1-1-2008

Investigation of model calibration issues in the safety performance assessment of Ontario highways

S.M. Morjina Ara Begum
Ryerson University

Recommended Citation

Begum, S.M. Morjina Ara, "Investigation of model calibration issues in the safety performance assessment of Ontario highways" (2008). *Theses and dissertations.* Paper 168.

INVESTIGATION OF MODEL CALIBRATION ISSUES IN THE SAFETY PERFORMANCE ASSESSMENT OF ONTARIO HIGHWAYS

by

S.M.Morjina Ara Begum, B.Sc.Engg.
Bangladesh University of Engineering and Technology (BUET)
Dhaka, Bangladesh

A thesis
presented to Ryerson University

in partial fulfilment of the
requirements for the degree of
Master of Applied Science
in the Program of
Civil Engineering

Toronto, Ontario, Canada, 2008

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.

Signature –

S.M.Morjina Ara Begum

Department of Civil Engineering

Ryerson University

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

Signature —

S.M.Morjina Ara Begum

Department of Civil Engineering

Ryerson University

BORROWER'S PAGE

Ryerson University requires the signature of all persons using or photocopying this thesis. Please sign below and give your address and date.

Name	Address	Signature	Date

INVESTIGATION OF MODEL CALIBRATION ISSUES IN THE SAFETY PERFORMANCE ASSESSMENT OF ONTARIO HIGHWAYS

Master of Applied Science, 2008

By

S.M.Morjina Ara Begum

Department of Civil Engineering

Ryerson University

Abstract

A set of Safety Performance Function (SPFs) commonly known as accident prediction models, were developed for evaluating the safety of Highway segments under the jurisdiction of Ministry of Transportation, Ontario (MTO). A generalized linear modeling approach was used in which negative binomial regression models were developed separately for total accidents and for three severity types (Property Damage Only accidents, Fatal and Injury accidents) as a function of traffic volume AADT. The SPFs were calibrated from 100m homogeneous segments as well as for variable length contiguous segments that are homogeneous with respect to measured traffic and geometric characteristics. For the models calibrated for Rural 2-Lane Kings Highways, the variables that had significant effects on accident occurrence were the terrain, shoulder width and segment length. It was observed that the dispersion parameter of the negative binomial distribution is large for 100m segments and smaller for longer segments. Further investigation of the dispersion parameter for Rural 2-Lane Kings Highways showed that the models calibrated with a separate dispersion parameter for each site depending on the segment length performed better than the models calibrated considering fixed dispersion parameter for all sites. For Rural 2-Lane Kings Highways, a model was calibrated with trend considering each year as a separate observation. The GEE (Generalized Estimating Equation) procedure was used to develop these models since it incorporates the temporal correlation that exists in repeated measurements. Results showed that integration of time trend and temporal correlation in the model improves the model fit.

ACKNOWLEDGMENTS

I would like to thank and express my deep gratitude to my thesis supervisor professor Dr.Bhagwant Persaud for giving me this opportunity and introducing me to the domain of road safety. His outstanding supervision made it possible for me to complete this thesis. I will always be grateful to him for arranging the financial assistance he gave me throughout the study and to the Ministry of Transportation Ontario's HIIFP programme and the Natural Sciences and Engineering Research Council of Canada for providing the fund for that assistance. I am also grateful to Mr. Craig Lyon, a Research Associate in the Department of Civil Engineering, for his help throughout the study. I would like to express my gratitude to Professor Dr Said M. Easa who made my admission possible to this program.

Finally, I wish to also thank Ms Susan Nichole of the Ministry of Transportation, Ontario for her assistance at the data preparation stage. Her quick turnaround in providing the data is greatly appreciated. I would also like to acknowledge the contributions of Dr. Mike Chapman and Dr. Jinyuan Liu for reviewing my thesis and giving some valuable suggestions. I am also grateful to my husband Zayedi and my dear sons Zahid and Farhan, who sacrificed so much for me to be engaged in and complete this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES PAGE

LIST OF ACRONYMS

AADT	Annual Average Daily Traffic
AIS	Accident information system
CURE	Cumulative Residuals
EB	Empirical Bayes
FI	Fatal and Injury (Collisions)
GEE	Generalized Estimating Equation
GLM	Generalized Linear Modeling
IHIS	Integrated Highway Information System
k	Dispersion Parameter, a parameter describing the relationship between mean and variance.
LHRS	Linear Highway Referencing System
MTO	Ministry of Transportation, Ontario
MPB	Mean Prediction Bias
MAD	Mean Absolute Deviation
MSE	Mean squared error
MSPE	Mean squared Prediction error
n	Number of years of collision used for the study.
PDO	Property Damage Only (collisions)
PSI	Potential for Safety Improvement
SAS	Statistical Analysis Software
SPF	Safety Performance Function

1 INTRODUCTION

Road safety problems have been a serious concern since the start of the automobile age. An enormous economic and human toll has been exacted as a result of the public's ongoing reliance on motor vehicles. As estimated by Transport Canada (2005 annual report TP 13347 E (10-2006)), each of Canada's 31.6 million people travelled an average of 16,000 kilometres on Canada's roadways during the year 2004. Unfortunately, this level of mobility comes with a price, since about 2,725 road users were killed and over 212,000 were injured in that year. More than 18,000 of these injury victims suffered serious injuries that kept them in hospital for at least 24 hours. The annual economic cost to Canadian society of injury-producing and property damage traffic collisions is estimated at between $11 and $27 billion, depending on the calculation method used. Within the framework of all transportation casualties, road transportation fatalities and injuries are unacceptably high. Collectively, motor vehicle crash victims accounted for more than 94% of those killed and 99% of those seriously injured in transportation-related occurrences. (Transport Canada, 2005 annual report TP 13347 E (10-2006)).

Deaths and injuries resulting from traffic collisions continue to be the biggest transportation safety problem in Canada. Traffic collisions remain one of the leading contributors to years of lost life among Canadians, due in large part to deaths among young people.

1.1 Background

Hauer (1997) has proposed that the objective measure of safety is the prevalence of accidents and their harms and that the subjective perception of safety is the feeling of security when one is on the road. In measuring safety, it is impractical to equate it with the count of accidents or with the derived measures of accident frequency and accident rate since the count of accidents changes from one period to another even when there is no change in any observable causal factor. The essential nature of accident count is that they are subject to randomness. So the only useful definition of safety is the "mean" or "average in the long run" that is behind the randomly fluctuating counts. The safety property of an entity is the number of accidents or the number of accident consequences, by kind and severity, expected to occur on the entity during a specified period (Hauer, 1997).

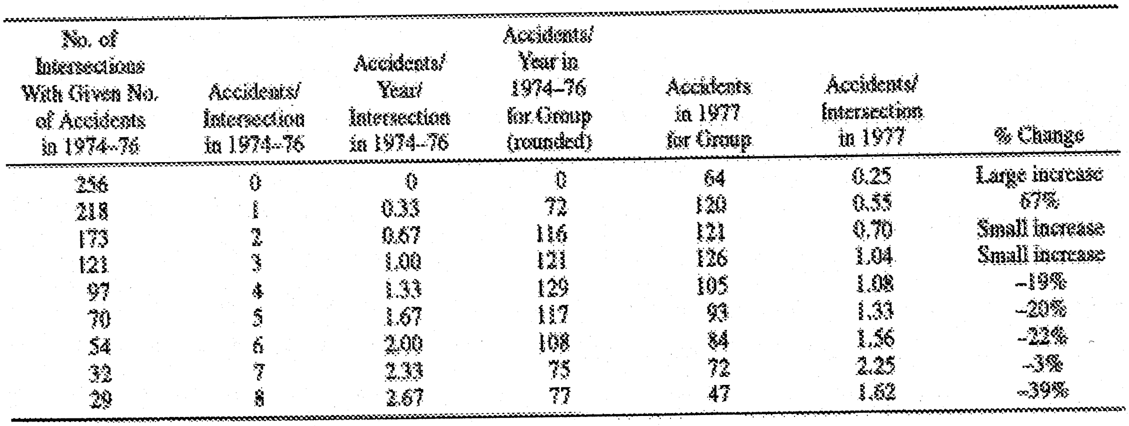
Efficient improvement of safety performance of highway sections requires an efficient Network Evaluation Procedure. The first step of this Network Evaluation Procedure is the Network screening process. Screening based on the collision counts and/or collision rates suffers from the regression-to-mean effect which could result in incorrect identification of elements as unsafe and vice-versa. Regression-to-mean phenomenon and the problems of using collision count and collision rates in safety analysis are discussed in the subsections to follow, which are excerpted from Persaud (2001).

1.1.1 Regression-to-Mean phenomenon (RTM)

An unusually high count at a roadway site is likely to decrease subsequently even if no improvement were implemented. This phenomenon is known as regression-to-mean. Following is an illustrative example of the problem of RTM as presented in Persaud (2001).

Table 1.1 shows 1,072 San Francisco intersections grouped according to the specific numbers of accidents occurring in 1974–1976. For the same intersections in each row, the average number of accidents per intersection for 1977 is also shown. Thus, for example, those 218 intersections that had exactly 1 accident in 1974–1976 recorded, in total, 120 accidents in 1977, for an average of 0.55 accidents per intersection (as shown in Table 1.1, column 6). There was no real change in safety at these intersections between 1974–1976 and 1977 in that accidents averaged over all intersections remained essentially constant over the years at approximately 1.1 accidents per intersection per year. Yet, as the table shows, intersections that had exactly 4 accidents in 1974–1976 (1.33/year) recorded, on average, 1.08 accidents in 1977, a decrease of 19 percent. Each group of intersections with 4 or more accidents in 1974–1976 (more than the average of 1.1 per year) recorded substantial reductions in accidents the following year; conversely, each group with 3 or fewer accidents (i.e., less than the average of 1.1 per year) experienced an increase. These changes have nothing to do with safety and are artefacts of the RTM phenomenon.

Table 1.1 Illustration of Regression-to-Mean phenomenon



No. of Intersections With Given No. of Accidents in 1974–76	Accidents/ Intersection in 1974–76	Accidents/ Year/ Intersection in 1974–76	Accidents/ Year in 1974–76 for Group (rounded)	Accidents in 1977 for Group	Accidents/ Intersection in 1977	% Change
256	0	0	0	64	0.25	Large increase
218	1	0.33	72	120	0.55	67%
173	2	0.67	116	121	0.70	Small increase
121	3	1.00	121	126	1.04	Small increase
97	4	1.33	129	105	1.08	–19%
70	5	1.67	117	93	1.33	–20%
54	6	2.00	108	84	1.56	–22%
32	7	2.33	75	72	2.25	–3%
29	8	2.67	77	47	1.62	–39%

To appreciate the magnitude of the problem, imagine that the 54 intersections with 6 accidents in 3 years (a total of 324 accidents in 3 years, or 108 per year) were treated at the end of 1974–1976 and recorded, for example, a total of 72 accidents in 1977. A conventional before and after comparison would estimate the treatment effect as a reduction of 108 – 72 = 36 accidents per year or 33.3 percent [= 100(36)/108]. Yet, as the last column in Table 1.1 shows, this would be a gross overestimate since the reduction due to RTM alone (and not to safety) would have been 24 accidents per year or 22 percent.

1.1.2 Problem with using the Collision Count

A common measure used in the network screening to rank sites for safety improvement is the use of collision count. In this method observed collision counts are used to do the ranking. It is the simplest of techniques and requires much less data than the more sophisticated techniques but suffers from the regression-to-mean bias in which an unusually high count is likely to decrease subsequently even if no improvement were implemented. Therefore, a site with such counts may not be in need of improvement. Conversely, a truly hazardous site may have a randomly low observed count and incorrectly escape detection as a result.

1.1.3 Problem with using the Collision Rate

In some jurisdictions, collision rate is used directly or indirectly as a hazard measure to identify locations for safety investigation. AADTs are used directly in the computation of this measure, i.e., collision rate = collision frequency/AADT (or some scalar multiple of this). If collision rates are based on the observed counts, then the regression-to-mean difficulty will still apply. In addition, there is an additional problem that renders this method of screening dubious. The problem, as the extensive literature on SPFs shows, is that the relationship between collision frequency and AADT is not linear. Figure 1.1, which depicts the SPF for injury collisions for two-lane rural roads in Ontario, illustrates the inherent nonlinearity and the difficulties with the linearity assumption. The relationship depicted is of the form:

$$Collisions/km/unit\ of\ time = a\,(AADT)^{b}$$

Where: *a and b* are regression coefficients calibrated from data.

A value of *b* = 1 would have indicated a linear relationship. The nonlinearity depicted in Figure 1.1 points to an inherent flaw in the use of collision rate as a measure of safety. Specifically, comparing collision rates of two entities at different traffic levels to judge relative safety may lead to erroneous conclusions. According to Figure 1.1, the collision rate (the slope of a line from the origin to a point on the curve) is expected to be lower at higher traffic volumes. Thus, saying that when two rates are equal they indicate equivalent levels of hazard may be completely false if different AADT levels are involved. The upshot of all this is that the use of collision rates to compare sites in regard to their safety levels is potentially problematic. When the slope of the collisions/AADT relationship is decreasing with increasing traffic volume levels, as is often the case, screening by collision rates will tend to identify low AADT sites for further investigation. The most valid basis of comparison using collision rates is for the relatively rare cases when the traffic volume levels are the same or when the relationship between collisions and AADT is linear.

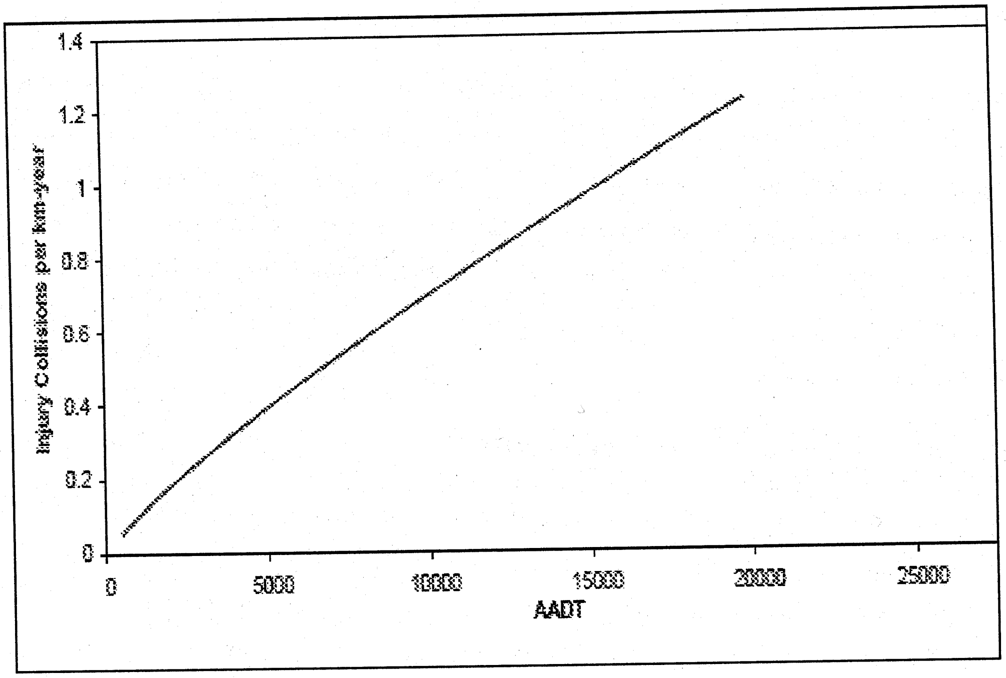


Figure 1.1 Safety Performance Function for Two-Lane Rural Arterial Highways in Ontario

1.1.4 Safety Performance Function Approach (Empirical Bayes Method)

To overcome the difficulties with the above conventional methods, a safety performance function approach has been developed using an empirical Bayes method to compensate for the random fluctuations in collision occurrence by combining the collision count of an element and its expected safety performance. It makes joint use of two clues to the safety of an entity: the collision record of that entity and the collision frequency expected at similar entities. The joint use of the two clues is implemented by a weighted average (Hauer, Harwood and Council, 2002). That is,

EB estimate of the expected accident frequency for an entity =

weight × accidents expected on similar entities + (1- weight) × count of collisions on this entity

where $0 \leq weight \leq 1$

This can be written as:

$$m = w\,(p) + (1-w)\,x \tag{1.1}$$

where *p* is *the* collision frequency expected on similar sites and is estimated from the safety performance function, a regression model with traffic and geometric factors as independent variables

x is the count of collision on an entity and

w is the weight estimated from the mean and the variance of regression estimate.

The EB estimator pulls the collision count towards the mean and thereby, accounts for the regression-to-mean bias as described in the preceding sections. To estimate the safety of a specific segment of, say, a rural two-lane road, one should use not only the collision counts for this segment, but also the knowledge of the typical collision frequency of such roads in the same jurisdiction. Hauer (2002) has suggested that "The time has come for the EB method to be the standard of professional practice; it should be used whenever the need to estimate road safety arises, whether in the search for sites with promise, the evaluation of the safety effects of interventions, or the assessment of potential safety savings due to site improvements".

