MODELLING THE SAFETY PERFORMANCE OF ROADWAYS FOR SPECIFIC COLLISION TYPES

by

Tony Chiu, B.Eng., Ryerson 2002

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the Program of

Civil Engineering

Toronto, Ontario, Canada, 2005

© Tony Chiu, 2005

UMI Number: EC53011

UMI®

Author's Declaration

I hereby declare that I am the sole author of this thesis.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

_____

I further authorize Ryerson University to reproduce this thesis by photo copying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____

Modelling the Safety Performance of Roadways for Specific Collision Types

Master of Applied Science, 2005

By Tony Chiu

Department of Civil Engineering
Ryerson University

Abstract

Safety performance functions (SPFs) are rarely developed for specific accident types because this can be very lengthy especially when relevant data are unavailable. Because of this constraint, a factor is applied along with the SPF for all accident types to estimate the safety for specific accidents types. This factor is the proportion of the individual collision type in the entire population of all accidents. However, there is no reason to believe that this factor is a constant which is independent of Annual Average Daily Traffic (AADT). Accordingly, a constant factor and the proportion model are applied to the SPF for all accident types combined to estimate the SPF for specific accident types on both rural road segments and Two-Way Stop-Controlled (TWSC) intersections. The validity of these factors are tested using the state-of-the-art network screening approaches. Furthermore, a detailed investigation on Property Damage Only (PDO) estimates is carried out on certain aspects of safety performance functions, using negative binomial regression. PDO estimates are then evaluated based on three different approaches.

Acknowledgements

Dedication

Dedicated to Him,

whose created the universe.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Acronyms

AADT	Annual Average Daily Traffic
CURE	Cumulative Residual
DES	Detailed Engineering Studies
EB	Empirical Bayes
FHWA	Federal Highway Administration (USA)
GLM	Generalized Linear Modelling
GOF	Goodness of Fit
HSIS	Highway Safety Information System
MAD	Mean Absolute Deviation
MPB	Mean Prediction Bias
MnDOT	Minnesota Department of Transportation
MSE	Mean Squared Error
MSPE	Mean Squared Prediction Error
NB	Negative Binomial
PDO	Property Damage Only
PSI	Potential for Safety Improvement
RTM	Regression to Mean
SPF	Safety Performance Function
TWSC	Two-Way-Stop-Controlled

CHAPTER ONE

INTRODUCTION

1.1 Introduction

In the past several decades, reducing the number of traffic accidents has become a major concern.   Motor vehicle accidents are recognized as a very serious matter in both urban and rural societies.   In the year 2002, Ontario alone reported 244,642 collisions, which resulted in 770 fatalities, 56,516 injuries, and 187,356 property damage collisions. (*17*)   Not only did these collisions cause the victims to suffer but they also burden our society with enormous financial costs.   Many countermeasures have been adopted to improve road safety.   Direct applications, such as vehicle design improvement, were implemented inside the vehicle to enhance the safety of both drivers and passengers. Roadway treatments such as modification of the lane widths and shoulders widths, and rescheduling the timing of traffic signals were applied to the roadways directly to improve road safety.   Other implementations include a graduated licensing program, an impaired driving program, and safety campaigns were phased in to reduce vehicular accidents.

In order to understand and prevent future accidents, one must employ appropriate methods to carefully examine and formulated treatment plans for an entity.   In this paper, statistical safety modelling is applied to rural roadways and Two-Way-Stop-Controlled (TWSC) intersections.   The intention of a statistical safety model is to estimate the expected accident frequency of an infrastructure based on its traits.   This method has been widely adopted in many research projects in the past few decades. (*1-15*)   Details

of statistical safety modelling will be discussed in Chapter Two. One common mistake in road safety analysis is to identify sites for treatment merely based on the accident counts for a short period of time. Targetting an entity with high accident counts over a certain period can be inefficient because accidents take place at random, a phenomenon that is further discussed in Chapter Two.

1.2 SafetyAnalyst

In the last decade, substantial research on statistical safety modelling has been done. However, it is still very challenging to develop a good statistical road safety model because it requires lots of information (i.e. accident counts, traffic flow, etc.). This information is costly and sometimes difficult to obtain. The notion of *SafetyAnalyst* was conceived by Federal Highway Administration (FHWA) in the United States. They realized the need of implementing a state-of-the-art safety management method in engineering practices. *SafetyAnalyst* comprises a set of tools that help engineers to facilitate and improve road safety using state-of-the-art safety management approaches. Currently, it is still under development, with a beta version to be released in year 2006. The *SafetyAnalyst* package is made up of six different modules:

1. Network Screening Tool - identifies sites with promise for safety improvements

2. Diagnosis Tool - analyzes the nature of safety problems at specific sites

3. Countermeasure Selection Tool - assists user in the selection of a proper treatment at specific site

4. Economic Appraisal Tool - assesses the economic implications of one or more of the selected countermeasures

5. Priority Ranking Tool - distinguishes sites which most urgently call for treatment

based on the benefit and cost estimates determined by the economic appraisal tool

6. Evaluation Tool - helps user to conduct a before-and-after study of a treatment

In a nutshell, the purpose of these modules is to assist engineers when they identify problematic sites and to help design an effective treatment for the target sites, using the most up-to-date and appropriate road safety management procedures.

1.3 Objectives of the Study

This thesis is a detailed investigation of some of the unresolved elements concerning the development of safety performance functions (SPFs) for accident prediction in the *SafetyAnalyst* tools. The objectives are three-fold:

i) Comparison of two methods for developing safety performance functions for specific accident types on road segments

ii) Modelling the proportions of specific collision types at TWSC intersections

iii) Detailed investigation of methods for estimating the expected number of Property Damage Only (PDO) accidents

In the first two parts, the objective is to investigate the accuracy of applying the proportion of specific accident type to the total collision model using two different approaches. Only traffic count is used for the covariate in this study because AADT usually explains more than 70% of the variation in accident occurrence. Another reason for doing so is that many jurisdictions do not necessary have the roadways information other than the traffic counts. Thus, it is important to understand the relationship between traffic count and accident frequency in order to give a clear definition in the software *SafetyAnalyst*. In part three of this thesis, a detailed investigation of PDO

estimates is carried out in certain aspects of safety performance functions. The reason for studying this issue is because PDO accident counts contain errors generated during the data collection stage. It is not clear whether calibrating models using the PDO data produce meaningful results. Thus, three different ways to estimate PDO accident are investigated in this thesis to clarify this issue.

1.4 Organization of the Thesis

This thesis is structured into seven chapters. Chapter One contains the background and objective of the research. Chapter Two documents the important findings from literature reviews. The HSIS database, which was used in the research, is described in Chapter Three. Chapters Four, Five and Six are the major findings of this endeavour. Conclusions and a number of recommendations are included in Chapter Seven.

CHAPTER TWO

LITERATURE REVIEW

2.1 Background and Introduction

The literature review focuses on statistical road safety modelling because the content of this thesis lies mainly in this area. In general, a statistical road safety model serves to estimate the safety of road elements based on the information of the roadways. At the beginning stage of statistical road safety modelling, there were no guidelines or rules on how to develop an appropriate model. Many models were developed using the conventional linear regression. However, this method lacks the distributional properties to adequately describe the random, discrete, non-negative and sporadic effects of the accident counts. Research conducted by Miaou and Lum (*11*) suggests that the Poisson distribution is best employed to describe the accident counts over a period of time at a site; its major limitation is that the variance of the accident data is constrained to be equal to the mean, which is interpreted as the true level of safety per unit of time at the site. If the variance of accident frequency is greater than its mean, overdispersion occurs. For this reason, the Poisson distribution does not accurately reflect the variation in accident frequency. As a consequence, the coefficients of the model would be biased and contain erroneous standard errors. These biased coefficients may result in the over- or under-prediction of accident frequency. Alternative ways to address the overdispersion issue are to use the Negative Binomial (NB) distribution or the double Poisson distributions. In this thesis, the NB distribution, which is commonly considered the

most appropriate approach in describing the accident counts in a population of entities, is employed in all the SPFs developed in order to overcome this weakness.

Besides the modelling techniques described above, the goodness of fit (GOF) of the model is of equal importance. In the past, many researchers used various statistical tests to examine the GOF for road safety models. However, the focal point of many of these statistical GOF tests was on the overall fit. Thus, the GOF for the entire wide range of a covariate was often overlooked. Recently, Hauer et al. (5) have developed a graphical method to solve this problem, by examining the residuals and evaluating the GOF for the entire range of a covariate. Detailed discussions for all of the above concepts are presented in this chapter.

2.2 Accident Prediction Model

The purpose of accident prediction models is to estimate the safety of road segments or intersections based on their traits. Accident prediction model is sometimes called safety performance function (SPF) and is developed by statistical regression analysis. In other words, all historical data on accidents, traffic and other characteristics of the road are represented in a model equation. The resultant equation (SPF) enables one to predict the number of accidents (the dependent variable) of an entity as a function of its traits (the independent variables). The generic form of an SPF is shown below:

$$\kappa = f(X_1, X_2, X_3, ..., X_n, \beta_1, \beta_2, \beta_3, ..., \beta_n) \qquad (2.1)$$

where κ is the predicted accidents per unit of time, X_i is the different road traits, and β_i is the parameters of the function. Model parameters are often optimized by the method of maximum likelihood for the case of generalized linear modelling.

In general, calibration of a SPF requires at least two variables: Annual Average

Daily Traffic (AADT) and accidents counts. It is possible to include other variables such as road geometry, human factors, weather conditions, and others, depending on the availability and precision of these data. Since a SPF is used to estimate the accident frequency as a function of road characteristics, it is important to categorize the entities into different subgroups with similar road trait prior to calibration. For example, one must first separate all the entities into different subgroups (i.e. 4-lane urban highway, TWSC intersection, etc.) if a researcher wishes to develop the SPFs for an urban area.

2.3   Model Formulation

One of the major challenges of developing a statistical accident prediction model is to select an adequate functional form. Broadly speaking, the process of developing a SPF is merely curve-fitting in which the modeller attempts to select a function which fits the data well. It is common to judge the validity of the model from the GOF tests. Some researchers have suggested choosing a function which contains as few parameters as possible. Nevertheless, there is no theory to support that this approach has value in road safety.

In calibrating an accident prediction model, measures must be taken to avoid deficiencies which result from the excluded variables and use of inappropriate functional form. Accordingly, researchers typically go through a series of exploratory data analyses in order to determine the adequate independent variables for the model and explain how these variables are entered into the model. In statistical road safety modelling, the exponential, quadratic, power and gamma functions are often used. Following the selection of a functional form, generalized linear modelling (GLM) is commonly used to estimate the parameters of the model. Software packages such as

SAS, GENSTAT, and GLIM are applied, which allow the modeller to specify the different error distributions for the model. In this thesis, parameters are optimized using the GENMOD procedure in SAS.

2.4 Negative Binomial Distribution

Accident counts are random, discrete, sporadic, and non-negative in nature, so they are best described by the Poisson distribution (*6-8, 11, 20*), which is given by:

$$P\{K \mid \kappa\} = \frac{\kappa^K e^{-\kappa}}{K!}$$ (2.2).

where K is the observed accidents in a unit of time at a site and κ is the mean accident. However, the variance of the accident data is restrained by the shortcoming of the Poisson distribution, namely,

$$\mathrm{var}(Y) = \kappa$$ (2.3).

When the variance of the accident data is greater than the mean, it is said to be overdispersed, which implies that the model coefficients and standard errors are misinterpreted. Thus, these model parameters cause the model to either over- or under-estimate accident likelihood. The overdispersion problem can be overcome by using the negative binomial distribution, defined by,

$$P(Y_i = y_i, \alpha, \beta) = \frac{(y_i + \phi + 1)!}{y_i!(\phi - 1)!} \alpha^{\phi} (1 - \alpha)^{y_i}$$ (2.4).

where Y_i is the number of accidents at i location and α, β and ϕ are the parameters of NB distribution. Miaou et al. and Hauer (*11, 25*) have mentioned that overdispersion could stem from several reasons in statistical road safety modelling.

i) Some road traits are excluded from road safety modelling since factors such as human factors, time of day, and vehicle type may not be available in the database.

ii) Data are subject to sampling errors. For example, AADT, which is the average of daily traffic based on a few days of counts throughout the year, may be subjected to estimation error.

iii) The roadway environment is different from site to site at the time data were gathered and recorded. For instance, weather, time of day, and light condition vary.

2.5 The Empirical Bayes Method

Traditionally, prediction of safety of an entity is based on its accident count alone. This results in a regression-to-mean (RTM) bias that occurs when one focuses on the entities where the accident counts seem to be high or low compared to the other sites. For example, treatment sites are often selected based on the high level of accident frequency. In such cases, the safety effect estimates based on a simple before-after comparison of these tend to exaggerate the effectiveness of the countermeasures. Safety estimates that are based entirely on accident counts can be imprecise. The Empirical Bayes (EB) method accommodates this problem with the traditional method of safety estimation. In general terms, the EB method combines the accident counts and the SPF together to obtain a weighted average of the two. This eliminates the RTM bias in addition to improving the precision of an estimate.

2.6 The Validity of the Statistical Road Safety Model

Besides the concerns which were discussed in the previous sections, the validation of the accident prediction model is equally important in order to produce a well-established SPF to fit a set of accident data. In the published literature, there isn't a generally acceptable set of guidelines to test the validity of a model. Wood (*25*) used the mean deviance and Chi-square ratio, Poch and Mannering (*20*) used the likelihood ratio, Miaou and Lum (*11*) adopted *t*-statistics, and Oh et al. (*16*) introduced several GOF measures which examined the external validity of accident prediction models. Recently, Hauer and Bamfo (*5*) have introduced a graphical method to test the GOF, which uses the plot of cumulative residuals against a covariate (e.g. AADT) and the $\pm 2\sigma$ standard deviations to identify how well the model is fit.

In this thesis, the mean deviance and the Chi-square ratio along with the several GOF measures introduced by Oh et al. are employed to test the overall GOF of the SPFs. The advantage of using these measures is that they are focused on the GOF of statistical models to independent data through the investigation of the residuals in different aspect. In addition, the graphical method by Hauer and Bamfo is used to test the GOF for the full range of the AADT variables.

2.6.1 Examine the Overall Goodness of Fits

Oh et al. (*16*) have introduced five measures to examine the GOF of accident prediction models. Much of the text below is extracted directly from this source.

(1) Pearson's Product Moment Correlation Coefficients

The Pearson's product moment correlation coefficient measures the linear correlation between the predicted and observed values. The Pearson's product moment correlation coefficient is given by:

$$r = \frac{\sum (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\left[\sum (Y_{i1} - \bar{Y}_1)^2 \sum (Y_{i2} - \bar{Y}_2)^2 \right]^{0.5}} \tag{2.5}$$

where r is Pearson's product moment correlation coefficient, Y_1 is the observed values, Y_2 is the predicted values, and $\bar{Y}$ is the mean of Y_i observations. When the linear correlation is equal to ± 1, the model has a perfect linear correlation. In contrast, a linear correlation of 0 indicates that the model completely lacks linear association. One would expect the correlation coefficient to be high if the model has an ideal linear correlation and vice versa.

(2) Mean Prediction Bias (MPB)

This statistic measures the magnitude and direction of the average model bias based on the validation data. It is expected that the magnitude of the mean prediction bias would be small when the calibrated model predicts the observed data well. Generally speaking, a negative MPB demonstrates that the model under-predicts crashes on average. Similarly, a positive MPB means that the model over-predicts crashes. The formula of MPB is shown below:

$$MPB = \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)}{n}$$ (2.6)

where n is the sample size of the validation data, $\hat{Y}i$ is the fitted values and Y_i is the observed values.

(3) Mean Absolute Deviation (MAD)

This statistic measures on average the mis-prediction of the model. It is different from MPB because the absolute sign in the equation remains the positive and negative prediction errors. A small value indicates the prediction of the model is very accurate. An ideal value would be zero which represents an immaculately perfect prediction. The equation of MAD is given by:

$$MAD = \frac{\sum_{i=1}^{n}|\hat{Y}_i - Y_i|}{n}$$ (2.7)

where n is the sample size of the validation data, $\hat{Y}i$ is the fitted values and Y_i is the observed values.

(4) Mean Squared Prediction Error (MSPE) and Mean Squared Error (MSE)

MSPE is used to measure the error relative to the validation data, whereas the MSE is employed to assess the model error in relation to the estimation data. MSE and MSPE are given by:

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_1 - p}$$ (2.8)

$$MSPE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_2}$$ (2.9)

12

where n_1 is the estimated data sample size, n_2 is the validated data sample size and p is the number of independent variables. By comparing these two statistical measures, one can determine whether the models are over-fitting or under-fitting the estimation data. An MSPE greater than its corresponding MSE suggests that the model may have been over-fit to the estimation data and the observed relationships may be misleading. It is also possible that a very important variable is omitted in the model or that the model is not well-defined.

2.6.2 Cumulative Residual Plot

The aforementioned GOF measures are only appropriate in determining the models overall GOF. However, an accident prediction model with a reasonable overall GOF may fit well in some ranges of AADT but not so well in the other ranges of AADT. Because this issue often arises in an accident prediction model, it is important to examine both the overall fitness and the GOF for the entire range of AADT to avoid misinterpretation of the statistical results.

In order to understand the overall GOF, it is important to investigate the residuals of the model. A residual is the difference between the predicted and the observed value of an entity. A graph of residuals versus AADT for Minnesota 2-lane undivided rural roadways used in this research is shown in Figure 2.1. The graph does not reveal much information because the residuals are scattered all over the graph. Hence, one cannot conclude that the model chosen would satisfy all ranges of AADT from this simple scatter plot.
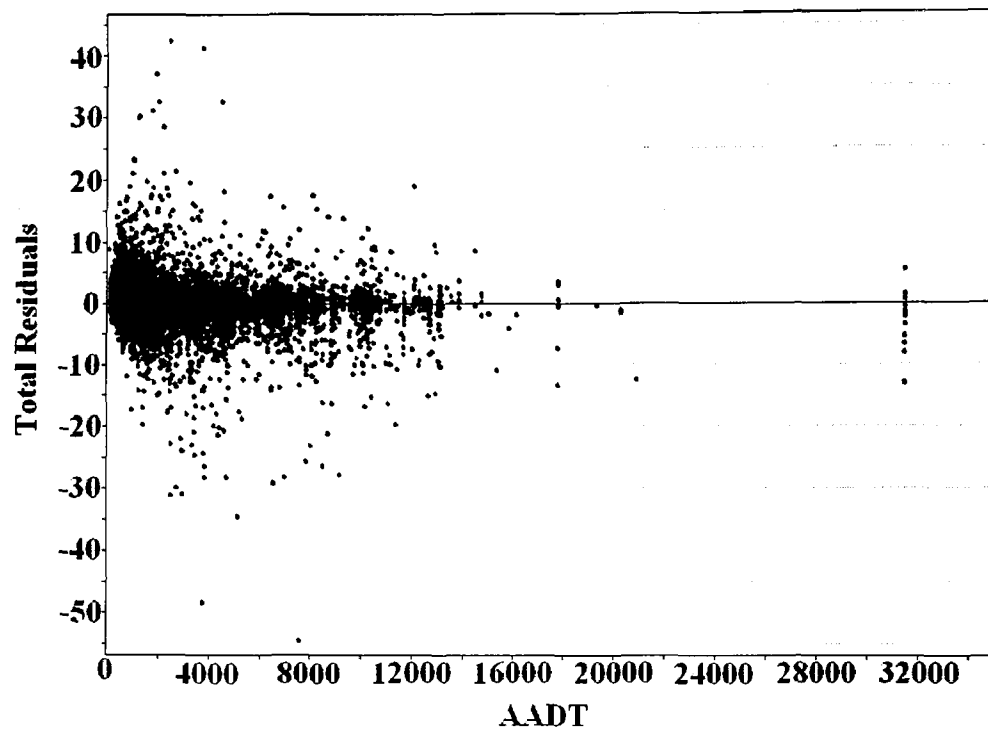
Figure 2.1 – Total Collision Residuals vs. AADT Plot for the Minnesota 2-lane Undivided Rural Roadways

To address this problem, Hauer et al. (*2, 5*) have developed a graphical method termed the cumulative residuals (CURE) procedure to examine the GOF for all ranges of AADT. In order to produce a CURE plot, the residuals of the dataset are calculated and then the dataset is sorted according to the AADT. The next step is to calculate the cumulative residuals. Once completed, cumulative residuals are plotted against the AADT. Figure 2.2 illustrates a CURE plot for Minnesota rural, 4-leg TWSC intersections.

CURE Plot for 4leg Angle Collision
Functional Form #3
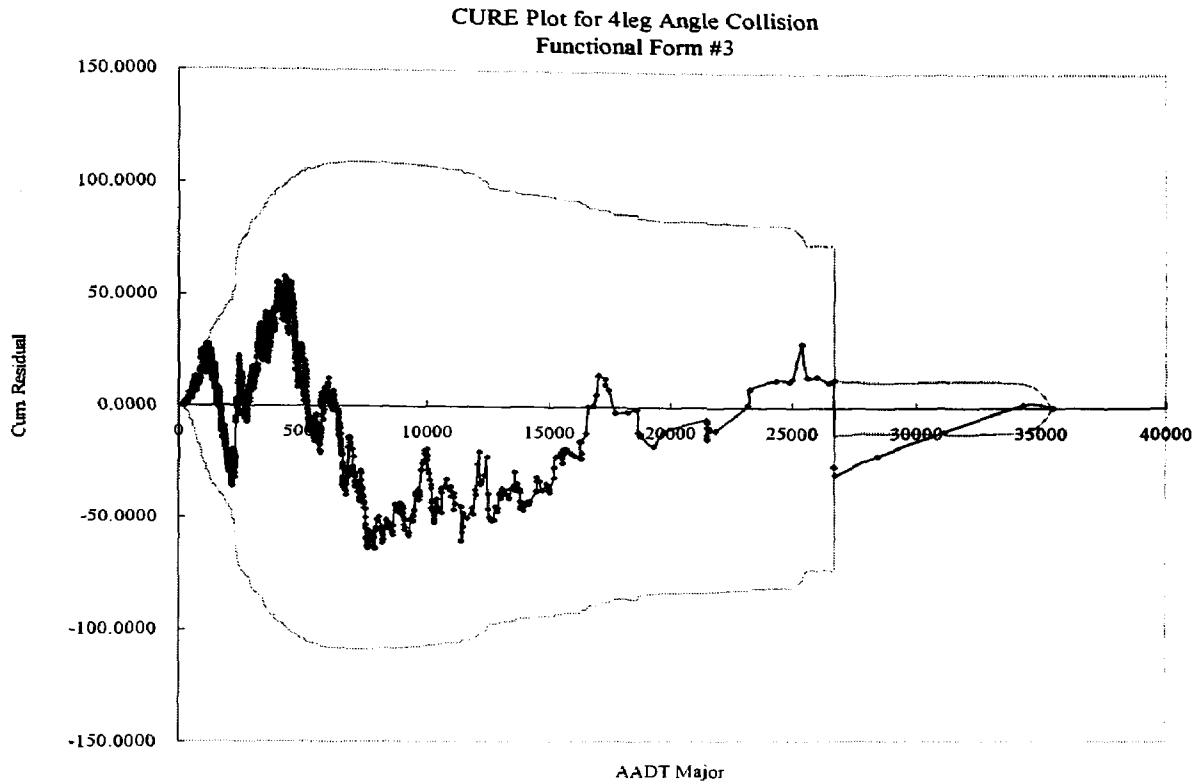
Figure 2.2 – CURE Plot of Angle Collision for Minnesota Rural 4-leg TWSC Intersections

The two dotted lines in Figure 2.2 are the $\pm 2\sigma$ standard deviation limits and it is a tool to help distinguish between what may be expected if the model fits perfectly. The two dotted lines are calculated based on the assumption that the sum of residuals for the model is approximately normally distributed with the mean equalling to zero. With this assumption, two normal probability density functions are established. First, the probability density function for the random walk reaches the point S(n)=s. This function can be written as:

$$P(\text{Current Position}) = \frac{1}{\sqrt{2\pi\sigma(n)}} e^{-\frac{s^2}{2\sigma^2(n)}} \qquad (2.10)$$

with the mean=0 and variance=σ^2(n). Second, the probability density function for return to S(n)=0 at the end point (N) and it is given by:

$$P(\text{Return to the End Point}) = \frac{1}{\sqrt{2\pi\sqrt{\sigma^2(N)-\sigma^2(n)}}} e^{-\frac{s^2}{2[\sigma^2(N)-\sigma^2(n)]}} \qquad (2.11)$$

with the mean=0 and variance=σ^2(N)-σ^2(n). Thus, the probability density function to describe S(n)=s and return to S(n)=0 at the end point is the product of the two normal probability functions and it can be written as,

$$P = \frac{1}{\sqrt{2\pi\sigma(n)}} e^{-\frac{s^2}{2\sigma^2(n)}} \times \frac{1}{\sqrt{2\pi\sqrt{\sigma^2(N)-\sigma^2(n)}}} e^{-\frac{s^2}{2[\sigma^2(N)-\sigma^2(n)]}} \qquad (2.12)$$

Equation 2.12 may be further simplified by representing,

$$\sigma^{*2}(n) = \left(1 - \frac{\sigma^2(n)}{\sigma^2(N)}\right) \qquad (2.13)$$

The exponential components of equation 2.12 can be rewritten as,

$$e^{-\frac{s^2}{2\sigma^2(n)}} \times e^{-\frac{s^2}{2[\sigma^2(N)-\sigma^2(n)]}} = e^{-\frac{s^2}{2\sigma^*(n)^2}} \qquad (2.14)$$

The remaining part of equation 2.12 can be rewritten as,

$$\frac{1}{\sqrt{2\pi\sigma(n)}} \times \frac{1}{\sqrt{2\pi\sqrt{\sigma^2(N)-\sigma^2(n)}}} = \text{constant} \frac{1}{\sqrt{2\pi\sigma^*(n)}} \qquad (2.15)$$

This combined equation 2.15 represents a normal probability density with the mean equal to 0 and the standard deviation equal to σ^*, which can be used to calculate the ±2σ limits shown in Figure 2.2.

Broadly speaking, one would expect the CURE plot to oscillate up and down around the zero mark when the model fits well in all ranges of variables (i.e. AADT in this case). Conversely, one would expect the CURE plot to be all above or below zero when the model is not well-fit. (2)   A few points should be emphasized when analysing the CURE plot.

a) The model should perhaps be modified, for example, by changing the functional form if for a substantial range of AADT the predicted accident from the model is larger than the observed values in that range and vice versa. For instance, the sudden decrease between AADT 4,000 to 5,000 indicates such a case.

b) Where there is a vertical jump in the CURE plot, there is an unusually large residual between the predicted accident and the observed accident count at that location. Hauer calls this sudden jump the 'outlier', which is illustrated in Figure 2.2 where the AADT is near 27,000.

c) The presence of a certain range of AADT where the random walk lies outside the $\pm 2\sigma$ boundaries (the dotted lines in Figure 2.2) indicates that the model is not predicting the observed value accurately within that range.

In summary, the CURE method provides an overview for GOF for all ranges of the covariate. If the model fits well, the cumulative residuals are expected to lie within the $\pm 2\sigma$ boundaries.

2.7 Chapter Summary

From the literature review in this chapter, one may realize that the calibration of SPF requires a lot of efforts, and there are many factors and concerns behind. Therefore, it is worthwhile to elaborate on the development of the SPF. Since the focus of this thesis is on the calibration and application of the SPF, an understanding of the procedures required to develop a meaningful SPF is essential.

