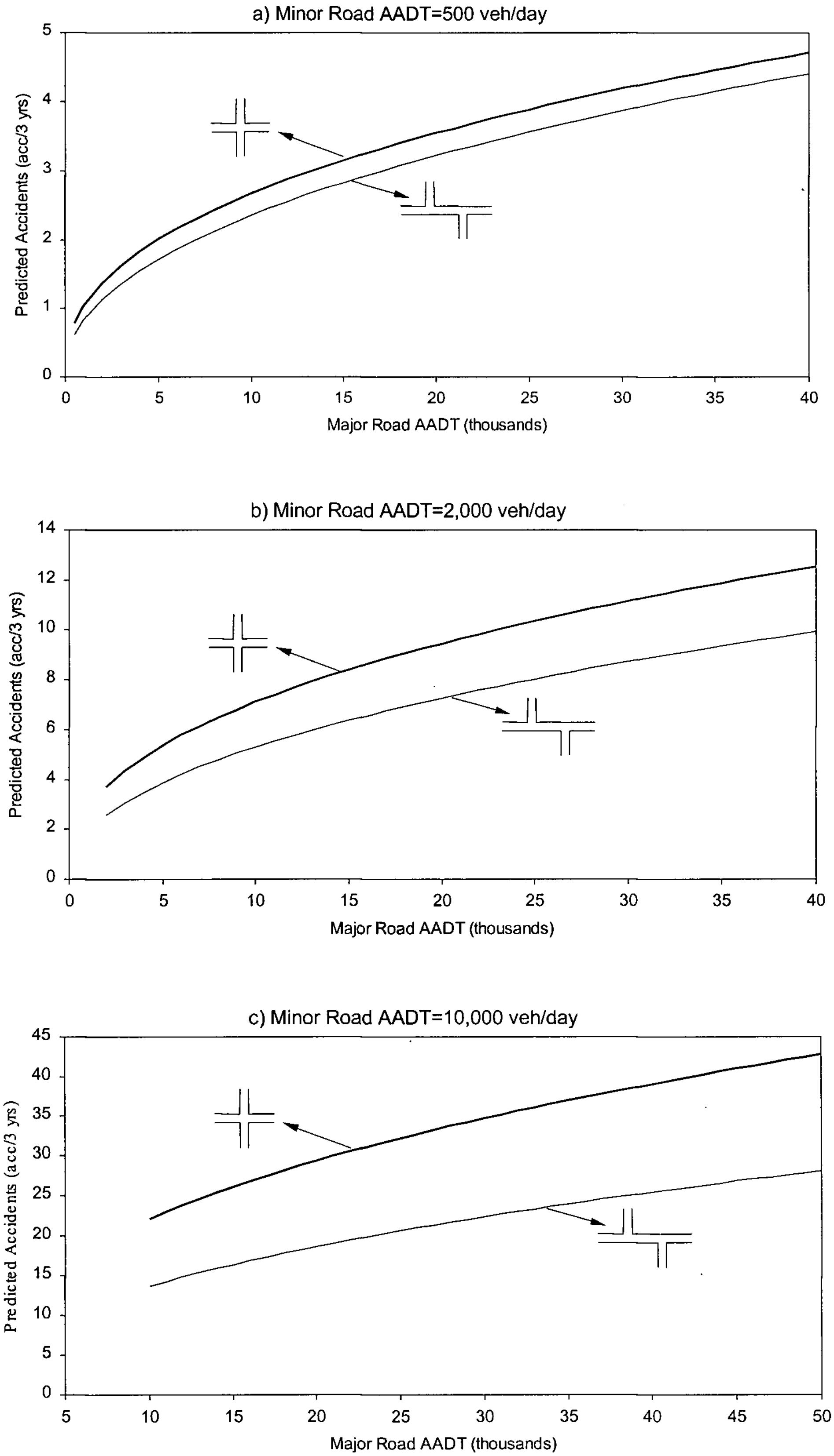


Figure 5.9 Staggered T vs 4-leg Intersections Safety Comparison

5.7 Conclusion

This chapter has shown five applications of accident prediction models. Most of these applications make use of the EB refinement methods, in order to reduce the regression to the mean phenomenon.

It has been shown that accident prediction models are useful in identifying accident prone locations (APLs) with a probabilistic confidence level by using both analytical and graphical methods. It is also possible to rank the APLs by two different criteria, difference and ratio, according to the particular objectives of the road's authorities.

In addition, accident prediction models can be used for evaluating the safety of a countermeasure, without having to define a reference group, because the GLIM models contains the characteristics of the location.

Finally, it was found that staggered T-intersections are safer than 4-leg intersections, a finding that should be taken into account by road planning authorities. These results agree with those found in the literature.

CHAPTER VI

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The main objective of this project is to develop accident prediction models for estimating the safety potential of urban unsignalized (T and 4-leg) intersections in the Greater Vancouver Regional District (GVRD) and Vancouver Island on the basis of their traffic characteristics. The models are developed using the generalized linear regression modeling (GLIM) approach, which addresses and overcomes the shortcomings associated with the conventional linear regression approach. The safety predictions obtained from GLIM models can be refined using the Empirical Bayes' approach to provide, more accurate, site-specific safety estimates. The use of the complementary Empirical Bayes approach can significantly reduce the regression to the mean bias that is inherent in observed accident counts.

This study made use of sample accident and traffic volume data corresponding to unsignalized (both T and 4-leg) intersections located in urban areas of the Greater Vancouver Regional District (GVRD) and Vancouver Island. The data included a total of 427 intersections located in the cities of Victoria, Surrey, Nanaimo, Coquitlam, Burnaby and Vancouver. The information available for each intersection included the total number of accidents in the 1993-1995 period, traffic volumes for both major and minor roads given in Average Annual Daily Traffic (AADT) and type of intersection (T or 4-leg).

Four categories of models were developed in this study: (1) models for the total number of accidents; (2) separate models for T and 4-leg intersections; (3) separate models for different regions (Vancouver Island, the Lower Mainland and Surrey); and (4) a model for Surrey including intersection control. Table 6.1 summarizes the models' results.

Models developed in this thesis used the negative binomial distribution approach, which has the advantage of explaining the dispersion characteristic of the observed data compared with the Poisson distribution. In addition, different tests showed that the maximum likelihood method yields the most appropriate parameters under the negative binomial distribution assumption.

Model Form	t-ratio		SD (dof)	κ	Pearson χ^2 (χ^2 test)*
Model for the total number of accidents $Acc/3\,yrs = 1.4929\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.3839}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.7044}$	a_o a_1 a_2	3.2 7.8 12.4	399 (424)	1.97	459 (472)
T-intersection model $Acc/3\,yrs = 0.9333\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.4531}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.5806}$	a_o a_1 a_2	-0.3 5.5 7.4	164 (183)	2.34	205 (214)
4-leg intersection model $Acc/3\,yrs = 1.6947\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.4099}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.7065}$	a_o a_1 a_2	3.6 6.8 9.1	230 (238)	2.17	251 (274)
Total Model with Intersection Type $Acc/3\,yrs = 0.5776\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.4221}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.6480}\times e^{0.5379\times Type}$	a_o a_1 a_2 b_1	-2.8 8.7 11.7 6.3	394 (423)	2.23	449 (471)
Vancouver Island model $Acc/3\,yrs = 1.3807\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.3042}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.5488}$	a_o a_1 a_2	2.6 6.0 9.0	302 (347)	2.92	383 (390)
Lower Mainland model $Acc/3\,yrs = 6.5929\times\left(\frac{AADT_{maj\,rd}}{1000}\right)^{0.2011}\times\left(\frac{AADT_{min\,rd}}{1000}\right)^{0.2864}$	a_o a_1 a_2	7.6 2.4 3.7	81 (74)	6.27	76 (94)

* Denotes significance at a 95-percent confidence level

Table 6.1 Summary of Accident Prediction Models

Five applications of accident prediction models were used in this thesis. Four of them related to the use of the Empirical Bayes refinement: identification of accident-prone locations, developing critical accident frequency curves, ranking the identified accident-prone locations and before and after safety evaluation. The fifth application provided a safety-planning example, comparing a 4-leg intersection to two staggered T-intersections.

It was shown that accident prediction models are useful in identifying accident prone locations (APLs) with a probabilistic confidence level by using both analytical and graphical methods. It is also possible to rank the APLs by two different criteria, difference and ratio, according to the particular objectives of the road's authorities.

In addition, accident prediction models can be used to evaluate the safety benefits of a countermeasure, without having to define a reference group, because the GLIM models contains the characteristics of the location.

Finally, it was found that staggered T-intersections are safer than 4-leg intersections, a finding that should be taken into account by road planning authorities. These results agree with previous researches made in the Scandinavian countries.

6.2 Recommendations for further research

This thesis has developed accident prediction models for urban unsignalized intersections that included independent variables such as traffic volumes and control type. It is recommended that these models be further refined by adding more variables such as:

Intersection control type: An attempt to develop this model was made in this thesis, but the results showed a poor fit. Therefore it is recommended to use a larger sample size to obtain a significant model, in order to assess the safety effect of intersection control type in a similar fashion that this thesis assessed the safety effect of T and 4-leg intersections.

Intersection Layout variables: Accident occurrence can be explained by several variables. Including intersection layout variables (e.g. number of lanes of each road, number of left and right turn lanes, pedestrian crosswalks, speed limit, etc) should enhance our understanding of the relationships between accident occurrence and geometric design.