The safety performance function (SPF) also called "accident prediction model" (APM) is a calibrated relationship between collision frequency and traffic volume and other characteristics of the element (lane width, shoulder width, shoulder type etc).

SPFs are calibrated from happenstance data by statistical procedures. In the past, collision counts were assumed to come from a Poisson distribution. But researchers have found that the collision counts used to calibrate SPFs are usually more widely dispersed than what would be consistent with the Poisson assumption. The Poisson distribution has been shown to be reasonable to model crash data at a given single site, but in reality, crash data over a series of sites often exhibit a large variance and a small mean, and display overdispersion with a variance-to-mean value greater than 1. For this reason, the negative binomial distribution, also known as the Poisson-Gamma distribution, has become the most commonly used probabilistic distribution for modeling accidents at a series of sites. The negative binomial distribution is considered to be able to handle overdisperson better than other distributions (Zhang et al., 2006)

Most accident prediction models are currently developed by negative binomial regression. The mathematical form to be used for the accident prediction model should in general satisfy the following two conditions:

- It must yield logical results i.e. it should not lead to the prediction of a negative number of accidents; it should ensure the prediction of zero accident frequency for zero values of the exposure variables.
- There must exist a known link function that can linearize this form for the purpose of coefficient estimation.

1.1.5 Dispersion Parameter

The 'dispersion parameter' is one of the parameters of the negative binomial distribution that is estimated in calibrating the Safety Performance Function (SPF). It is commonly denoted as 'k'. For road segments, the dispersion parameter is sometimes estimated per-unit-length i.e. the dimension of k is [1/km] or [1/mile]. The dimensions of k and length must be complementary. That is, if in the course of model calibration k is estimated per km, then length must be measured in kilometers thus k estimated per km = 0.622× k estimated per mile (Hauer, 2002).

The dispersion parameter is used in the calculation of the smoothing weight *w* of EB (Empirical Bayes) estimate and is given by:

$$w = \frac{1}{1 + k\mu} \tag{1.2}$$

Where

μ is the number of accidents/ (km-year) expected on similar sites

k is dispersion parameter

The variance of negative binomial distribution is given by the following equation:

$$Var(y) = \mu + k\mu^2 \tag{1.3}$$

where y is the random variable that represents the accident frequency at a given location at a specific period of time and

k is the dispersion parameter.

Equation1.1, 1.2 and 1.3 indicate that:

- The variance will increase as *k* increases and a large *k* makes the EB weight small resulting in the EB estimated accident frequency close to the accident count.

- The larger the value of *w*, the greater is the influence of the model prediction, *μ*, and hence there is less influence on the observed accident counts. This in turn suggests that the smaller the value of dispersion parameter the better a model is for a given set of data.

1.2 Purpose and Scope of the research

The research used data from Ontario provincial highways to explore a number of issues related to safety performance functions. Specifically the research investigated:

- The difference between models calibrated for fixed 100m segments and models based on aggregated segments of variable lengths
- The effect of AADT on accident frequency for various road classes in Ontario
- The difference in predictions between models with AADT as the only variable and models with additional explanatory variables
- The best form for the dispersion parameter of accident prediction models in empirical Bayes estimation
- The importance of estimating models to reflect time trends in accident occurrence and accounting for temporal correlation in data in estimating those models

A variety of accident prediction models were calibrated in the process. To calibrate the models the Generalized Linear Regression procedure was used.The assembly and statistical analysis of the data sets and the modeling was performed with the SAS software. SAS includes a variety of procedures for summarizing univariate and multivariate statistics and for modeling the relationship between a dependent variable such as number of accidents and covariates such as traffic volumes and highway geometric variables. The SAS GENMOD procedure was used to develop the models (SAS Institute Inc., 1999).

1.3 Brief description of thesis

This thesis is structured into ten chapters as follows:

- **Chapter 1: Introduction:** This chapter describes some backgrounds on safety measures of an entity and the purpose and scope of this study.
- **Chapter 2: Literature review:** This chapter reviews related materials in the scope of developing accident prediction models. Emphasis has been placed on road segment modeling.
- **Chapter 3: Data collection and preparation:** This chapter describes the features of the basic data provided by MTO and the step by step preparation of database to calibrate the SPFs (Safety performance functions)
- **Chapter 4: Data characteristics:** In this chapter univariate and bivariate statistics of the variables used in the study are provided.
- **Chapter 5: Research Methodology:** Details of the methodology used in this research to calibrate models and examine their goodness-of-fit are described in this chapter.
- **Chapter 6: Modeling:** This chapter describes calibration and comparison of "AADT-only" models from 100m segment and longer segments and the estimation of models for Rural 2-Lane Kings Highways with other available explanatory variables.
- **Chapter 7: Investigation of the negative binomial dispersion parameter:** In this chapter, the performance of models calibrated considering a fixed dispersion parameter for all sites and a varying dispersion parameter for each site that depends on the segment length are compared.
- **Chapter 8: Application of generalized estimating equations (GEE) procedure to calibrate models with trend:** This chapter describes the application of the GEE procedure to calibrate model with trend and the comparison of this model with the regular GLM models
- **Chapter 9: Conclusions and Recommendations.**

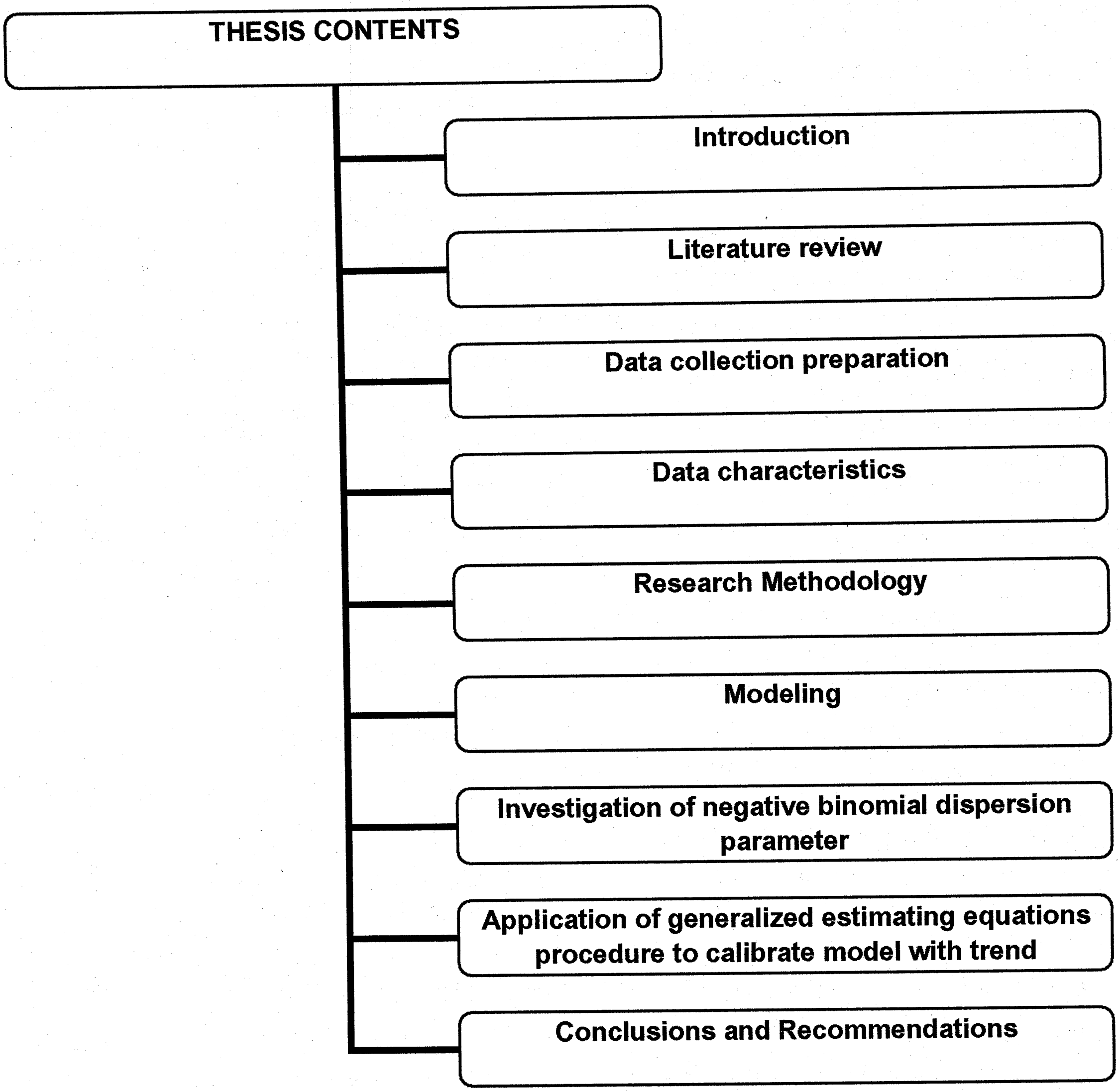

THESIS CONTENTS
Introduction
Literature review
Data collection preparation
Data characteristics
Research Methodology
Modeling
Investigation of negative binomial dispersion parameter
Application of generalized estimating equations procedure to calibrate model with trend
Conclusions and Recommendations

Figure 1.2 Thesis Contents

1.4 Final outcome

Safety performance functions (AADT-only) have been developed for freeway and highway segments. A full model was calibrated for Rural 2-Lane Kings Highways that also contains shoulder width, terrain and segment length as explanatory variables. An investigation showing the effect of the varying dispersion parameter on model prediction and EB estimates also forms a part of the research study. A time trend effect model for Rural 2-lane Kings Highway was calibrated using the Generalized Estimating Equation (GEE) procedure.

2 LITERATURE REVIEW

Researchers have developed SPFs and dispersion parameters for different types of entities – road segments, intersections, interchanges, roundabouts etc. and the results can be found in the literature. Some of that literature is reviewed below:

Persaud (1991) has developed multivariate statistical models using negative binomial regression to estimate the accident potential of Ontario road sections on the basis of its traffic and geometric variables. Readily available data at MTO was used for the analysis. Models were developed for three road classes, class1 (freeway), class 2(other, primary) and class 3(secondary, tertiary) of the Ministry's entire road network. The generalized linear modeling approach was used to estimate the accident potential. The model form used was:

$$E(m|T) = SCL \times a_1 \times AADT^{b_1}$$

Where

$E(m|T)$ = The underlying accident potential of a section.

T = set of traffic and geometric characteristics,

SCL = section length (km)

a_1 and b_1 = model parameters estimated by GLIM (Generalized linear interactive modeling)

For the freeways, traffic volume was found as the only explanatory variable to be included in the model since all attempts to incorporate geometric variables failed.

Class 2 road sections were categorized on the basis of road environment rural/urban, number of lane two-lane/multilane and divided/undivided. Six possible categories were obtained. Exploratory analysis revealed that the variation in speed and geometric factors was small within the individual category but significant from category to category. The final models for this class reflected the fact that use of road section category as a variable in the accident prediction model is sufficient to account for variation attributable to geometric and speed factors as all attempts to incorporate other variables in the model failed. The final model showed the AADT coefficient a_1 to vary with two categorical variables lane and environment whereas coefficient b_1 to vary also with the categorical variable divided/undivided. It also turned out that the estimated coefficients obtained in this way are the same as those that would have obtained by separately estimating models for each category separately.

For Class 3 road sections, the best explanatory variables were found to be surface width and surface type each with two levels giving four category combinations. Additional explanatory variables had insignificant effects. The final model term showed the coefficients a_1 and b_1 to vary with the two categorical variables narrow/wide pavement and low/high class surface.

Finally, an empirical Bayesian estimate was obtained by combining the predicted accident from the hypothesized regression model and observed accident count to estimate the accident potential of Ontario road sections.

The insight gained from this research was:

The empirical Bayesian procedure was observed to be superior to using the accident count or the regression predictions from variables alone but the benefit of empirical Bayesian estimation would be marginal in case of road sections which are long and have a relatively high traffic volume. This is because for these sections the prediction has a large variance which makes the weight *w* of empirical the Bayesian estimate small resulting in the accident potential estimate close to the accident count. So accident count can be considered as the reasonable estimator for accident potential in such cases.

Persaud and Dzbik (1993) used a generalized linear modelling approach to develop regression model estimates of Ontario freeway section's accident potential and these estimates were refined using the empirical Bayesian procedure. The approach was applied to both microscopic data (hourly accidents and hourly traffic) and macroscopic data (yearly accident data and average daily traffic).

For microscopic modeling each day was disaggregated by hour to derive data for each hour and for day and night. Accident and traffic data were obtained separately for collector and express lanes. For traffic data, hourly and seasonal variation factors and collector/express lane distribution factors were applied to the average daily traffic. The GLIM (Generalized Linear Interactive Modeling) computer package was used to obtain the regression models. Models were calibrated for uncongested hours (off-peak) and for total accidents and severe accidents.

The model form used was:

$$E\ (P) = a\ T^b$$

where *E (P)* is the accident potential per km-hour, *T* is the volume per unit of time and *a* and *b* are model parameters estimated by GLIM.

The prediction from macroscopic models indicated that for the same total traffic volume, four lane freeways have a lower accident risk than those with more lanes. Also the value of *b* suggests that the accident-volume regression prediction line has an increasing slope. The researchers suggest that this is an indication of increasing possibility of risky manoeuvres such as passing and lane changing, with higher ADT levels.

The prediction from microscopic models indicated that for a given traffic volume level, collector roadways have a higher accident potential than the express roadways. For these models, the slope of the regression prediction line was found to decrease with increasing traffic. The relationship between the quality of traffic operation and the accident risk was examined in the study. The results of the analysis indicated that:

- congestion is associated with higher risk of accidents than high-volume uncongested operation;
- the afternoon congested period has a higher accident risk than the morning rush period, but the difference is only significant for the express system;
- collector system congestion is associated with a higher accident risk than express system congestion.

Both the macroscopic and microscopic models were validated. The mean squared difference between the estimated and the observed accidents was used for the validation. It was assumed that the better estimate is the one with the smallest mean squared difference. The results show that the empirical Bayesian method appears to be best followed by the regression model prediction method.

Lord and Persaud (2000) have developed time effect accident prediction models to capture changes over time in traffic flow, weather, economic condition, accident reporting etc. They have used a Generalized Estimating Equation (GEE) procedure to obtain the coefficients of time effect models. The GEE procedure was applied to 4-leg signalized intersections in City of Toronto. Three different models were calibrated using simple model form that includes only the major and minor road flow:

$$E\{k\} = \alpha F_1^{\beta_1} F_2^{\beta_2} e^{(\beta_3 F_1)}$$

Model 1was calibrated from aggregated data i.e. using average AADT over the study period and total accidents with the regular GLM approach. Model 2 was calibrated from the data disaggregated by year and with the regular GLM. Model 3 was calibrated from disaggregated data but with the GEE procedure so this model incorporated both trend and temporal correlation. Results showed temporal correlation contributed to the standard error but still the coefficients were significant. The authors have mentioned the possibility of some coefficients to become insignificant due to their inflated variance. The performances of calibrated models were investigated by cumulative residual plots and the GEE model with trend was found to perform better than GLM models. The authors have suggested that model with trend is more suitable for before-and-after studies. These models will provide a better estimate for the after period to evaluate more efficiently the effect of treated site.

Sawalha and Sayed (2003) have discussed two important statistical issues related to modeling accidents using Poisson and negative binomial regression. They have introduced a procedure to develop parsimonious model (models that are not over fitted) and best fit model and also the procedure for outlier analysis in case of negative binomial regression. For the parsimonious model two criteria were examined to retain a variable in the model; the first was whether or not the Wald statistics is significant at 95% confidence level and the second one was whether or not the addition of the variable causes a significant drop in the scaled deviance. For best-fit-model only the first criterion was examined.

To develop the parsimonious model a reference model was created to compare the drop in deviance with the addition of a new variable in the model. The deviance drop of the two models was compared keeping same dispersion parameter for both. Variables were added one by one in the model and the drop of deviance was compared to the reference model. The variable which resulted in the highest drop of deviance was retained in the model. A new model was developed with this variable and the addition of rest of the variables was then examined in the same way considering this model as the new reference model.

Parsimonious and Best-fit-models were developed for urban arterials of British Columbia and it was observed that the best-fit-model contained more variables than the parsimonious model since it has used only one criterion, significant Wald statistics, for retaining variable in the model.

Hauer, Council, and Mohammedshah (2004) have developed multivariate statistical models to fit to data from the state of Washington to predict the frequency of nonintersection on-the-road and off-the-road accidents on urban four-lane undivided roads. A number of geometric variables were used in the modeling, including Lane width, Shoulder width, Shoulder type, Roadside Hazard Rating, Estimated Clear Zone Width; Vertical alignment; Horizontal alignment; Access; Two-way left-turn lanes (TWLTL);Parking control; Speed limit and Segment Length.