CHAPTER THREE

DATA

3.1 General Features of the HSIS Data

The data used in this research originate from the Highway Safety Information System (HSIS) database. The Federal Highway Administration (FHWA) developed the HSIS database by collecting information from nine states and subsequently converting this dataset into SAS format for research purposes. In this thesis, the Minnesota database is used to calibrate the safety performance functions for road segments and intersections. The database consists of four different files containing, respectively, information on accidents, traffic, roadlogs, and intersection characteristics.

3.2 Accident Database

The accident database contains case-by-case basic accident information, vehicle data that include the information on the associated vehicles, and occupant data that contain descriptive and injury information for all occupants involved in the accident. These data can be merged by using the accident report number and vehicle number. They can also be linked to the roadway or intersection file using the three common variables: route system, route number, and reference point.

In Minnesota, the law states that an accident must be reported to the police officer for investigation if there is injury in the crash or property damage over USD $1,000. Fifteen years of accident data are available in the HSIS database, and the ratio of vehicles per crash is between 1.83 and 1.87 for this duration.

The 1995-1999 Minnesota accident and road segments data were used for all the road segments studied, and all accident types are included. 32,400 total collisions, 12,832 fatal/injury collisions, and 19,544 property damage only (PDO) collisions were amassed during these five years on 2-lane undivided rural highways.

The 1995-1999 Minnesota intersection accident data are used in Chapter Five of this thesis and only rural 3-leg and 4-leg TWSC data are included. The database consists of 2,497 total accidents, 1,125 fatal/injury accidents, and 1,372 PDO accidents for 3-leg TWSC, and 6,218 total collisions, 2,823 fatal/injury collisions, and 3,395 PDO collisions for 4-leg TWSC.

3.3 Roadlog Data

The roadlog file consists of all the information of the roadway characteristics; however, the layout of the dataset is different from that of the accident database. Each record in the roadlog file contains information on a length of a segment that has similar road characteristics over that length. In other words, a new record is created whenever there is a change of road traits in a segment; the new starting point is referred to as the beginning reference point. The roadlog file includes information such as the surface type and width, shoulder and median information, lane information, etc. The database is updated daily according to the construction and maintenance plans which are submitted to the Minnesota Department of Transportation (MnDOT).

The roadlog file consists of approximately 134,000 miles of road systems: 12,000 miles of primary (urban) roadways, 33,000 miles of additional state-maintained rural roadways, and 89,000 miles of other non state-maintained county and local roads. Table 3.1 shows the breakdown of roadways from the 1996 HSIS roadlog file. In this research,

the 1995-1999 roadlog files are used, but the 1997 file is omitted because the original file received by HSIS from MnDOT cannot be linked to the crash data due to some errors in the dataset. Therefore, four years of road traits data are included in all the calibration and analysis processes for road segments, which cover a total of 37,358.37 miles of roadways. This number is slightly higher than the one presented in Table 3.1 because the 1999 road file is used as the reference instead of 1996. For intersection analyses, all five years are included.

Table 3.1 – Breakdown of 1996 Minnesota HSIS roadway mileage by category

Roadway Category	Mileage
Urban	
Urban freeways	345.49 mi
Urban freeways < 4 Lanes	3.59 mi
Urban multilane divided non-freeways	666.60 mi
Urban multilane undivided non-freeways	519.88 mi
Urban 2 lane highways	11,616.20 mi
Rural	
Rural freeways	706.10 mi
Rural multilane divided non-freeways	825.16 mi
Rural multilane undivided non-freeways	60.21 mi
Rural 2 lane highways	36,444.34 mi
Other (unpaved/county/local roads)	82,906.66 mi
Total	134,094.22 mi

The Minnesota HSIS database measures roadway length in miles. Therefore, all the equations in this thesis are presented in this unit. Because the only non-metric unit presented in this thesis is mile, all formulae can be simply transformed into metric unit using the conversion factor of 1.00 mi = 1.61 km.

3.4   Intersection File

The intersection file contains information on the major intersections in Minnesota including variables such as approach lane information, entering AADT, and intersection control type, among others, and it is maintained by the eight different MnDOT districts across the state.   Approximately 6,800 intersections are contained in the intersection file; only the rural 3-leg and 4-leg TWSC intersections are included in this thesis.

3.5   Traffic Data

The traffic file contains estimated traffic counts based on the yearly AADT, and is updated every two years.   This estimation is based on a series of counts from both the temporary and permanent count stations throughout the State, and is maintained by the eight different MnDOT district offices.   Similar to the accident file, the traffic file can be linked to the roadlog file using the reference point.   There are approximately 120 automatic traffic recording stations throughout the State, covering approximately 55% of the urban roadways and 45% of the rural roadways.   In addition, there are nearly 34,000 temporary count stations across the State, where 48 hours of counts are made.   In this research, only the sites which contain non-zero AADT for all years are included.

Table 3.2 – Statistical Summary of 2-lane Undivided Rural Roadways for Minnesota HSIS Database

Total Length (mi)	28,728.49
Total # of Sites	32,275
Mean Site Length (mi)	0.89
Mean AADT	1,443.56
Total Collisions	32,400
Fatal/Injury Collisions	12,832
PDO	19,544
Other Collisions	24

Table 3.3 - Statistical Summary of 3-leg & 4-leg TWSC Intersections for Minnesota HSIS Database

3 Leg	
Total # of Sites	2,033 (966 Used)
Mean Entering AADT	5,435 (7,068)
Total Collisions	2,497
Fatal/Injury Collisions	1,125
4 Leg	
Total # of Sites	2,572 (1,649 Used)
Mean Entering AADT	4,512 (5,582)
Total Collisions	6,218
Fatal/Injury Collisions	2,823

CHAPTER FOUR

COMPARISON OF TWO METHODS FOR DEVELOPING SAFETY PERFORMANCE FUNCTIONS FOR SPECIFIC ACCIDENT TYPES ON ROAD SEGMENTS

4.1 Introduction

A safety performance function (SPF) is a mathematical model which relates the accident frequency of a road to its traits. The SPF is very powerful because of its ability to be utilized in network screening and in evaluation of safety treatments. Typically SPFs are developed for all accident types combined but there is a need for SPFs for specific accident types (e.g., run-off-road, head-on) so that network screening and treatment evaluation in *SafetyAnalyst* (*22*) and similar efforts can be so disaggregated. However, calibrating these can be very time-consuming especially if relevant data are not available in sufficient quantities. Because of this difficulty, a factor is applied in the initial version of *SafetyAnalyst* to the SPF for all accident types combined to estimate the SPF for a specific accident type. The factor is the proportion of that accident type in the population of all accidents. However, it is not known how these simplistic SPFs models stack up against those that might be obtained by using data and negative binomial regression to directly estimate SPFs for the specific accident type. The goal of the work described in this chapter is to compare the two methods for estimating SPFs for specific accident types. The scheme for this methodology is illustrated in Figure 4.1. The two methods were applied to get SPFs to obtain the expected accident counts for the two-lane undivided road segments, and the HSIS database from the state of Minnesota was used. Various statistical measures along with the CURE plots were used to analyze the GOF for

the models as described in Chapter Two. The purpose was to determine which of the two methods more suitable in predicting accidents for a specific collision type.

```
            ┌────────────────────────┐
           /   Database:             /
          /    HSIS Minnesota       /
          \    Rural, 2-lane        \
           \   undivided roadways    \
            └────────────────────────┘
                        │
                        ▼
            ┌────────────────────────┐
            │ Divide accident data   │
            │ into 18 different      │
            │ collision types        │
            └────────────────────────┘
                        │
                        ▼
            ┌────────────────────────┐
            │ Calibrate SPF for      │
            │ Total Accident         │
            └────────────────────────┘

┌──────────────────────┐      ┌──────────────────────────┐
│ Calibrate SPFs for   │      │ Apply "Factors" to the   │
│ Specific Accident    │      │ SPF for total accident   │
│ Type                 │      │ to estimate the SPFs     │
│                      │      │ for specific accident    │
│                      │      │ type                     │
└──────────────────────┘      └──────────────────────────┘

            ┌────────────────────────┐
            │ Compare the two        │
            │ methods using relevant │
            │ statistical measures & │
            │ CURE Plots             │
            └────────────────────────┘
                        │
                        ▼
                   ◇ Which one is
                     better? ◇
```
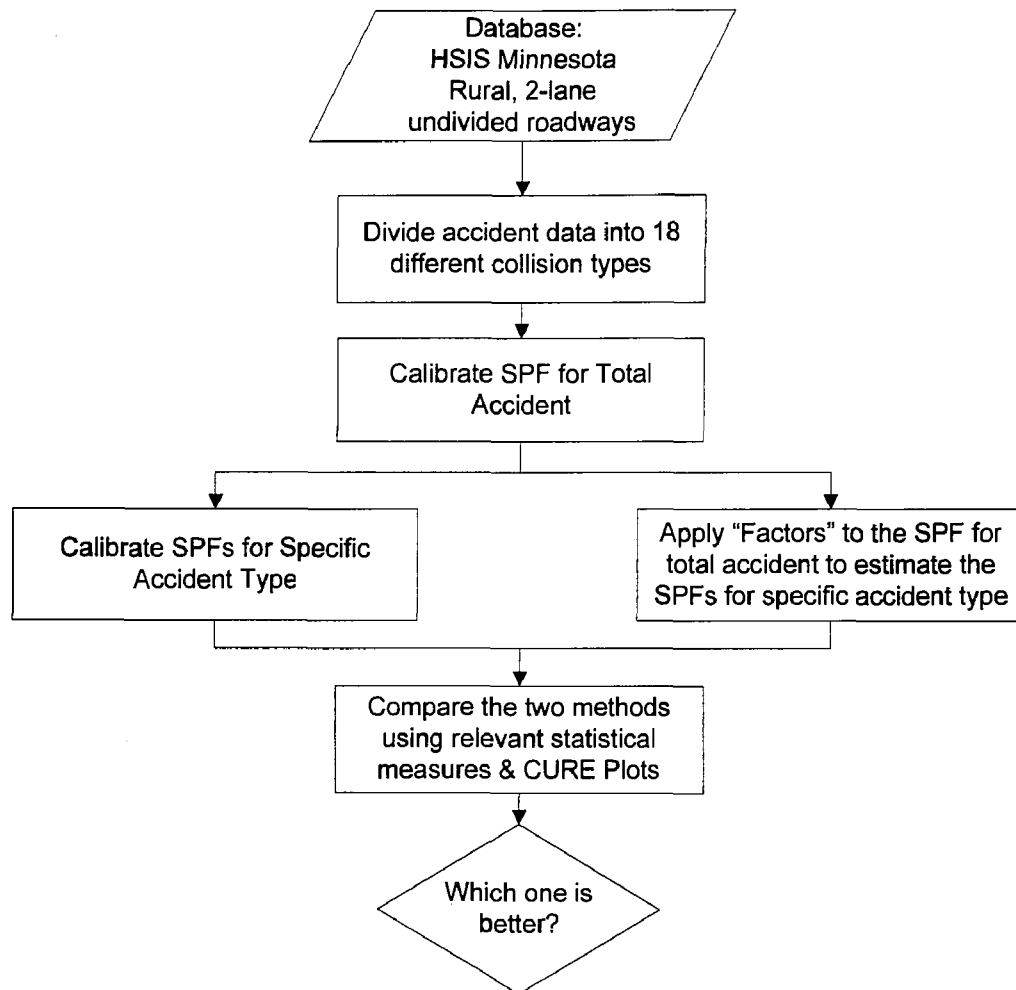
Figure 4.1 – Modelling Development Process

4.2 Data Assembly

Data from 1995 to 1999 taken from the HSIS Minnesota database, as described in the previous chapter, were used in the SPFs calibration process. The roadway database from HSIS classifies roadways based on their road geometry. Each year it may contain different numbers of road segments due to the development, separation, or combination of roadways. The rearrangement of the roadway database based on one particular year was needed in order to proceed further in the SPFs calibration. The year 1999 roadway classification was used as a reference base for the newly-merged database because it represents the most updated roadways condition in Minnesota. The aggregation work was done in the statistical software package SAS, using scripts to rearrange the database into a workable format. Collisions were then divided into different groups for the SPF calibration purpose.

The resulting database contains a total of 44,894 two-lane undivided rural roadway accidents which occurred in Minnesota between 1995 and 1999. However, not all of the 44,894 accidents were included in the SPF calibration process due to missing information such as AADT for one or more years at several sites. After eliminating all sites that which may potentially cause inaccuracy in the regression analysis, 32,275 sites were left for the calibration process. A statistical summary of the Minnesota two-lane undivided rural roadways is shown in Table 4.1.

Table 4.1 - Statistical Summary of 2-lane Undivided Rural Roadways in HSIS Minnesota Database

Total Length (mi)	28,728.49
Total # of Sites	32,275
Mean Site Length (mi)	0.89
Mean AADT	1,443.56
Total Collisions	32,400
Fatal/Injury Collisions	12,832
PDO	19,544
Other Collisions	24

Table 4.2 – The Eighteen Different Collision Types Used for Calibrating SPFs

Collision Type	Definition	# of Accidents
Total	Total Collision	32,400
1	Collision with parked motor vehicle	1,015
2	Collision with railroad train	48
3	Collision with bicyclist	143
4	Collision with pedestrian	230
5	Collision with animal	6,938
6	Collision with fixed object	6,362
7	Collision with other object	130
8	Other single vehicle collision	84
9	Overturn	7,803
10	Fire or Explosion	122
11	Other single vehicle non collision	1,074
21	Rear end	2,866
22	Head on	874
24	Angle	1,450
25	Sideswipe, same direction	1,039
26	Sideswipe, opposite direction	853
27	Other multiple vehicle collision	1,300
99	Unknown	69

4.3 SPFs Developed by Negative Binomial Regression Using Generalized Linear Modelling

As discussed in Chapter Two, the negative binomial model form is the most commonly adopted one in the development of SPFs. The advantage of using the negative binomial model is that the overdispersion can be represented through an overdispersion parameter k that is estimated iteratively in the model calibration process. The inverse of overdispersion parameter γ is such that:

$$\text{var}\{K\} = \frac{E\{K\}^2}{\gamma}$$ (4.1)

where $E\{K\}$ is the expected accidents and $var\{K\}$ is the variance. Note that variance decreases as γ increases; thus, the overdispersion parameter can be used to compare the goodness of fit for various models fitted with the same dataset. A large γ indicates a small variance and a better model, and vice versa. As for SAS, the output of overdispersion parameter from the GENMOD procedure is the inverse of γ.

Separate SPFs were developed for eighteen different collision types using the GENMOD procedure in SAS. Table 4.2 shows some details for the 18 different collision types. In view of the fact that the intention of this thesis is not to identify additional unknown variables but to observe which one of the two methods would produce better possible predictive results, AADT was the only independent variable used for calibrating the SPFs. Equation 4.2 shows the simple form with AADT as the independent variable.

$$\kappa = f(AADT) = \alpha(AADT)^\beta$$ (4.2)

where κ is the accidents/mile/year, AADT is the average annual daily traffic, and α and β are the coefficients estimated in the SPF calibration process. The results of the calibration are shown in Table 4.3.

Table 4.3 – Calibration Results of the AADT Model for all SPFs

Model Form: Accidents/mile/year $= \alpha(AADT)^\beta$

Collision Type	# of Accidents	β (S.E.)*	ln α (S.E.)	Overdispersion
Total	32,400	0.9718 (0.0079)	-8.0658 (0.0566)	0.6033
1	1,015	1.0687 (0.0443)	-11.6260 (0.3197)	17.8965
2	48	0.4144 (0.1637)	-10.2450 (1.1231)	246.5103
3	143	0.8503 (0.0735)	-12.6741 (0. 5454)	0.0038
4	230	1.0456 (0.0664)	-13.5219 (0.4919)	7.3936
5	6,938	0.9383 (0.0155)	-9.4727 (0.1117)	1.6576
6	6,362	0.8576 (0.0143)	-8.8982 (0.1026)	0.9947
7	130	1.3048 (0.0847)	-16.5162 (0.6764)	0.0001
8	84	1.0358 (0.1124)	-14.3892 (0.8257)	30.1253
9	7,803	0.6733 (0.0121)	-7.4282 (0.0852)	0.5724
10	122	1.1913 (0.0843)	-15.4071 (0.6597)	0.0128
11	1,074	Parameters not significant		
21	2,866	1.3464 (0.0181)	-13.4651 (0.1447)	0.0076
22	874	1.1627 (0.0313)	-13.2171 (0.2439)	0.0003
24	1,450	1.1834 (0.0288)	-12.6927 (0.2160)	2.3679
25	1,039	1.1096 (0.0284)	-12.6356 (0.2195)	0.0002
26	853	Parameters not significant		
27	1,300	1.0087 (0.0285)	-11.5385 (0.2110)	1.8179
99	69	Parameters not significant		

* S.E. denotes standard error

4.4 SPFs Developed by Applying a Factor to the SPF for all Accident Types Combined – The "Proportion" Method

Another approach for predicting accidents for each individual collision type is the "proportion" method.   This method estimates the proportion of each collision type in the population of all accidents and applies it to the model the total accident model for all accidents combined.   The following mathematical equation illustrates how the ratio is actually applied to the total collision model.

$$\kappa(accidents / mile / year) = (proportion) \times \alpha(AADT)^\beta \qquad (4.3)$$

where α and β are coefficients for the "total accident" model, estimated to be 0.000314 and 0.9718, respectively.   The overdispersion parameter k for this model was found to be 0.6033, which indicates that the model is reasonable.

SPF for Total Accidents
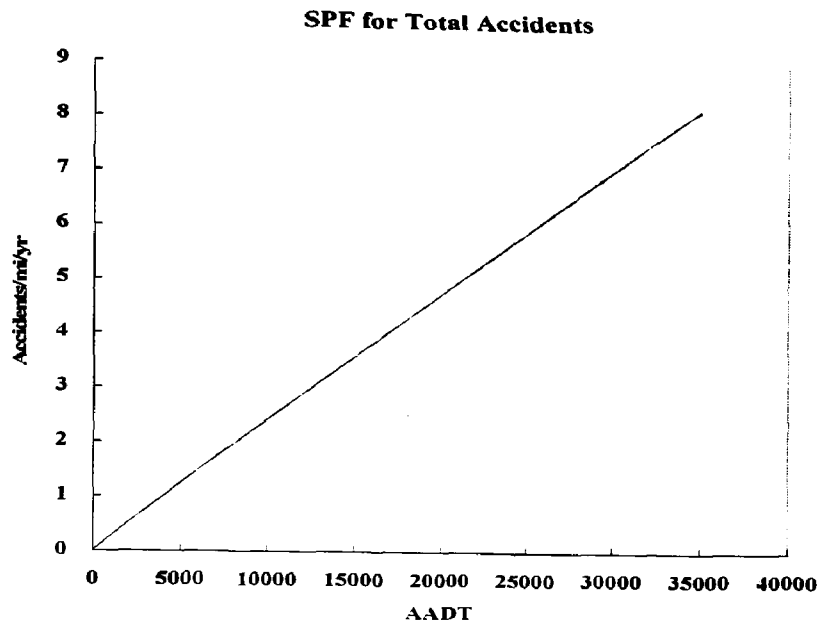
Fig. 4.2 – SPF for Total Accidents on Minnesota 2-lane Undivided Rural Roadways

Figure 4.2 illustrates the SPF for the total accidents on Minnesota two-lane undivided rural roadways. Table 4.4 shows the proportions of all collision types.

Table 4.4 – Proportion of Accidents by Types

Definition	Collision Type	# of Accident	Proportion
Collision with parked motor vehicle	1	1,015	0.031
Collision with railroad train	2	48	0.002
Collision with bicyclist	3	143	0.004
Collision with pedestrian	4	230	0.007
Collision with animal	5	6,938	0.214
Collision with fixed object	6	6,362	0.196
Collision with other object	7	130	0.004
Other single vehicle collision	8	84	0.003
Overturn	9	7,803	0.241
Fire or Explosion	10	122	0.004
Other single vehicle non collision	11	1,074	0.033
Rear end	21	2,866	0.089
Head on	22	874	0.027
Angle	24	1,450	0.045
Sideswipe, same direction	25	1,039	0.032
Sideswipe, opposite direction	26	853	0.026
Other multiple vehicle collision	27	1,300	0.040
Unknown	99	69	0.002

4.5 Comparison of the Results

The objective here is to compare the two methods for estimating SPFs for specific accident types to determine which one is better. To achieve the aforementioned objective, the goodness of fit statistics was used to determine the overall fit of the models. Notwithstanding that, these goodness of fit statistics only provide a basic measure of the overall fit of the model. They cannot specify whether the model is suitable for all practical values of a variable. For this reason, SPFs have in the past often failed to provide reasonably predictions for all ranges of AADT. The CURE procedure discussed in Chapter Two can address this issue by providing an assessment of how well the model fits in the full range of AADTs. Figure 4.3 shows the CURE plot for the total collision model and the two dashed curves are the $\pm 2\sigma$ boundaries.

The cumulative residuals plot reveal that the total collision model does not predict the crashes very accurately because the random walk of the cumulative residuals exceeds the $\pm 2\sigma$ boundaries (the dashed lines) in many ranges of AADT. The main reason why the cumulative residuals exceeds $\pm 2\sigma$ boundaries is that there are many short locations with zero accident in the database. These zero accident locations have caused the cumulative residuals to increase or decrease dramatically. Nevertheless, the peak values for the cumulative residual curve were found to be +750 and -1,700 with a total of 32,400 accident counts. With such high numbers of accidents, the peak values seem acceptable and one can conclude that the calibrated model is reasonable. The CURE plot indicates that the model overestimates crashes at very low AADT and between AADT 3,000-10,000 while it underestimates crashes between AADT 1,000-2,000 and those AADT is above 10,000.
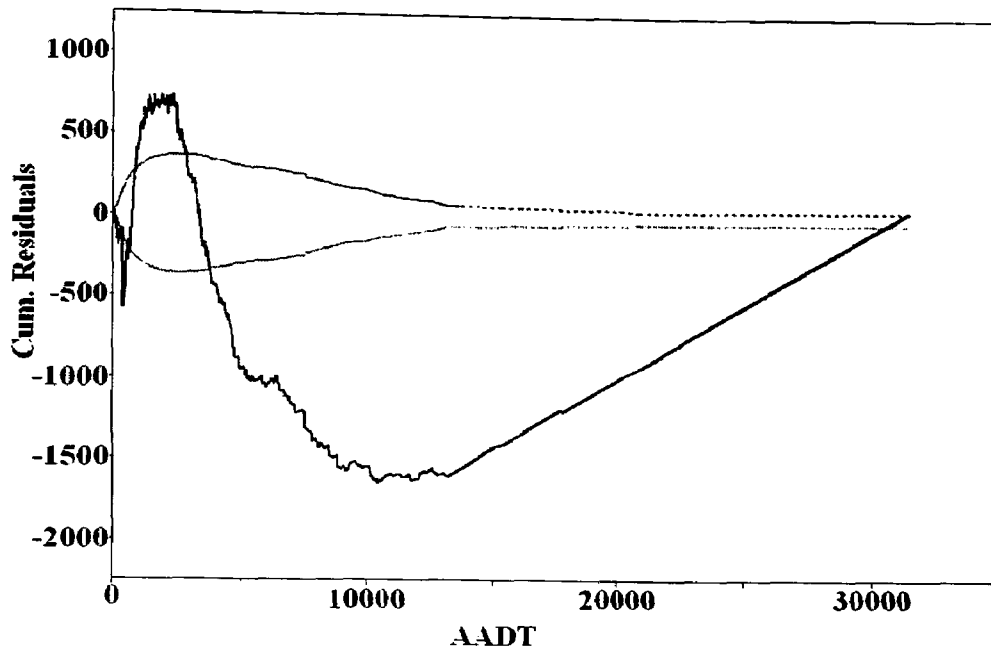
Fig. 4.3 – Cumulative Residuals Plot for Total Collision Model

The goodness of fit measures and CURE procedure were used to analyze all the safety performance functions for both methods. The goodness of fit statistics are displayed on Tables 4.5 and 4.6 for Minnesota rural, two-lane undivided roadways. The upper portion of Table 4.5 shows the parameters estimation generated from SAS with their standard errors indicated by the round brackets below the parameters. The lower portion of Table 4.5 shows the goodness of fit statistics for both methods.

All the accident types with less than 300 accident counts, along with the "other" single vehicle and other multi-vehicle collisions are excluded from the analysis because less accident counts would produce meaningless result and would be moot anyway since such types are unlikely to be of interest in SPF applications. For single vehicle accidents, only five types of collision were left for further analysis. Similarly, only six types of collision were left for multi-vehicle accidents. Tables 4.7 and 4.8 summarize the findings for the remaining collision types, for single vehicle and multi-vehicle accidents respectively.