Accident Type: In safety evaluation of countermeasures it may be necessary to look at individual accident types (e.g. rear-end, right angle, etc.) as opposed to the total number of accidents. Therefore, it is recommended that models for specific accident types be developed.

Finally, as explained earlier, there is a need for more research on ranking accident prone locations. This is very important in situations when the road authority has resources to address only a limited number of accident prone locations, it is important to focus on those with the highest potential of accident reduction or those which deviates from the normal safety levels for similar locations.

BIBLIOGRAPHY

Bélanger, C. (1994). "Estimation of safety of four-leg unsignalized intersections", Transportation Research Record, 1467, Transportation Research Board, National Research Council, Washington D. C., pp. 23-29.

Bonneson, J. A. and McCoy, P. T., (1993). "Estimation of safety at two-way stop-controlled intersections on rural highways", Transportation Research Record, 1401, Transportation Research Board, National Research Council, Washington D. C., pp. 83-89.

Bonneson, J. A. and McCoy, P. T., (1997). "Effect of median treatment on urban arterial safety: An accident prediction model", Transportation Research Record, 1581, Transportation Research Board, National Research Council, Washington D. C., pp. 27-36.

Brüde, U. and Larsson, J. (1988). "The use of prediction models for eliminating effects due to regression-to-the mean in road accident data", Accident Analysis and Prevention, Vol 20, No 4, pp. 299-310.

Famoye, F. (1997). "Parameter estimation for generalized negative binomial distribution", Communications-in-Statistics. Part B: Simulation and Computation, Vol. 26, No 1, pp 269-279.

Feng, S. and Sayed, T. (1997). "Accident prediction models for signalized intersections". The University of British Columbia, Department of Civil Engineering. Report prepared for the Insurance Corporation of British Columbia.

Hauer, E., Ng J. C. N. and Lovell J., (1988). "Estimation of safety at signalized intersections", Transportation Research Record, 1185, Transportation Research Board, National Research Council, Washington D. C., pp. 48-61.

Hauer, E. (1992). "Empirical Bayes approach to the estimation of 'unsafety': The multivariate regression method", Accident Analysis and Prevention, Vol 24, No 5, pp. 457-477.

Jovanis, P. P. and Chang H. L. (1986). "Modeling the relationship of accidents to miles traveled", Transportation Research Record, 1068, Transportation Research Board, National Research Council, Washington D. C., pp. 42-51.

Kulmala, R. and Roine, M., (1988). "Accident prediction models for two-lane roads in Finland", Conference on traffic safety theory and research methods proceedings, April, Session 4: Statistical analysis and models. Amsterdam: SWOV, pp. 89-103.

Kulmala, R., (1995). "Safety at rural three- and four-arm junctions. Development of accident prediction models", Espoo 1995, Technical Research Centre of Finland, VTT 233.

Lawless, J., (1987). "Negative binomial and mixed Poisson regression", The Canadian Journal of Statistics, Vol. 15 No 3, pp. 209-225.

McCullagh P. and Nelder J.A., (1983) "Generalized Linear Models", Chapman and Hall, New York.

Maher, M. J. and Summersgill, I., (1996). "A comprehensive methodology for the fitting of predictive accidents models", Accident Analysis and Prevention, Vol 28, No 3, pp. 281-296.

Maycock, G. and Hall, R. D. (1984). "Accidents at 4-arm roundabouts", Transport and Road Research Laboratory. TRRL Laboratory Report 1120.

Miau, S. and Lum, H., (1993). "Modeling vehicle accident and highway geometric design relationships", Accidents Analysis and Prevention, Vol. 25, No 6, pp. 689-709.

Mountain, L. and Fawaz, B., (1996). "Estimating accidents at junctions using routinely-available input data", Traffic Engineering and Control, Vol. 37, No 11, pp. 624-628.

Numerical Algorithms Group (NAG), (1994), "The GLIM system. Release 4 manual", The Royal Statistical Society.

Numerical Algorithms Group (NAG), (1996), "GLIM 4. Macro Library Manual, Release 2", The Royal Statistical Society.

Saccomanno, F. F. and Buyco, C. (1988). "Generalized loglinear models of truck accident rates". Paper presented at Transportation Research Board 67th annual meeting. Washington, D. C.

Satterthwaite, S. P., (1981). "A survey of research into relationships between traffic accidents and traffic volumes", Transport and Road Research Laboratory. TRRL Supplementary Report 692.

Stevens J., (1986). "Applied multivariate statistics for the social sciences" Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, N..J.

APPENDIX I

RESULTS OF OUTLIERS IDENTIFICATION

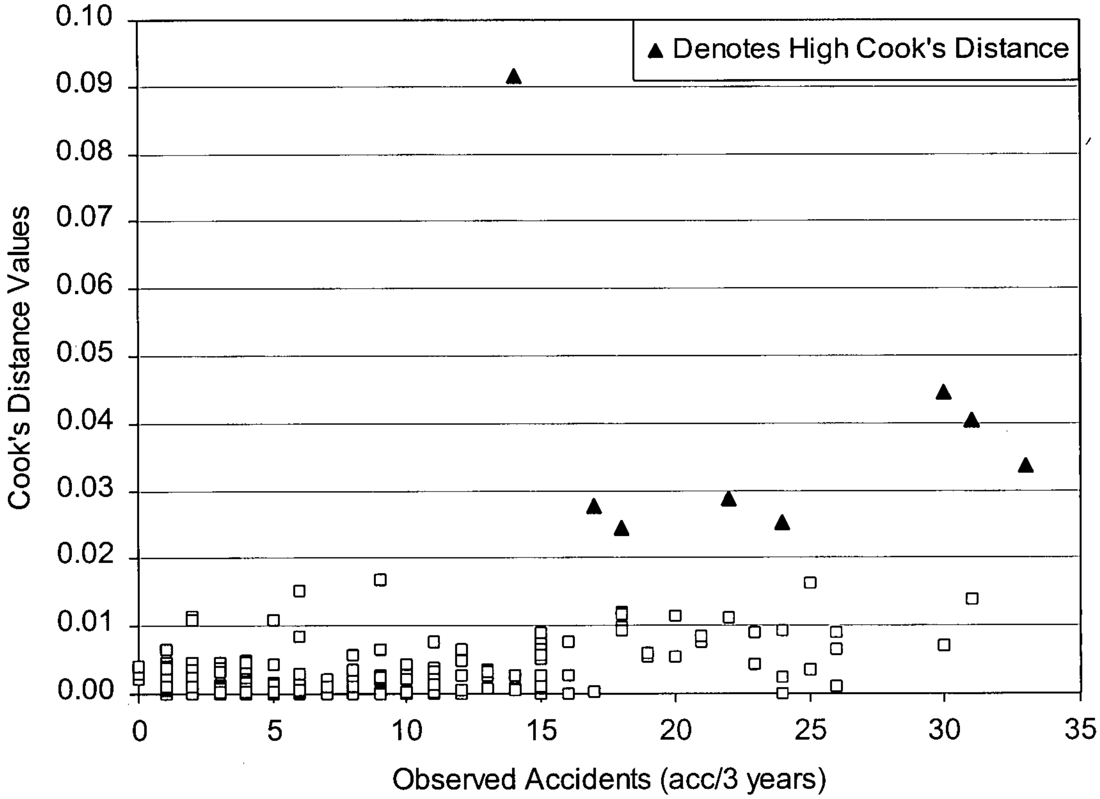


Figure AI-1 Identification of the Highest Cook's Distance Values for Total Model with Intersection Type

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	157	426	390.69	3.63	3.63	3.84
2	45	425	389.49	1.20	4.83	5.99
3	51	424	388.36	1.13	5.96	7.81
4	14	423	387.19	1.17	7.13	9.49
5	44	422	385.66	1.53	8.66	11.07
6	22	421	383.85	1.81	10.47	12.59
7	5	420	382.25	1.60	12.07	14.07
8	224	419	380.45	1.80	13.87	15.51