The researchers developed six separate multivariate statistical models for off-the-road and on-the-road property-damage-only (PDO), injury, and total accidents. Parameter estimation was done by maximum likelihood method. They have suggested modelling each side of the road separately because the two directions differ in many important geometric features like grade, number of access points, left-or right turning curves, etc. So to predict the number of accidents on a road segment it is required to add two model equations one for each side of the road. They have proposed a model structure as follows:

y = (scale parameter) × [(segment length for prediction) × (multiplicative portion) + (additive portion)

Where y is the annual number of accidents expected to occur on one side of road.

The multiplicative portion of the model was added to represent the influence of variables that are continuous along road segments or continuous on some part of segment for example, shoulder width and grade respectively. The additive portion of the model was added to represent the effect of point hazards for example driveways and short narrow bridges.

The Influence of variables in the model was examined by a ratio given as:

$$R(Variable\ value) = \frac{Recorded\ accidents\ on\ all\ segments\ with\ variable\ value}{Predicted\ accidents\ on\ all\ segments\ with\ variable\ value}$$

When the data showed an orderly relationship between the variable value and the ratio R (variable value), the variable was introduced and an appropriate functional form to represent it was chosen. Otherwise the variable was excluded from the model. Variables were introduced in the model in the order in which they increase a log-likelihood parameter.

Access points, roadside hazard rating and clear zone width, percent trucks, segment length (in model data) were found to have no effect on off-the-road accidents. For on-the-road accidents lane width was found to be associated with PDO accidents but not with injury accidents. Neither the number of intersection nor the number of commercial driveways was found to affect on-the-road accidents. But the number of residential driveways was found to affect these accidents and was introduced in the model as an additive component. In case of on-the-road accidents on examination of the R value, segment length variable was introduced in the model in the form $e^{\beta \times \text{(segment length in model data)}}$. The segment length variable was applied only to segments that are homogeneous in all traits i.e. each tangent, horizontal curve, or vertical curve was a separate segment.

The insights gained out of this research were:

- The contribution of a variable to the model can be established on the basis of the observed increase in log likelihood per parameter.
- There were several unexpected results like , accident frequency seems to diminish as speed limit increases or as the proportion of trucks increases;
- There were more on-the road accidents on tangent sections than on most horizontal curves;
- Roads with wider shoulders have more accidents.

The researchers raised the important question of whether or when any of the relationships in the model can be used as indicative of cause and effect, noting that the practice of obtaining accident modification factor from these multivariate models is questionable. Result also showed that, the relationship that holds for two-lane rural roads may not hold for urban four-lane roads. For example on four lane roads horizontal curves of moderate degree may be safer than tangent road sections.

3 RESEARCH METHODOLOGY

3.1 Model calibration

As identified by researchers the following elements are necessary for model development:

- An appropriate model form,
- An appropriate error structure,
- A procedure for selecting the model explanatory variables,
- Methods for assessing model goodness of fit.

These four items are discussed in the subsections to follow.

3.1.1 Model form

The mathematical form to be used for developing crash prediction models should satisfy the following two conditions to yield logical results (Sawalha and Sayed, 2003):

- it must not lead to the prediction of a negative number of accidents and
- it must ensure a prediction of zero accident frequency for zero values of the exposure variables; for road segments these exposure variables are segment length and annual average daily traffic (AADT).

Traditional linear regression such as the Least Square (LS) and Weighted least square (WLS) methods cannot be used to estimate the coefficients of crash prediction models since the count of accidents is discrete and non-negative and the variance of this count increases as flow increases.

To overcome the limitations of LS and WLS method, a generalized least square method is used to estimate the coefficients of crash prediction models. In this study Safety Performance Functions (SPFs), commonly known as crash prediction models, were developed using Generalized Linear Modeling. This is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions.

A generalized linear model (GLM) consists of the following three components (SAS Institute Inc., 2002):

- The linear component is defined as:
  $\eta_i = x_i' \beta$ (3.1)
- A monotonic differentiable link function *g*, which describes how the expected values of y_i (the response variable for the i^{th} observation) is related to the linear predictor η_i:
  $g(y_i) = x_i' \beta$ (3.2)
- The response variables y_i are independent for i=1, 2... and have a probability distribution from an exponential family. This implies that the variance of the response variable depends on the mean through a variance function *V*:

$$var(y_i) = \emptyset\, V(\mu_i)/\omega_i \qquad (3.3)$$

where φ is the dispersion parameter and ω_i is a known weight for each observation.

In order to use generalized linear regression in the modeling procedure, there must exist a known link function that can linearize this form to estimate the model coefficients. This condition is satisfied by a model form that consists of the product of powers of the exposure variables raised to some power multiplied by an exponential that incorporates the remaining explanatory variables. Such a model form can be linearized by the logarithmic link function. There exists many model forms but the most common one which has been employed for road sections in past research was used for this study:

$$E(Y) = L \times a_1 \times V^b \times exp \sum b_i\, x_i \qquad (3.4)$$

where *E(Y)* is the predicted accident frequency, *L* is the section length, *V* is the section AADT, x_i are the explanatory variables , and a_1, *b* & b_i are the model parameters estimated from the data in the generalized linear modeling procedure.

The GLM linear version of Equation 3.4 can be written as follows:

$$\ln [E(Y)] = \ln(L) + \ln(a_1) + b(logV) + \sum b_i\, x_i \qquad (3.5)$$

The Generalized linear model can be constructed by deciding on the response and explanatory variable for the data and choosing the appropriate response probability distribution and link function. In this project the response variable is a count and the distribution of response variable is considered to be negative binomial. So the appropriate link function was taken as the log link i.e. the linear predictor η = log (μ).

3.1.2 Error structure

The Generalized Linear Regression approach assumes that the error structure is either Poisson or negative binomial.

The advantage of using the negative binomial model is that Poisson distribution requires that the mean and variance to be equal. If this equality does not hold, then the data is said to be over or under dispersed. The Poisson distribution has been shown to be reasonable to model crash data at a given one site, but in reality, crash data over a series of sites often exhibit a large variance and a small mean, and display overdispersion with a variance-to-mean value greater than 1.

For this reason, the negative binomial distribution, also known as the Poisson-Gamma distribution, has become the most commonly used probabilistic distribution for modeling accidents at a series of sites. The negative binomial distribution is considered to be able to handle overdisperson better than other distributions.

The approach has the following theoretical basis (El-Basyouny and Sayed, 2006):

Let *Y* be the random variable that represents the accident frequency at a given location at a specific period of time with y being a certain realization of *Y*.

The mean of *Y*, denoted by Λ, is also a random variable. For Λ = λ, *Y* is Poisson distributed with mean λ:

$$p(Y = y|\Lambda = \lambda) = \frac{\lambda^{y} e^{-\lambda}}{y!} \qquad (3.6)$$

$$E\,(Y|\,\Lambda = \lambda) = \lambda \qquad (3.7)$$

$$Var\,(Y|\,\Lambda = \lambda) = \lambda \qquad (3.8)$$

If Λ is described by a gamma distribution with shape parameter k and scale parameter k/μ, its density function is

$$f_{\Lambda}(\lambda) = \frac{(\frac{k}{\mu})^{k} \lambda^{k-1} e^{-(\frac{k}{\mu})\lambda}}{\Gamma(k)} \qquad (3.9)$$

and the variance and mean are:

$$Var(\Lambda) = \frac{\mu^{2}}{k} \qquad (3.10)$$

$$E\,(\Lambda) = \mu \qquad (3.11)$$

And the distribution of Y around E (Λ) = μ is negative binomial. The probability density function of negative binomial distribution is given by:

$$P(Y = y) = \frac{\Gamma(k+y)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^{k}\left(\frac{\mu}{k+\mu}\right)^{y} \qquad (3.12)$$

and following are expected value and variance of this distribution

$$E(Y) = \mu \qquad (3.13)$$

$$\mathrm{Var}(Y) = \mu + \frac{\mu^2}{k} \qquad (3.14)$$

The term k shown in the above probability density function is usually defined as the "inverse dispersion parameter". The variance of the accident frequency is generally larger than its expected value which reflects the fact that accident data are generally over-dispersed. When k → ∞ the distribution of Λ is concentrated at a point and the negative binomial distribution becomes identical to the Poisson distribution.

3.1.3 Parameter estimation

The GENMOD procedure of the SAS software package was used to calibrate the SPF's .The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector β. The parameters are estimated numerically through an iterative fitting process. The dispersion parameter k is also estimated by a maximum likelihood method. The theoretical background is as follows:

This method provides maximum likelihood estimates of β and k. These are the values of β and k that maximize the log-likelihood function *l* (β, k.) given by:

$$l(\beta,k) = \sum_{i=1}^{n}\left[ln\Gamma(y_i + k) - ln\Gamma(k) - \ln(y_i\,!) - k ln\left(1+\frac{y_i}{k}\right) - y_i \ln\left(1+\frac{k}{y_i}\right)\right] \qquad (3.15)$$

The log-likelihood function depends on β through the terms $\hat{y}_i = \hat{E}\,(Y_i\,)$, which represent the model predictions or fitted values. The model form most commonly used in accident modeling given in equation (3.4) specifies the fitted values as:

$$\hat{y}_i = \exp\,(X_i'\,\beta) \qquad (3.16)$$

Where X_i' is the vector of explanatory variables corresponding to the ith observation. The method of maximum likelihood obtains β and k as the solution to the following two equations:

$$\frac{\partial l}{\partial \beta} = 0 \quad and \quad \frac{\partial l}{\partial k} = 0 \qquad (3.17)$$

Built-in link functions, accommodating probability distributions and associated variance functions are available in GENMOD procedure.

3.1.4 Selection of explanatory variables

Explanatory variables that have statistically significant model parameters contribute to the explanation of the variability of accident data and their inclusion in the model improves its fit to the data. Variables were added to the model equation one by one and the following two criteria are examined before making the decision on whether the variable should be retained in the model (Sawalha and Sayed, 2003):

3.1.4.1 Criterion 1

Whether or not the Wald statistics of the estimated parameters are significant at 95% confidence limit. That is one should be able to reject the null hypothesis that the parameter is zero.

3.1.4.2 Criterion 2

Whether or not the addition of a new variable to the model causes a significant drop in the scaled deviance at the 95 percent confidence level. This criterion represents an analysis of deviance procedure for comparing two nested models. This procedure is equivalent to carrying out a likelihood ratio test to determine whether the model containing the additional variable significantly increases the likelihood of the observed sample of accident data. The scaled deviance is asymptotically distributed with *n-p* (*n* is number of observations and *p* is the number model parameters) degrees of freedom, and therefore, owing to the reproductive property of the χ^2 distribution. If the addition of the variable causes a drop in scaled deviance exceeding $\chi 2_{0.05,1}$ (equal to 3.84) this criterion is met. But the analysis of deviance procedure should be conducted in different manners for Poisson and negative binomial regression models. In the case of Poisson regression models, the difference in scaled deviance between the model with *p +1* variables and the model with *p* variables is equal to the likelihood ratio test statistic, which compares the maximized likelihood functions of the sample of accident data under the two models. This statistic is defined as follows:

$$LRTS = 2ln\left(\frac{L_{p+1}}{L_p}\right) \qquad (3.18)$$

Where

LRTS= the likelihood ratio test statistics

L_{p+1} = the maximum likelihood of the accident data under the model *p+1* variable

L_p = the maximum likelihood of the accident data under the model *p* variable

The likelihood of the observing the accident data under a certain model is equal to the joint probability of each observation. It is therefore a function of the model parameters and the error structure assumed by the model. *LRTS* is a non-negative statistic and the minimum value it can have is zero in which case the additional variable contributes nothing to increase the likelihood of the sample of accident data. *LRTS* is used to test the null hypothesis $H_0 : \beta_{p+1} = 0$ where β_{p+1} is the parameter of the additional variable. A statistically significant *LRTS* leads to the rejection of the null hypothesis and the adoption of the model with *p +1* variable. A statistically insignificant *LRTS* means failing to reject the null hypothesis and concluding that the model with *p* variables is perfectly satisfactory. Statistical significance of *LRTS* can be determined from its sampling distribution. Researchers have found that if two nested models with degrees of freedom *df1* and *df2* are correct, the sampling distribution of their *LRTS* is an asymptotic χ^2 distribution with degrees of freedom equal to *(df1-df2)*. Therefore for the *LRTS* in Equation 3.18 to be statistically significant at the 95 percent confidence level it has to exceed $\chi 2_{0.05,1}$ = 3.84.In the case of Poisson regression models, the difference in scaled deviance between the model with *p +1* variables and the model with *p* variables is equal to their *LRTS,* meaning that this difference can be directly used to assess the usefulness of the additional variable. A drop in scaled deviance exceeding 3.84 can be taken as the basis for choosing the model with *p +1* variables.

In the case of negative binomial regression, models with different variables have different values of *k,* the dispersion parameter. The scaled deviances of two negative binomial regression models, with different values of *k* cannot be directly compared. The reason behind this is that the difference in scaled deviance between the two models is not equal to their likelihood ratio test statistic. In this case, conducting a meaningful analysis of deviance procedure for developing parsimonious models requires that the value of *k* for the model with *p* variables be imposed on the model with *p +1* variables. The drop in scaled deviance is then compared with $\chi 2_{0.05,1}$ (equal to 3.84) in order to assess the usefulness of the additional variable.

3.1.5 Model goodness-of-fit

The goodness-of-fit of Generalized Linear Models can be assessed by several statistical measures. Different researchers have used different criteria. Vogt and Bared (1998) suggested that the validity and practicality of estimated parameters can be first confirmed by applying engineering and intuitive judgement.

To assess model goodness of fit three measures, the ***Pearson*** $\chi 2$ statistic, the scaled deviance (Sawalha and Sayed, 2006) and the cumulative residual plots (Hauer and Bamfo, 1997) were used in this research. For a well-fitted model, both the scaled deviance and the Pearson $\chi 2$ should be significant compared with the critical value obtained from the $\chi 2$ distribution for the given degrees of freedom and level of confidence. The three measures are outlined below:

3.1.5.1 The *Pearson* $\chi 2$ statistics

The *Pearson* $\chi 2$ statistics is defined as:

$$Pearson\ \chi 2 = \sum_{i=1}^{n} \frac{[y_i - E(y_i)]^2}{Var(y_i)} \tag{3.19}$$

Where y_i is the observed number of accidents on section *i, E* (y_i) is the predicted number of accidents obtained from SPFs and *Var* (Y_i) is the variance of the accident frequency for section *i* .

3.1.5.2 The Scaled Deviance

The Scaled deviance is the likelihood ratio test statistic estimated as twice the difference between the maximized log-likelihoods of the studied models and the full or saturated model. The full model has as many parameters as there are observations, so that the model fits the data perfectly. It represents the maximum log-likelihood achievable under the given data. Therefore the full model provides a baseline for assessing the goodness of fit of an intermediate model with *p* parameters. For the negative binomial error structure the scaled deviance (*SD*) is:

$$SD = 2\sum_{i=1}^{n} \left[y_i ln\left(\frac{y_i}{E(y_i)}\right) - (y_i + k) ln\left(\frac{y_i + k}{E(y_i) + k}\right)\right] \tag{3.20}$$

Where y_i is the observed number of accidents on section *i,* and *E* (y_i) is the predicted number of accidents obtained from SPFs

3.1.5.3 CURE (Cumulative Residuals) plot

This is an examination of residuals after regression coefficients are estimated. The residual is defined as the difference between the accident count and the model prediction. The idea for using cumulative residuals is that it can provide potentially important information in the patterns when the usual plot of residuals does not show any systematic drift. It can be used for two purposes. The first to examine the whether the chosen functional form fits an explanatory variable along the entire range of its values represented in the data. The second is to ascertain whether an explanatory variable should be introduced in the model. The plot of cumulative residuals should oscillate around zero and end close to zero and not exceed the ±$2\sigma^{*}$ bounds calculated from the following equation:

$$\sigma^{*2} = \sigma^{2}(n)\left(1-\frac{\sigma^{2}(n)}{\sigma^{2}(N)}\right) \qquad (3.21)$$

where N is the total no of data points and n is an integer between 1 and N. The values of $\sigma^{2}(n)$ and $\sigma^{2}(N)$ can be estimated from cumulative squared residuals.

3.2 Network Screening (EB approach):

An empirical Bayes (EB) approach was followed in this study to perform network screening although this aspect was not an extensive part of this thesis. The objective was to compare the ranking of sites on the basis of EB estimates calculated in Chapter 7, considering a fixed dispersion parameter for all the sites and separate dispersion parameters for each site that depend on the segment length, and considering models with AADT as the only variable and models with additional variables..