31

Table 4.5 – Validation Statistics for Single Vehicle Accidents' Models (Goodness of Fit were calculated as per mile-year base)
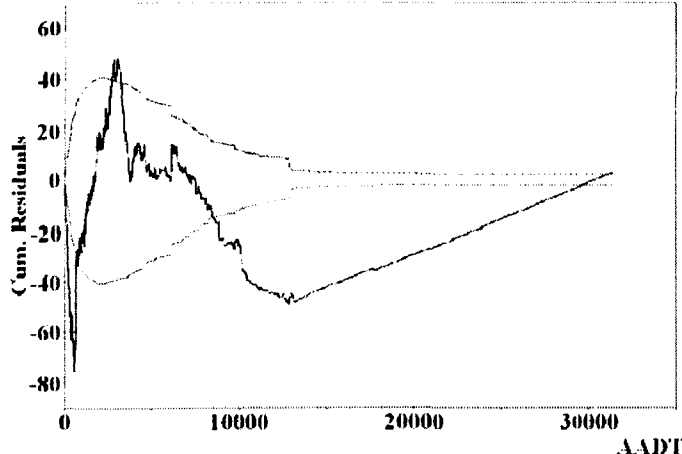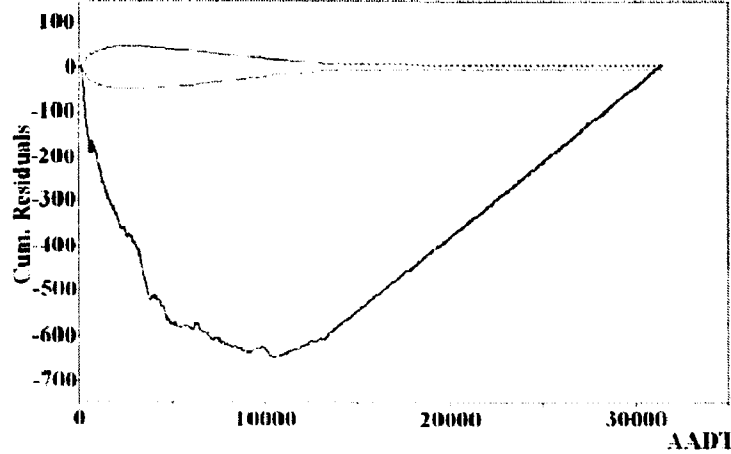
Model		Total Collision	Type 1 Collision w/ Parked Motor Vehicle	Type 5 Collision w/ Animal	Type 6 Collision w/ Fixed Object	Type 9 Overturn
# of Accidents		32.400	1.015	6.938	6.362	7.803
ln α (S.E.)		-8.0658 (0.0566)	-11.6260 (0.3197)	-9.4727 (0.1117)	-8.8982 (0.1026)	-7.4282 (0.0852)
β (S.E.)		0.9718 (0.0079)	1.0687 (0.0443)	0.9383 (0.0155)	0.8576 (0.0143)	0.6733 (0.0121)
Overdispersion (S.E.)	SPF	0.6033 (0.0160)	17.8965 (1.3213)	1.6576 (0.0668)	0.9947 (0.0496)	0.5724 (0.0325)
	Proportion	-	0.048	0.015	0.022	0.023
Proportion		1.000	0.031	0.214	0.196	0.240
Pearson's Correlation Coefficients	Proportion	-	0.086	0.519	0.529	0.579
	SPF	0.723	0.088	0.521	0.539	0.631
Mean Prediction Bias (MPB/Mile-Year)	Proportion	-	-0.024	0.010	-0.015	0.029
	SPF	-0.073	-0.013	0.000	-0.022	0.009
Mean Absolute Deviation (MAD/Mile-Year)	Proportion	-	0.046	0.120	0.133	0.119
	SPF	0.501	0.056	0.114	0.128	0.104
Mean Squared Prediction Error (MSPE/Mile-Year ^{2})	Proportion	-	1.330	1.481	6.566	1.352
	SPF	15.421	1.330	1.481	6.567	1.346
Mean Squared Error (MSE/Mile-Year ^{2})	Proportion	-	1.330	1.481	6.566	1.352
	SPF	15.422	1.330	1.481	6.567	1.346

Table 4.6 –Validation Statistics for Multi-Vehicles Accidents' Models (Goodness of Fit were calculated as per mile-year base)

Model		Type 21 Rear-End	Type 22 Head On	Type 24 Angle	Type 25 Sideswipe, Same Direction
# of Accidents		2,866	874	1,450	1,039
ln α (S.E.)		-13.4651 (0.1447)	-13.2171 (0.2439)	-12.6927 (0.2160)	-12.6356 (0.2195)
β (S.E.)		1.3464 (0.0181)	1.1627 (0.0313)	1.1834 (0.0288)	1.1096 (0.0284)
Overdispersion (S.E.)	SPF	0.0076 (0.0000)	0.0003 (0.0000)	2.3679 (0.2329)	0.0002 (0.0000)
	Proportion	0.040	0.500	0.080	0.182
Proportion		0.088	0.027	0.045	0.032
Pearson's Correlation Coefficients	Proportion	0.425	0.369	0.275	0.315
	SPF	0.442	0.374	0.281	0.311
Mean Prediction Bias (MPB/Mile-Year)	Proportion	-0.020	-0.002	-0.013	-0.016
	SPF	-0.017	-0.002	-0.009	-0.017
Mean Absolute Deviation (MAD/Mile-Year)	Proportion	0.075	0.020	0.043	0.038
	SPF	0.075	0.020	0.045	0.038
Mean Squared Prediction Error (MSPE/Mile-Year ^{2})	Proportion	0.447	0.021	0.117	2.519
	SPF	0.444	0.021	0.117	2.519
Mean Squared Error (MSE/Mile-Year ^{2})	Proportion	0.447	0.021	0.117	2.519
	SPF	0.444	0.021	0.117	2.519

Table 4.7a – Summary of Findings for Type 1: Collision with Parked Vehicle
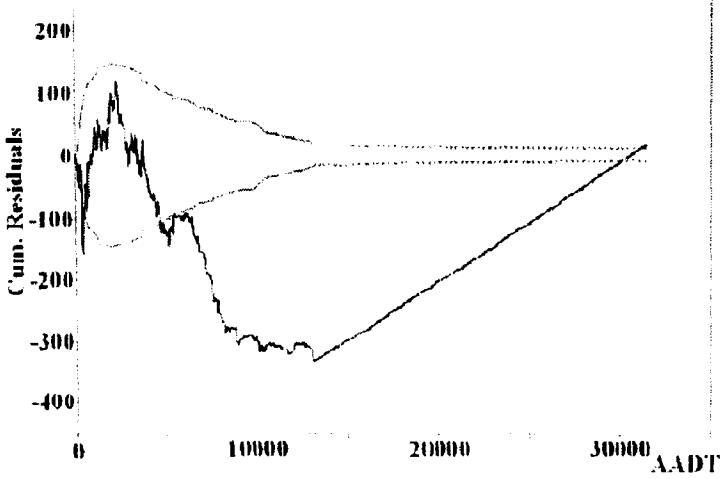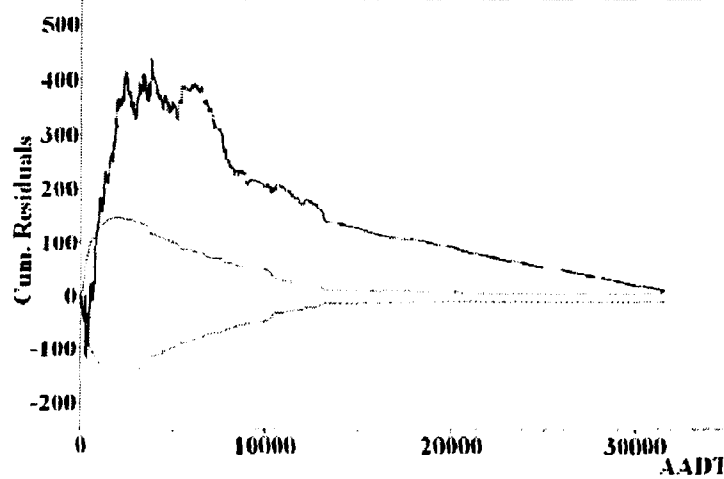
Observed Counts: 1,015

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	+48	0
Lower Peak:	-75	-640
Regions Exceed ±2σ	0-1000, 3000-3500 & >8000	Whole Range
Linear Correlation	Lack of Linear Correlation	Lack of Linear Correlation
Average Model Bias	Under-predict crashes	Under-predict crashes (Relatively Closer to Zero)
Mis-prediction	Slightly Better	-
Overdispersion	0.0480	17.8965

Table 4.7b – Summary of Findings for Type 5: Collision with Animal

Observed Counts: 6,938

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	+120	+440
Lower Peak:	-320	-120
Regions Exceed ±2σ	>4,800	>1,200
Linear Correlation	Average Linear Correlation	Average Linear Correlation
Average Model Bias	Over-predict crashes	Over-predict crashes (Relatively Closer to Zero)
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.0150	1.6576
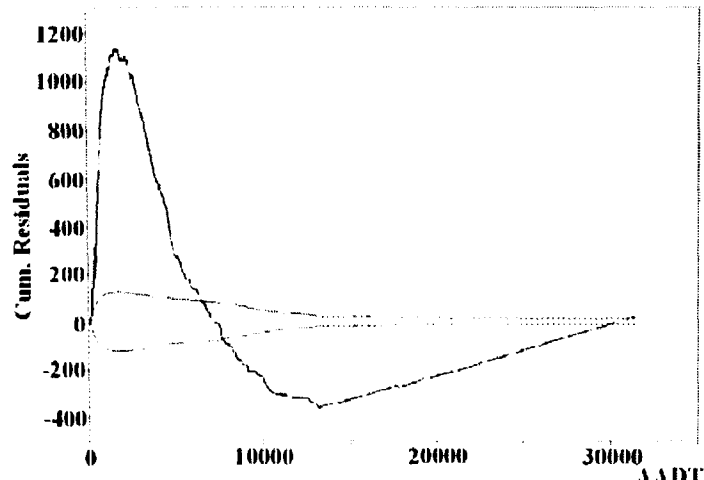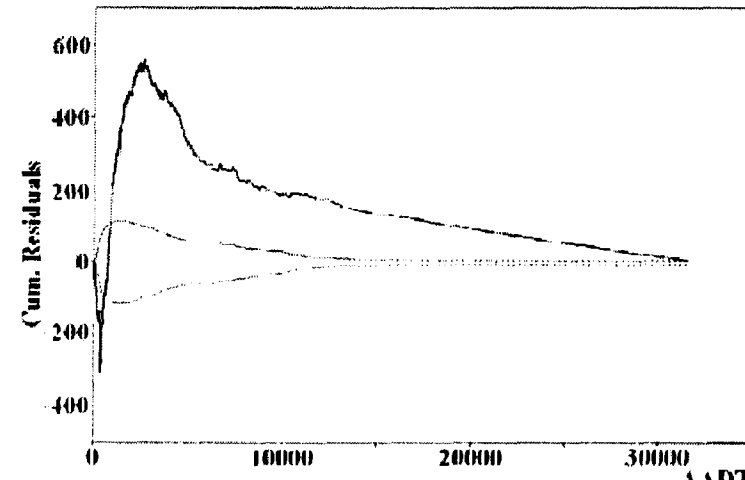
Table 4.7c – Summary of Findings for Type 6: Collision with Fixed Object

Observed Counts: 6,362

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	+420	+250
Lower Peak:	-320	-230
Regions Exceed ±2σ	1,000-4,000 & >6000	0-1,000, 1,500-4,000 & >8,000
Linear Correlation	Average Linear Correlation	Average Linear Correlation
Average Model Bias	Under-predict crashes (Relatively Closer to Zero)	Under-predict crashes
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.0220	0.9947

36

Table 4.7d - Summary of Findings for Type 9: Overturn

Observed Counts: 7,803

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	+1,100	+560
Lower Peak:	-380	-300
Regions Exceed ±2σ	0-6,500 & >7,500	0-600 & >1,000
Linear Correlation	Average Linear Correlation	Average Linear Correlation
Average Model Bias	Over-predict crashes	Over-predict crashes (Relatively Closer to Zero)
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.0230	0.5724

37

Table 4.8a – Summary of Findings for Type 21: Rear End

Observed Counts: 2,866
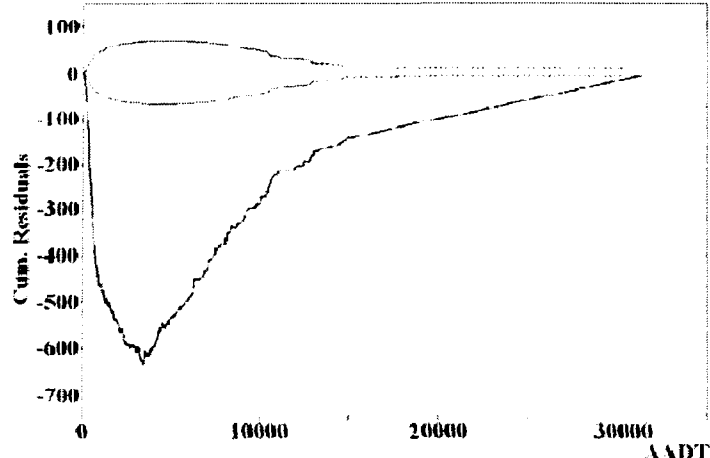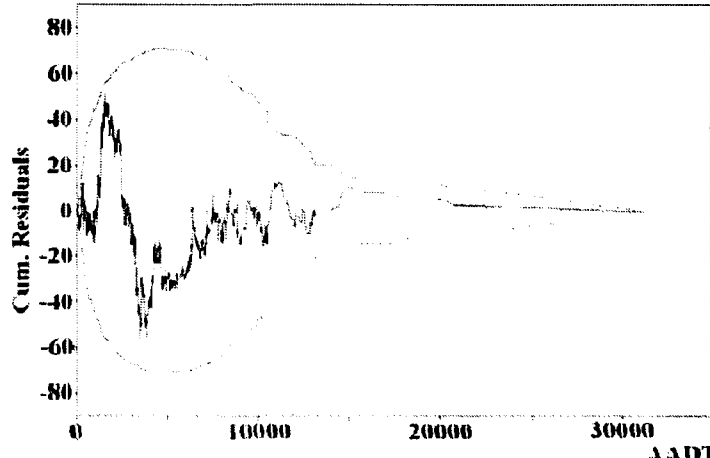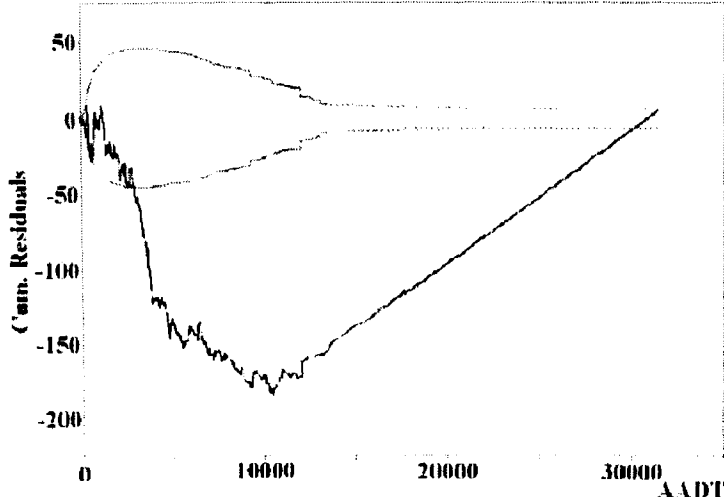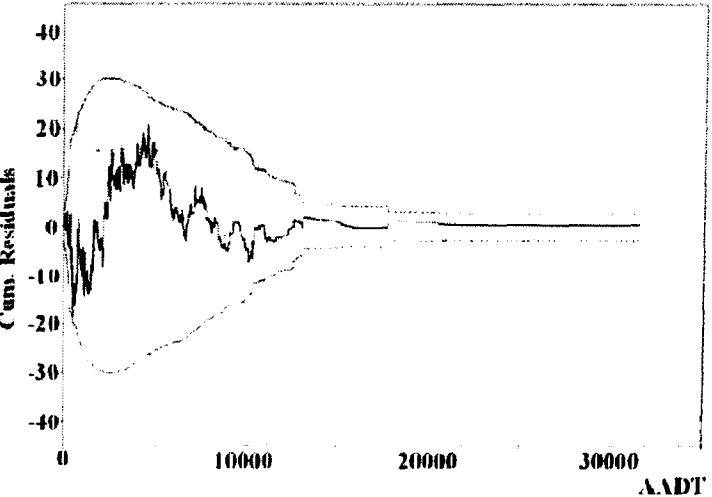
	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	0	+50
Lower Peak:	-640	-58
Regions Exceed ±2σ	Whole Range	None
Linear Correlation	Average Linear Correlation	Average Linear Correlation
Average Model Bias	Under-predict crashes	Under-predict crashes
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.0400	0.0076

38

Table 4.8b - Summary of Findings for Type 22: Head On

Observed Counts: 874

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	0	+20
Lower Peak:	-120	-20
Regions Exceed ±2σ	Whole Range	None
Linear Correlation	Low Linear Correlation	Low Linear Correlation
Average Model Bias	Under-predict crashes	Under-predict crashes
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.5000	0.0003

Table 4.8c - Summary of Findings for Type 24: Angle

Observed Counts: 1,450

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	0	+10
Lower Peak:	-170	-180
Regions Exceed ±2σ	Whole Range	>3,000
Linear Correlation	Low Linear Correlation	Low Linear Correlation
Average Model Bias	Under-predict crashes	Under-predict crashes
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.0800	2.3679

Table 4.8d - Summary of Findings for Type 25: Sideswipe (Same Direction)

Observed Counts: 1,039

	Proportion Method	SPF Method
CURE Plots		
Upper Peak:	0	+35
Lower Peak:	-110	-35
Regions Exceed ±2σ	Whole Range	500-700, 3,000-3,500 & 6,500-7,000
Linear Correlation	Low Linear Correlation	Low Linear Correlation
Average Model Bias	Under-predict crashes	Under-predict crashes
Mis-prediction	Similar Result	Similar Result
Overdispersion	0.1820	0.0002

41

4.6 Chapter Summary

The Minnesota rural, two-lane undivided roadway database divided accidents into eighteen different collision types, and each type was analyzed using its own SPFs by applying negative binomial regression and the "proportion" method. In general, not all of the SPFs are well-developed in terms of both overall goodness of fit and the CURE method. The reason for these poor results is that the models do not include other important independent variables except the AADT. Thus, the SPFs are expected to be somewhat less accurate.

Based on the sites and accident types considered in the comparison, the proposed methods along with the number of crashes for each of the accident types are shown in Tables 4.9 and 4.10. The selection of the proposed method is based on the GOF tests and CURE plot results. Tables 4.9 and 4.10 show no significant connection between accident counts and which method is selected. Generally speaking, the negative binomial regression generates more accurate results than the "proportion" method when a large number of accident counts are available. Nevertheless, one can still make use of the "proportion" method when a dataset is lack of accident counts for a specific collision type. In short, it is difficult to draw a definite conclusion based on these findings since they are based only on the Minnesota dataset. In order to draw a firm conclusion, similar tests on other databases are suggested. In the next Chapter, similar methodology was applied to the Minnesota TWSC intersections database.

Table 4.9 – Proposed Method for Single Vehicle Collision Types

Collision Type	Proposed Method	# of Accidents
Collision w/ Parked Motor Vehicle	Proportion	1,015
Collision w/ Animal	Proportion	6,938
Collision w/ Fixed Object	SPF	6,362
Overturn	SPF	7,803

Table 4.10– Proposed Method for Multi-Vehicle Collision Types

Collision Type	Proposed Method	# of Accidents
Rear-End	SPF	2,866
Head-On	SPF	874
Angle	SPF	1,450
Sideswipe, Same Direction	SPF	1,039

CHAPTER FIVE

MODELLING THE PROPORTIONS OF SPECIFIC COLLISION TYPES

5.1 Background

As discussed in Chapter Four, a constant factor is recommended in *SafetyAnalyst* to apply to the SPF for all accident types combined to estimate the SPF for a specific accident type. This factor is the proportion of that accident type in the population of all accidents. However, there is no reason to believe that this factor is a constant that is independent of AADT. This may in fact be one reason why the proportion method did not perform well when compared to direct calibration in the investigation described in Chapter Four. To improve this performance an individual model for this proportion factor may be desirable. The objective of this chapter is to investigate models for the proportion of specific collision types. The investigation is confined to accidents at TWSC intersections. The central modelling idea resembles that presented in Chapter Four, but the proportion of accidents of a specific type is used as a dependent variable instead of the accident frequency. The major difficulty with this endeavour is that there is limited previous research on modelling collisions by proportion. One of the main challenges is to determine the functional form. Related to this is the need to determine the error distribution of accidents' proportion for use in a generalized linear modelling approach. As was done in Chapter Four, this method is compared to that in which a constant factor is applied to the SPF for all accident types combined and the method in which a negative binomial regression model is estimated for each individual accident type. Finally, network screening is done for all three methods to investigate the

sensitivity of the ranked lists to the method used. The methodology of this chapter is illustrated in Figure 5.1.
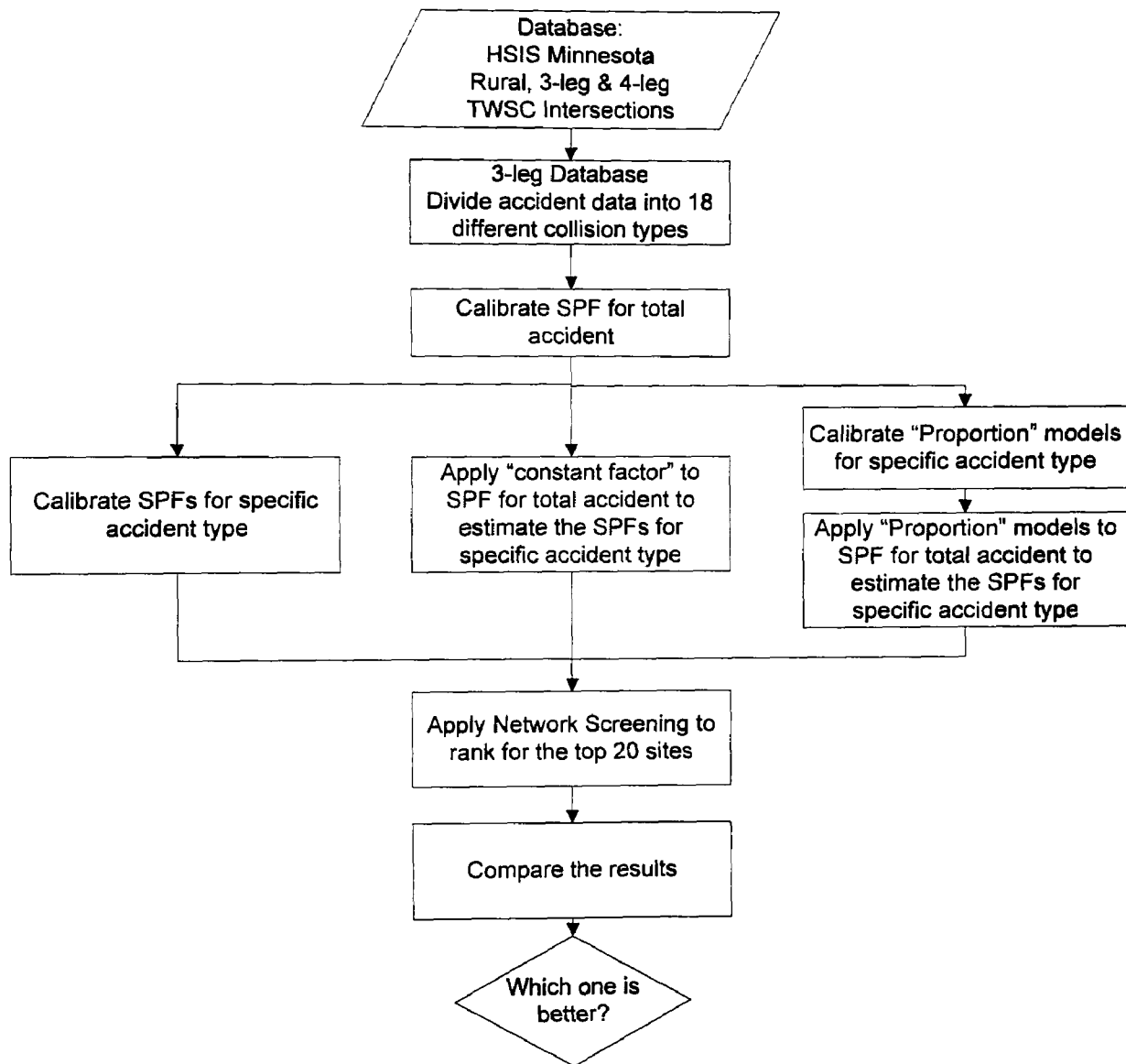
Figure 5.1 – Modelling Development Process

5.2 Data Assembly

In this part of the research, the Minnesota HSIS rural 3-leg and 4-leg TWSC intersection databases were used. Accidents were divided into 18 different collision types in the same manner as in Chapter Four. The statistical summaries of the Minnesota rural, 3-leg and 4-leg TWSC intersections are shown in Tables 5.1 and 5.2, respectively.

Table 5.1 - Statistical Summary of 3-leg TWSC Intersections for Minnesota HSIS Database

Total # of Sites	2,033 (966 Used)*
Mean Entering AADT	5,435 (7,068)*
Total Collisions	2,497
Fatal/Injury Collisions	1,125

*Please refer to Section 5.3.3 for details.

Table 5.2 - Statistical Summary of 4-leg TWSC Intersections for Minnesota HSIS Database

Total # of Sites	2,572 (1,649 Used)*
Mean Entering AADT	4,512 (5,582)*
Total Collisions	6,218
Fatal/Injury Collisions	2,823

*Please refer to Section 5.3.3 for details.

In this part of the thesis, only collision types with a minimum of 100 accidents were included for the modelling process because modelling types with too few accidents would reduce the accuracy of the results and would be moot since such types are not of interest in the SPF applications. Having eliminated all the collision types with less than 100 accidents, only three types of single vehicle accident and four types of multi-vehicle accident were left for further analysis. Tables 5.3 and 5.4 show the remaining collision types with 100 or more accidents, along with their proportion, for Minnesota rural, 3-leg and 4-leg TWSC intersections.

Table 5.3 –Collision Types with 100 or More Accident Counts for 3-leg TWSC Intersections

Collision Type	Definition	# of Accidents	Proportion
Single Vehicle Collision			
6	Collision with fixed object	296	0.1185
8	Other single vehicle collision	102	0.0408
9	Overturn	152	0.0609
Multi-Vehicle Collision			
21	Rear end	699	0.2799
24	Angle	742	0.2972
25	Sideswipe, same direction	133	0.0533
27	Other multiple vehicle collision	148	0.0593
Total	Total Collision	2,497	1.0000

Table 5.4 –Collision Types with 100 or More Accident Counts for 4-leg TWSC Intersections

Collision Type	Definition	# of Accidents	Proportion
Single Vehicle Collision			
8	Other single vehicle collision	529	0.0851
Multi-Vehicle Collision			
21	Rear end	1,081	0.1739
24	Angle	2,996	0.4818
25	Sideswipe, same direction	347	0.0558
Total	Total Collision	6,218	1.0000

The data show that single vehicle collisions seldom occur at TWSC intersections, which is quite logical because the majority of intersection-related accidents involve two or more vehicles.