Table AI-1 Identification of Outliers for Total Model with Intersection Type

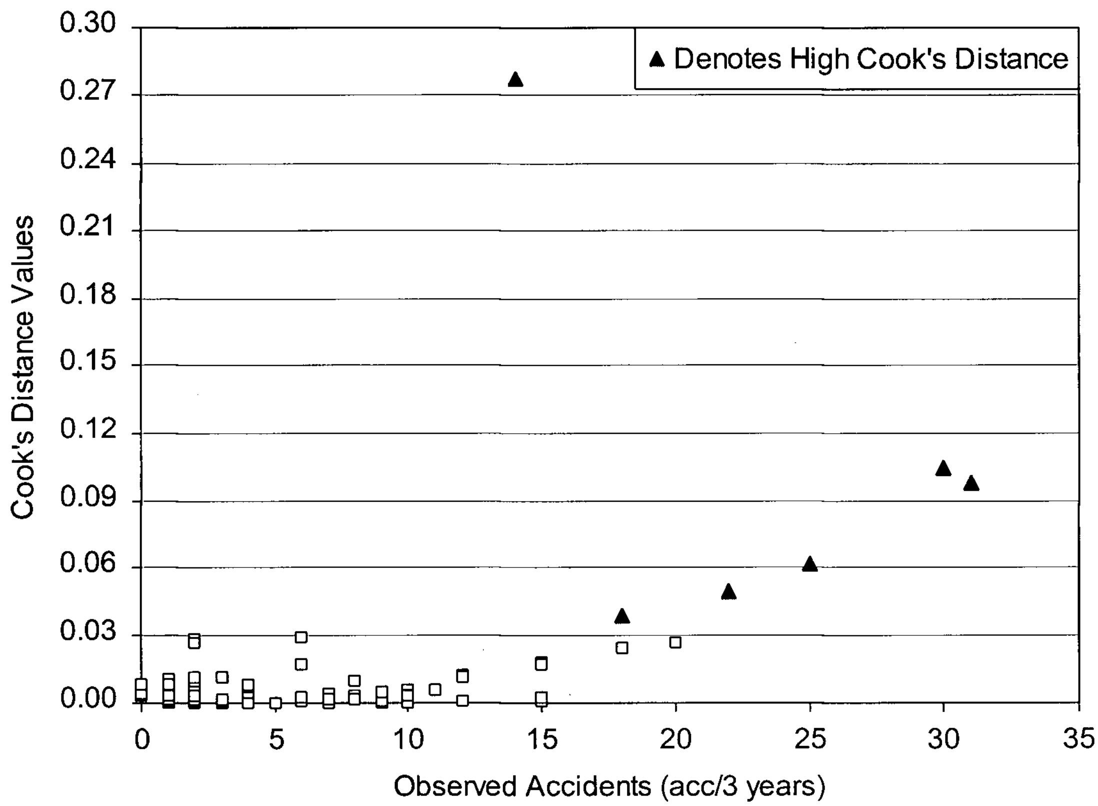


Figure AI-2 Identification of the Highest Cook's Distance Values for T-Intersection Model

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	66	185	160.61	3.00	3.00	3.84
2	9	184	159.54	1.07	4.07	5.99
3	14	183	158.15	1.39	5.46	7.81
4	53	182	157.21	0.94	6.40	9.49
5	8	181	156.03	1.18	7.58	11.07
6	103	180	154.62	1.41	8.99	12.59

Table AI-2 Identification of Outliers for T-Intersection Model

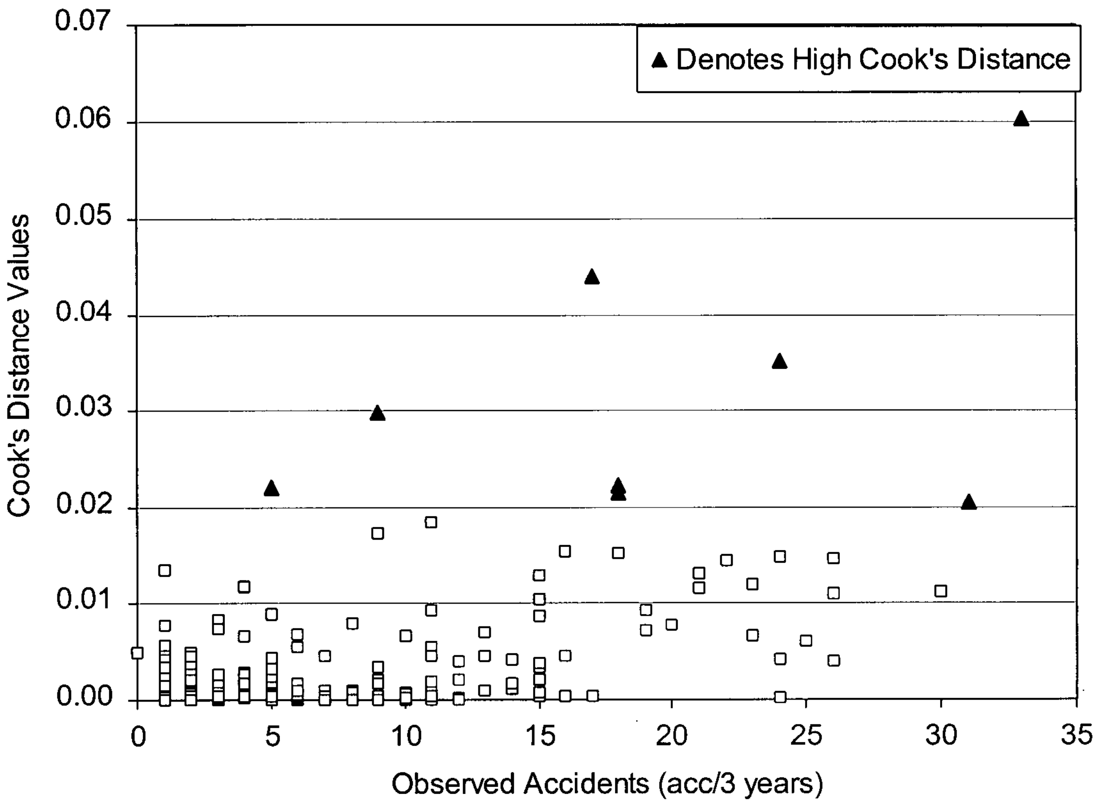


Figure AI-3 Identification of the Highest Cook's Distance Values for 4-leg Intersection Model

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	12	240	229.12	1.18	1.18	3.84
2	20	239	227.18	1.94	3.12	5.99
3	3	238	225.32	1.86	4.98	7.81
4	29	237	223.38	1.94	6.92	9.49
5	223	236	221.94	1.44	8.36	11.07
6	159	235	219.83	2.11	10.47	12.59
7	133	234	218.25	1.58	12.05	14.07
8	231	233	217.27	0.98	13.03	15.51

Table AI-3 Identification of Outliers for 4-Leg Intersection Model

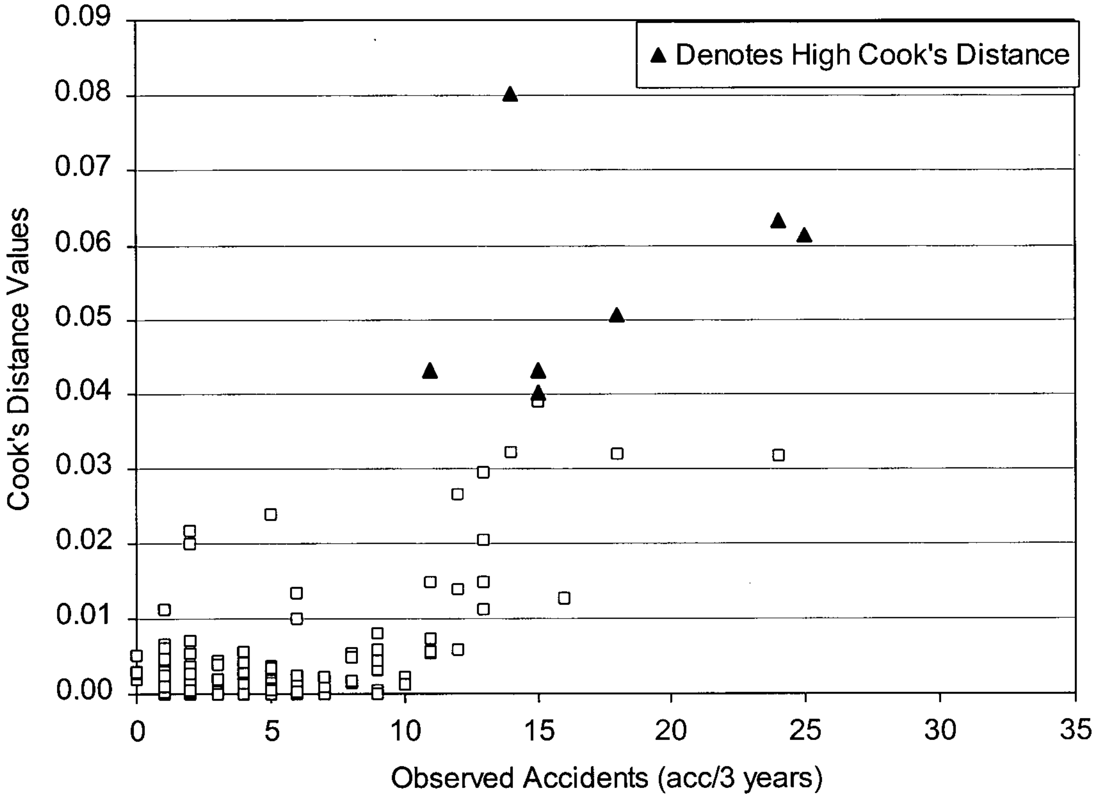


Figure AI-4 Identification of the Highest Cook's Distance Values for Vancouver Island Model

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	101	349	298.88	2.97	2.97	3.84
2	41	348	298.01	0.87	3.84	5.99
3	73	347	297.44	0.57	4.41	7.81
4	200	346	295.41	2.03	6.44	9.49
5	326	345	294.14	1.27	7.71	11.07
6	222	344	291.91	2.23	9.94	12.59
7	89	343	290.04	1.87	11.81	14.07