As already mentioned in Chapter 1, EB estimate makes joint use of two clues to the safety of an entity: the accident record of that entity and the accident frequency expected at similar entities and the joint use of the two clues is implemented by a weighted average. That is,

EB estimate of the expected accident frequency for an entity =

weight × accidents expected on similar entities + (1- weight) ×count of accidents on this entity

where 0≤weight≤1 (3.22)

Equation (3.22) can be written as:

$m = w\,(p) + (1\text{-}w)\,x$ (3.23)

where p is estimated from SPFs (negative binomial regression model) as described in Chapter 1

x is the count of collision on an entity and

w is the weight estimated from the dispersion parameter obtained through SPF calibration and is given by:

$$w = \frac{1}{1 + k\mu} \qquad (3.24)$$

Where μ is the model prediction accidents/km-year and

k is dispersion parameter estimated per km length for segments.

3.2.1 Step by step illustration of the procedure for deriving EB estimates (excerpted from Hauer, E, "Estimating Safety by the Empirical Bayes Method: A Tutorial" TRB 2002):

Step 1: Estimate accidents expected on similar sites

To estimate the safety of a specific segment of, say, a rural two-lane road, one should use not only the collision counts for this segment, but also the knowledge of the typical collision frequency of similar sites in the same jurisdiction. The average accident frequency of 'similar sites' and the variation around this average are brought into the EB procedure by the Safety Performance Function (SPF). The SPF is an equation giving an estimate of μ the average accidents/ (km-year) for road segments or accidents/year for intersections, as a function of some trait values (e.g., AADT, lane width . . .) and of several regression parameters.

To illustrate, consider the SPF: estimate of $\mu = 0.0224 \times AADT^{0.564}$ for a certain kind of road in a given jurisdiction. Here AADT plays the role of one trait value, no additional trait values are represented in the SPF, 0.0224 and 0.564 are the estimated regression parameters. If on a road of this kind AADT=4000 vehicles per day, then one should expect $0.0224 \times 4000^{0.564} = 2.41$ accidents/ (km-year). If a road segment is 1.8 km long, has an AADT of 4000, and recorded 12 accidents in the last six years then the estimate of the safety of this road is obtained as:

The expected safety of the similar road from the model calibrated with recorded count and characteristics of the road is $0.0224 \times AADT^{0.564}$ accidents/ (km-year) with the dispersion parameter obtained from the calibration = 0.18/km.

So road segments similar to the one in this example will be expected to have $0.0224 \times AADT^{0.564} = 2.41$ accidents/ (km-year), on average. Therefore segment of similar length would be expected to have $1.8 \times 2.41 \times 6 = 26$ accidents in six years.

Step 2: Estimate the weight

Since we are using six years of accident count following Equation 3.24:
$w = 1/ [1+0.18(2.41 \times 6)] = 0.277$

Step 3: Estimate the EB expected accident frequency

Using Equation 3.23, the estimate of the expected accident frequency for the specific road segment is $0.277 \times 26 + (1-0.277) \times 12 = 15.88$accidents in six year or $15.88/ (1.8 \times 6) = 1.47$Accidents/km-year.
Noted that 15.88 is between the average for similar sites (26) and the accident count for this site (12). The EB estimator pulls the accident count towards the mean and thereby accounts for the regression-to-mean bias.

4 DATA COLLECTION AND PREPARATION

4.1 Data Collection

The Ontario Ministry of Transportation provided the accident and inventory data for five regions -- Central, Eastern, Northeast, Northwest and Southwest -- under their jurisdiction. Data were made available in Excel Files in three formats for the period 2000 to 2004 separately for each region: a master file with collision counts, traffic volume and geometric variables; a detailed collision file and a Non-intersection collision.

The master file contained the following information for each 100m interval of the Highway sections identified by the corresponding LHRS (Linear Highway Referencing System) and Offset as described in section 4.1.1(III) bellow:

4.1.1 Master file

The collision counts associated with each 100m interval of the highway were provided by their severity type, namely fatal, Injury and PDO (property damage only) separately for each year of the study period.

Traffic volume data were provided in AADT, the total of flows in both directions. Each 100m segment was provided with volume information. There were locations with no volume data which happened to be at ramps or interchanges. So those segments were removed.

Geometric data were made available for the key LHRS point, i.e., LHRS and 0.0 offset. The 100m segments within each key LHRS have the same geometric characteristics. Information was provided on number of lanes, functional class, road environment, terrain, median width, median shoulder width, lane width, shoulder width, shoulder type, surface width, surface type and other variables.

Geometric data were extracted from MTO's Integrated Highway Information System (IHIS).Collision counts and traffic volume data were extracted from MTO's Accident Information System (AIS) Database which is comprised of the following data sources:

I. The 100 Meter Summary (AR) Database:
Each AR (accident record) shows the information on collisions and traffic volumes occurring within the specified 100m interval of highway (as mapped by the Linear Highway Reference System (LHRS).

II. The Motor Vehicle Accident (MVA) Database:
The MVA database provides the Accident Information System with detailed information taken directly from the accident report filed by the Police or by an Accident Self-Reporting Facility. The MVA records are associated to the AR records by their LHRS and Offset values.

III. The Linear Highway Reference System (LHRS) Database:

The LHRS is a unique five-digit number assigned to each basic reference point along each highway. The numbering system used along each highway increases in the basic direction of that highway (called the "cardinal direction"). Any intermediate point along the highway, between the established basic reference points, can be located by its offset (i.e. distance) measured in the basic direction of the highway from the nearest basic reference point in the cardinal direction. Each basic reference point is always identified by its corresponding LHRS number and zero-offset that may be combined e.g., LHRS_Offset = 100000000 for LHRS 10000 Offset 0.0.

IV. The Traffic Volume Information System (TVIS) Database.

The TVIS database provides historical traffic volume information that relates to past years of traffic activity over various roadway segments. Information is provided concerning the specific counting stations collecting volumetric traffic data as well as statistics that relate to the various traffic patterns recorded on that segment of roadway. The Traffic Volume Information System (TVIS) summarizes traffic volume information and provides reports on a historical database of traffic volume. This data are used to support analysis of historical collision rates through comparison with provincial and/or local averages and summarized traffic volumes. Historical and statistical information is received from a variety of counting devices located throughout the province at strategic locations.

The AR and MVA databases are specific to each region's road network with an overlap of a minimum of 1km between regions. This overlapped portion had been removed while providing the data for this study.

4.1.2 Detailed collision data

The detailed collision file contained the following information for each collision:

- **LHRS and Offset:** The **LHRS** number identifies the section where the accident occurred in and offset gives the position of the accident occurrence within the section.
- **Accident Severity**: Accidents are classed as fatal, non-fatal injury, property damage only,
- **Initial impact type:** Provides the first impact type of the collision. Impacts were described as angle, rear-end, side swipe etc.
- **Number of vehicles involved:** The file provides the information on each vehicle involved in the collision. Where more than one vehicle involved in a collision only the information for vehicle 1 was used, to avoid any double-counting.
- **Accident location:** This field gives the type of the location where the accident occurred. Locations were described as, at intersection, non-intersection, intersection related, at or near private drive, other road location, at railway crossing or parking lot etc.

4.1.3 Non-intersection Collision data

These files contained accident counts by severity for each 100m segments excluding those that occurred at intersections or were related to the intersections. Data were provided for the whole study period i.e. from 2000 to 2004.

MTO has included the following accidents as non-intersection accidents:

- NonIntSec: Collisions not occurred at intersections, underpasses, overpasses, tunnels, bridges, private drives or railway crossing and also not related to activity at a nearby intersection.
- PrvDrv: Collisions that occurred at or near Private drive which is not public roadway.
- RRxing: Collisions that occurred at railway crossing.
- UndPass: Collisions that occurred in a tunnel or on a roadway underneath of a structure
- OvrPass: Collisions that occurred on a bridge or on a roadway on a structure
- Oth-RdLoc: Collision that occurred on a public roadway.

4.2 Data Preparation

An extensive data preparation process was required to prepare the database for SPF calibration. The master files described above contained geometric information only for the key LHRS numbers. So the first task was populating all the segment cells within an LHRS with the geometric information provided. It was done electronically by programming in Excel to create a new variable for each geometric field.

Before starting the main preparation process there were some data issues that needed to be solved through discussion with MTO officials. These involved getting the missing data and removing the non-assumed segments since the data contained some segments which were not assumed as a provincial highways.

It was desired that two types of segment models to be developed:

- **Type 1: Modeling with minor intersection collisions and**
- **Type 2: Modeling with non-intersection collisions**

4.2.1 Data Assembly for Type 1 modeling

The thesis is focused only on road segment modeling, but the data files provided by MTO contained intersection and interchange segments. So the first key issue was to remove all these signalized/unsignalized intersections and interchange segments from the master files including the segments within the influence area so that the accident count does not include any intersection/interchange related accidents. MTO provided a complete list of major intersections for all regions. Interchanges were identified by searching for the word "IC" in the segment description column of the master files.

4.2.1.1 Intersection segments removal

As described earlier each LHRS + offset combination represents a 100m stretch of the roadway i.e. one segment. To remove the intersection segments it was required to define the collision radius. After discussion with MTO it was decided to use 100m radius for this study. Since the break between segments generally occurs in the centre of an intersection, two segments had to be removed at these locations– the one downstream of the intersection (intersection segment itself) and the one upstream of the intersection.

The major intersection list provided by MTO was merged to the master files in SAS and then the intersection segment and the upstream segments were identified creating two variables that allowed for the intersection segments to be removed electronically through a code developed in SAS.

4.2.1.2 Interchange segments removal

For the interchanges, influence length was considered as 1km on either side of the interchange segment. So it was required to remove 10 segments upstream and 9 segments downstream of the interchanges. No automatic process could be developed to remove these segments electronically. So removing the interchanges was tedious.

After removal of major intersection and interchange segments, some minor intersections were still remained in the data set including the accident counts on those segments. So models were calibrated from this data set for the Type 1 models.

4.2.1.3 Road Classification

For effective planning and management of provincial roads the Ministry has established a Highway inventory database whereby the provincial highways are divided into following classes based on a system related to the service provided by the Highway:

- The desirable Kings Highway System
- The desirable Secondary Highway System and
- Transfer candidate

For transportation planning and design purposes the Ministry has further grouped highways on two major considerations access and mobility which constitute the functional classification system .The major division in the highway functional classification are :

- Freeway-Urban and Rural
- Arterial-Urban and Rural
- Collector-Urban and Rural
- Local-Urban and Rural

For this study three functional classes-Arterial, Collector and Local were combined and grouped under Kings Highway and Secondary Highway.

After removal of intersections and interchanges the remaining 100m segments (containing minor intersections) were classified into eight groups as presented in Table 4.1.

Table 4.1: Road Classification

Main Class	Highway Class	Group	Road Environment	No of Lanes
Freeways	Complex Freeways	1	All	All
	Simple Freeways	2	All	4
	Simple Freeways	3	All	More than 4
Kings Highways	Kings Highways	4	Rural	Less than 4
	Kings Highways	5	Rural	4 or more
Kings Highways	Kings Highways	6	Urban	Less than 4
	Kings Highways	7	Urban	4 or more
Secondary Highways	Secondary Highways	8	All	All

4.2.2 Data Assembly for Type 2 modeling

Segments with non-intersection accidents were made available by MTO in Excel files separately for each region but these contained intersection and interchange segments as was in the case for the master files. But no extensive segment removal process was required to prepare this database. The intersection files were simply merged with the database prepared for Type 1 modeling in SAS retaining only the segments that exist in the Type 1 database. So the final database contained no intersection related accidents. MTO has included the following accidents as non-intersection accidents:

- NonIntSec: Collisions not occurring at intersections, underpasses, overpasses, tunnels, bridges, private drives or railway crossing and also not related to activity at a nearby intersection.
- PrvDrv: Collisions that occurred at or near private drive which is not public roadway.
- RRxing: Collisions that occurred at railway crossing.
- UndPass: Collisions that occurred in a tunnel or on a roadway underneath of a structure
- OvrPass: Collisions that occurred on a bridge or on a roadway on a structure
- Oth-RdLoc: Collision occurred on a public roadway.

To prepare the database for Type 2 (nonintersection collision modeling), the private driveway collisions (PrvDrv) were taken away from the nonintersection collision files provided by MTO.

4.2.3 Incorporating 2005 data with the databases

At the beginning of the study MTO has mentioned the possibility of providing data for one more year i.e. for 2005, which was made available at the end of data assembly. In order to have SPFs with more significant coefficients, it is essential to have data for as many years as possible. So it was decided to incorporate 2005 data with the previous databases. In the 2005 data, traffic volumes came in a separate file whereby volumes were only shown against road sections of varying length. So it was required to disaggregate the volumes by 100m segment which involved considerable manual work before doing the final link in SAS.

4.2.4 Segments Aggregation

For each group 100m segments were aggregated to form contiguous segments of variable length that were homogeneous with respect to measured traffic and geometric characteristics (AADT, shoulder width, speed limit, etc.) of variable length.

5. DATA CHARACTERISTICS

5.1 Univariate statistics

These statistics indicate the characteristics of the main variables of each dataset and are presented in the Tables 5.1 and 5.2. The summary univariate statistics for these variables indicate that most of them show a good range of values that will provide the much desired variation for the modeling.

5.2 Bivariate statistics

Bivariate statistics show the correlation between accident count and other highway variable and is measured by a correlation coefficient. A positive coefficient indicates that the accident count increases as the variable increases and vice versa. Vogt and Bared (1998) mentioned that the significant relationship is the one which shows that the two variables are correlated in the population from which the samples are selected. P-values less than 5% indicate significant relationship. The correlation coefficients between total collision and each independent variable available in the dataset of Rural 2-Lane kings highway (roadway group 4) were examined which is shown in Table 5.3.The intention of doing this analysis was to help interpret models with various combinations of variables for this roadway type.

From Table 5.3, it can be observed that all the variables are positively correlated with the total accidents. Among these segment length and posted speed exhibit the most pronounced positive correlations with the total number of accidents emphasizing the importance of these variables in explaining the variations in total accidents. The correlation matrix shows the interrelationship between the independent variables which indicates that the presence of one input variable in the model may mask the effect of another input variable if the independent variables by themselves exhibit a strong interrelationship. Even though this is a valid concern, all the variables chosen would be tried in developing model for Rural 2-Lane Kings Highways.

Table 5.1: Summary Univariate Statistics (100m segments)

Roadway type	Groups	Number of segments	Variables	Min	Max.	Mean	Frequency	Variance
Freeways	1	70	AADT	177560	367040	272443		
			Total accidents	0	76	12.24	857	267.98
			FI	0	24	3.39	237	20.21
			PDO	0	64	8.86	620	154.75
	2	7385	AADT	5530	93980	23451		
			Total accidents	0	61	1.36	10069	6.03
			FI	0	15	0.35	2601	0.573
			PDO	0	53	1.01	7468	3.96
	3	1687	AADT	18260	349920	84970		
			Total accidents	0	88	3.96	6687	44.70
			FI	0	24	0.851	1436	2.6
			PDO	0	69	3.11	5251	29.36
Kings Highways rural	4	69590	AADT	540	31220	4038		
			Total accidents	0	34	0.5	35278	1.02
			FI	0	10	0.12	8423	0.16
			PDO	0	28	0.39	26855	0.68
	5	924	AADT	5040	48400	19362		
			Total accidents	0	42	1.67	1542	8.30
			FI	0	7	0.40	366	0.66
			PDO	0	35	1.27	1176	5.74

Table 5.1: Summary Univariate Statistics (100m segments)

Roadway type	Groups	Number of segments (100m)	Variables	Min	Max.	Mean	Frequency	Variance
Kings Highways urban	6	534	AADT	5800	19940	9393		
			Total accidents	0	47	1.13	602	9.40
			FI	0	12	0.33	178	0.71
			PDO	0	35	0.79	424	5.90
	7	209	AADT	9150	40880	18843		
			Total accidents	0	29	1.86	388	11.90
			FI	0	6	0.44	92	0.77
			PDO	0	25	1.42	296	8.06
Secondary Highways	8	48884	AADT	28	7650	461		
			Total accidents	0	12	0.10	4788	0.13
			FI	0	3	0.018	876	0.019
			PDO	0	9	0.08	3912	0.104

Table 5.2: Summery Univariate Statistics (longer segments)

Roadway type	Groups	Number of segments	Variables	Min	Max.	Mean	Frequency	Variance
Freeways	1	8	AADT	177560	367040			
			Total accidents	8	327	129.37	1035	15338.82
	2	192	AADT	5530	93980			
			Total accidents	0	296	63.44	12180	3090.69
	3	94	AADT	18260	349920			
			Total accidents	0	576	87.72	8246	13745.22
Kings Highways rural	4	1355	AADT	540	31220			
			Total accidents	0	224	31.43	42585	822.19
	5	70	AADT	5040	48400			
			Total accidents	0	141	25.94	1816	796.37
Kings Highways urban	6	30	AADT	5800	19940			
			Total accidents	0	116	24.03	721	798.63
	7	17	AADT	9150	40880			
			Total accidents	0	112	27.17	462	1227.01
Secondary Highways	8	613	AADT	28	7650			
			Total accidents	0	121	8.91	5676	183.87

Table 5.3: Pearson's correlation matrix with p-values for independent variables (Rural 2-Lane Kings Highway)

	Total accident	AADT	Segment length	Posted speed	Averag e speed	Shoulder width	Lane Width
Total accident	1.0000						
AADT	0.05517 0.0436	1.00000					
Segment length	0.65286 <.0001	-0.37488 <.0001	1.00000				
Posted speed	0.12041 <.0001	-0.15773 <.0001	0.18543 <.0001	1.00000			
Average speed	0.09159 0.0008	-0.02559 0.3497	0.06163 0.0242	0.58304 <.0001	1.00000		
Shoulder width	0.08393 0.0021	0.43569 <.0001	-0.19999 <.0001	0.23787 <.0001	0.14687 <.0001	1.00000	
Lane Width	0.07263 0.0079	0.28773 <.0001	-0.10872 <.0001	0.26493 <.0001	0.09821 0.0003	0.45599 <.0001	1.00000

6 MODELING

6.1 "AADT-only" Models

"AADT-only" models could suffer from omitted variables bias, but they are still the most popular type of models developed and used by transportation safety analysts (Hauer, 1997, Persaud, 2001). This type of model is preferred over the models that include several other covariates because they can be easily re-calibrated to be applied to another jurisdiction (Persaud, 2002 and Lord, 2005).