5.3 Modelling Approach and Results

In the present context, the proper functional form of modelling collisions by proportion is not known. Instead, a graphical method is introduced to determine the proper functional form for modelling collisions by proportion. For this, scatter plots

were employed to identify the relationship between proportion of accidents and a candidate independent variable. Since this research is exploratory, AADT was the only such variable investigated, which seems reasonable because AADT typically explains more than 70% of the variation in accident occurrence. Next, various functions were selected to test the relationship between the dependent variable and AADT. To accommodate the need of generalized linear modelling (GLM), a suitable distribution was also necessary for the dependent variable (proportion of accident); this was achieved by using another graphical method known as the Normal Probability Plot, which enables modellers to assess the normality of the data. (*15, 24*)

5.3.1   Selection of Functional Forms

Suitable functional forms were selected after conducting a series of exploratory data analyses on the HSIS Minnesota rural, 3-leg and 4-leg TWSC intersections databases. Since the only independent variable used in this research is AADT, it is possible to use the scatter plots of proportion of accident vs. total entering AADT to explore several functional forms for the development of models. Nevertheless, it is unlikely that every collision type will make use of the same functional form equally well. For simplicity purposes, a common functional form was used for all collision types. As a departure point, gamma and power distributions were tested based on previous research indicating that these are the typical model forms used in accident modelling. The gamma and power functions are shown in Figure 5.2. The scatter plots for various collision types of Minnesota rural, 3-leg and 4-leg TWSC intersections are shown in Figures 5.3 and 5.4, respectively.
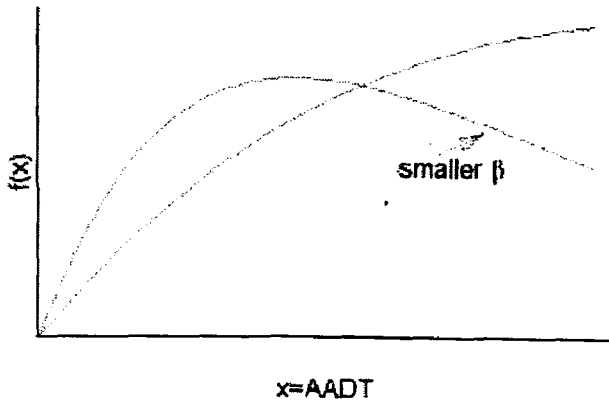
Figure 5.2a – Power Function

$$f(x)=\alpha x^{\beta}$$
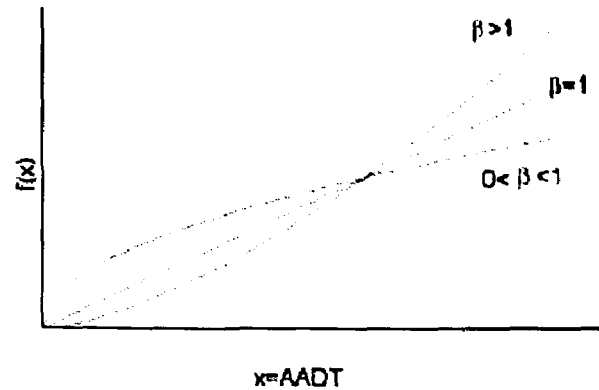
Figure 5.2b - Gamma Function

$$f(x)=\alpha x^{\beta_1} e^{\beta_2 x}$$

Because a proper function is the primary determinant of a reliable model, it should be able to clearly explain the relationship between the dependent variable (proportion of accident) and the independent variable (AADT). The data points in most of the scatter plots show signs of a reasonable fit to the two proposed functions. Therefore, both functions were selected to model collisions by proportion. The two chosen model forms are listed below:

$$\text{Proportion of Accident}=\alpha(AADT/1000)^{\beta} \qquad (5.1)$$

$$\text{Proportion of Accident}=\alpha(AADT/1000)^{\beta_1} e^{\beta_2(AADT/1000)} \qquad (5.2)$$

where AADT is the total entering AADT while α and β are the constants.
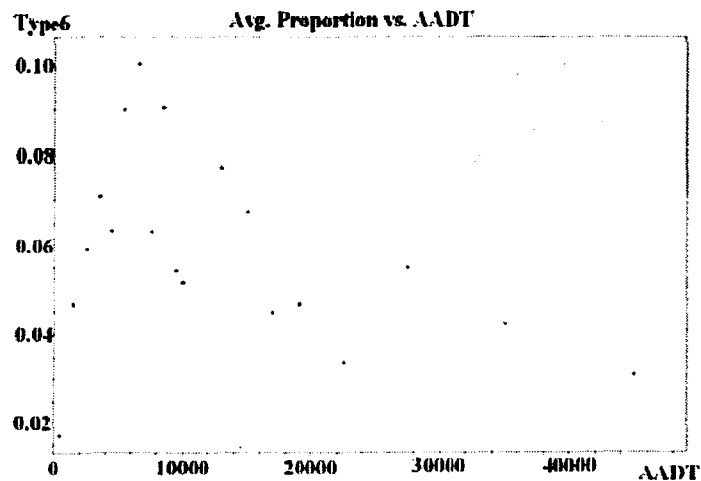
Figure 5.3a – Scatter plot for Collision with Fixed Object at Minnesota
Rural 3-leg TWSC Intersections
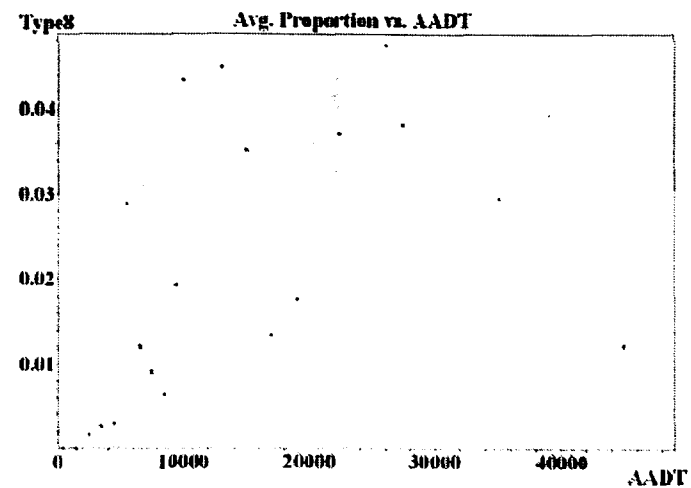
Figure 5.3b – Scatter plot for Other Single Vehicle Collisions at
Minnesota Rural 3-leg TWSC Intersections
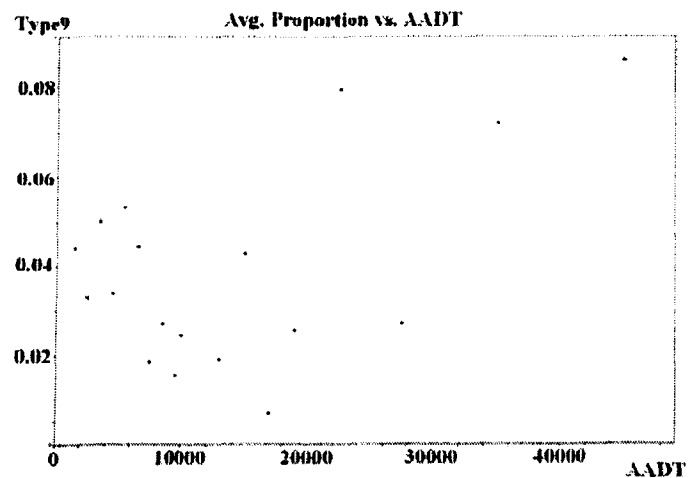
Figure 5.3c – Scatter plot for Overturn Collisions at Minnesota Rural
3-leg TWSC Intersections
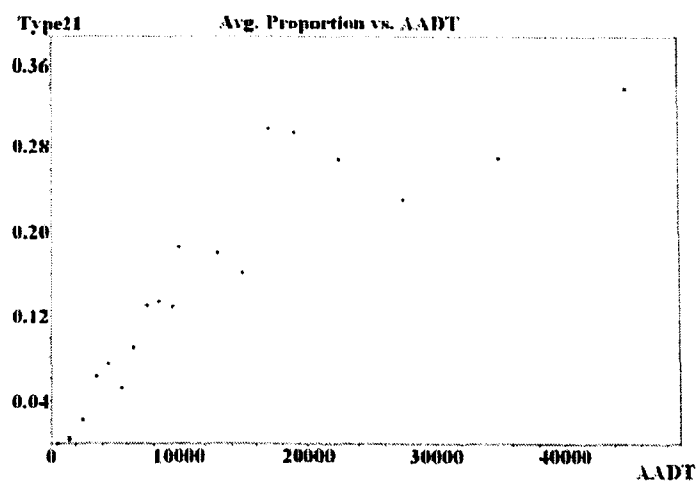
Figure 5.3d – Scatter plot for Rear-End Collisions at Minnesota Rural
3-leg TWSC Intersections

50
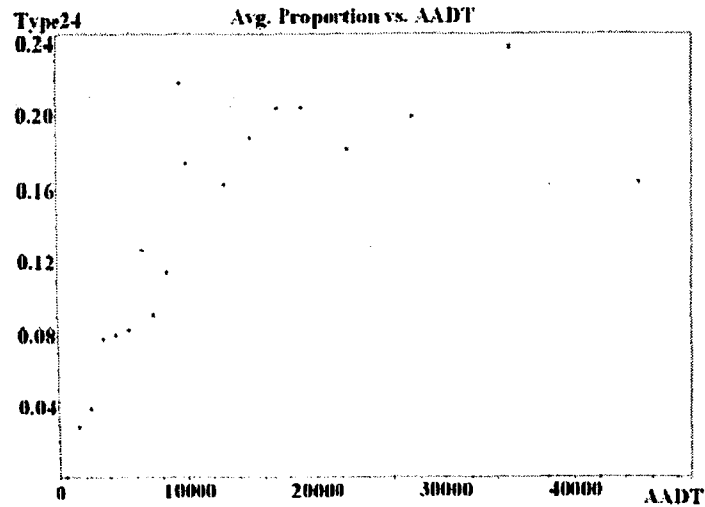
Figure 5.3e – Scatter plot for Angle Collisions at Minnesota Rural 3-leg TWSC Intersections
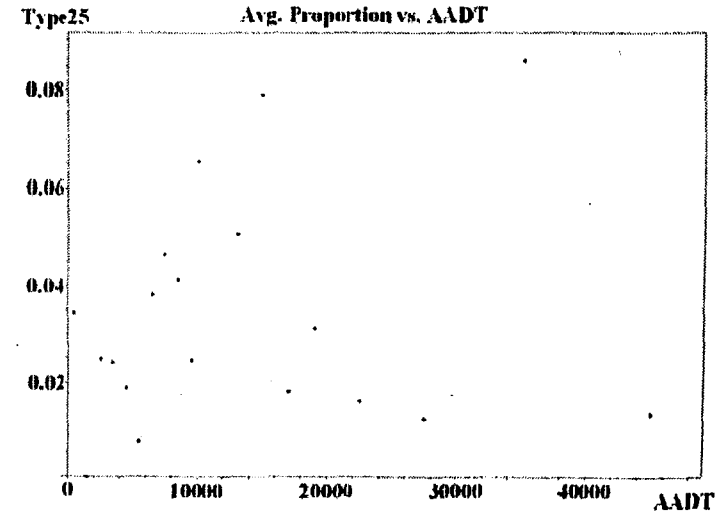
Figure 5.3f – Scatter plot for Sideswipe (Same Direction) Collisions at Minnesota Rural 3-leg TWSC Intersections
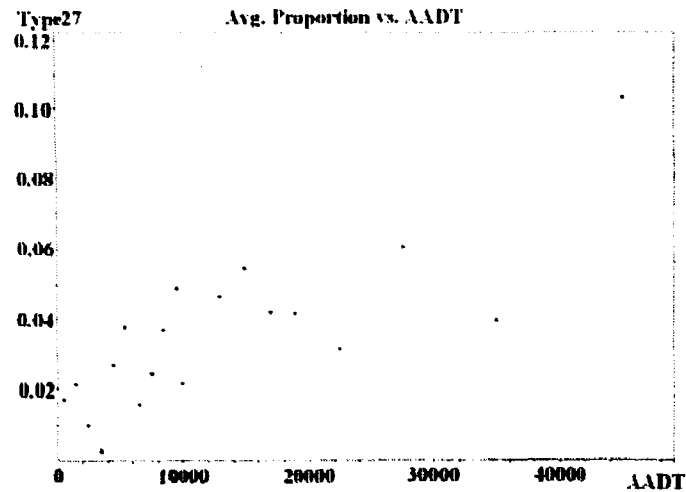
Figure 5.3g– Scatter plot for Other Multiple Vehicle Collisions at Minnesota Rural 3-leg TWSC Intersections
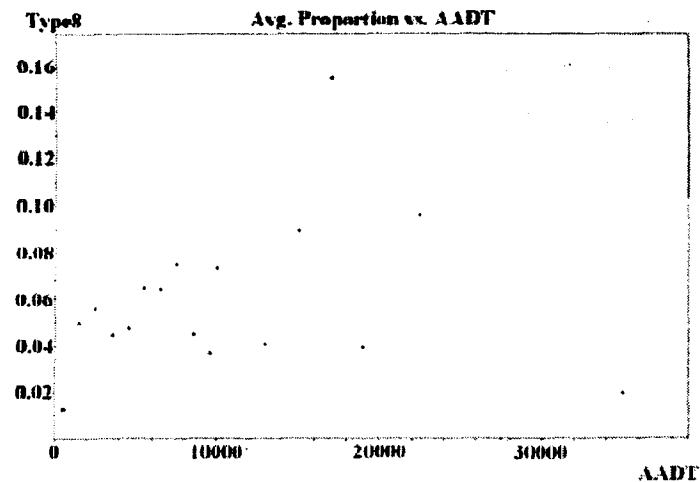
Figure 5.4a– Scatter plot for Other Single Vehicle Collisions at Minnesota Rural 4-leg TWSC intersections
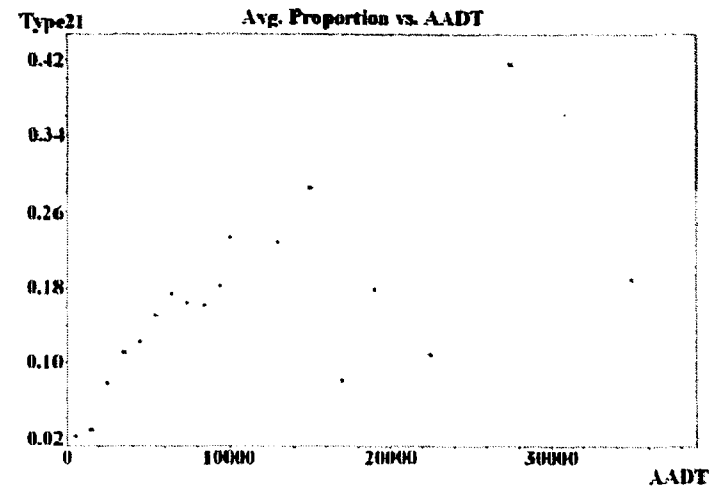
Figure 5.4b– Scatter plot for Rear-End Collisions at Minnesota Rural 4-leg TWSC Intersections
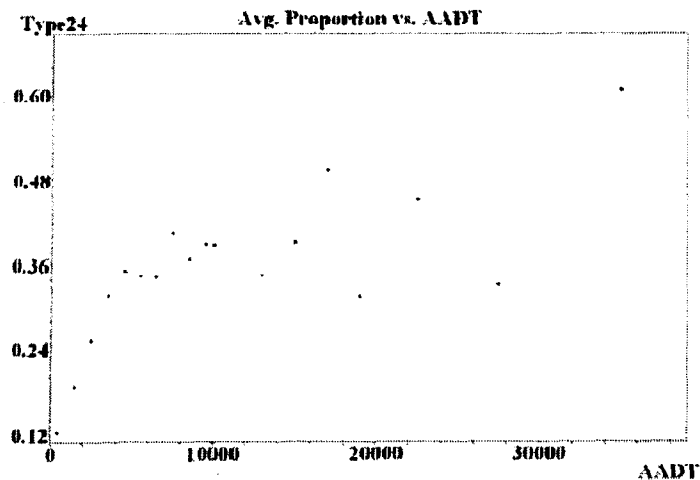
Figure 5.4c– Scatter plot for Angle Collisions at Minnesota Rural 4-leg TWSC intersections
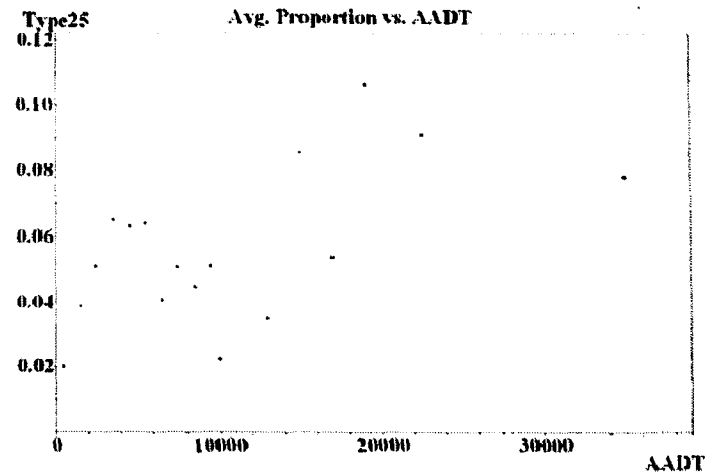
Figure 5.4d– Scatter plot for Sideswipe (Same Direction) Collisions at Minnesota Rural 4-leg TWSC intersections

52

5.3.2 Investigation of the Error Distribution for the Proportion of Specific Accident Types

In any generalized linear model, the response variable follows some kind of distribution (i.e. Poisson, binomial, gamma, or negative binomial). The distribution of the proportion of accidents of specific collision types in this study is not known. By default the normal distribution was considered. But before making this assumption, the normal probability plot was applied to assess the normality of the data. Detailed analyses are shown in Appendix A. The normal probability plots from Appendix A reveal that accidents' proportion follows the normal distribution reasonably well, which justifies the adoption of this assumption for the error distribution of the dependent variable.

5.3.3 Calibration Results

Following the determination of suitable model forms and a correct distribution for fitting a GLM, calibration procedures were performed in the SAS software. Maximum likelihood was used in all parameter estimation, using the GENMOD procedure in SAS, as described in Chapter Four. The two models for proportion of accidents, as specified in Section 5.3.1, were estimated as a function of AADT.

Throughout the calibration process, only the sites with at least one accident for all specific accident types were included. If sites with zero total accidents were included, the results would be meaningless because one cannot determine the accidents' proportion for the specific accident type by dividing zero with zero (i.e. 0 angle accident ÷ 0 total accident). The equations below describe the sum of accident proportions following equations 5.1 and 5.2 respectively,

$$\sum_{i=1}^{\infty} \alpha_i (AADT/1000)^{\beta_i} \cong 1 \tag{5.3}$$

$$\sum_{i=1}^{\infty} \alpha_i (AADT/1000)^{\beta_{1i}} e^{\beta_{2i}(AADT/1000)} \cong 1 \tag{5.4}$$

where i is the individual collision type. Having eliminated all the sites with zero accident for all specific accident types, only 966 and 1,649 sites were left for Minnesota rural, 3-leg and 4-leg TWSC intersections, respectively. This sum can also act as a check by adding all the proportion estimates from the models. Any sum greater or less than one indicates that a mistake has been made during the calibration process. The calibration results for Minnesota rural, 3-leg and 4-leg TWSC intersections, using both model forms, are shown in Tables 5.5, 5.6, 5.7, and 5.8.

Table 5.5 – Calibration Results: Minnesota 3-leg Rural TWSC Intersections
Model Form:
$$\text{Proportion of Accident} = \alpha(\text{AADT}/1000)^{\beta}$$

Collision Type	Par.	Estimate	S.E.	P-value	R^2	Mean Observed Proportion	Mean Predicted Proportion	Constant Proportion	Observed Counts
Type 6	lnα	-1.2471	0.0891	<.0001	0.049	0.1345	0.1360	0.1185	296
	β	-0.5111	0.0695	<.0001					
Type 8	lnα	-3.7008	0.2918	<.0001	0.028	0.0361	0.0363	0.0408	102
	β	0.2255	0.1360	0.0972					
Type 9	lnα	-1.7214	0.1239	<.0001	0.026	0.0763	0.0747	0.0609	152
	β	-0.6101	0.1110	<.0001					
Type 21	lnα	-1.9085	0.1104	<.0001	0.044	0.2646	0.2659	0.2799	699
	β	0.3347	0.0496	<.0001					
Type 24	lnα	-1.5352	0.1046	<.0001	0.004	0.2547	0.2548	0.2972	742
	β	0.1006	0.0535	0.0602					
Type 25	lnα	-2.4492	0.1914	<.0001	0.703	0.0642	0.0642	0.0533	133
	β	-0.1881	0.1173	0.1086					
Type 27	lnα	-2.8615	0.2565	<.0001	0.007	0.0613	0.0613	0.0593	148
	β	0.0424	0.1388	0.7597					
Sum						0.8916	0.8931		

Type 6 = Collision with Fixed Object
Type 8 = Other Single Vehicle Collision
Type 9 = Overturn
Type 21 = Rear End Collision
Type 24 = Angle Collision
Type 25 = Sideswipe, same direction
Type 27 = Other Multiple Vehicle Collision

R^2 = Coefficient of Determination
AADT = Total Entering AADT
C = Constant Proportion for All Sites

Table 5.6 – Calibration Results: Minnesota 3-leg Rural TWSC Intersections
Model Form:

$$\text{Proportion of Accident} = \alpha(\text{AADT}/1000)^{\beta_1} \, e^{\beta_2(\text{AADT}/1000)}$$

Collision Type	Par.	Estimate	S.E.	P-value	R^2	Mean Observed Proportion	Mean Predicted Proportion	Constant Proportion	Observed Counts
Type 6	lnα	-1.2098	0.1033	<.0001	0.056	0.1345	0.1331	0.1185	296
	$\beta1$	-0.2287	0.2221	0.3031					
	$\beta2$	-0.0873	0.0693	0.2079					
Type 8	lnα	-4.3216	0.5771	<.0001	0.006	0.0361	0.0364	0.0408	102
	$\beta1$	1.0849	0.6031	0.0721					
	$\beta2$	-0.120	0.0890	0.1772					
Type 9	lnα	-1.7352	0.1186	<.0001	0.030	0.0763	0.0757	0.0609	152
	$\beta1$	-0.6866	0.1199	<.0001					
	$\beta2$	0.0247	0.0129	0.0560					
Type 21	lnα	-2.0505	0.1531	<.0001	0.266	0.2646	0.2653	0.2799	699
	$\beta1$	0.4697	0.1092	<.0001					
	$\beta2$	-0.0129	0.0103	0.2105					
Type 24	lnα	-1.6268	0.1483	<.0001	0.006	0.2547	0.2544	0.2972	742
	$\beta1$	0.2509	0.1664	0.1317					
	$\beta2$	-0.0224	0.0240	0.3500					
Type 25	lnα	-2.5856	0.2697	<.0001	0.005	0.0642	0.0637	0.0533	133
	$\beta1$	0.2562	0.4528	0.5715					
	$\beta2$	-0.0909	0.0924	0.3254					
Type 27	lnα	-2.8306	0.2743	<.0001	0.001	0.0613	0.0614	0.0593	148
	$\beta1$	0.0056	0.1921	0.9769					
	$\beta2$	0.0042	0.0148	0.7761					
Sum						0.8916	0.8900		

Type 6 = Collision with Fixed Object
Type 8 = Other Single Vehicle Collision
Type 9 = Overturn
Type 21 = Rear End Collision
Type 24 = Angle Collision
Type 25 = Sideswipe, same direction
Type 27 = Other Multiple Vehicle Collision

R^2 = Coefficient of Determination
AADT = Total Entering AADT
C = Constant Proportion for All Sites

Table 5.7 – Calibration Results: Minnesota 4-leg Rural TWSC Intersections
Model Form:

$$\text{Proportion of Accident} = \alpha(\text{AADT}/1000)^{\beta}$$

Collision Type	Par.	Estimate	S.E.	P-value	R^2	Mean Observed Proportion	Mean Predicted Proportion	Constant Proportion	Observed Counts
Type 8	ln α	-2.3679	0.1108	<.0001	0.002	0.0784	0.0784	0.0851	529
	β	-0.1317	0.0756	0.0812					
Type 21	ln α	-2.3575	0.0887	<.0001	0.028	0.1583	0.1596	0.1739	1,081
	β	0.3481	0.0429	<.0001					
Type 24	ln α	-0.8202	0.0432	<.0001	0	0.4337	0.4337	0.4818	2,996
	β	-0.0109	0.0273	0.6888					
Type 25	ln α	-2.3778	0.1228	<.0001	0.002	0.0751	0.0752	0.0558	347
	β	-0.1568	0.0845	0.0634					
Sum						0.7455	0.7469		

Type 8 = Other Single Vehicle Collision
Type 21 = Rear End Collision
Type 24 = Angle Collision
Type 25 = Sideswipe, same direction

R^2 *= Coefficient of Determination*
AADT = Total Entering AADT
C = Constant Proportion for All Sites

Table 5.8 – Calibration Results: Minnesota 4-leg Rural TWSC Intersections
Model Form:

$$\text{Proportion of Accident} = \alpha(\text{AADT}/1000)^{\beta_1} e^{\beta_2(\text{AADT}/1000)}$$

Collision Type	Par.	Estimate	S.E.	P-value	R^2	Mean Observed Proportion	Mean Predicted Proportion	Constant Proportion	Observed Counts
Type 8	ln α	-2.3678	0.1122	<.0001	0.002	0.0784	0.0784	0.0851	529
	β1	-0.1325	0.1095	0.2264					
	β2	0.0002	0.0169	0.9927					
Type 21	ln α	-2.6411	0.1569	<.0001	0.039	0.1583	0.1584	0.1739	1,081
	β1	0.6759	0.1540	<.0001					
	β2	-0.0377	0.0197	0.0557					
Type 24	ln α	-0.8135	0.0445	<.0001	0	0.4337	0.4337	0.4818	2,996
	β1	-0.0251	0.0366	0.4922					
	β2	0.0023	0.0038	0.5446					
Type 25	ln α	-2.3745	0.1214	<.0001	0.002	0.0751	0.0752	0.0558	347
	β1	-0.1894	0.1112	0.0885					
	β2	0.0077	0.0148	0.6029					
Sum						0.7455	0.7457		

Type 8 = Other Single Vehicle Collision
Type 21 = Rear End Collision
Type 24 = Angle Collision
Type 25 = Sideswipe, same direction

R^2 *= Coefficient of Determination*
AADT = Total Entering AADT
C = Constant Proportion for All Sites

At rural 3-leg TWSC intersections, the P-values are slightly better for model (5.1) since they are significantly smaller than the P-values for model (5.2) in most cases. However, the P-values reveal that the models for same direction sideswipe collision (Type 25) and other multiple-vehicle collision (Type 27) are not so good. The observed counts in both Type 25 and Type 27 show fewer than 150 accidents, compared to other collision types where the accident counts are relatively higher, and this may be the reason for the poor P-values.

The P-values for rural 4-leg TWSC intersections are slightly better for model (5.1) but less significant for angle collision (Type 24) and same direction sideswipe collision (Type 25). In contrast to the 3-leg models, both collision types have a large number of accident counts but both models are insignificant. Thus, the functional forms do not describe the dataset well and perhaps other functional forms should be considered for Type 24 and Type 25.

Of the two functional forms, model (5.1) is a better choice for modelling proportion by type for rural 3-leg and 4-leg TWSC intersections. The P-values indicate that the model for 4-leg TWSC intersections is not significant for one type of multi-vehicle collision (Type 24) but is quite good for single vehicle collision. However, both models have very small coefficients of determination for all collision types, indicating a weak relationship. Given that the exploratory nature of this project was to compare this method with the other two approaches, and not to determine an ideal model to describe accidents' proportion for all collision types, model (5.1) was chosen for proceeding to the next stage.

5.4 Comparison of Networking Screening Methods for Angle Collisions at Rural Two-Way Stop-Controlled Intersections

In any traffic safety improvement study, one is interested in identifying sites which require attention for road safety treatments. For years, engineers have used the traditional screening techniques that rely on observed counts - a method now proven to be unreliable due to the fact that an accident is a random event. The randomness in collisions can cause misinterpretation of the safety of sites with relatively high or low accidents. In order to efficiently identify sites with promise of safety treatment, a screening method which is more reliable must be used.

The SPF screening approach, which had been documented in FHWA (*22*), was adopted and applied in this study. This approach is trustworthy but it requires extensive SPF development and the availability of reliable traffic volume data. Acquiring these can be costly, time consuming, and difficult. Therefore, an alternative is sought to overcome such shortfalls. This potential network screening approach known as the collision pattern recognition method or the method screening for high proportions of specific accident types was developed by Heydecker and Wu (*9*), which is based on the Bayes' theorem. The theoretical framework for both SPF and collision pattern recognition methods will be discussed in the following sections.

5.4.1 Potential for Safety Improvement (PSI) Method for Intersections

The purpose of network screening is to identify sites with promise for detailed engineering studies (DES). Traditionally, engineers have selected hazardous sites for road improvement purely based on the observation of high accident occurrences within a short period of time. Nevertheless, this traditional method is biased for safety

estimation for reasons explained earlier. Recently, two alternative ways to identify sites with promise which are based on EB framework were documented by FHWA. (*22*)

1) PSI based on Expected Accident Frequency

2) PSI based on Excess Accident Frequency

The chief advantage of these methods is that they remove the Regression-To-Mean (RTM) bias. However, the main disadvantage of these methods is that they require a database which contains reliable information of AADT as well as SPFs.

Specifically, the first method ranks sites based on the EB expected accident frequency, X, whereas the second method ranks sites using the difference between the EB expected accident frequency, X, and the SPF predicted accident frequency, κ:

$$Excess = X - \kappa \qquad (5.5).$$

The procedures for calculating the PSI are very straightforward and both methods follow similar procedures except for the fact that there is an additional step in calculating the excess accident frequency. Stepwise, the procedures for the PSI methods are listed below.