Table AI-4 Identification of Outliers for Vancouver Island Model

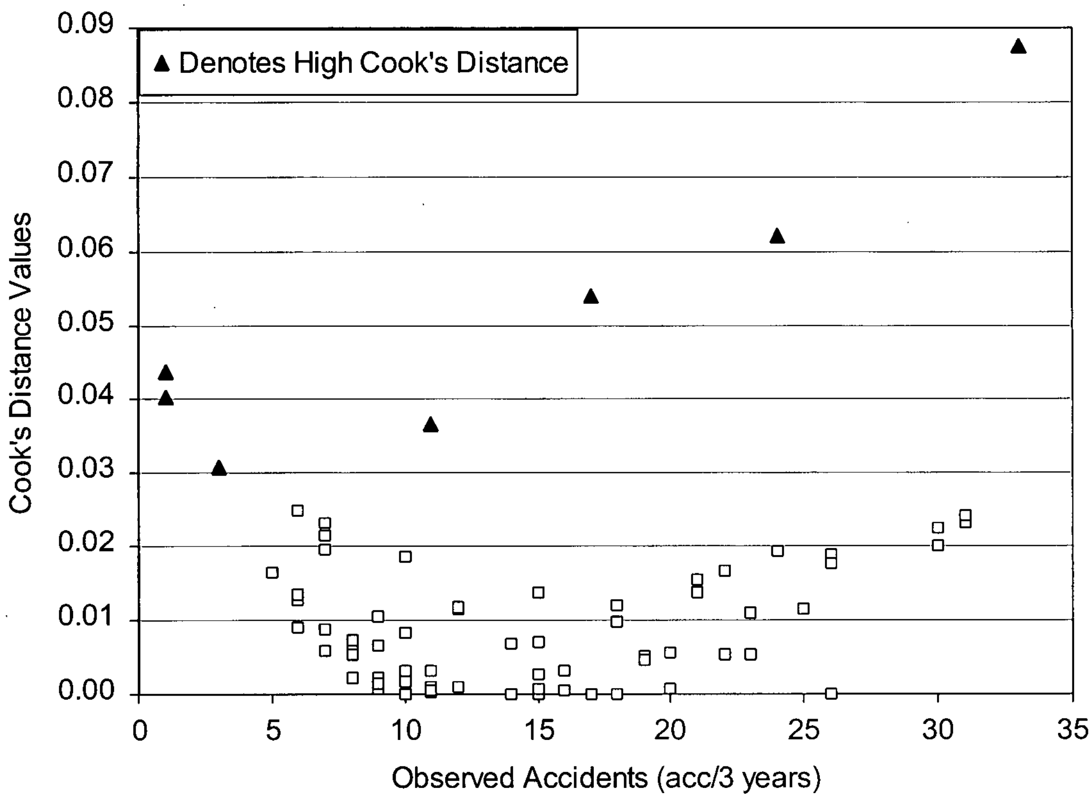


Figure AI-5 Identification of the Highest Cook's Distance Values for Lower Mainland Model

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	14	76	80.08	0.95	0.95	3.84
2	5	75	78.81	1.27	2.22	5.99
3	22	74	77.47	1.34	3.56	7.81
4	74	73	73.99	3.48	7.04	9.49
5	66	72	70.52	3.47	10.51	11.07
6	10	71	69.49	1.03	11.54	12.59
7	57	70	67.94	1.55	13.09	14.07

Table AI-5 Identification of Outliers for Lower Mainland Model

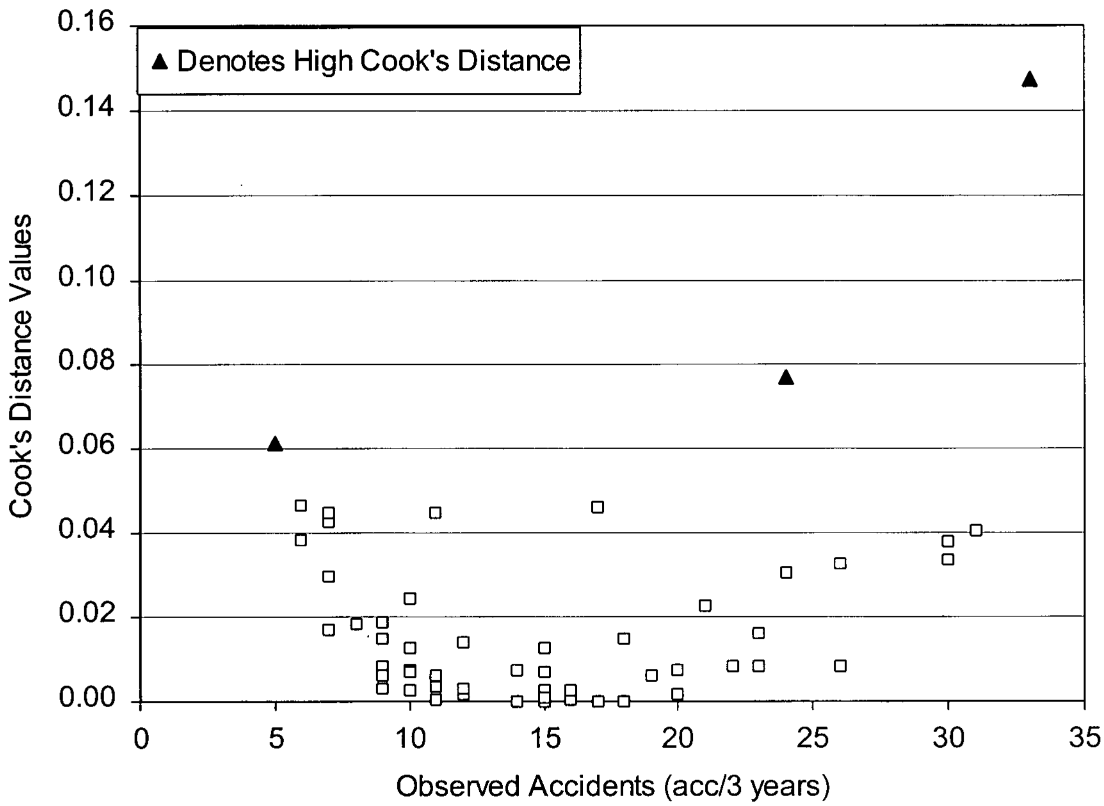


Figure AI-6 Identification of the Highest Cook's Distance Values for Surrey Total Model

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	14	55	54.60	1.12	1.12	3.84
2	5	54	53.14	1.45	2.58	5.99
3	4	53	52.26	0.88	3.45	7.81

Table AI-6 Identification of Outliers for Surrey Total Model

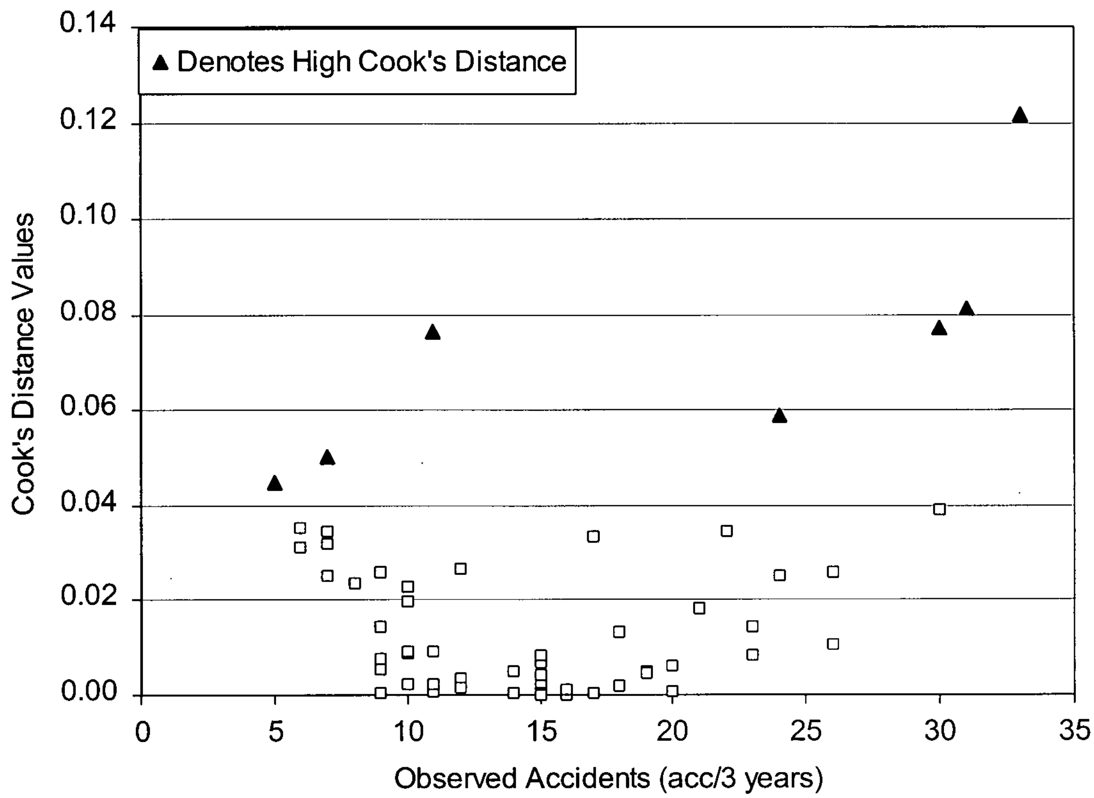


Figure AI-7 Identification of the Highest Cook's Distance Values for Surrey Total Model with Intersection Control Type