Type 1 (with minor intersection accidents) and **Type 2** (with non-intersection accidents) "AADT-only" negative binomial regression models were developed for all eight roadway classes following the form bellow:

$$E(Y) = n \times L \times a \times V^b \tag{6.1}$$

Where *E(Y)* is the predicted no of accidents

n is the number of years being used in the study

L is the section length in km,

V is the section AADT (annual average daily traffic)

a and *b* are regression coefficients estimated from data in the generalized linear modelling procedure.

Models were developed from aggregated contiguous segments that are homogeneous with respect to measured traffic and geometric characteristics (AADT, shoulder width, speed limit, etc.) for all roadway classes except complex freeway for which 100m segments were used.

Model calibration was performed with the SAS software using the GENMOD procedure that considers a negative binomial error structure. *n*×*L* was treated as offset variable. The SAS GENMOD procedure estimated the model coefficients and negative binomial dispersion parameter using a maximum likelihood method.

Total accident, Fatal and Injury (FI) accidents and Property Damage Only (PDO) accidents were considered as the dependent variable represented by *E(Y)* in Equation 6.1. A total of 49 regression models were developed. Models for PDO and FI accidents are shown in Tables 6.1.1 and 6.1.2.

Table 6.1.1 Type 1: Non Freeway Models for minor intersection accidents

Model Form: Accident/Year=Segment Length × a × AADT^{b}

Roadway Groups	Type of Accidents And p value of the parameters	Model Parameters		Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	Dispersion Parameter
		a	b			
Rural Kings Highways 2-lane	PDO	0.003661069	0.6588	1.86	1.186	0.1982
	p-value	<.0001	<.0001			
	FI	0.000316591	0.8055	1.41	1.126	0.1544
	p-value	<.0001	<.0001			
Rural Kings Highways >2-lane	PDO	9.25266E-06	1.2626	1.02	1.16	0.2398
	p-value	<.0001	<.0001			
	FI	3.51249E-05	1.0133	1.12	1.22	0.1303
	p-value	<.0001	<.0001			
Urban Kings Highways 2-lane	PDO	7.84221E-12	2.8835	2.87	1.23	0.869
	p-value	<.0001	<.0001			
	FI	4.26809E-09	2.0745	1.68	1.11	0.408
	p-value	<.0001	<.0001			
Urban Kings Highways > 2-lane	PDO	3.66965E-07	1.5994	0.96	1.04	0.3912
	p-value	<.0001	<.0001			
	FI	8.20625E-07	1.4074	1.09	1.32	0.3799
	p-value	0.0003	0.0003			
Secondary Highways	PDO	0.001597844	0.7661	1.21	1.02	0.4844
	p-value	<.0001	<.0001			
	FI	0.00022397	0.8356	1.45	0.93	0.2369
	p-value	<.0001	<.0001			

Table 6.1.2 Type 2: Models for nonintersection accidents

Model Form: Accident/Year=Segment Length × a × AADT^b

Roadway Groups	Type of Accidents	Model Parameters		Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	Dispersion Parameter
		a	b			
Complex freeway	PDO	0.00092289	0.7854	1.04	1.19	1.4607
	p-value	0.3164	0.1592			
	FI	1.57354E-05	1.0354	1.04	1.14	1.0711
	p-value	0.0974	0.0522			
Simple Freeway 4-lane	PDO	0.001264473	0.7509	4.51	1.19	0.3745
	p-value	<.0001	<.0001			
	FI	0.000208661	0.8113	3.56	1.25	0.2163
	p-value	<.0001	<.0001			
Simple Freeway >4-lane	PDO	0.000632899	0.8113	1.13	1.13	0.4138
	p-value	<.0001	<.0001			
	FI	1.99417E-05	1.0043	1.45	1.16	0.3803
	p-value	<.0001	<.0001			

Table 6.1.2 Type 2: Models for nonintersection accidents continued

Model Form: Accident/Year=Segment Length × a × AADT^{b}

Roadway Groups	Type of Accidents	Model Parameters		Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	Dispersion Parameter
		a	b			
Rural Kings Highways 2-lane	PDO	0.008861848	0.5356	1.15	1.1	0.1688
	p-value	<.0001	<.0001			
	FI	0.0008611	0.6638	1.05	1.07	0.158
	p-value	<.0001	<.0001			
Rural Kings Highways >2-lane	PDO	8.21294E-06	1.2535	1.01	1.20	0.2968
	p-value	<.0001	<.0001			
	FI	1.18795E-05	1.0967	1.03	1.23	0.1036
	p-value	<.0001	<.0001			
Urban Kings Highways 2-lane	PDO	1.68811E-06	1.4699	1.47	1.22	0.1749
	p-value	<.0001	<.0001			
	FI	1.38862E-06	1.3915	1.08	1.04	0.1438
	p-value	<.0001	<.0001			
Urban Kings Highways > 2-lane	PDO	1.44891E-06	1.4348	0.77	1.05	0.4359
	p-value	0.0002	<.0001			
	FI	8.57097E-08	1.607	1.34	1.37	0.3541
	p-value	<.0001	0.0001			
Secondary Highways	PDO	0.001533351	0.76	1.07	0.98	0.4787
	p-value	<.0001	<.0001			
	FI	0.000221586	0.8211	1.10	0.87	0.2328
	p-value	<.0001	<.0001			

Figure 6.1.1 plots the Safety Performance functions (SPFs) for total accidents calibrated from 100m segment for complex freeways and from longer segments for all other roadway classes. For complex freeways the number of aggregated longer segments were too few to calibrate models. From Figure 6.2.3, it is evident that, except for Urban 2-Lane and 4-Lane Kings Highways and Rural 4-lane Kings Highways, the slope of the regression line decreases as the flow increases indicating perhaps the influence of decreasing speed as flow increases for these classes. For classes that have SPFs with increasing slopes, there seems to be an increasing possibility of risk with higher AADT levels.

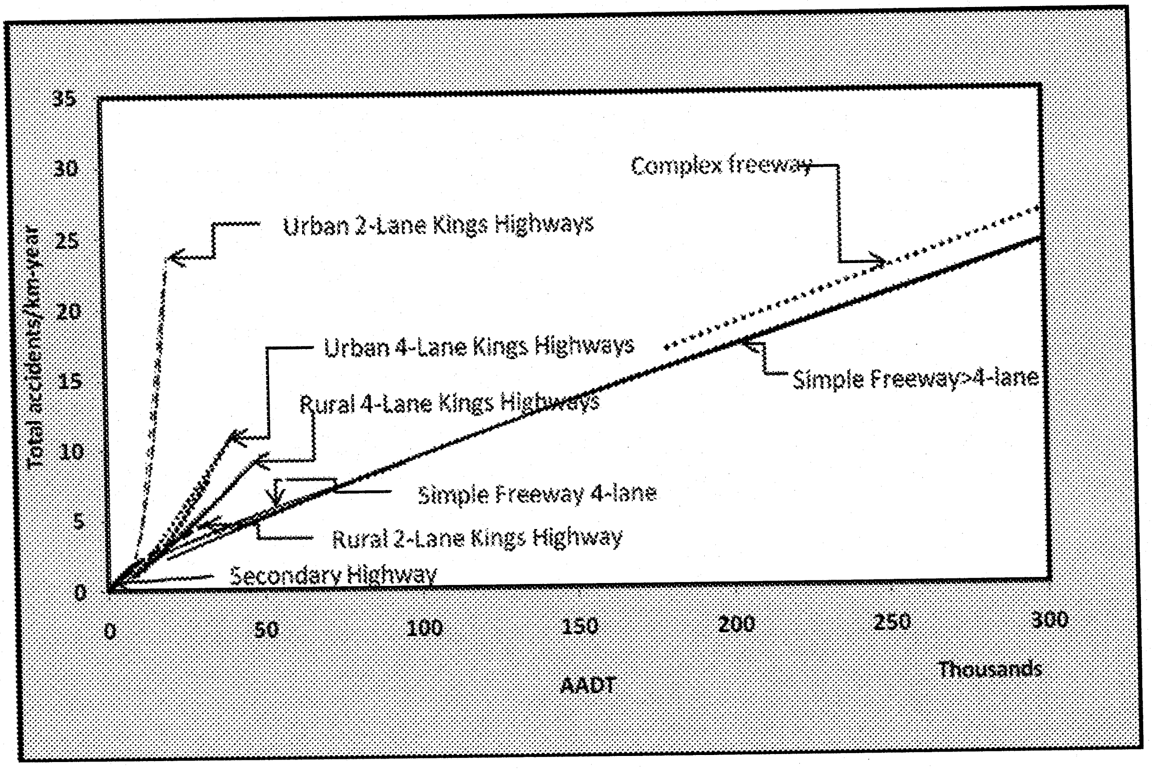


Figure 6.1.1 Plots of SPFs for total accidents for all Ontario roadway classes calibrated from longer segments

The Goodness of fit of the calibrated models was assessed by the values of Pearson chi-square and scaled deviance as described in Chapter 3. Values for each model are shown in Table 6.1.1 and 6.1.2. For most of the models these two values were around 1 fulfilling the required criteria. To supplement these two tests, CURE (cumulative residual) plots were also used to assess model goodness of fit. Figure 6.1.2 shows the cure plot of PDO (property damage only) accidents for complex freeway obtained from 100m segment model. An oscillation around zero shows a good fit as is the fact that the CURE curve is well within the 2STD band.

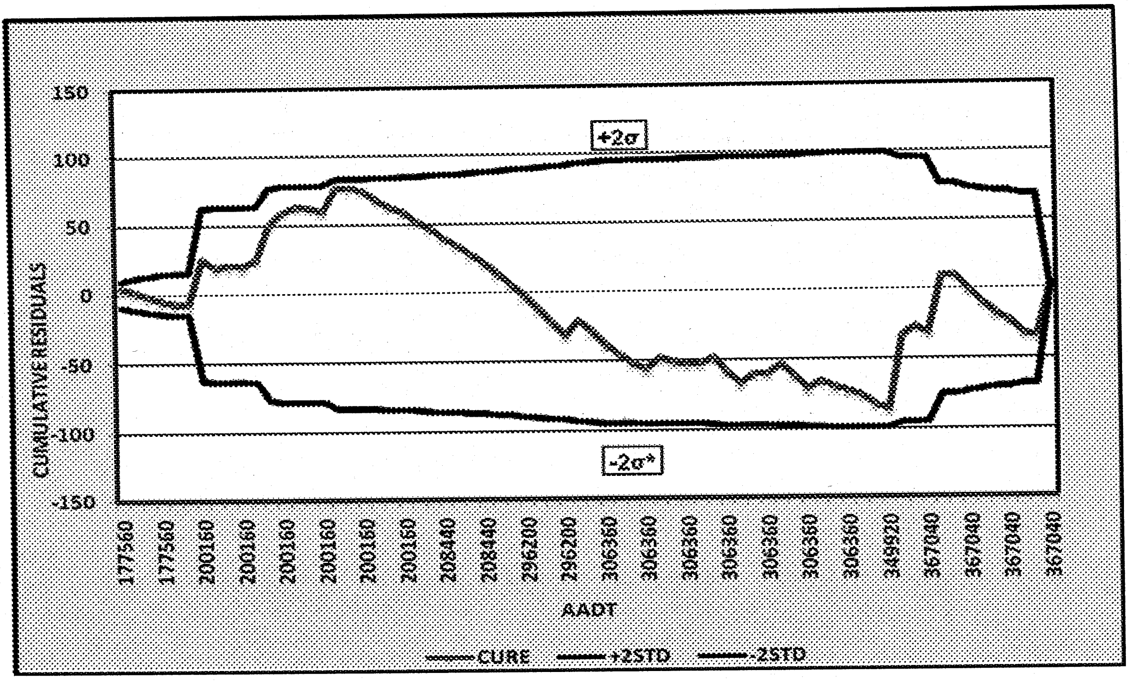


Figure 6.1.2 CURE plot for PDO accidents of Complex freeway

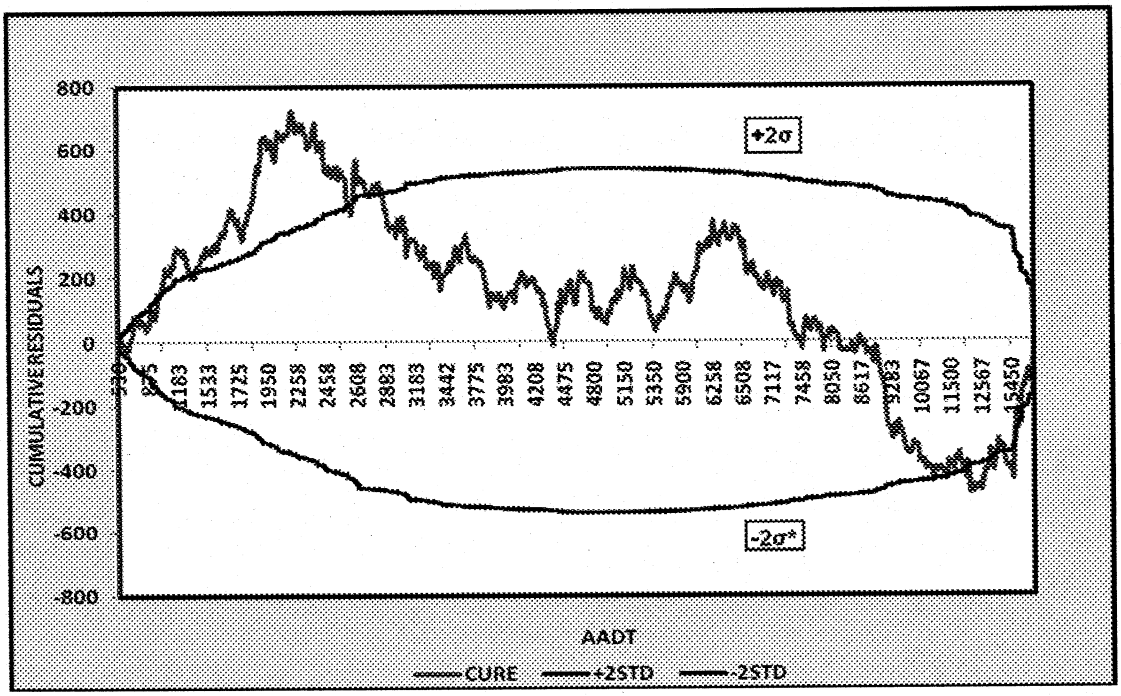


Figure 6.1.3 CURE plot for total accidents of Rural 2-Lane Kings Highways

Figure 6.1.4 shows the Safety Performance Functions (SPFs) obtained from “AADT-only” model for Rural 2-Lane Kings Highways.