Step 1: Determine the Safety Performance Function and estimate the SPF model parameters. Calculate the predicted accidents for each year (y = 1, 2, ..., Y)

$$\kappa_{y(Type)} = SPF_{(Type)} \{\text{Major AADT, Minor AADT}\} \qquad (5.6)$$

where κ_y is the predicted accident frequency with unit accidents/year, and the subscript *Type* denotes the collision types.

Step 2: Use the model in Step 1 to calculate the yearly correction factors, C_y, for years $(y = 1, 2, ..., Y)$.

$$C_{y(Type)} = \frac{\kappa_{y(Type)}}{\kappa_{1(Type)}}$$

(5.7)

Step 3: Calculate the EB weight.

$$w_{Type} = \frac{1}{1 + d_{Type} \sum_{y=1}^{Y} \kappa_{y(Type)}}$$

(5.8)

where d is the overdispersion parameter and w is the EB weight for the target collision type.

Step 4: Calculate the EB adjusted expected number of accidents, X_1, for the target type of collision during Year 1:

$$X_{1(Type)} = w_{Type} \kappa_{1(Type)} + (1 - w_{Type}) \frac{\sum_{y=1}^{Y} K_{y(Type)}}{\sum_{y=1}^{Y} C_{y(Type)}}$$

(5.9)

where K_y is the observed accident count in year y.

Step 5: Calculate X_y for all other years using the yearly correction factor, C_y, for the target collision type.

$$X_{Y(Type)} = X_{1(Type)} C_{Y(Type)}$$

(5.10)

The variance of X_y is obtained by.

$$Var(X_{Y(Type)}) = X_{Y(Type)} (1 - w_{Type}) \frac{C_{Y(Type)}}{\sum_{y=1}^{Y} C_{y(Type)}}$$

(5.11)

61

If sites were indeed ranked based on expected accident frequency, the results obtained from Step 5 would then be used for ranking. The expected accidents $X_{y(Type)}$ are ranked in descending order. If sites were to be ranked based on the excess accident frequency, the investigator should then follow Step 6.

Step 6: Calculate the excess accident frequency and the variance of the target collision type:

$$Excess_{Y(Type)} = X_{Y(Type)} - \kappa_{Y(Type)} \tag{5.12}$$

and

$$Var(Excess_{Y(Type)}) = Var(X_{Y(Type)}) + \frac{1}{d_{Type}}(\kappa_{Y(Type)})^2 \tag{5.13}$$

where *Excess* is the excess accident frequency. The sites are ranked using the results obtain from Step 6, PSI $= X_{y(Type)} - \kappa_{y(Type)}$.

5.4.2 Method of Identifying Sites with High Proportions of Specific Accident Types

As discussed above, there are two shortfalls for the PSI method. Because it requires an SPF and traffic volumes, Heydecker and Wu (*9*) have developed an alternative screening approach which can avoid this need. The Bayesian method is now used to determine the probability of a specific collision types (i.e. angle collision) by combining two sources of information - the range of frequency of occurrence for the specific collision type, and the counts of specific accident type.

Mathematically, the proportion of a collision type at a site i can be represented by the binomial distribution:

$$f(x_i \mid n_i, \theta) = \binom{n_i}{x_i} \theta_i^{x_i} (1 - \theta_i)^{n_i - x_i} \tag{5.14}$$

where $\binom{n}{x}$ is the binomial coefficient, defined by:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.15}$$

where x_i is the count of target accident, with total accident n_i and the mean proportion θ_i at a site. The variation between different sites is assumed to follow the Beta distribution:

$$g(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \tag{5.16}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{5.17}$$

where α and β are the parameters of the prior distribution (Eq. 5.16) which can be determined by fitting the observations data from the reference group (x_i and n_i) into the Beta distribution.

There are several reasons for choosing the beta distribution. First, the beta distribution can cover wide range of shapes. Second, it is a conjugate prior for the binomial distribution parameter θ which makes the integration process of the Bayes' theorem more convenience. Conjugate prior refers to the selection of prior distribution which makes the posterior distribution to have the same algebraic form as the prior. In other words, the posterior distribution acts as an update version of the prior distribution. The mean and variance of the Beta distribution are given by:

$$E(\theta) = \frac{\alpha}{\alpha + \beta} \qquad\qquad (5.18)$$

$$Var(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{E(\theta)(1-\mu)}{\alpha + \beta + 1} \qquad\qquad (5.19)$$

The binomial distribution, which explains the proportion of a specific accident type at a site can be combined with the beta distribution, which describes the variation between similar sites, to form an unconditional Binomial-Beta distribution.

$$h(x_i \mid n_i, \alpha, \beta) = \binom{n_i}{x_i} \frac{B(\alpha + x_i, \beta + n_i - x_i)}{B(\alpha, \beta)} \qquad\qquad (5.20)$$

Note that the parameter θ, which is unknown, is eliminated after the combination.

It is now possible to combine the prior Beta distribution (Eq. 5.16) with the site specific collision data (n_i, x_i) for each site to obtain the posterior distribution of θ using the Bayes' theorem, which is defined as:

$$g_a(\theta \mid n_i, x_i, \alpha, \beta) = \frac{f(x_i \mid n_i, \theta) \cdot g_b(\theta \mid \alpha, \beta)}{h(x_i \mid n_i, \alpha, \beta)} \qquad\qquad (5.21)$$

where the subscripts a denotes after (posterior) and b denotes before (prior). The adjusted posterior Beta distribution becomes

$$g_a(\theta \mid \alpha + x_i, \beta + n_i - x_i) = \frac{\theta^{\alpha + x_i - 1}(1 - \theta)^{\beta + n_i - x_i - 1}}{B(\alpha + x_i, \beta + n_i - x_i)} \qquad\qquad (5.22).$$

To further simplify Equation 5.22, two new parameters α' and β' are introduced:

$$\alpha' = \alpha + x_i \qquad\qquad (5.23)$$

and

$$\beta' = \beta + n_i - x_i \qquad\qquad (5.24)$$

By substituting Equations 5.23 and 5.24 into Eq. 5.22,

$$g(\theta \mid \alpha', \beta') = \frac{\theta^{\alpha'-1}(1-\theta)^{\beta'-1}}{B(\alpha', \beta')} \qquad (5.25)$$

It is clear that Equation 5.25 is now the updated beta distribution, which is very similar to Equation 5.16. In fact, one of the reasons for choosing Beta prior at the outset is that the posterior distribution will then result in an updated version of the prior distribution. Mathematically, this is known as the conjugate priors. The posterior estimate of θ_i is obtained by

$$E(\theta_i) = \frac{\alpha'}{\alpha' + \beta'} \qquad (5.26)$$

and the variance of each site is given by

$$Var(\theta_i) = \frac{\alpha' \beta'}{\left[(\alpha' + \beta')^2 (\alpha' + \beta' + 1) \right]} \qquad (5.27).$$

To rank the sites using these results, a 'pattern score' is introduced. As for network screening, sites are ranked in descending order of the pattern score, defined as the probability of expected value of θ_i being greater than the median value of θ_m from the beta prior distribution. A pattern score of one indicates that the given site is experiencing a relatively high proportion of a certain accident type. Conversely, a pattern score of zero implies the opposite.

$$P(\theta_i > \theta_m) = 1 - B(\theta_m, \alpha', \beta') \qquad (5.28)$$

where θ_m is written as

$$\int_{\theta_m}^{1} g(\theta \mid \alpha, \beta) d\theta = \pi \qquad (5.29)$$

In this thesis, and following the lead of Heydecker and Wu, π is assumed to be 0.5, which is the median value of the beta prior distribution.

The posterior distribution can be estimated by the method of maximum likelihood –

a very simple idea that yields complicated equations. The likelihood function for Eq.

5.20 can be expressed as:

$$L_i = \binom{n_i}{x_i} \frac{B(\alpha + x_i, \beta + n_i - x_i)}{B(\alpha, \beta)} \tag{5.30}.$$

The log-likelihood function is used to simplify this equation:

$$\log(L_i) = \log\binom{n_i}{x_i} + \log[B(\alpha + x_i, \beta + n_i - x_i)] - \log[B(\alpha, \beta)] \tag{5.31}$$

Parameters α and β are estimated by maximizing the log-likelihood function. Note that

the first component on the right hand side of the Equation 5.31 can be removed since for

each site it is a constant. By applying the beta function into Equation 5.31, the

components on the right hand side of Equation 5.31 become:

$$\log[B(\alpha + x_i, \beta + n_i - x_i)] = \log[\Gamma(\alpha + x_i)] + \log[\Gamma(\beta + n_i - x_i)] \\ - \log[\Gamma(\alpha + \beta + n_i)] \tag{5.32}$$

and

$$\log[B(\alpha, \beta)] = \log[\Gamma(\alpha)] + \log[\Gamma(\beta)] - \log[\Gamma(\alpha + \beta)] \tag{5.33}.$$

Substituting equations 5.32 and 5.33 into Eq. 5.31, the log-likelihood function becomes,

$$\log(L_i) = \log[\Gamma(\alpha + x_i)] + \log[\Gamma(\beta + n_i - x_i)] - \log[\Gamma(\alpha + \beta + n_i)] \\ - \log[B(\alpha, \beta)] - \log[\Gamma(\alpha)] - \log[\Gamma(\beta)] + \log[\Gamma(\alpha + \beta)] \tag{5.34}$$

This expression can be maximized using the Solver tool in Microsoft Excel to obtain the

beta prior estimates of α and β. *(14)*

5.5 Application to Minnesota HSIS Data

Before proceeding to the next step, a recapitulation of the three different approaches to estimating SPFs would be useful. They are represented by the capital letters A, B and C.

A) Using data and negative binomial regression to directly estimate the SPF for a specific accident type.

B) Applying a constant factor to the SPF for all accident types combined to estimate the SPF for the specific accident type.

C) Applying a proportion model to the SPF for all accident types combined to estimate the SPF for the specific accident type.

Both PSI screening methods were applied to the angle collision data for Minnesota rural, 3-leg and 4-leg TWSC intersections to test the accuracy and sensitivity the results from the three approaches to developing the SPFs. The expected accident frequency for the PSI method was established as the reference to compare with the other two screening methods the PSI based on the Excess Accident Frequency and the High Proportion Screening Method. Statistics for both the databases and the angle collisions are listed in Tables 5.9 and 5.10 respectively.

Table 5.9 – Summary of Minnesota TWSC Databases

Type of Site	# of Sites	Years	# of Accidents
Rural, 3-leg TWSC Intersections	2,033	1995-1999	2,497
Rural, 4-leg TWSC Intersections	2,572	1995-1999	6,218

Table 5.10 – Summary of Angle Collision for Minnesota TWSC Intersections

	Observed Target Accidents	Observed Proportion
Rural, 3-leg TWSC Intersections	742	0.2972
Rural, 4-leg TWSC Intersections	2,996	0.4818

Three most common forms of SPFs for intersections were selected to test the fitting of accident data. Both GOF statistics and CURE plots were used to select the best fitting model to proceed with the networking screening study. The SPF parameter estimates for all three models are shown in Tables 5.11 and 5.12, for 3-leg and 4-leg as for total and angle accidents, respectively.

Table 5.11 – SPF Parameter Estimates for Minnesota Rural 3-leg TWSC Intersections

Collision Type	Par.	Estimate	S.E.	P-value	k	Deviance	Chi-Square	Observed Counts
Model I: $Acc/yr = \alpha(AADT_{Major} + AADT_{Minor})^{\beta}$								
Total Accident	$\ln\alpha$	-8.7944	0.3540	<.0001	1.1798	0.9153	1.1084	2,497
	β	0.8687	0.0416	<.0001				
Angle	$\ln\alpha$	-11.8868	0.6492	<.0001	2.6614	0.5252	0.9976	742
	β	1.0820	0.0753	<.0001				
Model II: $Acc/yr = \alpha(AADT_{Major} \, AADT_{Minor})^{\beta}$								
Total Accident	$\ln\alpha$	-7.3506	0.2984	<.0001	1.2228	0.9189	1.1490	2,497
	β	0.4098	0.0205	<.0001				
Angle	$\ln\alpha$	-10.4125	0.5383	<.0001	2.5884	0.5289	1.0598	742
	β	0.5315	0.0364	<.0001				
Model III: $Acc/yr = \alpha(AADT_{Major})^{\beta_1} (AADT_{Minor})^{\beta_2}$								
Total Accident	$\ln\alpha$	-8.7565	0.3398	<.0001	1.1123	0.9186	1.1194	2,497
	$\beta 1$	0.7103	0.0406	<.0001				
	$\beta 2$	0.2325	0.0279	<.0001				
Angle	$\ln\alpha$	-11.9309	0.6260	<.0001	2.4334	0.5283	1.0174	742
	$\beta 1$	0.8486	0.0722	<.0001				
	$\beta 2$	0.3499	0.0483	<.0001				

Table 5.12 – SPF Parameter Estimates for Minnesota Rural 4-leg TWSC Intersections

Collision Type	Par.	Estimate	S.E.	P-value	k	Deviance	Chi-Square	Observed Counts
Model I: $\text{Acc/yr}=\alpha(\text{AADT}_{\text{Major}}+\text{AADT}_{\text{Minor}})^{\beta}$								
Total Accident	lnα	-9.2752	0.2562	<.0001	0.83	1.0248	1.1344	6,218
	β	1.0198	0.0307	<.0001	90			
Angle	lnα	-10.0197	0.3510	<.0001	1.42	0.8633	1.1005	2,996
	β	1.0217	0.0420	<.0001	56			
Model II: $\text{Acc/yr}=\alpha(\text{AADT}_{\text{Major}}\text{AADT}_{\text{Minor}})^{\beta}$								
Total Accident	lnα	-8.7473	0.2162	<.0001	0.70	1.0131	1.0370	6,218
	β	0.5583	0.0151	<.0001	80			
Angle	lnα	-10.0197	0.3023	<.0001	1.14	0.8552	1.0367	2,996
	β	0.5954	0.0209	<.0001	67			
Model III: $\text{Acc/yr}=\alpha(\text{AADT}_{\text{Major}})^{\beta_1}(\text{AADT}_{\text{Minor}})^{\beta_2}$								
Total Accident	lnα	-9.1266	0.2347	<.0001	0.69	1.0160	1.0552	6,218
	β1	0.6614	0.0290	<.0001	34			
	β2	0.4826	0.0232	<.0001				
Angle	lnα	-10.0161	0.3242	<.0001	1.14	0.8554	1.0365	2,996
	β1	0.5911	0.0393	<.0001	71			
	β2	0.5986	0.0327	<.0001				

The CURE plots for the total accidents models are shown in Figures 5.5 and 5.6 for both 3-leg and 4-leg TWSC intersections. Likewise, the CURE plots for angle collision models are shown in Figures 5.7 and 5.8, respectively. The statistics and CURE plots reveal that Model III is the best choice to model accident frequency for both 3-leg and 4-leg rural TWSC intersections. Thus, Model III was chosen to combine with the PSI method for network screening.

Figure 5.5 – The CURE Plots of Total Accidents for Three Different Functional Forms (Minnesota Rural 3-leg TWSC Intersections)

CURE Plot for 3leg Total Accident
Functional Form #1

CURE Plot for 3leg Total Accident
Functional Form #2

CURE Plot for 3leg Total Accident
Functional Form #3
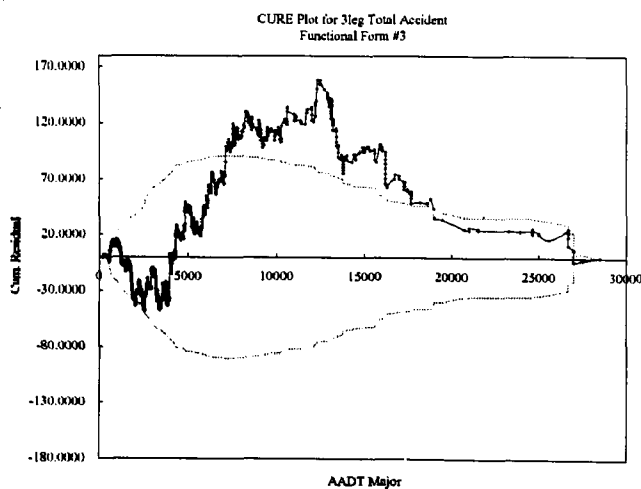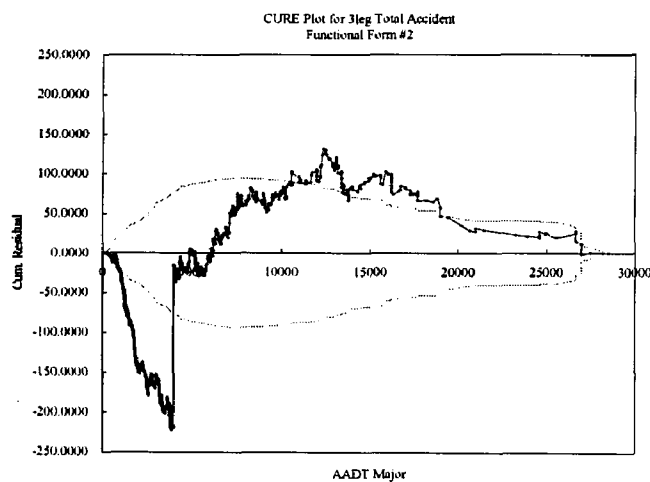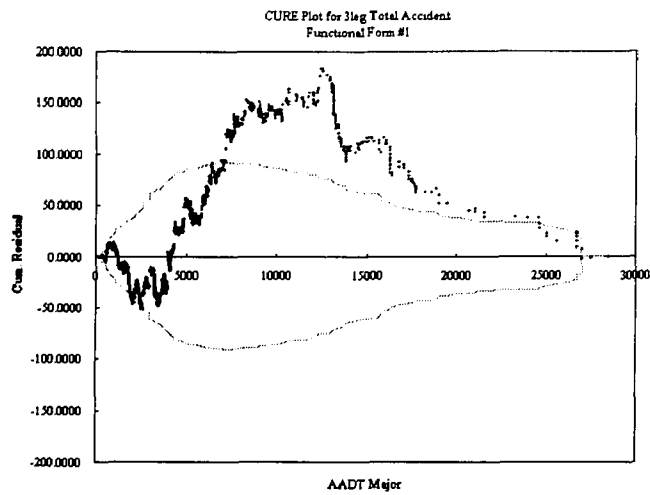
Figure 5.6 – The CURE Plots of Total Accidents for Three Different Functional Forms (Minnesota Rural 4-leg TWSC Intersections)

CURE Plot for 4leg Total Accident
Functional Form #1

CURE Plot for 4leg Total Accident
Functional Form #2

CURE Plot for 4leg Total Accident
Functional Form #3
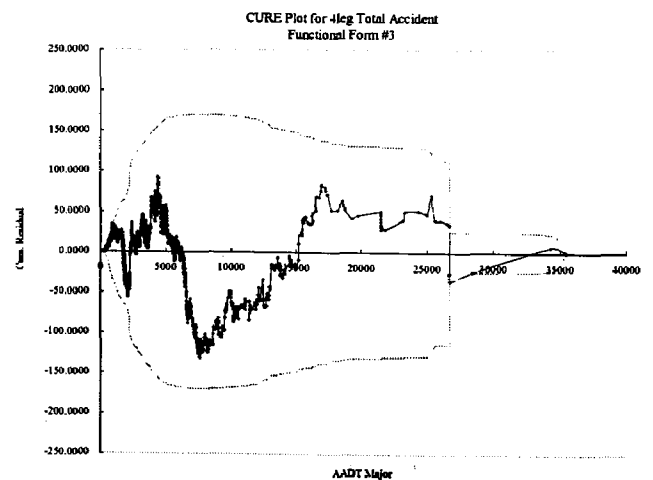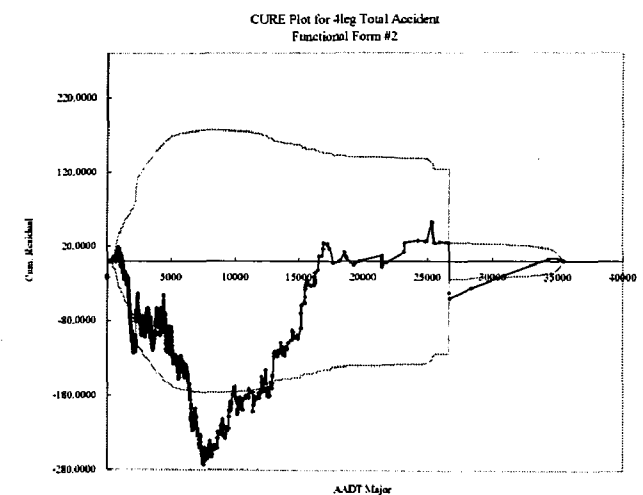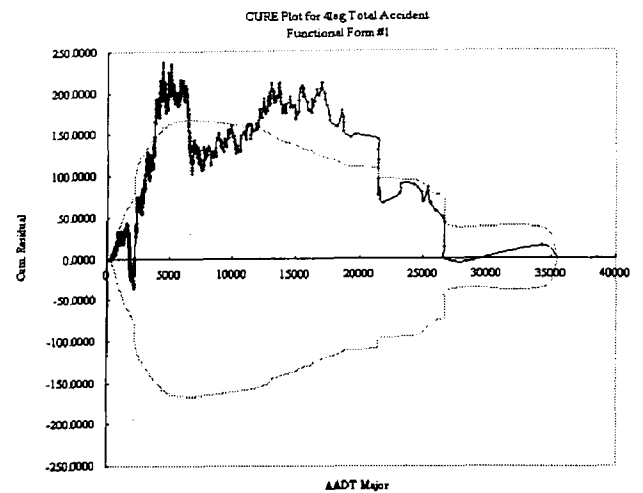
Figure 5.7 – The CURE Plots of Angle Accidents for Three Different Functional Forms (Minnesota Rural 3-leg TWSC Intersections)
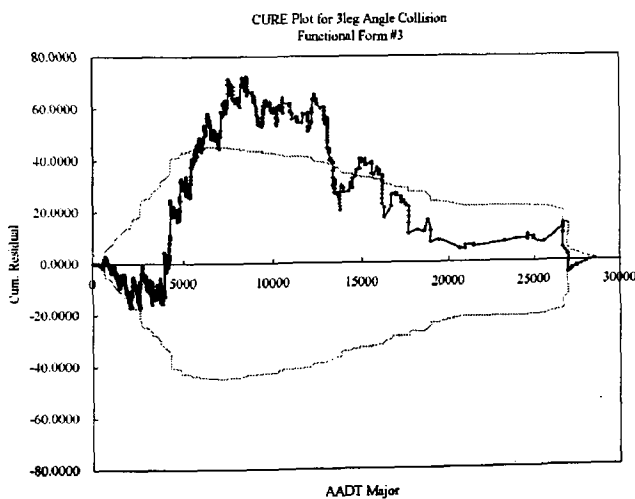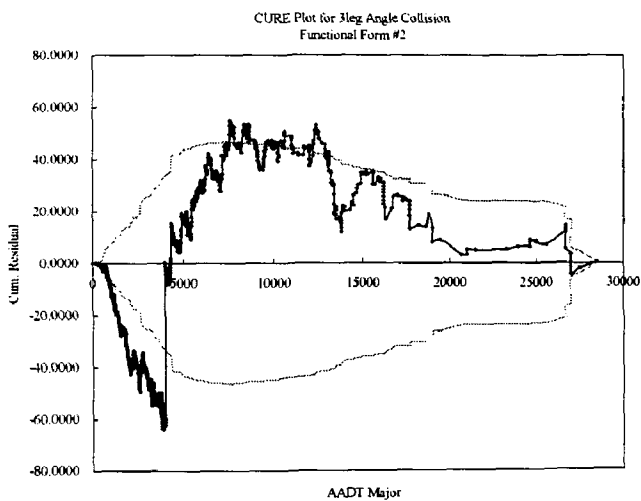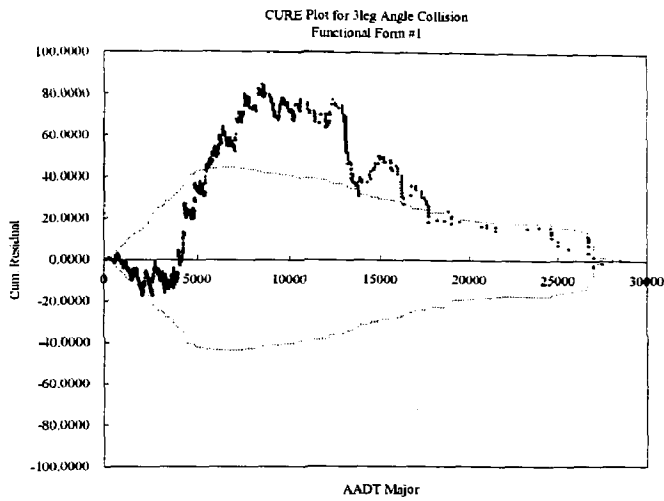
Figure 5.8 – The CURE Plots of Angle Accidents for Three Different Functional Forms (Minnesota Rural 4-leg TWSC Intersections)
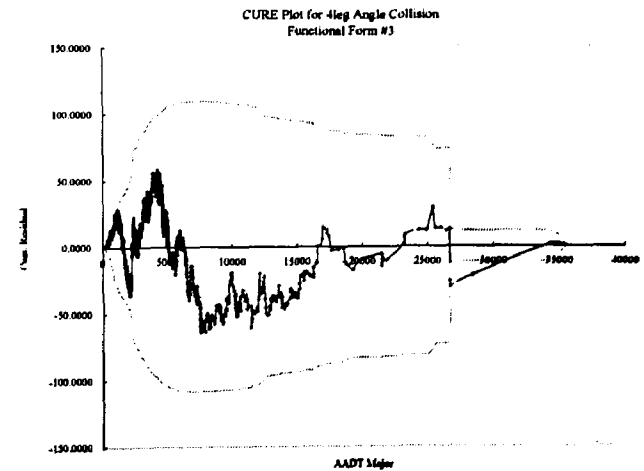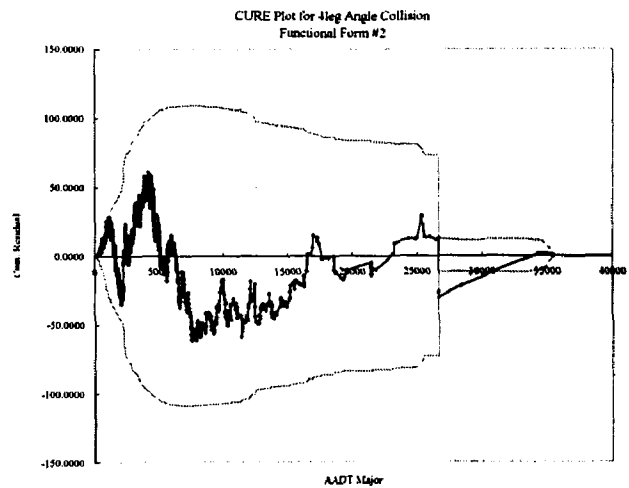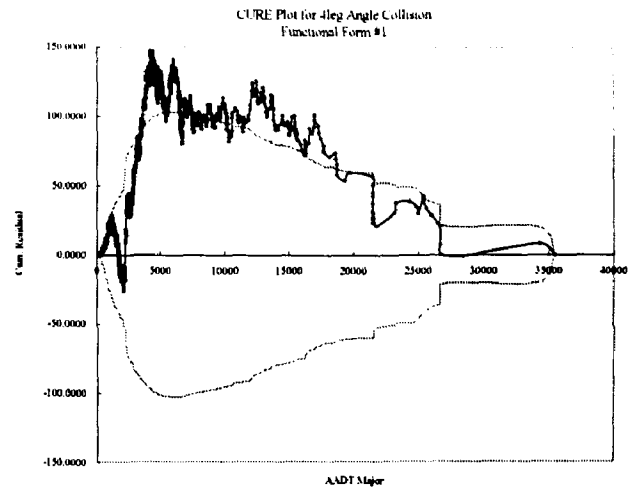
For each dataset, the two PSI methods and the high proportion method were applied to screen for the top 20 sites that require attention. The beta prior distributions were calibrated for angle collision using the method of maximum likelihood. Table 5.13 shows the beta prior estimates for Minnesota rural, 3-leg and 4-leg TWSC intersections. The results of the top 20 sites ranked for both screening methods are listed in Appendix B.