Rank Cook's Distance	Intersection Number	Sample Size	Scaled Dev.	SD Drop	Cumulative SD Drop	χ^2
1	14	55	54.40	1.08	1.08	3.84
2	51	54	53.40	1.00	2.08	5.99
3	45	53	52.40	1.00	3.08	7.81
4	10	52	51.49	0.90	3.98	9.49
5	5	51	50.15	1.35	5.33	11.07
6	49	50	49.35	0.80	6.13	12.59
7	4	49	48.40	0.95	7.08	14.07

Table AI-7 Identification of Outliers for Surrey Total Model with Intersection Control Type

APPENDIX II

PREDICTION RATIO VS. ACCIDENT FREQUENCY

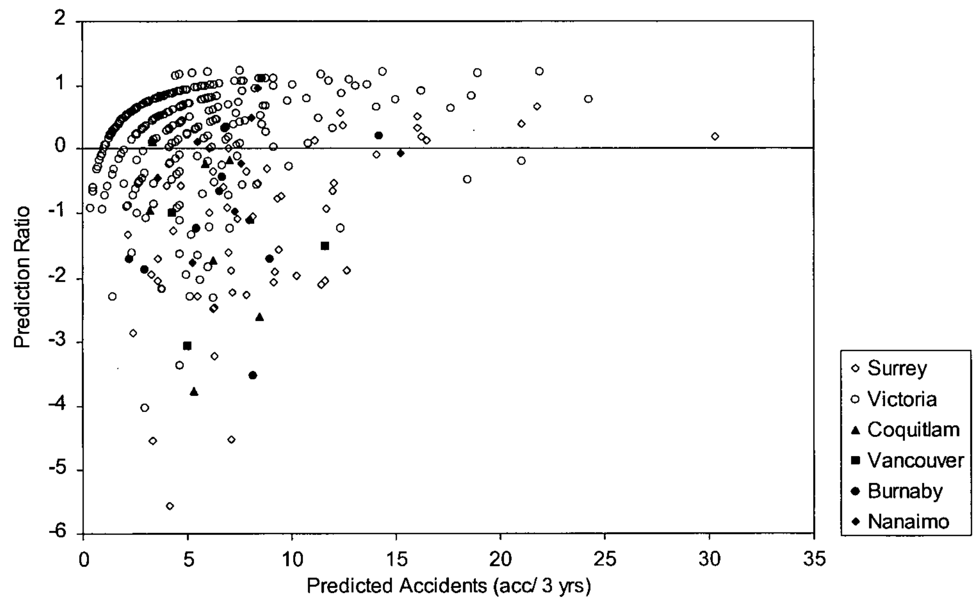


Figure AII-1 AR vs. Accident Frequency for Total Model

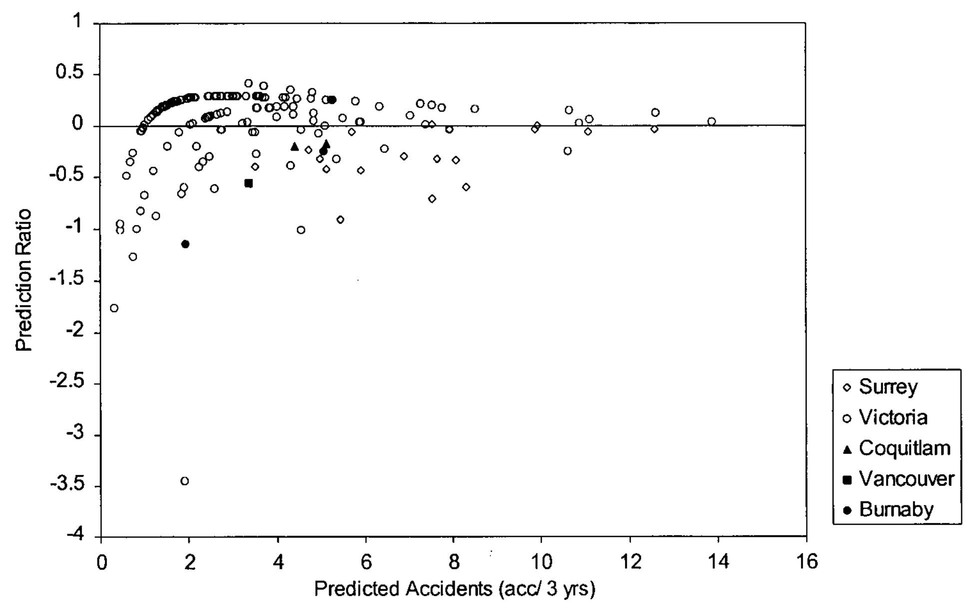


Figure AII-2 AR vs. Accident Frequency for T-intersection Model

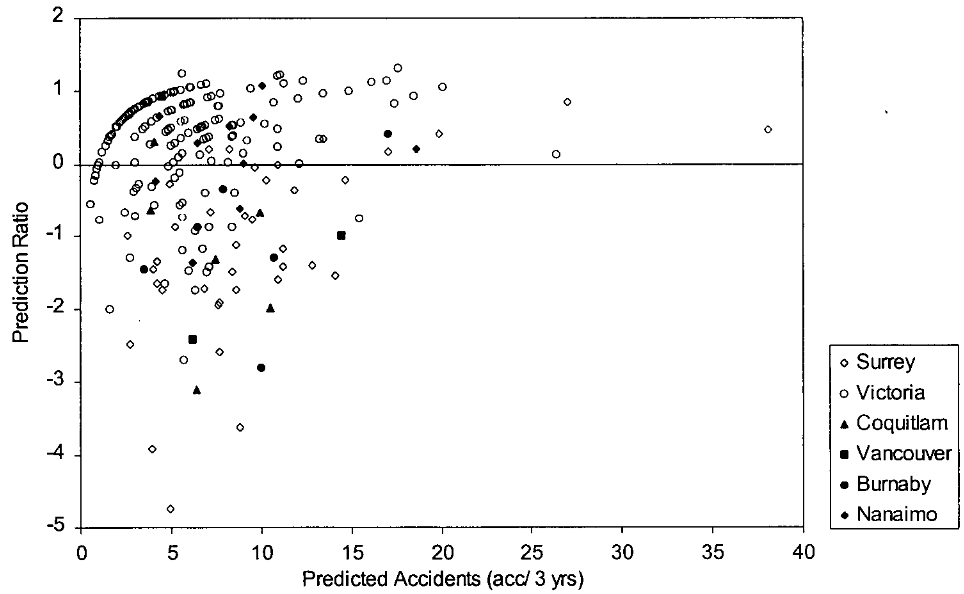


Figure AII-3 AR vs. Accident Frequency for 4-leg intersection Model

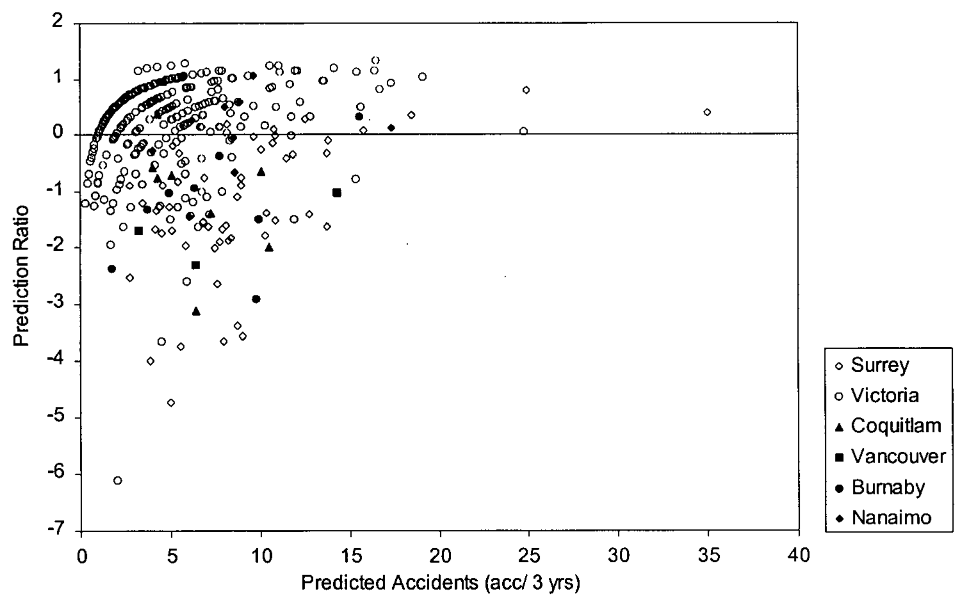


Figure AII-4 AR vs. Accident Frequency for Total Model Including Intersection Type