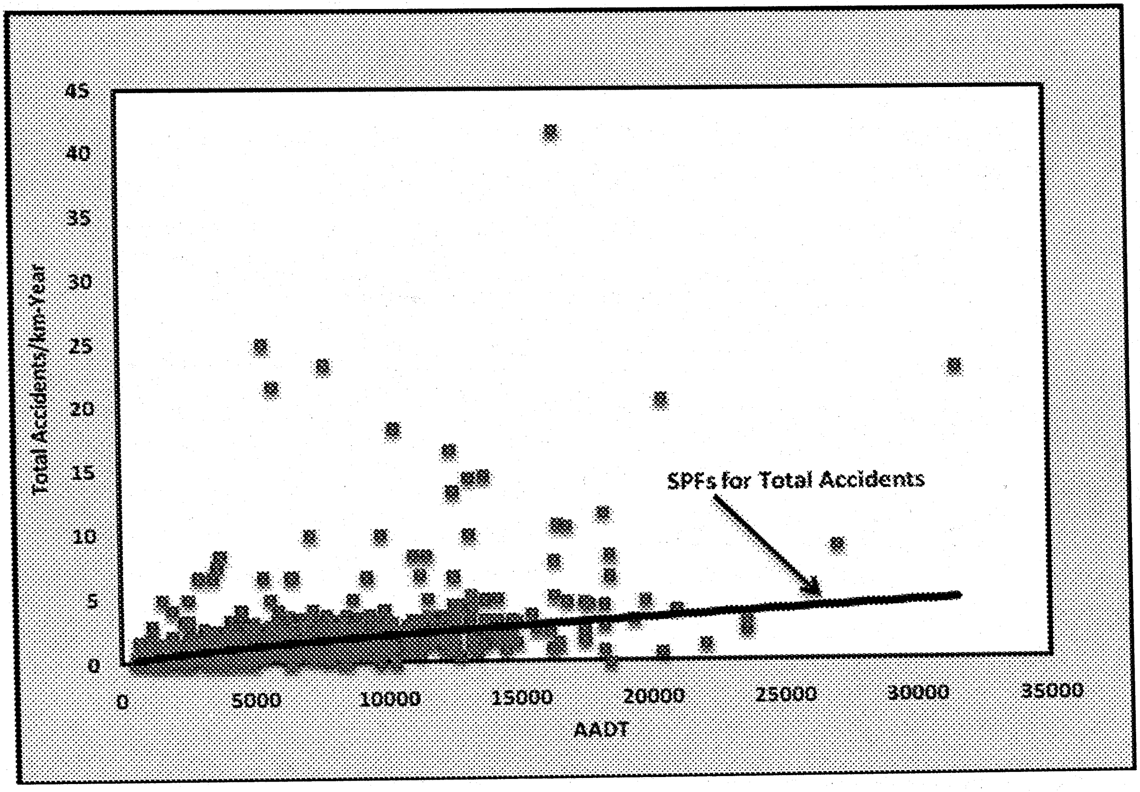


Figure 6.1.4 Scatter plot of total observed accidents and Safety Performance Functions (SPFs) of Rural 2-Lane Kings Highway calibrated from longer segments

6.2 A Comparison between constant length shorter segment models and variable length longer segment models

In this part of thesis “AADT-only” negative binomial regression models were calibrated from 100m constant length homogeneous segments to compare the outcomes of these models with those calibrated in the previous section from longer segments.

The model coefficients and dispersion parameter estimated from the two modeling approaches for Rural and Urban 2-lane Kings Highways are presented in the following two tables:

Table 6.2.1 Comparison of estimated coefficients and dispersion parameters obtained from 100m segment and aggregated segments for Rural 2-lane Kings Highways

	100m segment			Aggregated segment		
Accident severity	Coefficients		Dispersion	Coefficients		Dispersion
	a	b	k	a	b	k
PDO	0.004229	0.6347	1.0275	0.00366	0.6588	0.1982
FI	0.00036	0.7879	1.2089	0.000317	0.8055	0.1544
Total	0.004245	0.6663	0.9579	0.003423	0.6992	0.1831

Table 6.2.2 Comparison of estimated coefficients and dispersion parameters obtained from 100m segment and aggregated segments for Urban 2-lane Kings Highways

	100m segment			Aggregated segment		
Accident severity	Coefficients		Dispersion	Coefficients		Dispersion
	a	b	k	a	b	k
PDO	0.00000485	1.3415	1.0689	0.00000073	1.5648	0.22
FI	0.00000194	1.3558	0.4047	0.00000080	1.4544	0.1189
Total	0.00000497	1.3798	0.843	0.00000089	1.582	0.2022

It is evident from the Tables 6.2.1 and 6.2.2 that the model coefficients as well as dispersion parameters depend on whether equal length shorter segments or aggregated longer segments are modelled. The coefficient “a” decreases and “b” increases as longer segments are used. Also shorter segments have larger dispersion parameter than longer aggregated segments.

Recall that with the EB (Empirical Bayes) estimate, the expected accident frequency of an entity is given as:

$m = w(p) + (1\text{-}w)\, x$

where p is estimated from SPFs as described in Chapter 1

x is the count of collision on an entity and

w is the weight estimated from the dispersion parameter obtained from model calibration and is given by:

$$w = \frac{1}{1 + k\mu}$$

where μ is the model prediction accidents/km-year and

k is dispersion parameter estimated per km length for segments.

The above expression for w shows that a large dispersion parameter makes the EB weight small resulting in the accident potential close to the accident count. Alternately a small dispersion parameter makes w large and the larger the value of w, the greater is the influence of the model prediction, μ, and hence the smaller is the influence of the observed accident counts. This in turn indicates that the smaller the value of dispersion parameter the better a model is for a given set of data.

So models calibrated from longer segments are better since they resulted in small dispersion parameter. Further study on dispersion parameter is provided in Chapter 8.

Figures 6.2.1 and 6.2.2 represent the plot the Safety Performance Functions (SPFs) for total accidents calibrated from 100m segments and longer segments for Rural 2-Lane Kings Highways and Urban 2-Lane Kings Highways. It is evident that for Rural 2-Lane Kings Highway the slope of the regression line decreases as the flow increases but this is in contrast to the SPF plots for Urban 2-Lane Kings Highways which shows increasing slopes. It is possible that the SPF plot for Urban 2-Lane Kings Highways is reflecting the increasing possibility of risk with higher AADT levels. It is also evident from the plots that the SPFs calibrated from the 100m segments predicts fewer accidents than do the SPFs from the longer segments.

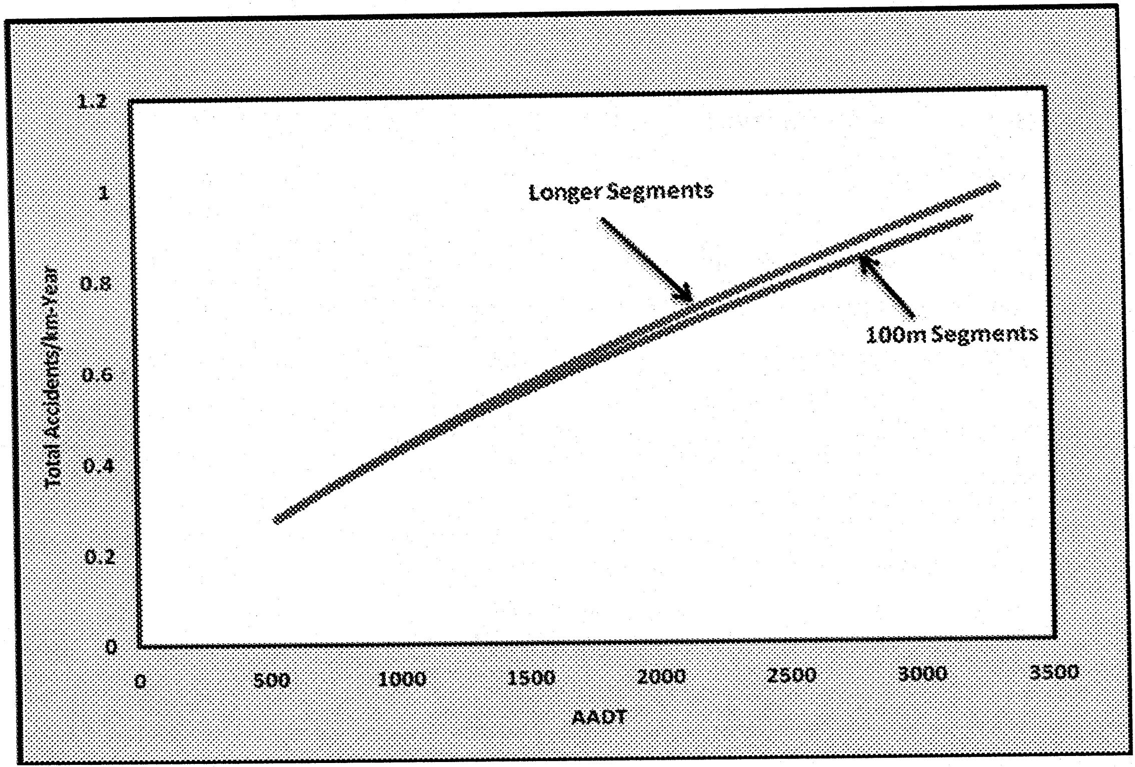


Figure 6.2.1 Safety Performance Functions (SPFs) for total accidents for Rural 2-Lane Kings Highways calibrated from 100m segment and longer segments

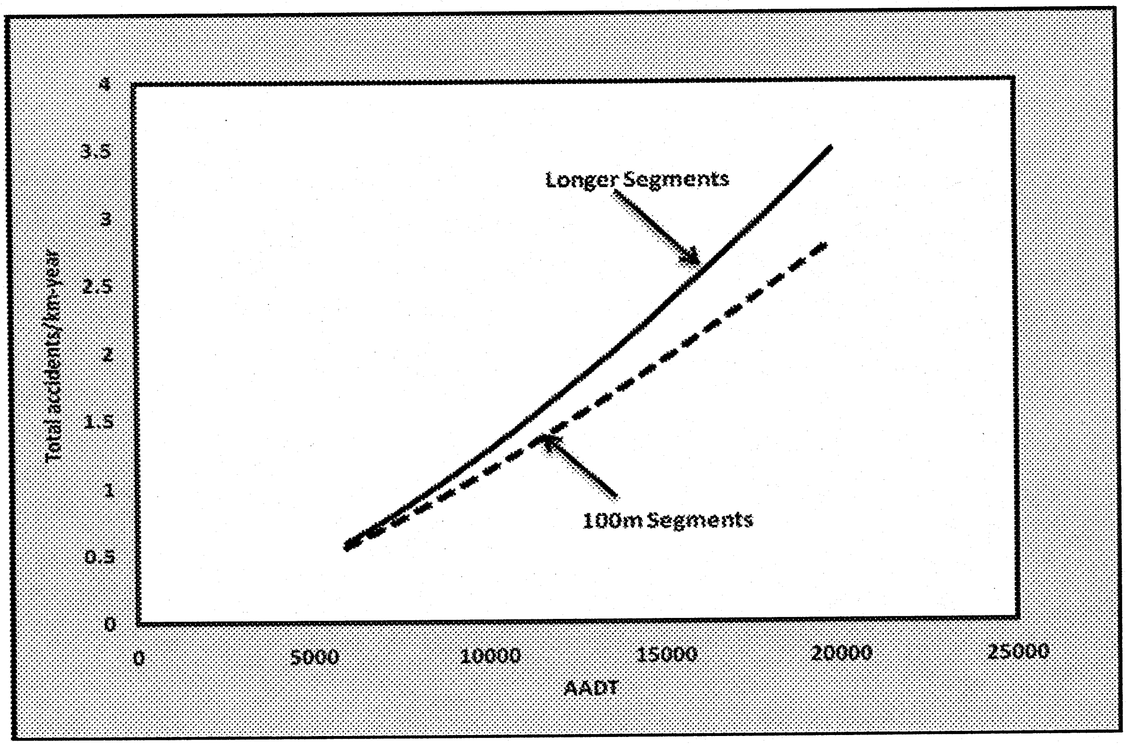


Figure 6.2.2 Safety Performance Functions (SPFs) for total accidents for Urban 2-Lane Kings Highways calibrated from 100m segment and longer segments

6.3 Estimation of models with additional variables for Rural 2-Lane Kings Highways

"AADT-only" models have already been developed for this road class in section 6.1.In this part of thesis inclusion of the additional explanatory variables available in the datasets is examined to develop a full model for this road class:

Seg_length	Segment Length
surftype	Surface Type
post_dspeed	Posted speed
avg_speed	Average speed
shldwidth	Shoulder Width
lane_width	Lane Width
Terrain	Terrain

For predictive models it is logically necessary to add segment length as a multiplier in the model equation as is shown in Equation 6.1 which makes the prediction proportional to segment length. However the relationship between the segment length and accident frequency might not be linear. As noted by several researchers' segment length as a variable in the model could be a surrogate for other unmeasured variables. So segment length was also taken as an explanatory variable in this investigation of additional variables.

In this investigation except for the segment length variable, all other variables were treated as a categorical rather than as a continuous. Treating variables as categorical provides an opportunity to identify unusual or unexpected relationships between dependent and independent variables. Two criteria, a drop in scaled deviance and the *p-value* were examined before making the decision on retaining a variable in the model. These two criteria were explained in detail in Chapter 3.

Table 6.3.1 shows the relevant statistics of the data used in this part of study.

Table 6.3.1 Relevant statistics of database for 1338 Rural 2-Lane Kings Highway Segments used for full models

Variables	N	Mean	Std Dev	Sum	Minimum	Maximum
Total accidents	1338	31.68	28.74	42385	0	224
AADT	1338	5545	3966		530	31550
Posted speed (kph)	1338	83.27	6.85		0	90
Average speed (kph)	1338	99.07	9.03		0	114
Shoulder Width (m)	1338	1.23	0.26		0	3.90
Lane Width (m)	1338	3.61	0.16		2.23	4.25
Segment length (km)	1338	5.18	5.18	6925	0.1	43.3

In Figure 6.3.1 variables are sorted in order of their contribution to the log-likelihood value. As expected, AADT contributes the most. The next important variable is the terrain of the segment. This is followed by segment length. The contribution of Lane width and shoulder width are similar. Posted speed contributes much more than the average operating speed. The least contribution was seen for the surface type variable. The contribution of each non-AADT variable was examined with AADT in the model.

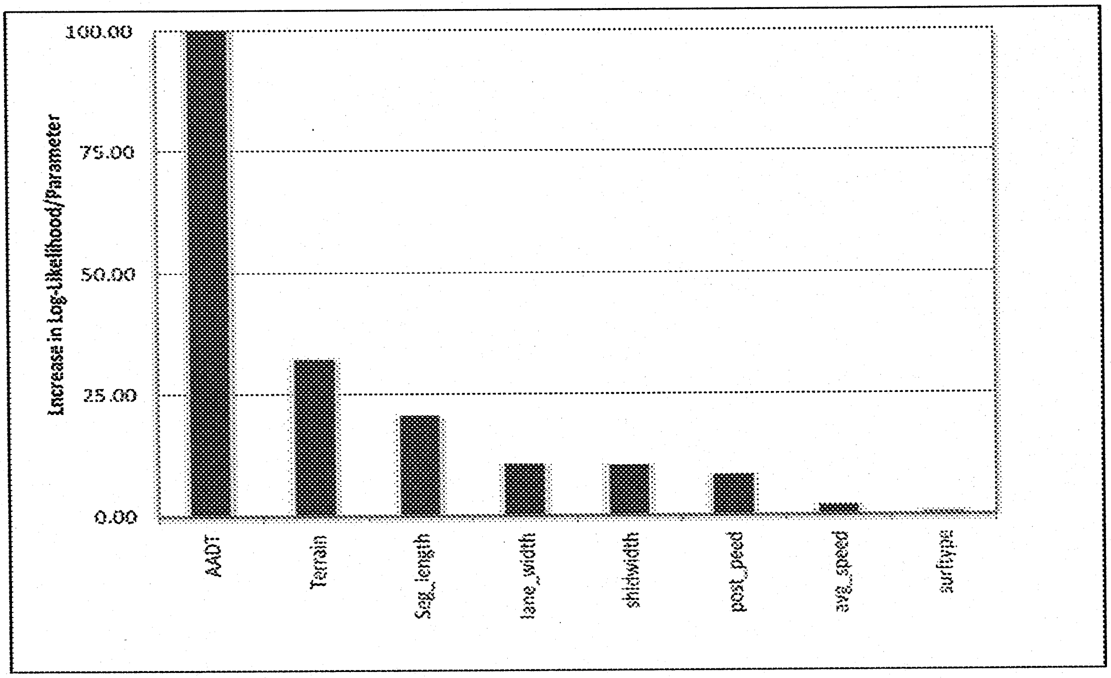


Figure 6.3.1 Contribution of the variables to the log likelihood value

Following Hauer et al. (2004) variables were added one by one to the "AADT-only" model on the basis of their contribution to the log-likelihood value as shown in Figure 6.3.1 and retained in or discarded from the model according to the criteria mentioned above. The "AADT-only" model was of the following form:

$$E(Y) = n \times L \times a \times V^{b} \qquad (6.1)$$

where *E(Y)* is the predicted no of accidents, n is the number of years being used in the study

L is the segment length in km, *V* is the segment AADT

a and b are regression coefficients estimated from data in the generalized linear modelling procedure.

The linear version of Equation 6.1 can be written as:

$$E(Y) = exp\ [ln\ (nL) + ln\ (a) + bln\ (AADT) \qquad (6.2)$$

Variables were added to the model one by one. Significance and inclusion of each variable is discussed in the subsections to follow.