Table 5.13 – Maximum Likelihood Beta Prior Estimates for Minnesota Rural, TWSC Intersections

Type of Site	Observed Proportion	Total Sites	α	β	$E(\theta)$	$var(\theta)$
Rural, 3-leg TWSC Intersections	0.2972	2,033	1.6721	4.3581	0.2773	0.0285
Rural, 4-leg TWSC Intersections	0.4818	2,572	4.8955	5.5944	0.4667	0.0217

5.6 Comparison of the Screening Results

Using the PSI method based on expected accident frequency as a reference point for a screening method, the results of the other two methods were sorted by types as listed on Tables 5.14 and 5.15, for 3-leg and 4-leg, respectively.

As for 3-legged rural TWSC intersections, the PSI method based on expected accident frequency shows that applying a constant factor with SPF for total collision yielded 16 of 20 sites which were also ranked top 20 by using the SPF for angle collision from the reference group. As well, applying the proportion model with SPF for total collision also yielded 17 of 20 similar sites. Likewise, the PSI method based on the excess accident frequency reveals that using a constant factor or proportion model with the SPF for total collision both yielded 15 of 20 sites which were also ranked in the top 20 using the SPF for angle collision from the reference group.

Table 5.14 – Ranking Results by Different Types of Method for 3-leg TWSC Intersections

Database   : HSIS, Minnesota
Accident Type   : Angle Collision
# of Sites Ranked   : 2,033

	AADT			Observed Counts		Expected Accident Frequency			Excess Accident Frequency			High Proportion MLE
						Method 1	Method 2	Method 3	Method 1	Method 2	Method 3	
Site No.	Major	Minor	Total Entering	Total	Angle	Rank						
281	4,300	3,755	8,055	19	15	1	1	1	1	1	1	1
2017	16,750	2,294	19,044	9	6	2	4	3	6	6	6	9
1355	13,825	1,177	15,002	17	8	3	2	2	2	2	2	21
97	18,800	1,285	20,085	8	6	4	5	4	7	7	7	5
646	14,400	3,922	18,322	12	5	5	9	9	11	12	11	70
277	10,300	2,592	12,892	15	7	6	6	6	4	5	4	28
1371	11,700	4,081	15,781	14	5	7	10	10	12	15	13	104
103	26,700	267	26,967	12	6	8	8	7	8	8	8	29
488	13,000	5,140	18,140	7	4	9	17	16	25	26	24	42
489	12,900	12,900	25,800	3	3	10	21	20	116	59	57	22
1078	8,275	1,255	9,530	16	9	11	3	5	3	3	3	7
84	7,125	2,307	9,432	16	8	12	7	8	5	4	5	15
202	7,750	6,644	14,394	7	5	13	16	15	13	19	18	12
497	14,500	1,441	15,941	8	5	14	11	11	14	16	14	17
2018	14,900	1,071	15,971	7	5	15	13	13	17	17	16	12
201	16,700	679	17,379	9	5	16	15	14	18	18	17	27
102	12,300	709	13,009	14	6	17	12	12	10	10	10	45
107	26,700	679	27,379	6	3	18	26	22	52	53	48	71
1848	10,275	3,063	13,338	6	4	19	23	21	23	28	26	25
645	14,722	2,722	17,444	5	3	20	30	28	50	47	43	64
# of Similar Sites Ranked by the Reference Group				-		16/20	17/20	15/20	15/20	15/20	15/20	8/20

Method 1: Using SPF for angle collision
Method 2: Combine a constant proportion with SPF for total collision
Method 3: Combine a proportion model with SPF for total collision
MLE: Maximum Likelihood Estimation

Table 5.15 – Ranking Results by Different Types of Method for 4-leg TWSC Intersections

Database : HSIS, Minnesota
Accident Type : Angle Collision
# of Sites Ranked : 2,572

Site No.	AADT			Observed Counts		Expected Accident Frequency			Excess Accident Frequency			High Proportion MLE
	Major	Minor	Total Entering	Total	Angle	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3	
						Rank						
247	25.400	612	26.012	32	19	1	1	1	1	3	1	56
122	16.225	3.445	19.670	27	16	2	3	2	14	15	13	72
146	4.825	1.760	6.585	29	19	3	2	3	2	1	2	17
925	10.600	1.988	12.588	24	16	4	5	5	5	5	5	21
1405	12.100	1.000	13.100	19	17	5	4	4	4	4	4	1
391	8.700	3.140	11.840	19	15	6	7	7	11	10	9	3
744	7.100	6.393	13.493	22	14	7	10	10	26	19	16	50
1413	2.263	1.880	4.143	32	20	8	6	6	3	2	3	28
270	23.200	1.752	24.952	22	14	9	8	8	21	28	23	50
250	23.300	1.615	24.915	24	14	10	9	9	19	23	20	99
222	17.000	1.366	18.366	18	14	11	12	11	12	14	14	6
261	21.550	1.936	23.486	20	13	12	14	14	32	32	30	38
1335	7.325	1.632	8.957	24	15	13	13	13	7	6	6	52
1888	13.600	2.188	15.788	21	13	14	15	15	22	22	21	62
257	16.600	597	17.197	24	15	15	11	12	6	7	7	52
706	15.250	2.591	17.841	24	12	16	17	17	42	46	33	613
118	12.000	874	12.874	18	14	17	16	16	10	11	11	6
700	12.950	1.154	14.104	24	12	18	20	19	17	20	22	613
1407	19.750	1.414	21.164	14	11	19	22	21	36	50	37	14
1757	2.358	1.825	4.183	22	15	20	19	20	8	9	10	20
# of Similar Sites Ranked by the Reference Group				-		19/20	19/20	14/20	14/20	14/20		7/20

Method 1: Using SPF for angle collision
Method 2: Combine a constant proportion with SPF for total collision
Method 3: Combine a proportion model with SPF for total collision
MLE: Maximum Likelihood Estimation

Similar outcomes were found for 4-leg rural TWSC intersections. For one thing, the PSI method based on expected accident frequency shows that applying a constant factor with the SPF for total collisions yielded 19 of 20 sites which were also ranked in the top 20 by using the SPF for angle collision from the reference group. Applying the proportion model with SPF for total collision also yielded 19 of 20 similar sites. Likewise, the PSI method based on the excess accident frequency suggests that applying a constant factor or proportion model with the SPF for total collision both yielded 14 of 20 sites which were also ranked in the top 20 using the SPF for angle collision from the reference group. In addition, the ranking results obtained from the high proportion method seem to be very different from the results obtained from the PSI methods, which yielded 8 of 20 and 7 of 20 sites for 3-leg and 4-leg TWSC intersections respectively. It is not surprising to observe a difference in ranking using the high proportion method because the method itself does not depend on AADT.

Based on the sites and accident types considered for comparison, the ranking results suggest that whether the factor (proportion of accident type) is constant does not influence the network screening results significantly. Also, the ranking results obtained by applying a constant factor to the SPF for total collision is very close to the one determined by the SPF for angle collision.

CHAPTER SIX

DETAILED INVESTIGATION OF METHODS FOR ESTIMATING THE EXPECTED NUMBER OF PROPERTY DAMAGE ONLY ACCIDENTS

6.1 Background and Introduction

Safety performance functions have been widely used in many applications for years. More often than not, an SPF provides a good measure of safety. However, using an SPF to estimate PDO accidents may be generate uncertainty due to several reasons. First, an SPF for PDO accidents may be based on accident count data with relatively large errors developed during the data collection stage. This is because these counts are based on police records, which may be quite inaccurate. Invariably, police reports may only include a portion of the PDO accidents because not all PDO collisions are reported to the police. Second, the laws state that an accident must be reported to the police only if the damage sustained is over a certain amount. For example, Minnesota requires drivers to report a collision to the police if the damage caused is over $1,000 USD. Therefore, the PDO data do not include records under this amount. As a result, the collision data for PDO accidents is unreliable due to the inconsistency in reporting because of inflation, and biased in estimating PDO accidents that is over $1,000 USD. In *SafetyAnalyst*, the expected number of PDO accidents is estimated by subtracting the difference between SPF estimates for total accidents and injury accidents. This method of calculating PDO estimates may seem inaccurate if the PDO data consist of the errors noted. Because of these difficulties, there is a need of detailed investigation on PDO modelling to assess whether modelling PDOs using the method suggested in *SafetyAnalyst* is indeed reliable.

6.2 Empirical Bayes Procedure

For this investigation, the EB method, proposed by Hauer, is used to estimate safety. The EB procedure is described below.

Step 1: Calibrate the SPF for total accidents and determine the SPF predicted accidents, κ, for each year (y=1, 2... Y).

$$\kappa_y = \alpha (AADT)^\beta \tag{6.1}$$

Step 2: Calculate the yearly correction factor, C_y.

$$C_y = \frac{\kappa_y}{\kappa_1} \tag{6.2}$$

Step 3: Calculate the EB weight, w, using the SPF predicted accidents (κ_1, κ_2... κ_Y).

$$w = \frac{1}{1 + k \sum_{y=1}^{Y} \kappa_y} \tag{6.3}$$

where k is the overdispersion parameter of the model.

Step 4: Determine the EB-adjusted expected number of accidents, X_1, for year 1.

$$EB\{X_1\} = w\kappa_1 SL_{sub} + (1-w) \frac{\sum_{y=1}^{Y} K_y}{\sum_{y=1}^{Y} C_y} \tag{6.4}$$

Step 5: Calculate the EB-adjusted expected number of accidents for y=Y, the final year for which the data exist for the site.

$$EB\{X_Y\} = X_1 C_Y \tag{6.5}$$

The EB procedure describes above was used in both schemes in this study.

6.3 Methodology

The objective for this chapter is to compare the accuracy of two modelling techniques for PDO accidents for road segments. Generally speaking, the negative binomial regression and "yearly multipliers" methods follow similar calculation and validation procedures, but are different in the calibration process.

On one hand, the negative binomial method uses all four years of data to obtain different alphas for each year. Throughout the calibration process, the parameter 'year' was set as a class variable. This is illustrated by Equation 6.6,

$$\kappa(accidents \, / \, mile) = \alpha_{year}(AADT)^{\beta} \tag{6.6}$$

where α_{year} is the yearly alphas and β is a constant. Different alphas are generated from SAS for each year, which are then applied to derive the Empirical Bayes (EB) PDO estimates. On the other hand, the "yearly multipliers" method uses the four years' accident data to calculate the ratios of the observed and the predicted accident counts for each year. These yearly multipliers are then applied to the SPFs calibrated from all four years of data, such that:

$$\kappa(accidents \, / \, mile) = [Multiplier] \times [\alpha(AADT)^{\beta}] \tag{6.7}$$

where α and β are the constant. These yearly multipliers act like the alphas in part one, and are implemented with the SPFs calibrated from all years of data to derive the EB estimates of PDO accidents. Finally, comparisons of the two methods are made based on the accuracy of the results.

6.3.1 Scheme One: Negative Binomial Regression Using Generalized Linear Model

The GENMOD procedures in SAS were performed to calibrate the negative binomial SPFs for total, injury, and PDO accidents respectively and the variable 'year' was set as a class variable. Different yearly alphas were generated from SAS and these yearly alphas were then applied to the SPFs to obtain the EB estimates of PDO accidents. Three methods are introduced to arrive at the PDO estimation:

A) The SPF estimates for PDO accidents were calculated by applying the corresponding yearly alpha:

$$\text{SPF}_{\text{PDO}} = \text{Predicted } \{\kappa\}_{\text{PDO}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{PDO}} \tag{6.8}$$

where κ is the expected accident frequency. Subsequently, the EB procedure was then performed to obtain the EB estimates.

B) The PDO estimates were computed by subtracting the SPF for injury accidents from SPF for total accidents:

$$\text{SPF}_{\text{Total}} - \text{SPF}_{\text{Injury}} = \text{Predicted } \{\kappa\}_{\text{PDO}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{PDO}} \tag{6.9}$$

where κ is the expected accident frequency. Then EB procedure was then performed to obtain the EB estimates.

C) The PDO estimates for both total and injury accidents were determined from the corresponding SPFs:

$$\text{SPF}_{\text{Total}} = \text{Predicted } \{\kappa\}_{\text{Total}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{Total}} \tag{6.10}$$

$$\text{SPF}_{\text{Injury}} = \text{Predicted } \{\kappa\}_{\text{Injury}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{Injury}} \tag{6.11}$$

Next, the EB estimates for total and injury accidents were calculated individually. Last, the EB estimates for PDO accidents were determined by subtracting the EB estimates of injury accidents from the EB estimates of total accidents.

$$EB\{\kappa\}_{Total} - EB\{\kappa\}_{Injury} = EB\{\kappa\}_{PDO} \qquad (6.12)$$

These EB predictions for PDO accidents were obtained from two years of data for the three methods and were compared with the observed accidents from the two subsequent years using various GOF measures. The scheme for this methodology is illustrated in Figure 6.1.
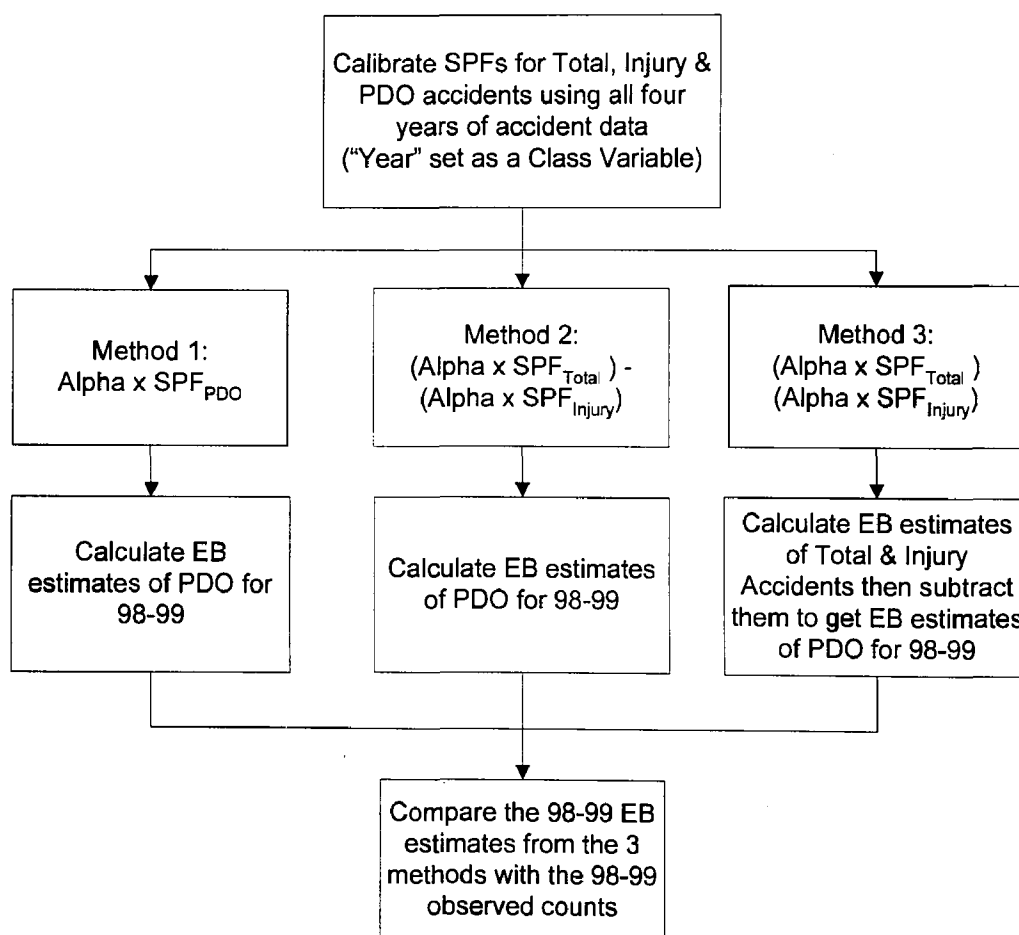
Fig. 6.1 – Methodology for the Class Variable Regression

6.3.2 Scheme Two: Apply Multiplier to the SPF – "Yearly Multipliers" Method

The second method involves the calibration and use of the total, injury and PDO SPFs. These SPFs were calibrated using the GENMOD procedure in SAS and all four years of accident data. Then, yearly multipliers were calculated for total, injury and PDO accidents as the value of the observed accident counts divided by the predicted accidents as shown in Equation 6.6.

$$\text{Multiplier}_Y = \frac{\displaystyle\sum_{i=1}^{n} \text{Observed Counts}}{\displaystyle\sum_{i=1}^{n} \text{Expected Accidents}} \tag{6.13}$$

where i is the total number of sites and Y is the year (i.e. 1995). Again the dataset was subdivided into the two earlier years and two later years. The first two years were used for prediction purposes and the last two years of the data served as the observed accident counts for comparison. These multipliers were then applied to the SPFs to obtain the EB estimates of PDO accidents for the two later years. Based on the EB estimates for PDO accidents, comparisons of the following three different methods were made.

A) The SPF estimates for PDO accidents were calculated by applying the predetermined yearly multipliers:

$$(\text{Multipier}) \times \text{SPF}_{\text{PDO}} = \text{Predicted } \{\kappa\}_{\text{PDO}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{PDO}} \tag{6.14}$$

where κ is the expected accident frequency. Then the EB procedure was performed to obtain the EB estimates.

B) The SPF estimates for PDO accidents were computed by subtracting the SPF estimates for injury accidents from the SPF estimates for total accidents:

$$(\text{Multipier}) \times \text{SPF}_{\text{Total}} - (\text{Multipier}) \times \text{SPF}_{\text{Injury}} = \text{Predicted } \{\kappa\}_{\text{PDO}}$$

$$\text{Predicted } \{\kappa\}_{\text{PDO}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{PDO}} \tag{6.15}$$

where κ is the expected accident frequency. As before, the EB procedure was used to determine the EB estimates.

C) First, the SPF estimates for the total and injury accidents were calculated respectively by applying the multipliers with the SPFs:

$$(\text{Multipier}) \times \text{SPF}_{\text{Total}} = \text{Predicted } \{\kappa\}_{\text{Total}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{Total}} \tag{6.16}$$

$$(\text{Multipier}) \times \text{SPF}_{\text{Injury}} = \text{Predicted } \{\kappa\}_{\text{Injury}} \xrightarrow{\text{EB Procedure}} \text{EB}\{\kappa\}_{\text{Injury}} \tag{6.17}$$

where κ is the expected accident frequency. The EB estimates for total and injury accidents were calculated individually. Then, the EB estimates for PDO accidents were determined by subtracting the EB estimates of injury accidents from the EB estimates of total accidents.

$$\text{EB}\{\kappa\}_{\text{Total}} - \text{EB}\{\kappa\}_{\text{Injury}} = \text{EB}\{\kappa\}_{\text{PDO}} \tag{6.18}$$

The PDO estimates obtained from the two years of data for the above three methods were compared with the observed counts for the two subsequent years using various goodness of fit statistics. The scheme for this method is shown in Figure 6.2.

```
                    ┌─────────────────┐
                    │ Calibrate SPFs for │
                    │ Total, Injury & PDO │
                    │ accidents using all four │
                    │ year of accident data │
                    └─────────────────┘
                             │
                             ▼
                    ┌─────────────────┐
                    │ Calculate Yearly │
                    │ Multipliers for 95-99 │
                    └─────────────────┘
```
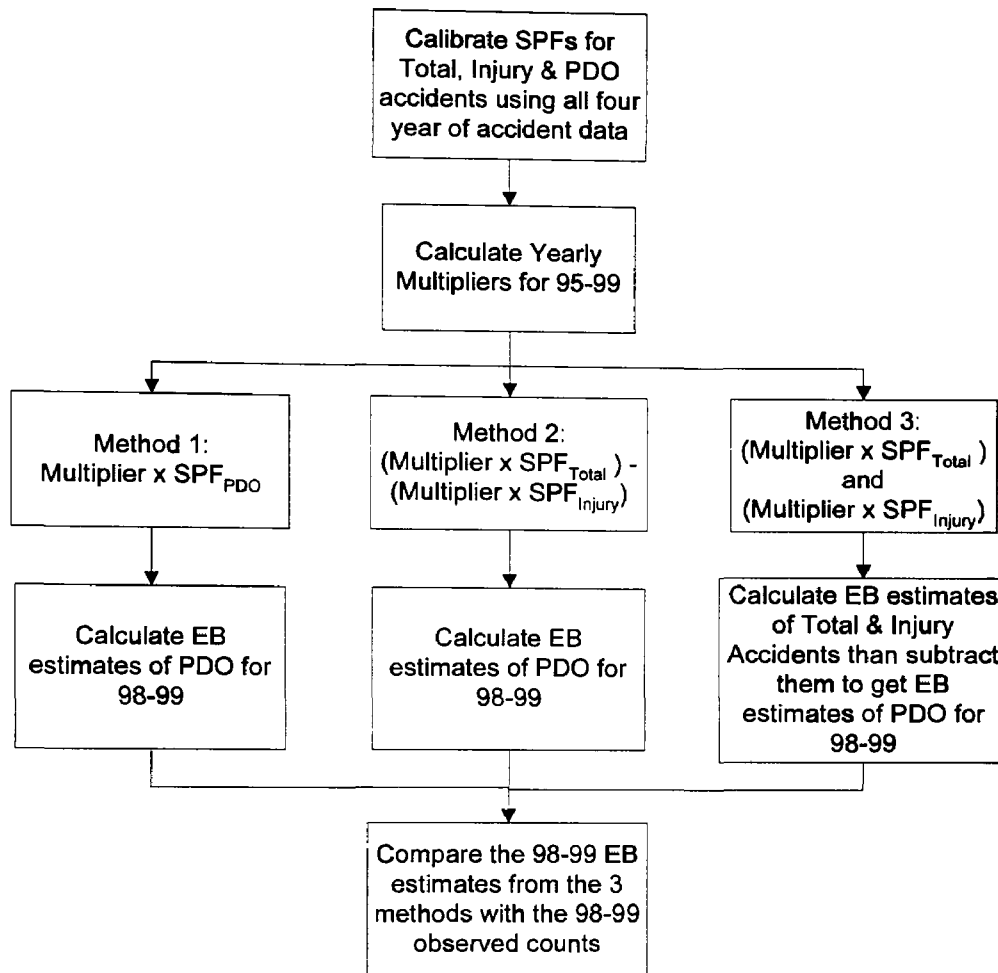
Fig. 6.2 – Methodology for the Yearly Multipliers Method

Method 1:
Multiplier x SPF$_{PDO}$

Method 2:
(Multiplier x SPF$_{Total}$) - (Multiplier x SPF$_{Injury}$)

Method 3:
(Multiplier x SPF$_{Total}$) and (Multiplier x SPF$_{Injury}$)

Calculate EB estimates of PDO for 98-99

Calculate EB estimates of PDO for 98-99

Calculate EB estimates of Total & Injury Accidents than subtract them to get EB estimates of PDO for 98-99

Compare the 98-99 EB estimates from the 3 methods with the 98-99 observed counts

6.4 Application to Minnesota Rural 2-lane Undivided Road Segments

The negative binomial regression and "yearly multipliers" schemes were applied following similar procedures and the three methods to derive EB estimates for PDO accidents, using the HSIS Minnesota rural 2-lane undivided road segments database. Four years of accident and traffic data, from 1995-1996 to 1998-1999, were used to estimate SPFs and the corresponding EB estimates. The Minnesota database was divided into two. The 1995-1996 accident data were used to obtain the EB estimates, whereas the 1998-1999 accident data were treated as observed counts to compare with the EB estimates based on the 1995-1996 data.

6.4.1 Development of Safety Performance Functions for Scheme One

Safety performance functions were developed from the HSIS database adopting the GENMOD procedures in SAS. The calibration methods follow the same GENMOD procedures as described in the previous chapter. In the same manner, SPFs were calibrated in the simplest functional form with AADT as the only independent variable; however, this time the variable 'year' was used as a class variable, as shown in Equation 6.12.

$$\kappa(accidents\ /\ mile) = f(AADT) = \alpha_{year}(AADT)^{\beta} \tag{6.19}$$

SPFs for total, injury and PDO accidents were calibrated and the SPF parameter estimates are shown on Table 6.1.

Table 6.1 – Scheme One: SPF Parameter Estimates for Minnesota Rural 2-lane Undivided Highways

	$\ln(\alpha_{1995})$ (S.E.)	$\ln(\alpha_{1996})$ (S.E.)	$\ln(\alpha_{1998})$ (S.E.)	$\ln(\alpha_{1999})$ (S.E.)	β (S.E.)	k
Total	-7.4128 (0.0224)	-7.3050 (0.0220)	-7.5030 (0.0223)	-7.6315 (0.0511)	0.8877 (0.0066)	0.6841
Injury	-7.6838 (0.0319)	-7.6143 (0.0314)	-7.7768 (0.0318)	-7.9557 (0.0726)	0.7973 (0.0093)	0.6214
PDO	-8.2216 (0.0281)	-8.0910 (0.0275)	-8.3117 (0.0280)	-8.4423 (0.0650)	0.9283 (0.0084)	0.8777

* S.E. stands for standard error

These parameters were used to calculate the expected accidents for each site. The estimates of the overdispersion parameter, k, were used in the EB estimation for each of three methods introduced in Section 6.3. The EB procedure is outlined in Section 6.2.

6.4.2 Development of Safety Performance Functions for Scheme Two

Again, SPFs for total, injury, and PDO collisions with AADT as the only covariate were calibrated using the GENMOD procedures in SAS. The estimated parameters are

shown in Table 6.2. Note that the estimates of β are quite different from those in Table 6.1.

Table 6.2 – Scheme Two: SPF Parameter Estimates for Minnesota Rural 2-lane Undivided Highways

Model Form: $\kappa(accidents / mile / year) = \alpha(AADT)^{\beta}$

	ln α (S.E.)	β (S.E.)	k
Total	-8.0658 (0.0566)	0.9718 (0.0079)	0.6033
Injury	-8..2071 (0.0713)	0.8588 (0.0099)	0.4752
PDO	-8.8996 (0.0702)	1.0176 (0.0097)	0.7608

* S.E. stands for standard error

The yearly multipliers calculated by dividing the observed accident counts in each year by the expected accident counts derived from the SPFs are presented in Table 6.2. Table 6.3 displays the yearly multipliers for each of the collision types.

Table 6.3 – Yearly Multipliers for Individual Collision Type

	1995	1996	1998	1999
Total	1.0295	1.1415	0.9288	0.6686
Injury	1.0902	1.1650	0.9850	0.7371
PDO	1.0051	1.1422	0.9086	0.6467

The SPFs and the yearly multipliers were used to calculate the expected accident frequency for the first two years, while the overdispersion parameter, k, was used to derive the EB estimates as shown in Section 6.2.

6.5   Interpretations of the Results

The overall GOF measures for each method for both schemes were calculated and are listed in Tables 6.4 and 6.5, respectively.