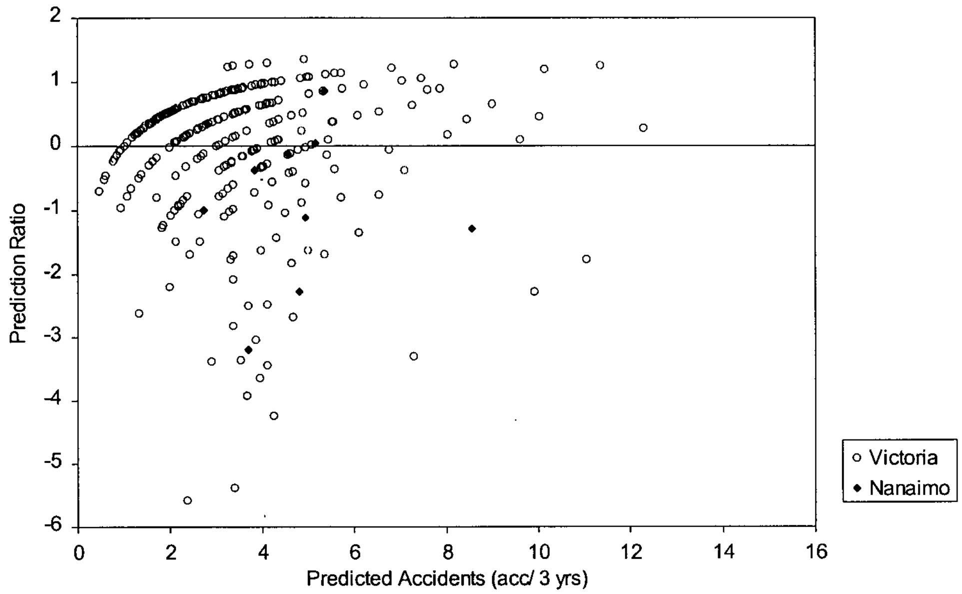


Figure AII-5 AR vs. Accident Frequency for Vancouver Island Model

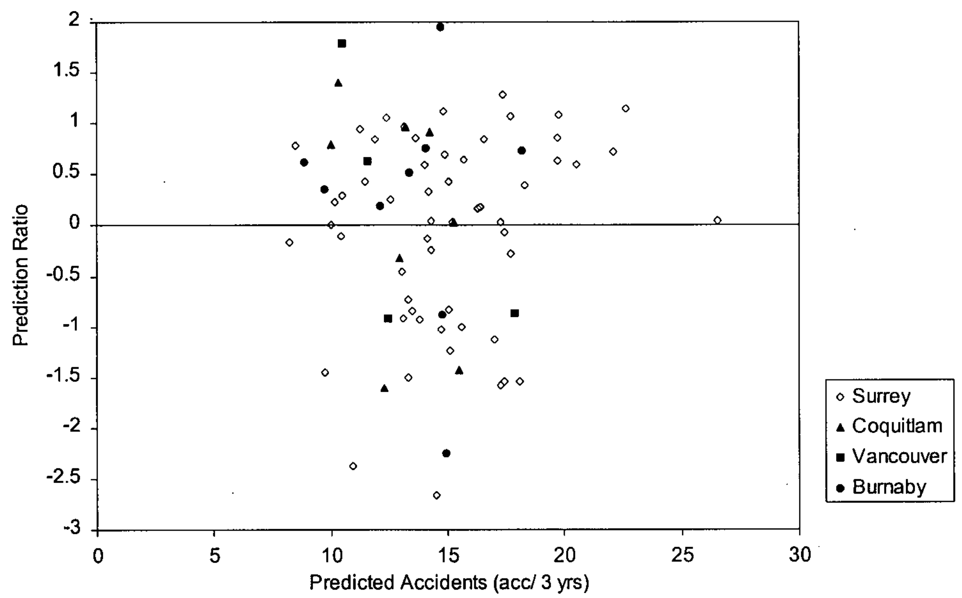


Figure AII-6 AR vs. Accident Frequency for Lower Mainland Model

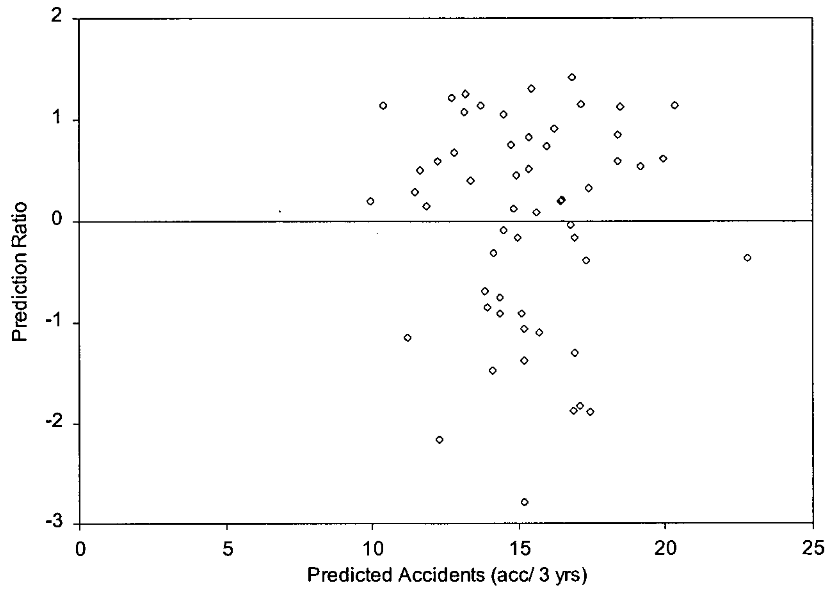


Figure AII-7 AR vs. Accident Frequency for Surrey Total Model

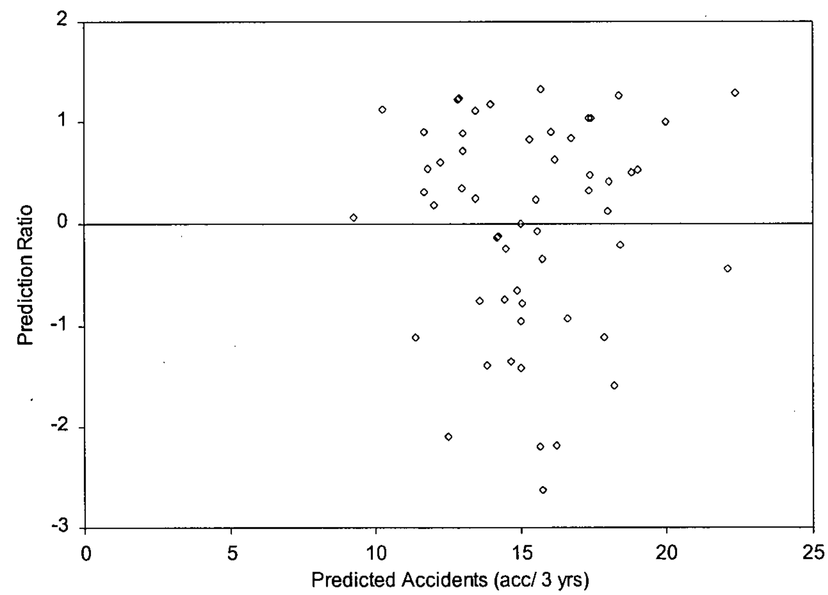


Figure AII-8 AR vs. Accident Frequency for Surrey Model with Control Type

APPENDIX III

GLIM SESSION FOR ESTIMATING APM

```
[o] GLIM 4, update 8 for IBM etc. 80386 PC / DOS  on 28-Oct-1997 at 09:33:59
[o]                  (copyright) 1992 Royal Statistical Society, London
[o]
[i] ? $C ACCIDENT PREDICTION MODELS FOR UNSIGNALIZED INTERSECTIONS$
[i] ? $C TOTAL MODEL$
[i] ? $Units 427$
[i] ? $Data V1 V2 Total Total_3yr Type$
[i] ? $Dinput 'unsigall.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  1576.8 at cycle 4
[o]     residual df =   424
[o]
[o]           estimate        s.e.     parameter
[o]     1       0.3943     0.07586      1
[o]     2       0.4067     0.02814      LV1
[o]     3       0.6086     0.02636      LV2
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
[e] !***********************************************************
[e] !************************************************************
[e] $echo off$
[h]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  398.83 (change =   -1178.) at cycle 3
[o]     residual df =  424    (change =       0 )
[o]
[o]       ML Estimate of THETA =     1.966
[o]                  Std Error = (  0.1828)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    4697. on 424 df
[o]   2 x Full Log-likelihood  =   -2138.
[o]
[o]           estimate        s.e.     parameter
[o]     1       0.4007      0.1233      1
[o]     2       0.3839     0.04940      LV1
[o]     3       0.7044     0.05676      LV2
[o] scale parameter 1.000
[o]
[i] ? $C END OF TOTAL MODEL$
[i] ? $C TOTAL MODEL INCLUDING INTERSECTION TYPE$
```

```
[i] ? $Yvar Total_3yr $Error P $Link L$
[w] -- model changed
[i] ? $Fit LV1+LV2+Type $D E$
[o] scaled deviance =  1435.8 at cycle 4
[o]     residual df =   423
[o]
[o]           estimate        s.e.     parameter
[o]      1     -0.5266      0.1114      1
[o]      2      0.4336     0.02801      LV1
[o]      3      0.5937     0.02721      LV2
[o]      4      0.5268     0.04564      TYPE
[o] scale parameter 1.000
[o]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  394.32 (change =   -1042.) at cycle 2
[o]     residual df =  423    (change =       0 )
[o]
[o]       ML Estimate of THETA =      2.228
[o]                  Std Error = (   0.2170)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    4734. on 423 df
[o]   2 x Full Log-likelihood  =   -2100.
[o]
[o]           estimate        s.e.     parameter
[o]      1     -0.5488      0.1920      1
[o]      2      0.4221     0.04843      LV1
[o]      3      0.6480     0.05534      LV2
[o]      4      0.5379     0.08504      TYPE
[o] scale parameter 1.000
[o]
[i] ? $C END OF TOTAL MODEL INCLUDING INTERSECTION TYPE$
[i] ? $C MODEL FOR T INTERSECTIONS$
[i] ? $Units 186$
[i] ? $Data V1 V2 Total Total_3yr$
[i] ? $Dinput 'unsigt.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  485.10 at cycle 4
[o]     residual df =  183
[o]
[o]           estimate        s.e.     parameter
[o]      1     -0.3980      0.1675      1
[o]      2      0.5809     0.05957      LV1
[o]      3      0.5902     0.04250      LV2
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
[e] !***********************************************************
[e] !************************************************************
[e] $echo off$
```

```
[h]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  163.61 (change =   -321.5) at cycle 4
[o]     residual df =  183    (change =      0  )
[o]
[o]       ML Estimate of THETA =     2.345
[o]                  Std Error = (  0.3825)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    995.9 on 183 df
[o]   2 x Full Log-likelihood  =   -805.6
[o]
[o]           estimate        s.e.     parameter
[o]     1    -0.06907      0.2149      1
[o]     2      0.4531     0.08263      LV1
[o]     3      0.5856     0.07892      LV2
[o] scale parameter 1.000
[o]
[i] ? $C END OF T INTERSECTION MODEL$
[i] ? $C FOUR LEGGED INTERSECTION MODEL$
[i] ? $Units 241$
[i] ? $Data V1 V2 Total Total_3yr$
[i] ? $Dinput 'unsig4.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  942.29 at cycle 4
[o]     residual df =  238
[o]
[o]           estimate        s.e.     parameter
[o]     1      0.6422     0.08496      1
[o]     2      0.3884     0.03194      LV1
[o]     3      0.5944     0.03557      LV2
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
[e] !**********************************************************
[e] !***********************************************************
[e] $echo off$
[h]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  230.30 (change =   -712.0) at cycle 3
[o]     residual df =  238    (change =      0  )
[o]
[o]       ML Estimate of THETA =     2.172
[o]                  Std Error = (  0.2647)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    3740. on 238 df
```

```
[o]   2 x Full Log-likelihood   =   -1293.
[o]
[o]            estimate        s.e.     parameter