6.3.1 Terrain

In the database the variable terrain was used to distinguish between Flat and Rolling terrain. This variable was added to the "AADT only" models of the form shown in Equation 6.1 as e^{cNTC} resulting in the following form:

$$E(Y) = exp\ [ln\ (nL) + ln\ (a) + bln\ (AADT) + c_1\ NTC_i\,(1) + c_2\ NTC_i\,(2)] \qquad (6.3)$$

where $NTC_i\ (j)$ are the indicator variable associated with j^{th} level of terrain and c_1 and c_2 are the coefficients; other variables and coefficients are as described before.

$$NTCi\,(j) = \begin{cases} 1\ if\ NTC = j \\ 0\ if\ NTC \neq j \end{cases} \qquad (6.4)$$

As discussed in Chapter 3, negative binomial regression models with different variables have different *k*-values. However, scaled deviance of two negative binomial models with different *k* values cannot be compared. So the *k* value obtained from the "AADT only "model was imposed to the model with AADT and terrain and the drop in deviance was then compared with $\chi 2_{\,0.05,1}$ (equal to 3.84) in order to assess the usefulness of this variable. Drops were found to exceed that value in the case of all of the accident severity models (i.e., for PDO, FI and Total accidents). Table 6.3.2 shows the results of this analysis.

Table 6.3.2 Drop in deviances for the variable terrain

	PDO		FI		Total Accidents			
Accident Severity	k-value	Deviance	k-value	Deviance	k-value	Deviance	Terrain coefficients	
Model with AADT	0.1958	1602.47	0.1561	1524.38	0.1819	1608.71	C_1	C_2
Model with AADT and Terrain	0.1958	1536.94	0.1561	1477.97	0.1819	1537.16	-0.1618	0
Drop in deviance		65.53		46.41		71.55		

The *p-value* of the estimated coefficients were all <.0001, indicating that this variable is highly significant. So this variable was considered for inclusion in the model. The estimated coefficients indicate that crash frequency is less in FLAT terrain. The actual effect of this variable is assessed from the value of $e^{-0.1618}$ which suggests that flat terrains would be expected to have 15% less accidents than rolling terrains.

6.3.2 Segment length

For predictive models it is common to add segment length as a multiplier in the model equation as is shown in Equation 6.1 which makes the prediction proportional to segment length. But the relationship between the segment length in the model and accident frequency might not necessarily be linear. As noted by several researchers, the use of segment length as a variable in the model is a surrogate for other unmeasured variables. So an attempt was made to include this as a variable in the model equation. It was considered as a continuous variable and added to the existing model which already contained the variables AADT and terrain. Two forms $e^{d \times \text{segment length}}$ and $(\text{segment length})^d$ were considered as follows:

$$E(Y) = \exp [\ln (nL) + \ln (a) + b\ln (AADT) + c_i NTC_i + d\,(\text{segment length})] \tag{6.5}$$

$$E(Y) = \exp [\ln (nL) + \ln (a) + b\ln (AADT) + c_i NTC_i + d\ln (\text{segment length})] \tag{6.6}$$

Where *d* is the coefficient for segment length and the other variables and coefficients are as described before.

Results for the form $e^{d \times \text{segment length}}$

The final model developed in the previous stage has a new *k* value which was imposed to the model with the variables AADT, terrain and Segment length. Deviance was calculated using the SAS GENMOD procedure. A significant drop is evident between the two models. Table 6.3.3 shows the drop in deviances for three accident severities. The drop observed for PDO and Total Accidents exceed 3.84 which indicate that the segment length variable needs to be included in the model equation for PDO and Total accidents. It appears from the insignificant drop in deviance for FI accidents that segment length is not associated with this accident severity. (Note that the other form for inclusion of segment length in the model, which is presented later, does show a strong relationship between segment length and FI accidents.

For the current form the *p-value* of the parameters were found highly significant (<.0001).

Table 6.3.3 Drop in Deviances for variable "segment length" with form $e^{d \times \text{segment length}}$

Models	PDO		FI		Total accidents	
	k-value	Deviance	k-value	Deviance	k-value	Deviance
Model with AADT and Terrain	0.1833	1593.2	0.1474	1505.88	0.1694	1601.46
Model with AADT, terrain and Segment length	0.1833	1574.11	0.1474	1504.04	0.1694	1590.95
Drop in deviance		19.05		1.84		18.35

Table 6.3.4 shows the variation in each coefficient and dispersion parameter with the addition of a new variable in the model. Addition of the segment length variable has changed all the parameters of the existing variables in the model to some extent.

Table 6.3.4 Parameter estimates of the successive models "only AADT", "AADT & Terrain", and "AADT, Terrain & Segment length" for total accidents with form $e^{d\times \text{segment length}}$.

Models	Intercept	AADT	Terrain Category		Segment Length	Dispersion Parameter
			Flat	Rolling		
AADT only	0.002923	0.695				0.1819
AADT & Terrain	0.002638	0.7164	-0.1618	0		0.1694
AADT, Terrain & Segment length	0.004298	0.6675	-0.1586	0	-0.1334	0.1687

Result for the form (segment length)^d :

More significant results were obtained with this form. The drop in deviances was more than that obtained with the previous form and was significant for all accident severities. Drop in deviances are shown in Table 6.3.5. Parameter estimates were also highly significant at 95% confidence limit (p<.0001).

Table 6.3.5: Drop in Deviances for variable "segment length" with form (segment length)^d

Models	PDO		FI		Total Accidents	
	k-value	Deviance	k-value	Deviance	k-value	Deviance
Model with AADT and Terrain	0.1833	1593.2	0.1474	1505.88	0.1694	1601.46
Model with AADT, terrain and Segment length	0.1833	1521.09	0.1474	1487.33	0.1694	1513.58
Drop in deviance		72.11		18.55		87.88

Table 6.3.6 shows that an inclusion of this variable in this form has changed the parameter estimates of AADT and intercept more than with the previous form. The decrease in the dispersion parameter with inclusion of this new variable also indicates that this variable is important. The negative sign of the coefficient of segment length indicates that shorter segments have more accidents than longer segments which, at first glance, is intuitively wrong. But in this case segment length was also taken as an offset variable in the model, as can be seen in Equations 6.4 and 6.5. So the resultant coefficient of segment length becomes (1-0.136) =0.864 i.e. the term becomes (Segment length)$^{0.864}$ which is intuitively right as it indicates that predicted accidents are more with longer segments.

Table 6.3.6 Parameter estimates of the successive models "only AADT"," AADT & Terrain", and "AADT, Terrain & Segment length" for total accidents with form (segment length)^{d}

Models	Intercept	AADT	Terrain Category		Segment Length	k-value of the model for total accident
			Flat	Rolling		
AADT only	0.002923	0.695				0.1819
AADT & Terrain	0.002638	0.7164	-0.1618	0		0.1694
AADT, Terrain & Segment length	0.0067	0.6284	-0.1631	0	-0.1361	0.1647

The coefficient of segment length is significantly different from 1. This is a reflection of the fact that the lengths of road segments can be correlated with various other causal variables not represented in the model equation and the relationship between accident frequency and segment length is not linear. The following two cure plots shown in Figures 6.3.2 and 6.3.3 evident the improvement in model fit with segment length as a variable in the model equation.

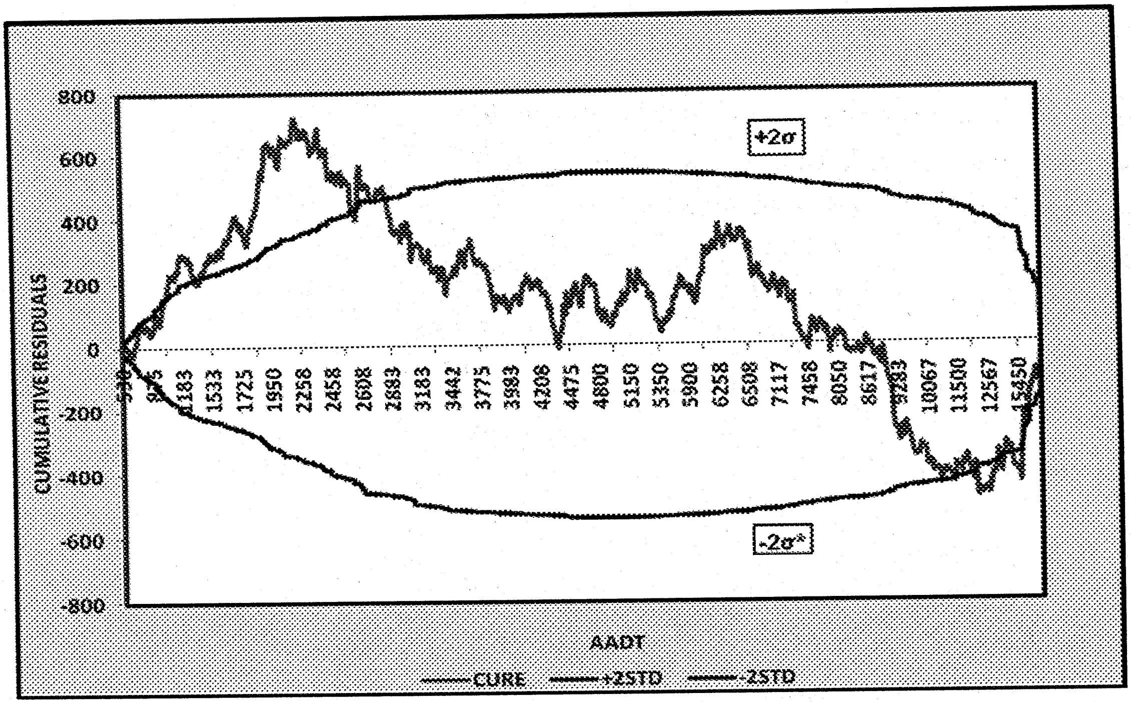


Figure 6.3.2 CURE plot with variable AADT in the model equation for Rural 2-Lane Kings Highways

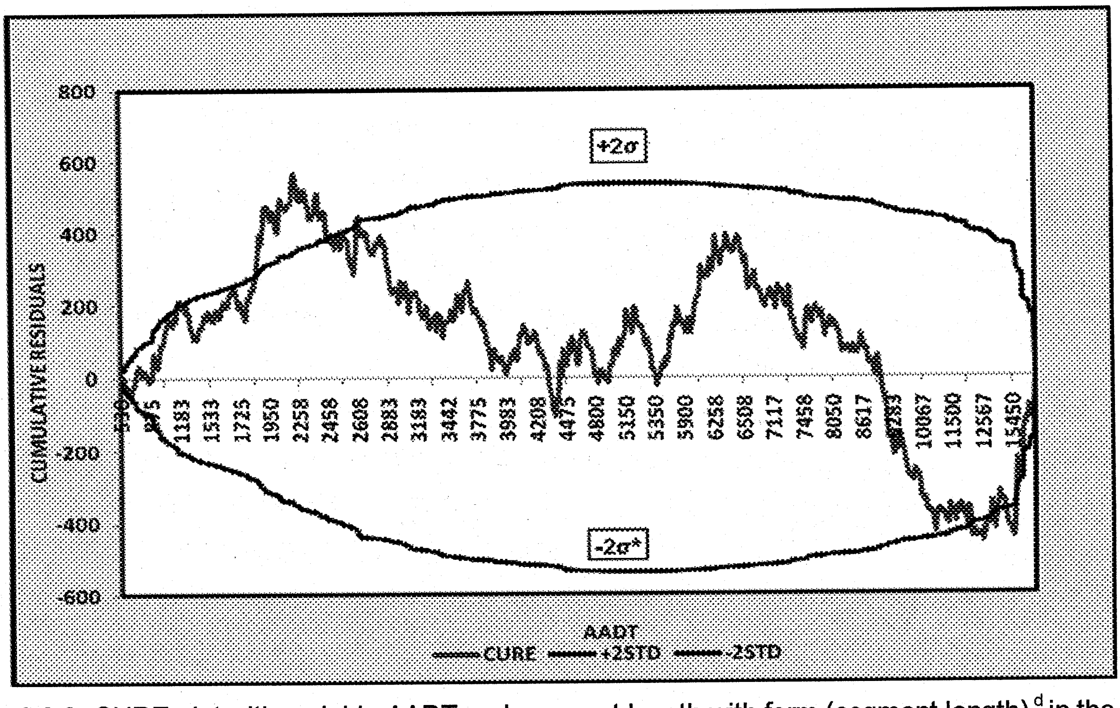


Figure 6.3.3 CURE plot with variable AADT and segment length with form (segment length)^{d} in the model equation for Rural 2-Lane Kings Highways

From the investigation it is clear that segment length can be retained in the model as a variable of the form shown in equation 6.4 i.e. (segment length)^{d}.

6.3.3 Lane width

To represent the effect of lane width, this variable was divided into three categories and added to the model equation which already contained the variables AADT, Terrain and segment length. Table 6.3.7shows the categories of lane width variable.

Table 6.3.7 Lane width category

Lane width Category	Range of lane width (m)
LWC_1	Less than 3.65(12ft)
LWC_2	=3.65
LWC_3	More than 3.65

The model equation for adding lane width was:

$$E(Y) = \exp [\ln (NL) + \ln (a) + b\ln (AADT) + c_i\, NTC_i + d\ln (\text{segment length}) + e_j\, LWC_j] \qquad (6.7)$$

Where LWC_j are the indicator variable of lane width and e_j are the coefficients.

LWC_1=1 if lane width is less than 3.65m otherwise 0
LWC_2=1 if lane width is equals to 3.65m otherwise 0
LWC_3=1 if lane width is more than 3.65m otherwise 0

Deviances of the present model were compared with the one developed in the previous stage keeping the k-value constant. The drop in deviances for the PDO and FI model are insignificant. For total accidents the drop is much lower compared to the drops observed in the previous two stages although it was above the drop of 3.84 required for significance (as shown in table 6.3.8). Parameter estimates of the entire lane width category were also not found significant at the 95% confidence limit (p-values are 0.0092 for LWC_1 and 0.7993 for $LWC_{2)}$. The change in dispersion parameter was also marginal. All these facts lead to the weak relationship between accident and lane width. So this variable was discarded from the model.

Table 6.3.8 Drop in deviances for variable “lane width” for all accident severities.

Model	PDO		FI		Total Accidents		k-value of the final model for total accident
	k-value	Deviance	k-value	Deviance	k-value	Deviance	
Model with AADT, terrain and Segment length	0.179	1539.64	0.1466	1487.85	0.1647	1537.34	0.1647
Model with AADT, terrain ,Segment length and Lane width	0.179	1536.39	0.1466	1485.09	0.1647	1531.6	0.1644
Drop in deviance		3.25		2.76		5.74	

Recall from the contributions of variables to the log-likelihood value shown in Figure 6.3.1, this lane width was the third most important variable. But when other variables, terrain and segment length are included in the model the significance of lane width has been reduced as is evident from the above drop in deviance and *p-value.*

6.3.4 Shoulder width

Shoulder width available in the data varied from 0m to 3.9m. These widths were grouped into four categories and added to the model equation which already contained variables AADT, terrain, and segment length. Table 6.3.9 shows the categories of shoulder width.

Table 6.3.9 Shoulder width categories

Shoulder width Category(SWC)	Range of shoulder width (m)
SWC_1	Less than 2.5
SWC_2	2.5-3
SWC_3	= 3
SWC_3	More than 3

The model equation for adding shoulder width was:

$$E(Y) = \exp [\ln (NL) + \ln (a) + b\ln (AADT) + c_i NTC_i + d\ln (\text{segment length}) + e_j SWC_j] \qquad (6.8)$$

where SWC_j are the indicator variables for shoulder width category, e_j are the coefficients and j =1,2...4. The values of the indicator variables are 1 and 0 as described before for lane width category.

Significant drops in deviances are observed for all accident severities. Parameter estimates were also found significant at 95% confidence limit. Drop in the dispersion parameter is also a sign of a better model. So this variable was considered to be included in the model.

Table 6.3.10 Drop in deviances for shoulder width variable

Model	PDO		FI		Total Accidents		k-value of the model for total accident
	k-value	Deviance	k-value	Deviance	k-value	Deviance	
Model with AADT, terrain, Segment length	0.1787	1535.06	0.1462	1484.82	0.1644	1530.55	0.1644
Model with AADT, terrain ,Segment length, and shoulder width	0.1787	1525.68	0.1462	1478.53	0.1644	1520.98	0.1621
Drop in deviance		9.38		6.29		9.57	

6.3.5 Speed limit

In the Rural 2-Lane Kings Highways database speed limit had values of 50kph, 60kph, 70kph, 80kph and 90kph. Most of the values were 80kph. Some segments had missing values and were excluded from the database. These values were grouped into three main categories for the variable SLC: <80kph, =80kph and >80kph and added to the model equation in the form $e^{g \times SLC}$ which already contained variables AADT, terrain, and segment length and shoulder width.