Table 6.4 - Validation Statistics for Scheme One* (Negative Binomial Regression)

Method	MPB	MAD	MSE	MPSE	Predicted Accidents	Observed Counts
A	0.121	0.762	469.88	469.86	8,352	7,575
B	0.109	0.770	383.04	383.03	8,561	7,575
C	0.103	0.770	339.37	339.36	8,673	7,575

*Goodness of Fit measures were calculated as per mile-year base

Method A: Using calibrated PDO SPF estimate in EB procedure

Method B: Subtracting injury SPF estimate from total accidents SPF estimate to get PDO SPF estimate used for EB PDO estimate

Method C: Subtracting EB injury estimate from EB total estimate to get EB PDO estimate

Table 6.5 - Validation Statistics for Scheme Two* ("Yearly Multipliers" Method)

Method	MPB	MAD	MSE	MPSE	Predicted Accidents	Observed Counts
A	0.148	0.743	708.08	708.06	7,975	7,575
B	0.145	0.745	679.43	679.41	7,928	7,575
C	0.141	0.744	642.58	642.56	7,836	7,575

*Goodness of Fit measures were calculated as per mile-year base

Method A: Using calibrated PDO SPF estimate in EB procedure

Method B: Subtracting injury SPF estimate from total accidents SPF estimate to get PDO SPF estimate used for EB PDO estimate

Method C: Subtracting EB injury estimate from EB total estimate to get EB PDO estimate

The validation statistics show that both schemes are slightly over-predicting crashes. The GOF measures are very similar in both cases, and the predicted values generated from both methods are very reliable. Generally speaking, the expected accident counts suggest that the "yearly multipliers" technique predicts accidents more accurately compared to the negative binomial regression, as shown in Table 6.5, where the predicted accidents closely resemble the observed counts. As for the three different methods, the results from both schemes suggest that method A (using a calibrated PDO SPF estimate in the EB procedure) predicts accidents more accurately compared to methods B and C.

6.6   Chapter Summary

Despite the similarity of the values of the expected and observed accidents as calculated using the "yearly multipliers" method, one cannot hastily conclude that the multiplier method is as good as the negative binomial regression. This is because adopting multipliers in the SPFs is somewhat forcing the sum of predicted accidents to equal the sum of observed counts because the ratios were actually determined using all four years of the same dataset.

That notwithstanding, the results obtained from the directly calibrated PDO model (Method A) are more satisfactory in both schemes compared to the other two methods. Simply put, it is recommended that the SPF for PDO accidents be calibrated, at least for the dataset explored. Nevertheless, it is suggested to explore the same issue on other datasets before one can generalize the results.

CHAPTER SEVEN

CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions and Recommendations for Future Work

Typically SPFs are developed for total, injury, or PDO accidents. However, there is a demand for development of SPFs for specific accident types so that network screening and treatment evaluation in *SafetyAnalyst* and similar tasks can be disaggregated by these accident types. In light of the fact that calibrating SPFs for individual collision types is not trivial and is somewhat difficult in the absence of relevant information, inclusion of a factor (accidents' proportion) in *SafetyAnalyst* to apply with the SPF for total collision has been recommended.

A comparison of the two methods of negative binomial regression and "proportion" methods, which are used in developing SPFs for specific accident types on road segments, shows no significant pattern in the findings. Therefore, it is impractical to generalize which method works better in each case. It seems worthwhile to calibrate appropriate SPFs for individual collision types when there is a large sample of accident counts. Conversely, the "proportion" method seems to be a more logical approach in predicting accidents when the database lacks sufficient accident counts. Further studies of this issue are suggested for databases for other jurisdictions and entity types, to identify patterns that may be generalizable.

A careful study of the proportion method for developing SPFs for specific accident types reveals that generalized linear modelling is capable of modelling collisions by proportion. It is reasonable to assume that the error structure for the dependent variable

follows the normal distribution and that the functional form follows either the power or gamma function. Application of network screening based on angle collisions at TWSC intersections suggests that whether the proportions of specific accident types remain constant or not does not significantly influence of the ranking. In other words, applying a constant factor or proportion model would not drastically change the results of the ranking. Based on the current dataset, one can conclude that it is reasonable to apply a constant factor to the SPF for total collision for network screening purposes. Moreover, it is suggested that a similar application for other collision types on the same database as well as datasets from other jurisdictions be repeated in order to draw a more definitive conclusion.

Lastly, the detailed investigation of methods for estimating the expected number of PDO accidents demonstrates that it is logical and practical to model PDO accidents despite the errors during the data collection stage. In general, both 'yearly multiplier' and negative binomial regression methods are accurate ways of estimating PDO accidents. Although one cannot generalize which method is more accurate in safety prediction, it is recommended that the SPF for PDO accidents be calibrated based on the dataset used. Notwithstanding that, further explorations on other datasets are advisable before one can make a more definitive statement.

APPENDIX A

STATISTICS BACKGROUND FOR ROAD SAFETY

A.1 Normal Probability Plot

Normal probability plot is a tool to assess the normality of the data. Generally speaking, the probability plot is a graphical method for determining whether a set of data is conformed by the hypothesized distribution (the distribution of accident proportion, in our case). In the present context, one is interested in showing whether the proportion of specific accident type follows the normal distribution.

The concept of normal probability plot is very simple and convenience. It works well for both large and small sample sizes. A stepwise procedure is presented below.

1) The data is disaggregated such that similar AADT are grouped together. In the present context, an interval of five hundreds is used.

2) Each group is ranked from smallest to largest, x_1, x_2, ..., x_n. Since there are too many groups in this case, only ten groups are randomly selected from various collision types and applied to this test.

3) The ordered observations x_n are plotted against their cumulative frequency. The cumulative frequency for normal distribution is calculated using Equation A.1,

$$\text{Cumulative Frequency} = \frac{j}{(n+1)} \tag{A.1}$$

where j is the rank order and n is the total numbers of observation.

4) If the proposed distribution adequately describes the data, the plotted points are expected to lie approximately on a straight line. Otherwise, the plotted points

would scatter all over the place.

Figure A.1 shows the normal probability plots for the ten randomly selected groups.
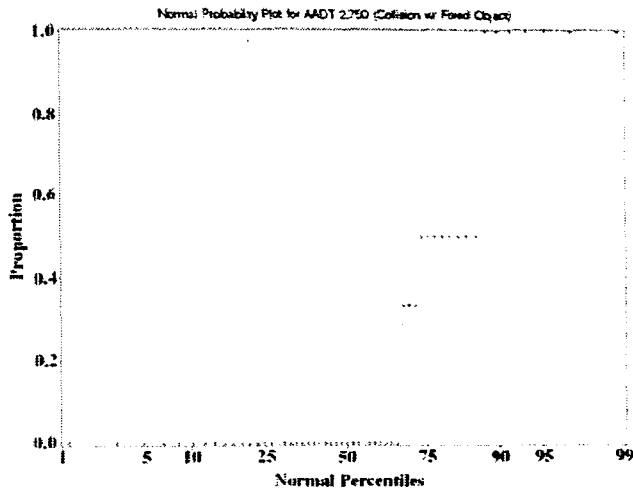
**Figure A.1a – Normal Probability Plot for AADT
2,750 for Collision with Fixed Object**
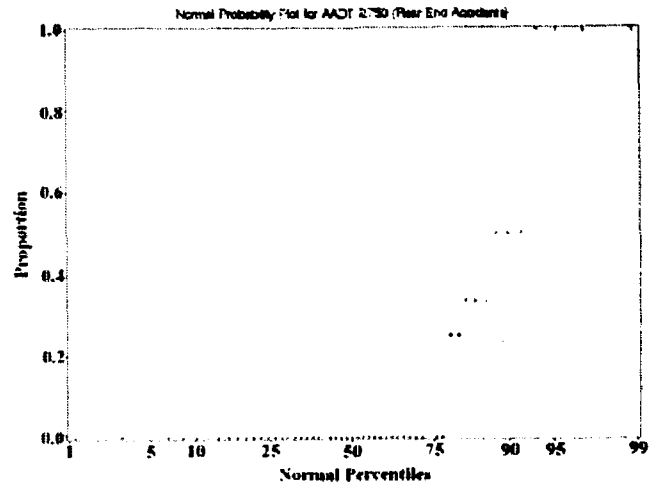
**Figure A.1b – Normal Probability Plot for AADT
2,750 for Rear-End Accidents**
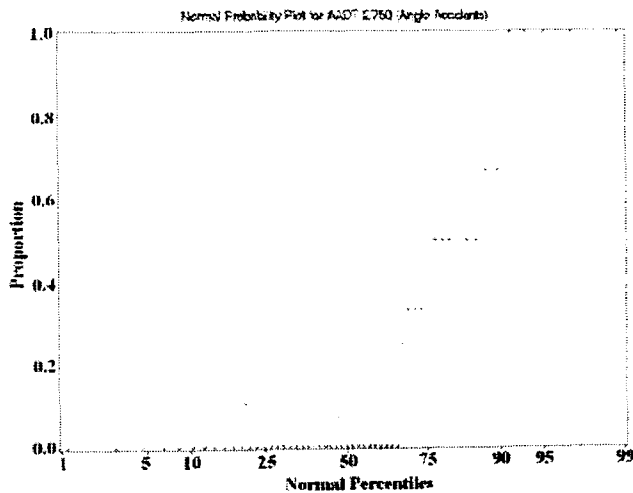
**Figure A.1c – Normal Probability Plot for AADT
2,750 for Angle Accidents**
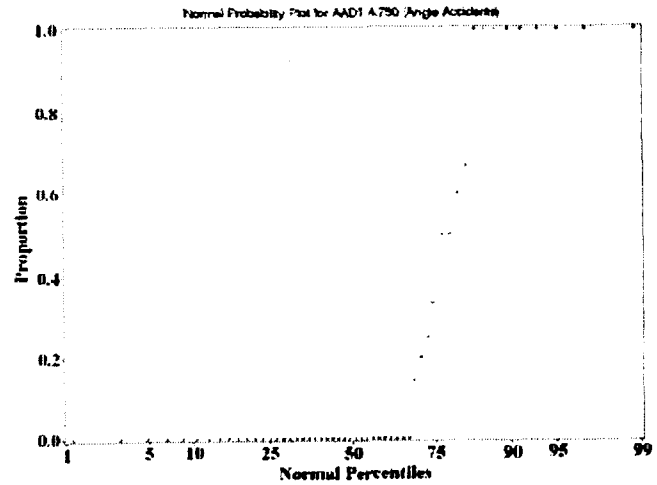
**Figure A.1d – Normal Probability Plot for AADT
4,750 for Angle Accidents**

91
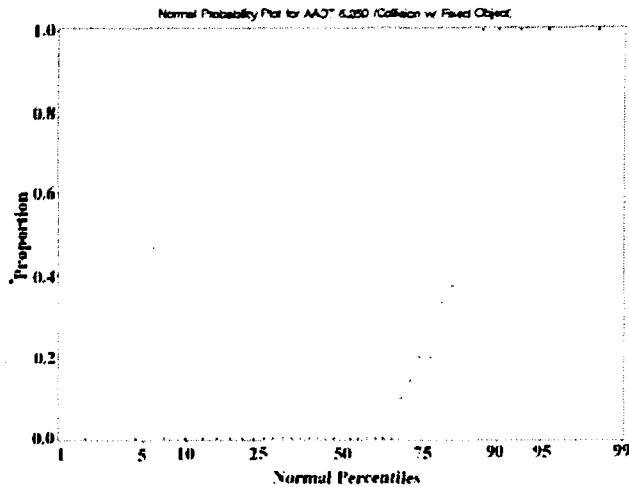
Figure A.1e – Normal Probability Plot for AADT 6,250 for Collision with Fixed Object
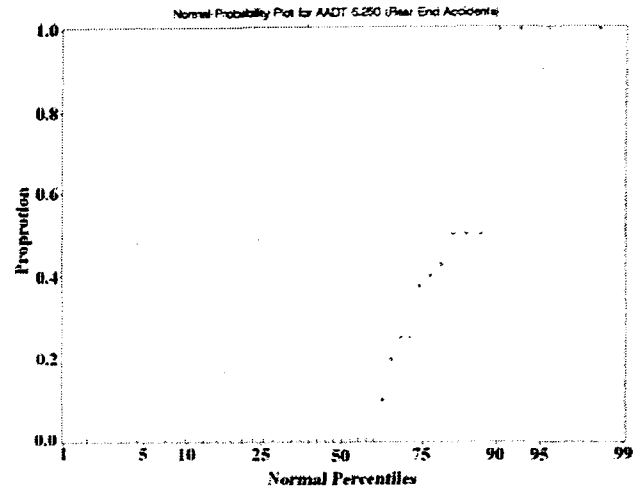
Figure A.1f – Normal Probability Plot for AADT 6,250 for Rear-End Accidents
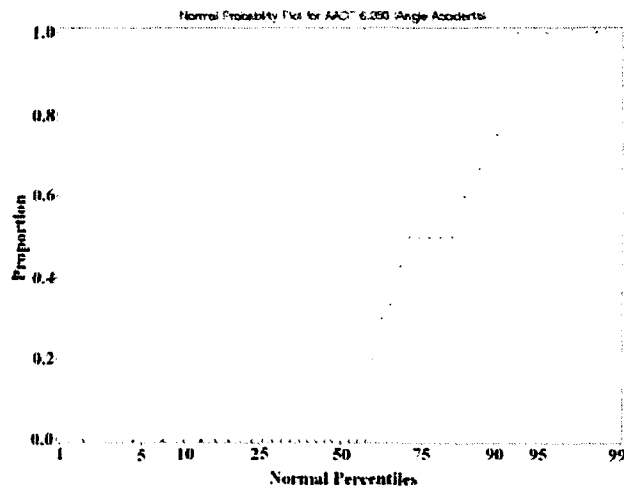
Figure A.1g – Normal Probability Plot for AADT 6,250 for Angle Accidents
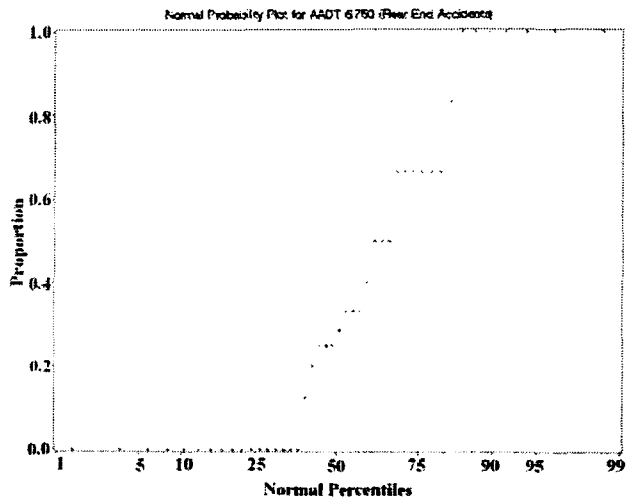
Figure A.1h – Normal Probability Plot for AADT 6,750 for Rear-End Accidents
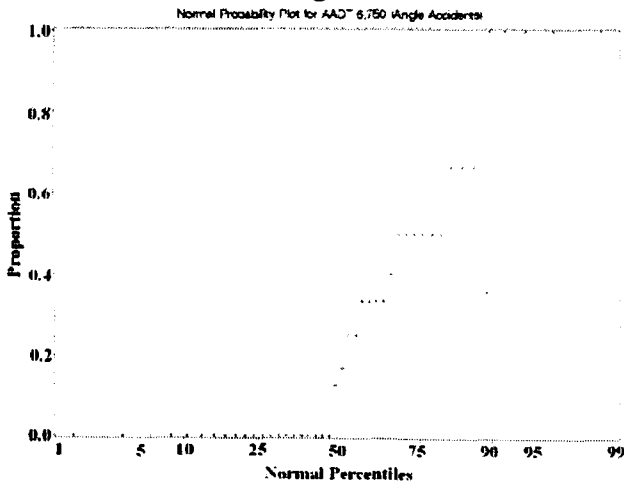
Figure A.1i – Normal Probability Plot for AADT 6,750 for Angle Accidents
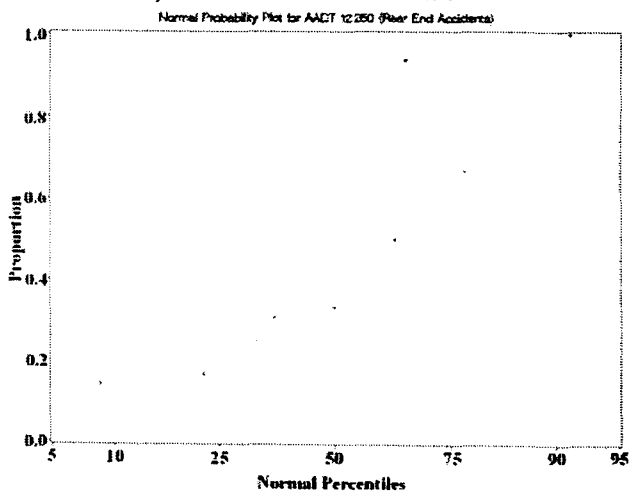
Figure A.1j – Normal Probability Plot for AADT 12,250 for Rear-End Accidents

All of the graphs reveal that the plotted points fall approximately along a straight line, meaning that the normal distribution adequately describes the accidents' proportion. Therefore, the assumption of choosing normal distribution as the response variable for generalized linear model is indeed valid.

APPENDIX B

RESULTS AND FINDINGS

B.1 A Detailed Interpretation of the Results for Chapter Four

| Proportion Method | SPF Method |
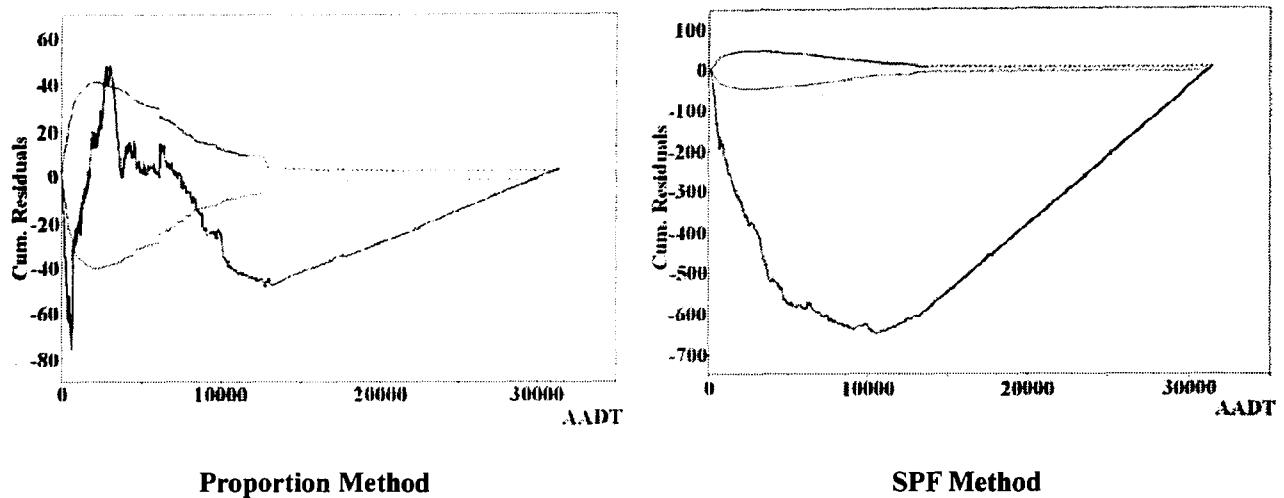
Figure B.1 - Cumulative Residuals vs. AADT for Collision Type 1

Type 1: Collision with Parked Motor Vehicle

The validation statistics in Table 4.5 shows that both models are very similar in terms of the overall fitness. Pearson's correlation coefficients explain that both models lack of linear correlation. The MPB/mi-yr statistics show that the SPF model has a smaller average model bias and MAD/mi-yr statistics shows that the "proportion" model has a slightly better fit. Finally, the MSE/mi-yr^2 and MSPE/mi-yr^2 illustrate that the two models have a similar fit relative to the validation data. The overdispersion parameter was found to be 17.8965, which indicates that the SPF model is not well established. The plot of cumulative residuals against AADT for collision with parked motor vehicle is shown in Figure B.1. The CURE plot for the "proportion" method in Figure B.1 shows

an oscillation of cumulative residuals around zero, an indication that the "proportion" model fits better compared to the SPF model. The "proportion" model seems to be overestimating crashes in low and mid ranges AADTs and underestimating between AADT 1,000 and 3,500 and AADT from 13,000 onwards. Moreover, there are a few locations in this model where the AADT exceeds the $\pm 2\sigma$ boundaries. The peak values were found to be +48 and -75 with 1,015 accidents in total; therefore, the fit is ideal.

Proportion Method   **SPF Method**
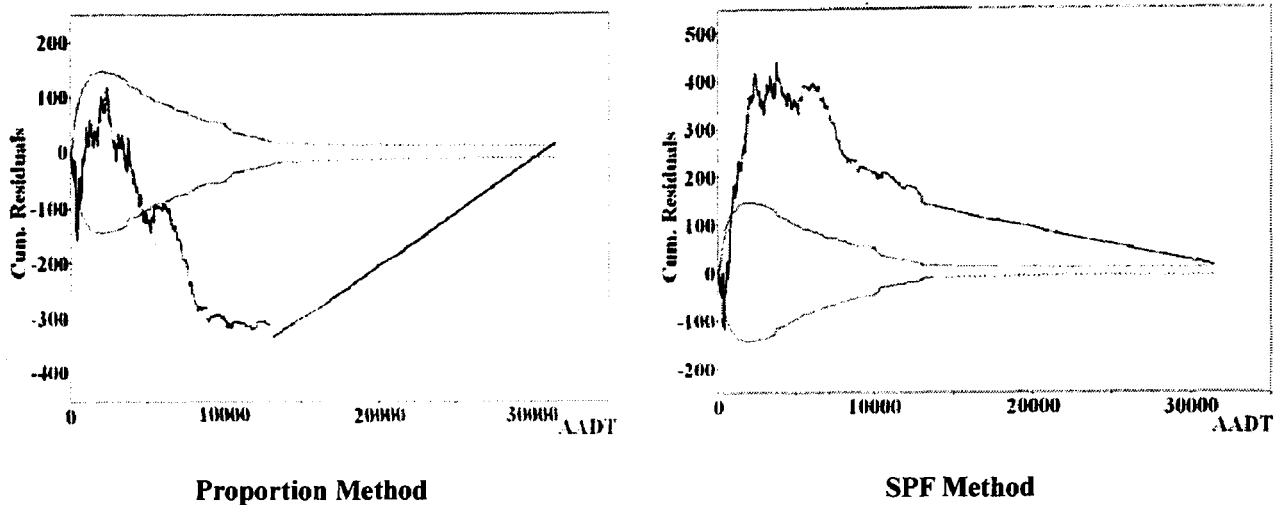
Figure B.2 - Cumulative Residuals vs. AADT for Collision Type 5

Type 5: Collision with Animal

Pearson's correlation coefficients show a fairly linear correlation for both models.   While a

comparison of MPB/mi-yr shows that the "proportion" model has somewhat over-predicted

crashes on average, the figure for SPF method was found to be approximately zero,

suggesting that the model is of little average bias.   Other GOF statistics shows marginal

differences between the two models.   The overdispersion parameter for the SPF was found

to be 1.6576, which indicates that the goodness of fit to the dataset is fair.   In Figure B.2,

the CURE plot for the proportion model shows an encroachment of random walk exceeding

the lower boundary between AADT 5,000 and 30,000.   The peak values were found to be

+120 and -320 with a total of 6,938 accidents, which is satisfactory for that particular number

of accident counts.   The random walk in the SPF method illustrates an intrusion of the upper

boundary between AADT 2,500 and 30,000.   The peak values of cumulative residuals were

found to be +420 and -110, not as exact as the proportion method.   In brief, the "proportion"

model shows a better predictability.   However, the extrusion of the cumulative residuals

curve between AADT 5,000 and 30,000 suggests a need of detailed investigation or even

perhaps the use of a better functional form.

Proportion Method

SPF Method
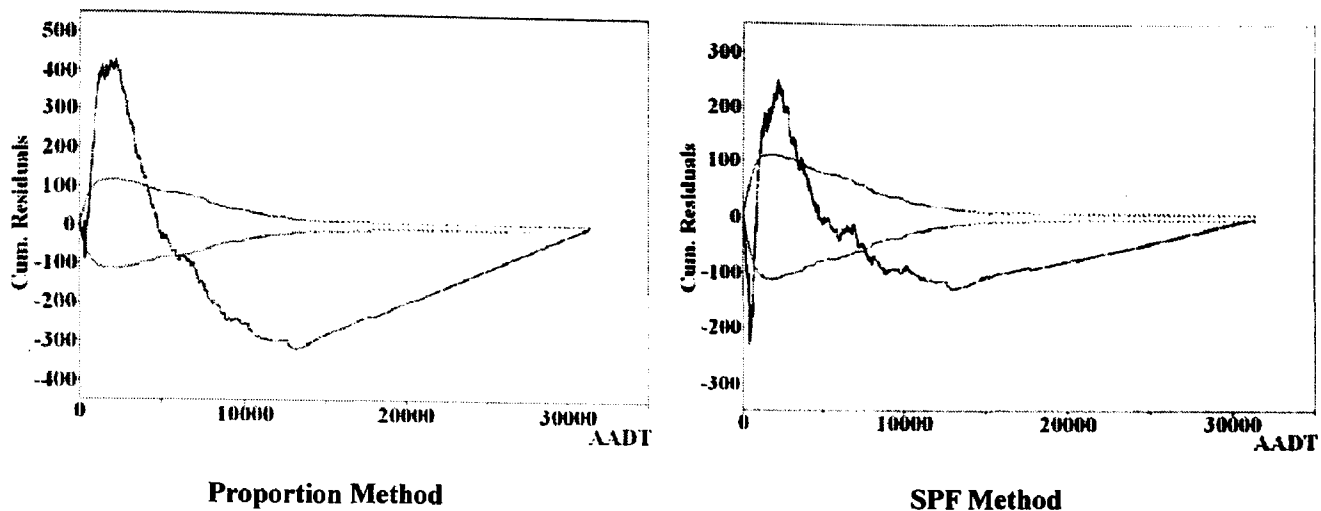
Figure B.3 - Cumulative Residuals vs. AADT for Collision Type 6

Type 6: Collision with Fixed Object

The results of the goodness of fit statistics are alike for both models, with the exception of

which the MPB/mi-yr figures show an improvement when predicting observed data in the

"proportion" model. Pearson's correlation coefficients suggest that a moderate linear

correlation exists in both models. The overdispersion parameter was found to be 0.9947.

This small value indicates that the variance of the accident prediction model is small; thus,

the SPF is satisfactory. The CURE plots shown in Figure B.3 illustrate that both models

fit poorly for all ranges of AADT. Generally speaking, the SPF model over-predicts

crashes in low and mid ranges AADTs and under-predicts crashes between AADT 1,000

and 3,000 and AADT above 13,000. In contrast, the cumulative residuals' curve exceeds

the $\pm 2\sigma$ boundaries in the SPF model but the largest accumulative residuals are about $\pm$

240 with a total of 6,362 accidents in the data. To a certain extent, this is acceptable,

especially when compared to the "proportion" model.