[o]      1       0.5275      0.1481      1
[o]      2       0.4099     0.06025      LV1
[o]      3       0.7065     0.07740      LV2
[o] scale parameter 1.000
[o]
[i] ? $C END OF FOUR LEGGED INTERSECTION MODEL$
[i] ? $C ISLAND MODEL$
[i] ? $Units 350$
[i] ? $Data V1 V2 Total Total_3yr$
[i] ? $Dinput 'island.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  734.32 at cycle 4
[o]     residual df =  347
[o]
[o]            estimate        s.e.     parameter
[o]      1       0.2872     0.09310      1
[o]      2       0.3231     0.03632      LV1
[o]      3       0.5240     0.03783      LV2
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
[e] !************************************************************
[e] !************************************************************
[e] $echo off$
[h]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  301.85 (change =   -432.5) at cycle 2
[o]     residual df =  347    (change =      0   )
[o]
[o]       ML Estimate of THETA =      2.920
[o]                  Std Error = (   0.3861)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood        =    985.9 on 347 df
[o]   2 x Full Log-likelihood   =   -1473.
[o]
[o]            estimate        s.e.     parameter
[o]      1       0.3226      0.1224      1
[o]      2       0.3042     0.05025      LV1
[o]      3       0.5488     0.06123      LV2
[o] scale parameter 1.000
[o]
[i] ? $C END OF ISLAND MODEL$
[i] ? $C MAINLAND MODEL$
[i] ? $Units 77$
[i] ? $Data V1 V2 Total Total_3yr$
[i] ? $Dinput 'mainland.txt'$
```

```
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  249.82 at cycle 4
[o]     residual df =   74
[o]
[o]           estimate        s.e.     parameter
[o]     1       1.912      0.1410      1
[o]     2       0.2036     0.04688     LV1
[o]     3       0.2474     0.04259     LV2
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
[e] !************************************************************
[e] !*************************************************************
[e] $echo off$
[h]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  81.024 (change =   -168.8) at cycle 3
[o]     residual df =  74     (change =      0  )
[o]
[o]       ML Estimate of THETA =     6.265
[o]                  Std Error = (   1.486)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    3870. on 74 df
[o]   2 x Full Log-likelihood  =   -505.9
[o]
[o]           estimate        s.e.     parameter
[o]     1       1.886      0.2478      1
[o]     2       0.2011     0.08382     LV1
[o]     3       0.2864     0.07817     LV2
[o] scale parameter 1.000
[o]
[i] ? $C END OF MAINLAND MODEL$
[i] ? $C TOTAL MODEL FOR SURREY$
[i] ? $Units 56$
[i] ? $Da V1 V2 Total_3yr TControl$
[i] ? $Dinput 'unsry.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  151.51 at cycle 3
[o]     residual df =   53
[o]
[o]           estimate        s.e.     parameter
[o]     1       2.148      0.1532      1
[o]     2       0.1529     0.05082     LV1
[o]     3       0.1720     0.05030     LV2
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
```

```
[e] !**********************************************************
[e] !***********************************************************
[e] $echo off$
[h]
[i] ? $Numer theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  55.716 (change =   -95.80) at cycle 3
[o]     residual df =  53     (change =     0   )
[o]
[o]       ML Estimate of THETA =      8.893
[o]                  Std Error = (    2.639)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    3007. on 53 df
[o]   2 x Full Log-likelihood  =   -360.5
[o]
[o]           estimate        s.e.     parameter
[o]      1       2.133      0.2439      1
[o]      2      0.1516     0.08195      LV1
[o]      3      0.1907     0.08343      LV2
[o] scale parameter 1.000
[o]
```
[i] ? $C END OF SURREY TOTAL MODEL$
[i] ? $C MODEL FOR SURREY INTERSECTION INCLUDING TYPE OF CONTROL$
```
[i] ? $Units 56$
[i] ? $Da V1 V2 Total_3yr TControl$
[i] ? $Dinput 'unsry.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2+TControl $D E$
[o] scaled deviance =  149.55 at cycle 3
[o]     residual df =   52
[o]
[o]           estimate        s.e.     parameter
[o]      1       2.191      0.1571      1
[o]      2      0.1647     0.05173      LV1
[o]      3      0.1958     0.05310      LV2
[o]      4    -0.05703     0.04071      TCONTROL
[o] scale parameter 1.000
[o]
[i] ? $Input %plc 80 NEGBIN.MAC$
[e] !**********************************************************
[e] !***********************************************************
[e] $echo off$
[h]
[i] ? $Number theta=0$
[i] ? $Use negbin theta $D E$
[o] scaled deviance =  55.476 (change =   -94.07) at cycle 3
[o]     residual df =  52     (change =     0   )
[o]
[o]       ML Estimate of THETA =      9.096
[o]                  Std Error = (    2.714)
[o]
```

```
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood        =    3008. on 52 df
[o]   2 x Full Log-likelihood   =   -359.4
[o]
[o]            estimate         s.e.     parameter
[o]      1        2.185       0.2491       1
[o]      2       0.1645      0.08265       LV1
[o]      3       0.2256      0.08752       LV2
[o]      4     -0.06994      0.06764       TCONTROL
[o] scale parameter 1.000
[o]
[i] ? $C END OF SURREY MODEL$
[i] ? $Stop
```