Proportion Method   **SPF Method**
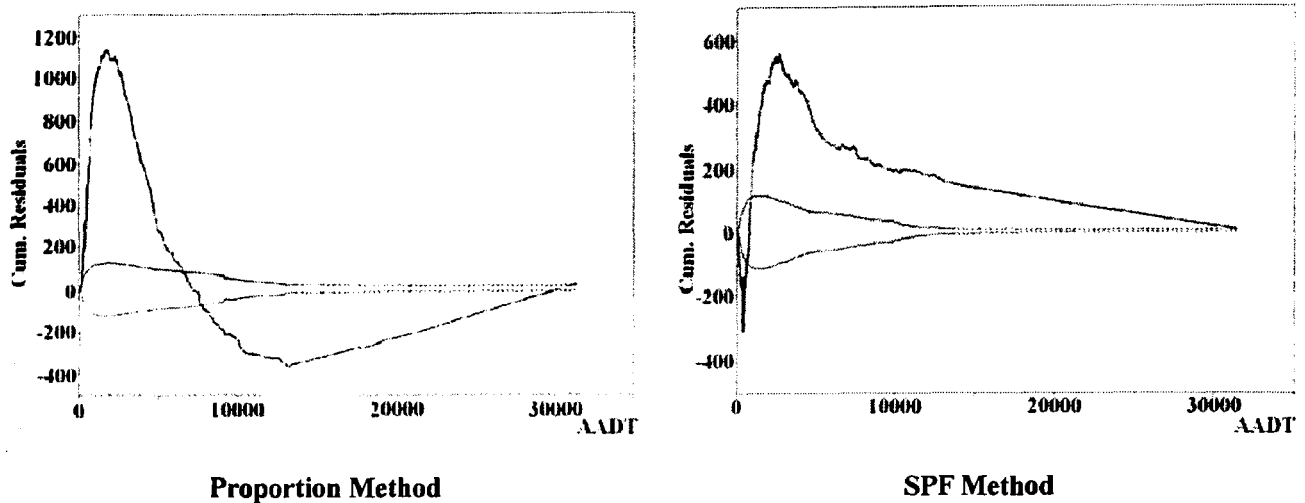
Figure B.4 - Cumulative Residuals vs. AADT for Collision Type 9

Type 9: Overturn

All the validation statistics in Table 4.6 suggest that the SPF method does a better job in predicting accident counts. Pearson's correlation coefficient shows a satisfactory linear correlation between the observed and calibrated data for the SPF method. Also, the small value of overdispersion parameter suggests that the variance of the prediction is very small, which is to say that the SPF model has a better overall GOF. The CURE plots in Figure B.4 show that the cumulative residuals' curve continuously exceeds the $\pm 2\sigma$ boundaries for both models. The models are therefore insignificant in almost all ranges of AADT. Nevertheless, based on 7,803 accident counts in the dataset, the peak values for the SPF method are approximately +550 and -300, comparatively acceptable in this case. One can conclude that the SPF method works better for the overturn model.

Proportion Method · SPF Method
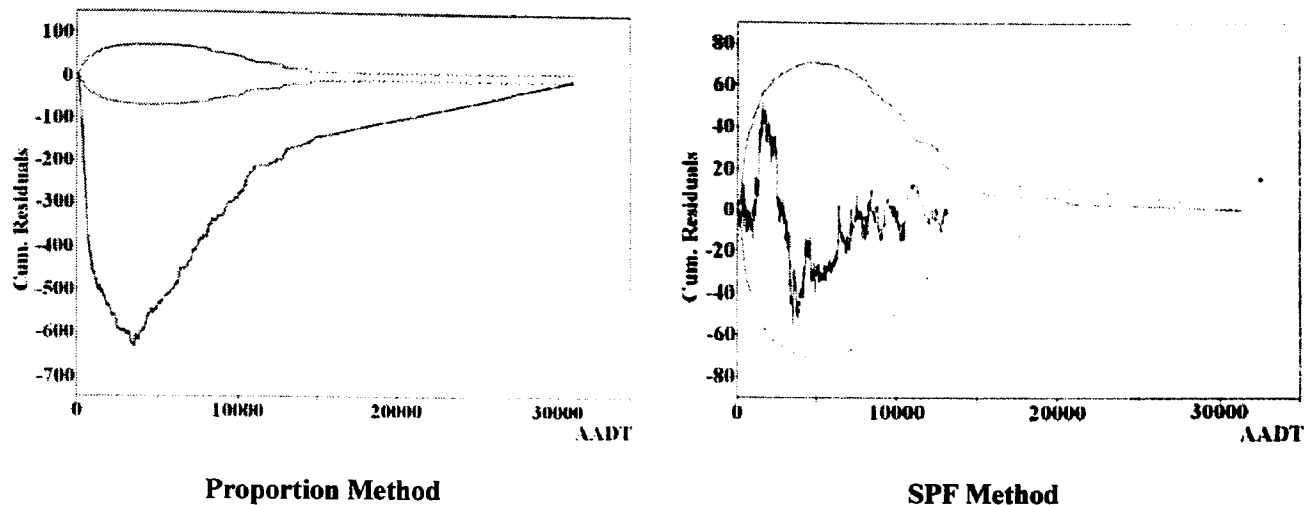
Figure B.5 - Cumulative Residuals vs. AADT for Collision Type 21

<u>Type 21: Rear-End</u>

In terms of goodness of fit statistics, Pearson's correlation coefficients show moderate linear correlation in both models. All validation statistics show that the overall fitting for both models are quite alike. A small overdispersion parameter demonstrates an excellent fit for the SPF. The CURE plots in Figure B.5 show a big discrepancy between the two methods. Whereas the proportion model completely lacks fit in all ranges of AADT, the SPF method shows a very impressive fitting for all ranges of AADT since the random walk reveals no encroachment which exceeds the $\pm 2\sigma$ boundaries. Generally speaking, the SPF overestimates crashes between AADT 0 and 1,000 and between 2,000 and 4,000, but it underestimates crashes between AADT 1,000 and 2,000 and 4,000 onwards. Hence, one can conclude that the SPF model is ideal for the purpose of accident prediction.

| Proportion Method | SPF Method |
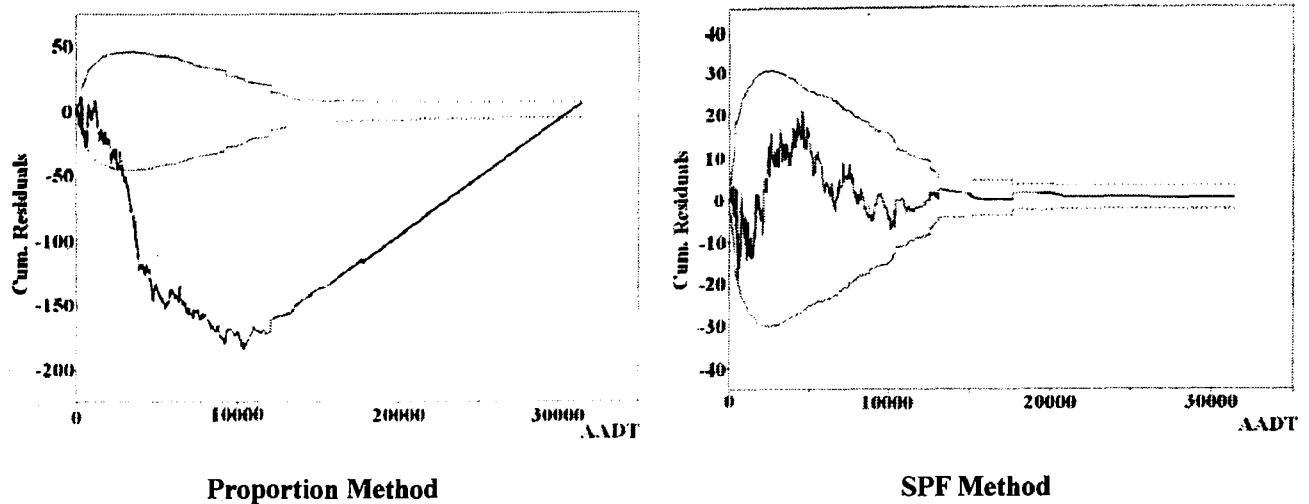
Figure B.6 - Cumulative Residuals vs. AADT for Collision Type 22

Type 22: Head On

Based on the GOF measures, one can conclude that both models have very good fits while Pearson's correlation coefficients suggest that both models lack linear correlation. The overdispersion parameter was found to be 0.0003, and which means that the variance in the predicted accidents is very small. In terms of CURE plots, the SPF method shows an excellent fit in all ranges of AADT and the fluctuations of random walk are all bounded within the $\pm 2\sigma$ boundaries. The consistent upward drift in the low AADT indicates that the model underestimates crashes, whereas the consistent downward drift in the mid- and high-ranges AADT is an indication of overestimation. On the contrary, the CURE plot shows that the proportion method is completely lack of fit in all ranges of AADT. Apparently, the SPF method is a better one for head-on collisions.

Proportion Method   SPF Method
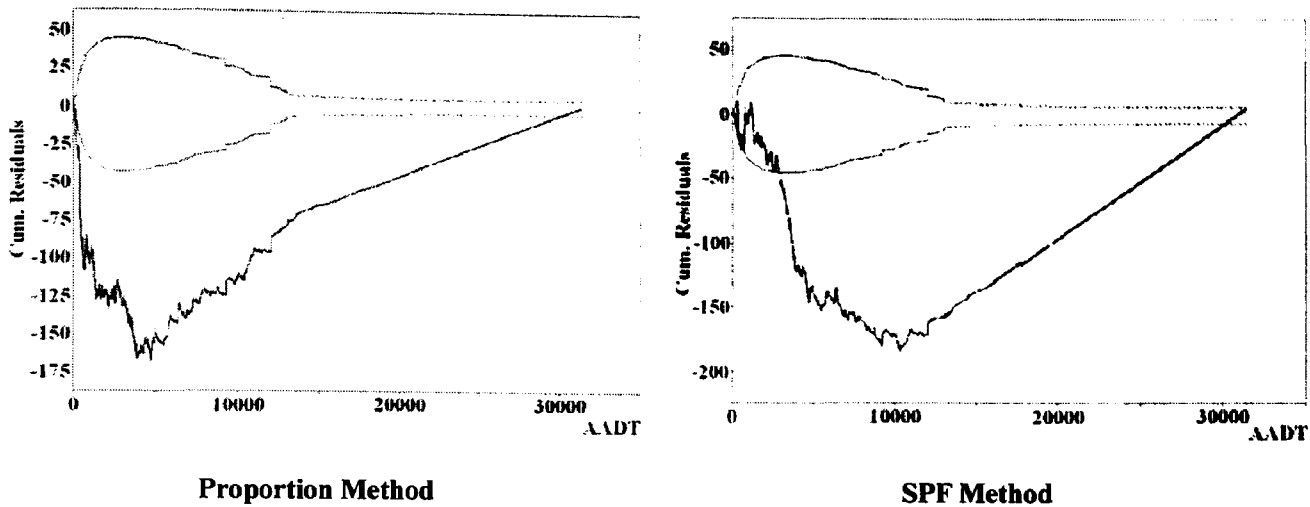
Figure B.7 - Cumulative Residuals vs. AADT for Collision Type 24

Type 24: Angle

The validation statistics reveal a good overall fit for both models, except for the correlation coefficients, which indicate that both models lack linear association. The overdispersion parameter denotes a high variance of the SPF. As a result, the prediction is not ideal. The CURE plots show that both models completely lack fit in all ranges of AADT since the fluctuations of random walk are all below zero. As the proportion model seems to overestimate crashes between AADT 0 and 5,000 and under-predict crashes above 5,000, the SPF method seems to overestimate crashes between AADT 0 & 10,000 and under-predict crashes above. In brief, the SPF method is slightly better than the proportion method. Nevertheless, both models are poorly developed and further investigation is required for a more accurate measurement.

Proportion Method   **SPF Method**
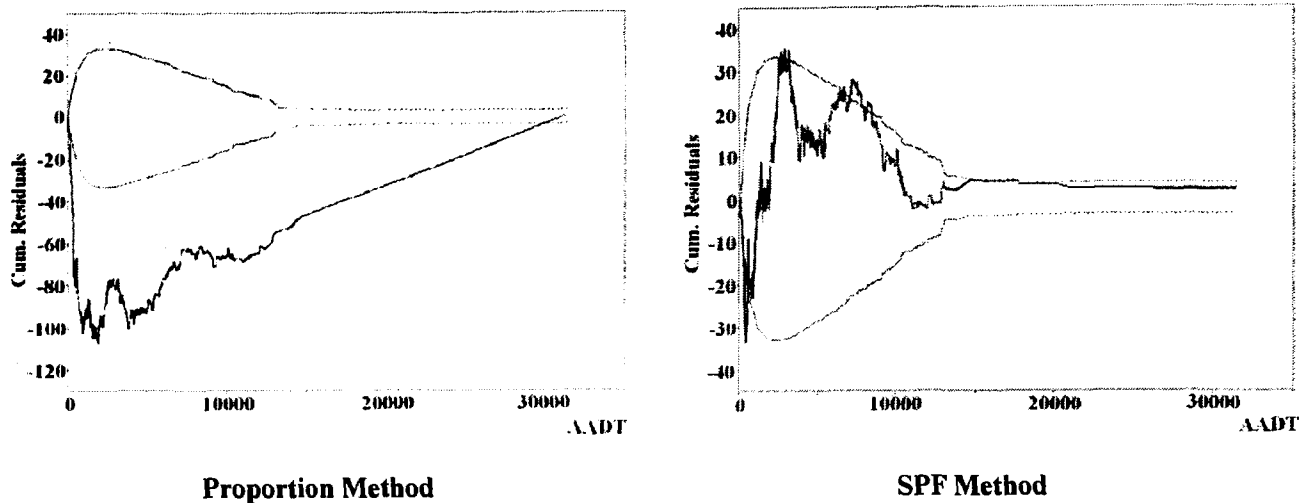
Figure B.8 - Cumulative Residuals vs. AADT for Collision Type 25

Type 25: Sideswipe, Same Direction

The validation statistics indicate an excellent overall fit for both models, but Pearson's correlation coefficients denote that both of the models lack linear correlation. The overdispersion parameter suggests that the SPF method is more exact because a small overdispersion implies only a small variance of the prediction. CURE plots suggest that the SPF model will predict accidents more accurately in all ranges of AADT. Unlike the proportion model, the fluctuation of random walk does not oscillate around zero. However, the cumulative residuals in the SPF method display some ranges of AADT which exceed the $\pm 2\sigma$ boundaries. Thus, the model requires some detailed investigation for those ranges of AADT. In general, SPF seems to be underestimating crashes between AADT 500 and 3,000 and between 5,000 and 8,000; other ranges seem to be overestimating crashes. The peak values of the CURE plot for the SPF method were found to be +35 and -34 respectively. Hence, the SPF method is still adequate even though the cumulative residuals exceed the $\pm 2\sigma$ boundaries.

B.2   Network Screening Results for Chapter Five

Table B.1 – Top 20 Rankings for Minnesota Rural 3-leg TWSC Intersections

Screening Criterion   :   Excess Accident Frequency

Accident Type   :   Angle Collisions

No of Sites Ranked   :   2,033

Rank	Method 1			Method 2			Method 3		
	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)
1	281	0.61	0.07	281	0.97	0.09	281	0.91	0.07
2	1355	0.45	0.18	1355	0.66	0.12	1355	0.65	0.12
3	1078	0.40	0.09	1078	0.64	0.08	1078	0.61	0.07
4	277	0.39	0.19	84	0.57	0.08	277	0.55	0.10
5	84	0.37	0.10	277	0.56	0.11	84	0.55	0.07
6	2017	0.36	0.36	2017	0.51	0.16	2017	0.51	0.16
7	97	0.35	0.30	97	0.51	0.15	97	0.50	0.15
8	103	0.32	0.19	103	0.49	0.13	103	0.49	0.13
9	1352	0.28	0.05	1352	0.48	0.05	1352	0.44	0.04
10	102	0.28	0.10	102	0.44	0.08	102	0.43	0.07
11	646	0.28	0.40	22	0.41	0.07	646	0.40	0.15
12	1371	0.27	0.29	646	0.40	0.15	22	0.40	0.06
13	202	0.26	0.21	657	0.40	0.06	1371	0.39	0.12
14	497	0.26	0.21	909	0.40	0.04	497	0.38	0.10
15	657	0.26	0.09	1371	0.40	0.13	657	0.38	0.05
16	22	0.26	0.08	497	0.39	0.11	2018	0.38	0.10
17	2018	0.25	0.18	2018	0.38	0.10	201	0.38	0.09
18	201	0.25	0.16	201	0.38	0.10	202	0.37	0.09
19	909	0.22	0.03	202	0.38	0.10	909	0.37	0.03
20	1613	0.22	0.05	1613	0.37	0.05	1613	0.35	0.04

Table B.2 – Top 20 Rankings for Minnesota Rural 4-leg TWSC Intersections

Screening Criterion   :   **Excess Accident Frequency**

Accident Type   :   **Angle Collisions**

No of Sites Ranked   :   **2,572**

Rank	Method 1			Method 2			Method 3		
	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)
1	247	2.67	1.58	146	2.96	0.86	247	2.93	1.32
2	146	2.63	1.00	1413	2.93	0.61	146	2.93	0.78
3	1413	2.57	0.63	247	2.88	1.51	1413	2.85	0.56
4	1405	2.33	1.24	1405	2.58	1.09	1405	2.60	0.97
5	925	2.08	1.96	925	2.29	1.38	925	2.35	1.21
6	257	1.99	1.00	1335	2.23	0.92	1335	2.24	0.82
7	1335	1.99	1.18	257	2.21	0.98	257	2.24	0.86
8	1757	1.86	0.54	1329	2.19	0.56	1329	2.16	0.51
9	1329	1.84	0.50	1757	2.14	0.49	391	2.11	1.30
10	118	1.82	1.02	391	2.03	1.50	1757	2.09	0.45
11	391	1.79	2.46	118	2.03	0.91	118	2.05	0.81
12	222	1.67	2.06	1750	2.00	0.48	1750	1.96	0.44
13	1750	1.66	0.43	1311	1.90	0.39	122	1.91	2.41
14	122	1.51	5.09	222	1.80	1.54	222	1.89	1.32
15	1311	1.51	0.30	122	1.75	2.89	1311	1.84	0.35
16	973	1.47	1.11	2393	1.70	0.33	744	1.76	1.64
17	700	1.44	1.32	1120	1.68	0.45	973	1.68	0.75
18	1120	1.43	0.46	973	1.65	0.85	1120	1.66	0.41
19	250	1.43	3.30	744	1.65	1.93	2393	1.65	0.30
20	2393	1.39	0.29	700	1.58	1.04	250	1.65	1.92

Table B.3 – Top 20 Rankings for Minnesota Rural 3-leg TWSC Intersections

Screening Criterion   :   Expected Accident Frequency

Accident Type   :   Angle Collisions

No of Sites Ranked   :   2,033

Rank	Method 1			Method 2			Method 3		
	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)
1	281	0.75	0.03	281	1.09	0.07	281	1.01	0.06
2	2017	0.71	0.06	1355	0.86	0.08	1355	0.84	0.08
3	1355	0.69	0.05	1078	0.78	0.06	2017	0.77	0.08
4	97	0.67	0.05	2017	0.78	0.08	97	0.75	0.08
5	646	0.65	0.06	97	0.76	0.08	1078	0.73	0.05
6	277	0.63	0.04	277	0.76	0.07	277	0.73	0.07
7	1371	0.59	0.05	84	0.72	0.06	103	0.72	0.07
8	103	0.57	0.04	103	0.72	0.07	84	0.68	0.05
9	488	0.56	0.05	646	0.67	0.07	646	0.66	0.07
10	489	0.56	0.06	1371	0.63	0.06	1371	0.62	0.06
11	1078	0.56	0.03	497	0.60	0.06	497	0.59	0.06
12	84	0.54	0.03	102	0.60	0.05	102	0.58	0.05
13	202	0.53	0.04	2018	0.59	0.06	2018	0.57	0.05
14	497	0.53	0.04	1352	0.58	0.04	201	0.57	0.05
15	2018	0.50	0.03	201	0.58	0.05	202	0.56	0.05
16	201	0.48	0.03	202	0.57	0.05	488	0.55	0.06
17	102	0.46	0.02	488	0.56	0.06	1352	0.53	0.03
18	107	0.45	0.04	22	0.56	0.04	22	0.52	0.04
19	1848	0.45	0.03	657	0.54	0.04	657	0.50	0.04
20	645	0.45	0.04	909	0.50	0.03	489	0.49	0.06

Table B.4 – Top 20 Rankings for Minnesota Rural 4-leg TWSC Intersections

Screening Criterion : **Expected Accident Frequency**

Accident Type : **Angle Collisions**

No of Sites Ranked : **2,572**

Rank	Method 1			Method 2			Method 3		
	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)	Site No.	X_Y	Var (X_Y)
1	247	3.58	0.64	247	3.91	0.78	247	3.84	0.75
2	122	3.47	0.70	146	3.53	0.63	122	3.47	0.72
3	146	3.27	0.53	122	3.51	0.74	146	3.45	0.60
4	925	3.17	0.59	1405	3.38	0.65	1405	3.32	0.62
5	1405	3.12	0.54	925	3.31	0.66	925	3.26	0.64
6	391	3.07	0.58	1413	3.28	0.52	1413	3.18	0.49
7	744	3.02	0.60	391	3.15	0.63	391	3.10	0.61
8	1413	2.99	0.42	270	3.07	0.64	270	3.03	0.63
9	270	2.99	0.59	250	3.06	0.64	250	3.02	0.62
10	250	2.97	0.58	744	3.02	0.62	744	2.98	0.60
11	222	2.83	0.53	257	2.98	0.57	222	2.92	0.58
12	261	2.80	0.55	222	2.96	0.60	257	2.92	0.54
13	1335	2.77	0.48	1335	2.95	0.56	1335	2.89	0.53
14	1888	2.71	0.52	261	2.86	0.60	261	2.82	0.58
15	257	2.69	0.45	1888	2.79	0.57	1888	2.75	0.55
16	706	2.59	0.51	118	2.78	0.52	118	2.72	0.50
17	118	2.54	0.43	706	2.63	0.54	706	2.59	0.53
18	700	2.33	0.41	1329	2.62	0.44	1329	2.54	0.41
19	1407	2.31	0.44	1757	2.50	0.40	700	2.43	0.46
20	1757	2.29	0.33	700	2.47	0.48	1757	2.42	0.37

Table B.5 – Top 20 Ranking for Minnesota Rural 3-leg and 4-leg TWSC Intersections
Screening Criterion: High Proportion Method
Accident Type: Angle Collisions
No of Sites Ranked (3-leg): 2,033
No of Sites Ranked (4-leg): 2,572

4 leg			3 leg		
Rank	Site no	Score	Rank	Site no	Score
1	1405	0.9991	1	281	1.0000
2	1750	0.9915	2	1352	0.9997
3	391	0.9900	3	657	0.9948
4	1695	0.9881	4	552	0.9932
5	1974	0.9852	5	22	0.9902
6	118	0.9846	5	97	0.9902
6	222	0.9846	7	1078	0.9900
8	1795	0.9822	8	986	0.9865
9	2393	0.9766	9	2017	0.9830
10	229	0.9762	10	341	0.9816
10	489	0.9762	10	1002	0.9816
12	226	0.9723	12	202	0.9763
13	1329	0.9714	12	2018	0.9763
14	1407	0.9685	14	548	0.9729
15	2440	0.9681	15	84	0.9698
16	1756	0.9650	16	1302	0.9666
17	146	0.9627	17	497	0.9617
18	95	0.9625	17	1015	0.9617
18	1370	0.9625	19	348	0.9581
20	1757	0.9558	19	909	0.9581

REFERENCES

1. Hadayeghi, A., Shalaby, A.S., and Persaud, B. Investigation of the Temporal Transferability of Zonal Level Accident Prediction Models. Presentation at TRB Annual Conference, Washington D.C., 2004.

2. Hauer, E., Harwood, D.W., Council, F.M., and Griffith, M.S. Estimating Safety by the Empirical Bayes Method: A Tutorial. Transportation Research Record 1784, pp. 126-131. National Academies Press, Washington, D.C., 2002.

3. Hauer, E. Re-estimation of Models for Two-lane Rural Road Segments. http://www.roadsafetyresearch.com, 2001.

4. Hauer, E., F.M. Council and Y. Mohammedshah. Safety Models for Urban Four-lane Undivided Road Segments. Presentation and publication at the 84th Annual Meeting of the TRB 2004.

5. Hauer, E. and Bamfo, J. Two tools for finding what function links the dependent variable to the explanatory variables. ICTCT 97, Lund, Sweden, pp. 1-19. 1997.

6. Hauer, E. Statistical Road Safety Modeling. Submitted for presentation at TRB Annual Conference, Washington D.C., 2004.

7. Hauer, E. Overdispersion in Modeling Accidents on Road Sections and in Empirical Bayes Estimation. Accident Analysis and Prevention, 33, 799-808, 2001

8. Hauer, E. Observational Before-After Studies in Road Safety. Pergamon, Oxford, 1997.

9. Heydecker, B., and Wu, J. Using the Information in Road Accident Records. Proceedings, 19th PTRC Summer Annual Meeting, London, September 1991.

10. Lord, D. and Bonneson, J. A. Calibration of Predictive Models for Estimating the Safety of Ramp Design Configurations. Submitted for presentation and publication at the 84th Annual Meeting of the TRB 2005.

11. Miaou, S.P., and Lum, H. Modeling Vehicle, Accidents and Highway Geometric Design Relationships. Accident Analysis and Prevention, 25(6), 689-709, 1993.

12. Miaou, S. P. Measuring the Goodness-of-Fit of Accident Prediction Models, Publication No. FHWA-96-040, Federal Highway Administration, Washington D.C., 1996.

13. Miaou, S. P. The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson and Negative Binomial Regressions, Accident Analysis and Prevention, 1994.

14. Mollett, C.J. Developing a Traffic Safety Improvement Program: A review and comparison of different screening network approaches. Canadian Multidisciplinary Road Safety Conference XIV, June 2004.

15. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.tl.nist.gov/div898/hankbook/, August 2005.

16. Oh, J., Lyon, C., Washington, S., Persaud, B., and Bared, J. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. Presented at TRB Annual Conference, Washington D.C., 2003.

17. Ontario Road Safety Annual Report 2002. Ministry of Transportation, pp.21, April 2004.

18. Persaud, B., Lyon, C., Haq, A., and Kodama, S. Development of Safety Performance Functions for Signalized Intersections in a Large Urban Area and Application to Evaluation of Left Turn Priority Treatment. Presented at TRB Annual Conference, Washington D.C., 2005

19. Persaud, B., Lord, D., and Palmisano, J. Calibration and Transferability of Accident Prediction Models for Urban Intersections. Transportation Research Record 1784. Transportation Research Board, National Research Council, Washington, D.C., pp. 57-64, 2002.

20. Poch, M. and Mannering, F. Negative Binomial Analysis of Intersection Accident Frequencies. Journal of Transportation Engineering, Vol. 122, No. 2, ASCE., 1996

21. Sayed, T., Navin, F. and Abdelwahab, W. A countermeasure-based approach for identifying and treating accident prone locations. Canadian Journal Civil Engineering, 24, 683-691, 1997.

22. SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites, Federal Highway Administration, December 2002.

23. SAS Online Documentation, http://v8doc.sas.com/sashtml.

24. Thomas, A.R., Jr. and Brian, L.J., Statistics Department, The Pennsylvania State University, 1976.

25. Wood, G.R. Generalised Linear Accident Models and Goodness of Fit Testing. Accident Analysis and Prevention, 34, 417-427, 2002.