APPENDIX IV

GLIM SESSION FOR DIFFERENT NEGATIVE BINOMIAL METHODS

```
[o] GLIM 4, update 8 for IBM etc. 80386 PC / DOS  on 09-Nov-1997 at 01:14:42
 [o]                  (copyright) 1992 Royal Statistical Society, London
 [o]
 [i] ? $!TOTAL MODEL ESTIMATION OF NEGATIVE BINOMIAL DISTRIBUTION PARAMETERS!
 [i] ? $!METHOD OF MAXIMUM LIKELIHOOD (NAG, 1996)!
 [i] ? $units 427$
 [i] ? $data v1 v2 total total_3yr type$
 [i] ? $dinput 'unsigall.txt'$
 [i] ? $calc lv1=%log(v1) : Lv2=%log(V2)$
 [i] ? $Yvar total_3yr $error P $link L$
 [i] ? $Fit LV1+LV2 $D E$
 [o] scaled deviance =  1576.8 at cycle 4
 [o]     residual df =   424
 [o]
 [o]           estimate        s.e.     parameter
 [o]      1      0.3943     0.07586      1
 [o]      2      0.4067     0.02814      LV1
 [o]      3      0.6086     0.02636      LV2
 [o] scale parameter 1.000
 [o]
 [i] ? $input %plc 80 NEGBIN.MAC$
 [e] !**********************************************************
 [e] !  Author:  John Hinde, MSOR Department, University of Exeter
 [e] !           jph@msor.ex.ac.uk
 [e] !  Version: 1.1 GLIM4 February 1996
 [e] !
 [e] !  Main Macros:
 [e] !   NEGBIN  Fits a negative binomial distribution for
 [e] !           overdispersed count data. For details on the
 [e] !           negative binomial distribution see Lawless (1987)
 [e] !           Canadian J. of Stats, 15, 209-225.
 [e] !           The overdispersion parameter theta can be fixed
 [e] !           or estimated, using an inner loop embedded
 [e] !           within the model fitting process. If the
 [e] !           specified parameter value is zero, estimation
 [e] !           is performed using either maximum likelihood (default),
 [e] !           the expected value of the chi-squared statistic
 [e] !           as in Breslow, N.E. (1984) Applied Statistics
 [e] !           33, p38-44, or the mean deviance.
 [e] !
 [e] !           Prior to using this macro the following model
 [e] !           aspects need to be declared:
 [e] !
 [e] !           y-variate:      use $YVAR <yvariate>
 [e] !
 [e] !           model formulae: this will be taken from the last fit
 [e] !                           directive, or can be explicitly set using
```

```
 [e] !                          $TERMS <model formula>
 [e] !
 [e] !            link function:  set using $LINK
 [e] !                            permissible values i, l, s
 [e] !
 [e] !      Formal arguments:
 [e] !            theta   (obligatory) scalar for negative binomial
 [e] !                    parameter estimate
 [e] !                    if theta=0 estimation is performed
 [e] !                    if theta/=0 used as fixed value in negative
 [e] !                                binomial fit
 [e] !            method  (optional) Scalar controlling estimation method when
 [e] !                    appropriate
 [e] !                    1 = maximum likelihood (default if theta=0)
 [e] !                    2 = mean chi-square estimation
 [e] !                    3 = mean deviance estimation
 [e] !                    4 = use fixed value of theta (default if theta/=0)
 [e] !            tol     (optional) Scalar specifies tolerance criterion to
 [e] !                    control convergence of iteration on theta.
 [e] !                    Defaults to 0.0001.
 [e] !                    If tol<=0 then convergence criterion is set to %cc,
 [e] !                    the system convergence criterion.
 [e] !
 [e] !      Output:
 [e] !            Displays the negative binomial deviance, the degrees of
freedom
 [e] !            for the fitted regression model, the estimate of theta, its
 [e] !            standard error when using maximum likelihood estimation,
 [e] !            and values of the log-likelihood. The deviance provides a
 [e] !            goodness-of-fit measure for a negative binomial
 [e] !            distribution with the current value of theta.
 [e] !            When theta is fixed deviance differences can be used to
 [e] !            assess the importance of model terms.
 [e] !            To compare models with different values of theta the
 [e] !            log-likelihood must be used.
 [e] !            In particular, this applies for comparisons with
 [e] !            the standard Poisson model (theta=infinity)
 [e] !            The log-likelihoods are those for the negative binomial
 [e] !            distribution, the full version including the y! terms.
 [e] !
 [e] !      Side Effects:
 [e] !            On exit from the macro the model is still defined with
 [e] !            a negative binomial variance function. Submodels can then
 [e] !            be fitted directly with $FIT directives. This will work
 [e] !            fine following a fixed parameter fit, but should be
 [e] !            used with caution if theta was estimated - use of $RECYCLE
 [e] !            could help things in this case.
 [e] !
 [e] !      Example of use:
 [e] !            $yvar y $link l $terms 11$
 [e] !            $number theta=0 $
 [e] !            $use negbin theta$
 [e] !
 [e] !   NB_OUT   Can be used after subsequent $FIT directives to obtain
 [e] !            output given by NEGBIN, i.e. the estimate of theta, its
```

```
[e] !           standard error for maximum likelihood fits and the
[e] !           log-likelihood values.
[e] !
[e] !     Formal arguments:
[e] !           theta  (obligatory) scalar for negative binomial
[e] !                  parameter estimate
[e] !
[e] !     Example of use:
[e] !           $yvar y $link l $terms 1l$
[e] !           $number theta=0 $
[e] !           $use negbin theta$
[e] !           $recy $fit -1l$
[e] !           $use nb_out$
[e] !
[e] !
[e] !     To delete macros and global variables, type
[e] !           $delete #d_negbin d_negbin $
[e] !
[e] !**********************************************************
[e] $echo off$
[f] ** identifier expected but not found, at [80 NEGBIN.]
[f]
[h] The $INP directive expected an identifier but found the character .
instead.
[h] Check the syntax of the directive.
[h]
[i] ? $number theta=0$
[i] ? $number mode=1$
[i] ? $use negbin theta mode $D E$
[w] -- model changed
[w] -- model changed
[o] scaled deviance =  398.83 (change =   -1178.) at cycle 3
[o]     residual df =  424    (change =       0 )
[o]
[o]       ML Estimate of THETA =     1.966
[o]                  Std Error = (   0.1828)
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood       =    4697. on 424 df
[o]   2 x Full Log-likelihood  =   -2138.
[o]
[o]           estimate        s.e.     parameter
[o]     1      0.4007      0.1233      1
[o]     2      0.3839     0.04940      LV1
[o]     3      0.7044     0.05676      LV2
[o] scale parameter 1.000
[o]
[i] ? $!METHOD OF MEAN CHI-SQUARE (NAG, 1996)!
[i] ? $number theta=0 : mode=2$
[i] ? $use negbin theta mode $D E$
[w] -- model changed
[w] -- model changed
[o] scaled deviance =  370.24 (change =   -1207.) at cycle 3
```

```
[o]     residual df =  424    (change =       0 )
[o]
[o]  Mean Chi-squared estimate of THETA =     1.764
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood        =    4696. on 424 df
[o]   2 x Full Log-likelihood  =   -2139.
[o]
[o]           estimate        s.e.     parameter
[o]      1      0.4030      0.1269      1
[o]      2      0.3827     0.05104      LV1
[o]      3      0.7058     0.05905      LV2
[o] scale parameter 1.000
[o]
[i] ? $!METHOD OF MEAN DEVIANCE (NAG, 1996)!
[i] ? $number theta=0 : mode=3$
[i] ? $use negbin theta mode $D E$
[w] -- model changed
[w] -- model changed
[o] scaled deviance =  424.00 (change =   -1153.) at cycle 2
[o]     residual df =  424    (change =       0 )
[o]
[o]  Mean Deviance estimate of THETA =     2.154
[o]
[o]  NOTE: standard errors of fixed effects do not
[o]        take account of the estimation of THETA
[o]
[o]   2 x Log-likelihood        =    4696. on 424 df
[o]   2 x Full Log-likelihood  =   -2139.
[o]
[o]           estimate        s.e.     parameter
[o]      1      0.3991      0.1201      1
[o]      2      0.3850     0.04803      LV1
[o]      3      0.7023     0.05490      LV2
[o] scale parameter 1.000
[o]
[i] ? $!TOTAL MODEL ESTIMATION OF NEGATIVE BINOMIAL DISTRIBUTION PARAMETERS!
[i] ? $!FOLLOWING THE METHOD OF MOMENTS PROPOSED BY KULMALA(1995) AND
MAYCOCK!
[i] ? $!AND HALL (1984)!
[i] ? $Units 427$
[i] ? $Da V1 V2 Total Total_3yr Type$
[i] ? $Dinput 'unsigall.txt'$
[i] ? $Calc LV1=%log(V1) : LV2=%log(V2)$
[i] ? $Yvar Total_3yr $Error P $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o] scaled deviance =  1576.8 at cycle 4
[o]     residual df =   424
[o]
[o]           estimate        s.e.     parameter
[o]      1      0.3943     0.07586      1
[o]      2      0.4067     0.02814      LV1
[o]      3      0.6086     0.02636      LV2
```

```
[o] scale parameter 1.000
[o]
[i] ? $Number k=1.733$
[i] ? $MACRO NEGBIN!
[i] $MAC? $Calc %va=%fv+(%fv**2)/k$
[i] $MAC? $Calc %di=2*(%yv*%log(%yv/%fv)-(%yv+k)**%log((%yv+k)/(%fv+k)))$
[i] $MAC? $ENDMAC$
[i] ? $Edit 74 Total_3yr 0.0001 : 118 Total_3yr 0.0001 : 228 Total_3yr
0.0001$
[w] -- change to data values affects model
[i] ? $Edit 238 Total_3yr 0.0001 : 375 Total_3yr 0.0001$
[i] ? $Yvar Total_3yr $Error Own NEGBIN $Link L$
[i] ? $Fit LV1+LV2 $D E$
[o]    deviance =  365.66 at cycle 5
[o] residual df =  424
[o]
[o]           estimate        s.e.     parameter
[o]      1      0.4033      0.1184      1
[o]      2      0.3825     0.04765      LV1
[o]      3      0.7061     0.05520      LV2
[o] scale parameter 0.8624
[o]
[i] ? $Number k=1.8474$
[i] ? $Yvar Total_3yr $Error Own NEGBIN $Link L$
[w] -- model changed
[i] ? $Fit LV1+LV2 $D E$
[o]    deviance =  382.23 at cycle 5
[o] residual df =  424
[o]
[o]           estimate        s.e.     parameter
[o]      1      0.4019      0.1190      1
[o]      2      0.3832     0.04779      LV1
[o]      3      0.7054     0.05513      LV2
[o] scale parameter 0.9015
[o]
[i] ? $Number k=1.8472$
[i] ? $Yvar Total_3yr $Error Own NEGBIN $Link L$
[w] -- model changed
[i] ? $Fit LV1+LV2 $D E$
[o]    deviance =  382.20 at cycle 5
[o] residual df =  424
[o]
[o]           estimate        s.e.     parameter
[o]      1      0.4019      0.1190      1
[o]      2      0.3832     0.04779      LV1
[o]      3      0.7054     0.05513      LV2
[o] scale parameter 0.9014
[o]
[i] ? $Stop
```