The model equation with this variable was:

$$E(Y) = exp\ [ln\ (NL) + ln\ (a) + bln\ (AADT) + c_i\, NTC_i + dln\ (segment\ length) + e_j\, SWC_j + g_m\, SLC_m] \qquad (6.9)$$

where SLC_m are the indicator variables for speed limit category, g_m are the coefficients for m=1,2 and 3. The value of the indicator variables are 1 and 0 as described for lane width category.

No significant drop in deviance was observed for total accidents. Parameter estimates of the categories <80kph and =80kph were also found insignificant. The dispersion parameter obtained with inclusion of this variable was equal to the one obtained without this variable. So this variable was not required for inclusion in the model.

6.3.6 Average travel speed

In the database this variable ranged in values from 60kph to 113kph. Most of the values were 100kph. So the variable was grouped into three categories: <100kph, =100kph and >100kph.

Deviances of the present and previous model were recorded. No significant drop was observed for the total accident model and the p-values of the parameters were also not found significant at the 95% confidence level. No change in dispersion parameter was evident. So this variable was not required for inclusion in the model.

6.3.7 Surface type

In the database the surface type variable distinguished between A/C (Asphalt on concrete), HCB (High Class Bituminous Pavement), CONC (Concrete) and NONE. To represent the effect of surface type in the model, a set of categorical variables were added to the model, giving the following model equation:

Model Equation:

$$E(Y) = \exp [\ln (nL) + \ln (a) + b\ln (AADT) + c_i\, NTC_i + d\ln (\text{segment length})$$

$$+ e_j\, SWC_j + g_j\, ST_j] \qquad (6.10)$$

Where ST_j are the indicator variables for surface type category and g_j are the coefficients for j = 1,2...4. The value of the indicator variables are 1 and 0 as described before.

No significant drop in deviance was observed for total accidents and the parameter estimates of the surface type category were highly insignificant. The dispersion parameter was almost unchanged from when this variable was excluded. So this variable was discarded from the model.

6.4 Final model equation for Rural 2-lane Kings Highway

From the analysis presented above, the following was the final model form:

$$E(m) = aAADT^{b} \times (segment\ length)^{c} \times e^{(d_i NTC_i + e_j SWC_j)}$$

Where

E (m) = Predicted number of accidents per km-year

NTC_i =indicator variable for terrain; NTC_1=1 if flat terrain otherwise 0

NTC_2=1 if rolling otherwise 0.

SWC_j =indicator variable for Shoulder Width; SWC_1=1 if shoulder width <2.5m otherwise 0

SWC_2=1 if shoulder width is between 2.5-3m otherwise 0.

SWC_3=1 if shoulder width is equals to 3m otherwise 0

SWC_4=1 if shoulder width is more than 3m otherwise 0.

a, b, c, d_i and e_j are the model parameters. The actual effect of the categorical variables in the model is determined by taking the exponential function of the parameters [e.g.exp (d_i)]

Model parameters are shown in Table 6.4.1 for PDO, FI and Total accidents. Figure 6.4.1 shows the CURE plot of the final model for total accidents. It can be observed from the CURE plot that this model better fits the data compared to "AADT-only" model.

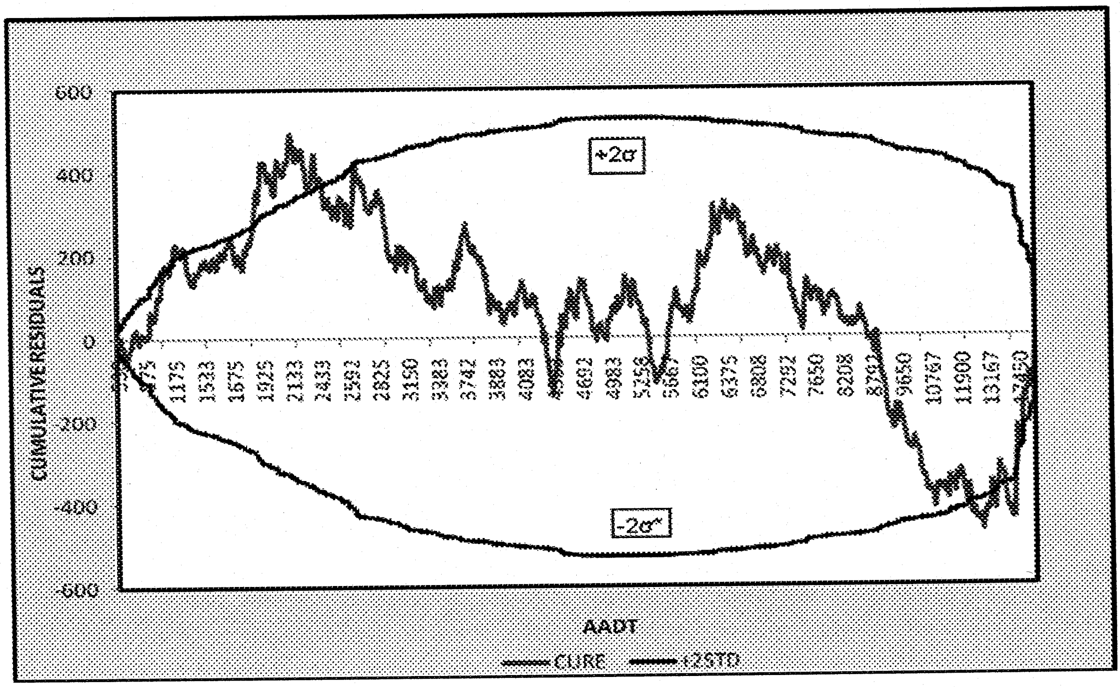


Figure 6.4.1 CURE plot of the final full model for Rural 2-Lane Kings Highways for total accidents

Table 6.4.1 Parameter estimates of full model for Rural 2-Lane Kings Highways

Model Form: Accident/km-year, $E(m) = aAADT^{b} \times (segment\ length)^{c} \times e^{(d_i NTC_i + e_j SWC_j)}$

Type of Accidents	Parameter estimates									Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	Dispersion Parameter
				Terrain indicator variable		Shoulder width indicator variable						
				NTC_1	NTC_2	SWC_1 (<2.5)	SWC_2 (2.5-3m)	SWC_3 (3m)	SWC_4 (>3m)			
	Ln(a)	b	c	d_1	d_2	e_1	e_2	e_3	e_4			
PDO(Property Damage Only)	-4.5985	0.5967	-0.15	-0.1486	0	-0.2023	-0.1998	-0.266	0	1.41	1.14	0.1789
p-value	<.0001	<.0001	<.0001	<.0001	-	0.0272	0.0299	0.0043	-			
FI(Fatal & Injury)	-7.4227	0.7799	-0.1007	-0.14	0	-0.1494	-0.2230	-0.2288	0	1.22	1.10	0.1439
p-value	<.0001	<.0001	<.0001	<.0001	-	0.1339	0.0261	0.0243	-			
Total Accidents	-4.6764	0.6371	-0.15	-0.1473	0	-0.188	-0.207	-0.258	0	1.46	1.14	0.1635
p-value	<.0001	<.0001	<.0001	<.0001	-	0.0297	0.017	0.0032	-			

7 INVESTIGATION OF THE NEGATIVE BINOMIAL DISPERSION PARAMETER

Until recently negative binomial models has been developed assuming a fixed dispersion parameter for all the sites. But recent research has shown that the dispersion parameter can potentially depend on the exposure covariates such as traffic flow and segment length. (Miaou and Lord, 2003, Geedipally, Srinivas and Lord, 2008). This means that each site to which an accident prediction model applies should have a unique dispersion parameter.

This part of the research investigated the need for a separate dispersion parameter for each segment as a function of segment length and the effect of these varying dispersion parameters on model prediction and EB estimates. The Rural 2-Lane Kings Highways database was used for this study. The database contained 1355 segments of variable length. The effect was examined on two model forms "AADT-only" and the full model developed in Section 6.3 as shown below:

$$E(m) = aAADT^{b} \tag{7.1}$$

$$E(m) = aAADT^{b} \times (segment\ length)^{c} \times e^{(d_i NTC_i + e_j SWC_j)} \tag{7.2}$$

Where ,

$E(m)$ is the predicted no of accidents/km-year

AADT is the annual average daily traffic volume

NTC_i and SWC_j are categorical variables for variable Terrain and shoulder width.

a ,b, c, d_i and e_j are regression coefficients

The following model was calibrated to obtain separate dispersion parameter for each site:

$$k = 1/dL \tag{7.3}$$

where k is the dispersion parameter

L is the segment length in km and

d is the model coefficient estimated using the maximum likelihood method.

The model calibration was performed in SAS. An initial value for d was assumed in the beginning and through an iterative maximum likelihood process developed in SAS the final value was obtained.

An alternate functional form for obtaining the separate dispersion parameter for each site was also investigated. This investigation is described later in this Chapter.

To assess the performance of the models with fixed and varying dispersion parameter, the following five performance measures were examined (Washington et al. 2003):

- **Pearson's Product Moment Correlation Coefficients between Observed and Predicted Crash Frequencies**

  Pearson's product moment correlation coefficient, usually denoted by *r*, is a measure of the linear association between the two variables Y_1 and Y_2 that have been measured on interval or ratio scales. A different correlation coefficient is needed when one or more variables are ordinal. Pearson's product moment correlation coefficient is given as:

$$r = \frac{\sum (Y_{i1} - \overline{Y_1})(Y_{i2} - \overline{Y_2})}{\left[\sum (Y_{i1} - Y_1)^2 \sum (Y_{i2} - Y_2)^2\right]^{1/2}} \qquad (7.4)$$

  where $\overline{Y}$ = the mean of the Y_i observations.

  A model that predicts observed data perfectly will produce a straight line plot between observed (Y_1) and predicted values (Y_2), and will result in a correlation coefficient of exactly 1. Conversely, a linear correlation coefficient of 0 suggests a complete lack of a linear association between observed and predicted variables. The expectation during model validation is a high correlation coefficient. A low coefficient suggests that the model is not performing well and that variables influential in the calibration data are not as influential in the validation data. Random sampling error, which is expected, will not reduce the correlation coefficient significantly.

- **Mean Prediction Bias (MPB)**

  The MPB is the sum of predicted accident frequencies minus observed accident frequencies in the validation data set, divided by the number of validation data points. This statistic provides a measure of the magnitude and direction of the average model bias as compared to validation data. The smaller the average prediction bias, the better the model is at predicting observed data. The MPB can be positive or negative, and is given by:

$$MPB = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - Y_i \right)}{n} \qquad (7.5)$$

  Where n = validation data sample size; and
  $\hat{Y}_i$ = the fitted value for i^th observation of *Y*

  A positive MPB suggests that on average the model over predicts the observed validation data. Conversely, a negative value suggests systematic under prediction. The value of MPB provides the magnitude of the average bias.

- **Mean Absolute Deviation (MAD)**

MAD is the sum of the absolute value of predicted validation observations minus observed validation observations, divided by the number of validation observations. It differs from MPB in that positive and negative prediction errors will not cancel each other out. Unlike MPB, MAD can only be positive.

$$MAD = \frac{\sum_{i=1}^{n} \left| \hat{Y}_i - Y_i \right|}{n} \tag{7.6}$$

where n = validation data sample size, and

$\hat{Y}_i$ = the fitted value for i^{th} observation of Y

The MAD gives a measure of the average magnitude of variability of prediction. Smaller values are preferred to larger values.

- **Mean Squared Prediction Error (MSPE) and Mean Squared Error (MSE)**

MSPE is the sum of squared differences between observed and predicted crash frequencies, divided by sample size. MSPE is typically used to assess error associated with a validation or external data set. MSE is the sum of squared differences between observed and predicted crash frequencies, divided by the sample size minus the number of model parameters. MSE is typically a measure of model error associated with the calibration or estimation data, and so degrees of freedom are lost (p) as a result of producing $\hat{Y}_i$, the fitted value.

$$\text{MSE} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}{n_1 - p} \tag{7.7}$$

$$\text{MPSE} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}{n_2} \tag{7.8}$$

where

n_1 = estimation data sample size; and

n_2 = validation data sample size.

A comparison of MSPE and MSE reveals potential over fitting or under fitting of the models to the estimation data. An MSPE that is higher than MSE may indicate that the models may have been over fit to the estimation data, and that some of the observed relationships may have been spurious instead of real. This finding could also indicate that important variables were omitted from the model or the model was misspecified. Finally, data inconsistencies could cause a relatively high value of MSPE. Values of MSPE and MSE that are similar in magnitude indicate that validation data fit the model similar to the estimation data and that deterministic and stochastic components are stable across the comparison being made. Typically this is the desired result.

7.1 Investigation of the varying dispersion parameter with "AADT-only" model of Rural 2-Lane Kings Highways

Using the maximum likelihood method and the model form shown in Equation 7.3 the following relationship was obtained between dispersion parameter and segment length:

$$k = 1/1.75L \tag{7.9}$$

Where *k* is the dispersion parameter and *L* is the segment length in km. The relation indicates that *k* decreases as segment length increases.

Using Equation 7.9, a unique dispersion parameter was obtained for each site. Two different models were compared, one considering a fixed dispersion parameter for all the sites and the other a varying dispersion parameter for each site. The estimates of the model parameters obtained from the two approaches are shown in Table 7.1.1.

Table 7.1.1 Parameter estimates of the "AADT-only" Models with fixed and varying dispersion parameter

Model Form: $Accidents/km\text{-}Year = a \times AADT^{b}$

Models	Parameter Estimates		Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	Dispersion Parameter
	ln(a)	b			
Model 1:Model with Fixed Dispersion	-5.6773	0.6992	1.97	1.19	0.1831
p-value of parameter	<.0001	<.0001			
Model 2: Model with Varying dispersion	-5.3681	0.6549	1.31	1.1	-
p-value of parameter	<.0001	<.0001			

Parameter estimates of the two models are different. This indicates that the prediction would also be different for the two models, as illustrated in Figure 7.1.1.

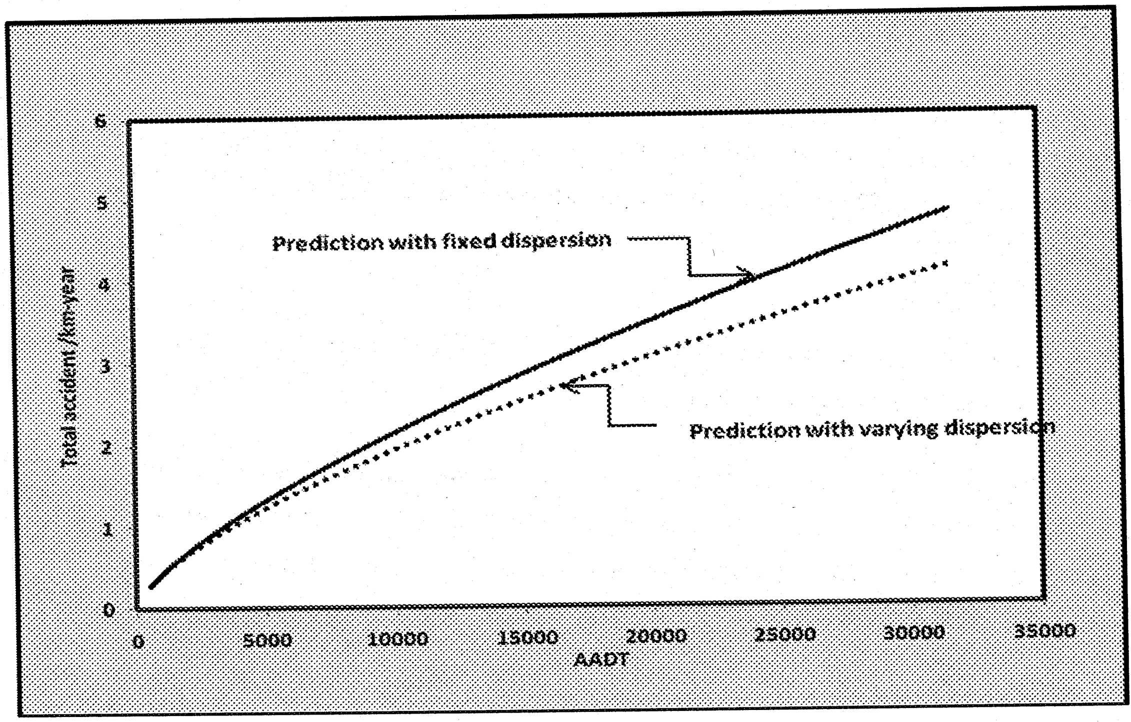


Figure 7.1.1 Model predictions with fixed and varying dispersion parameter for "AADT-only" model for Rural 2-Lane Kings Highways

The values of the goodness of fit criteria (Pearson chi-square, scaled deviance and p-value) as shown in Table 7.1.1 indicate that both the models are valid. But to find the best model that predicts the observed data most accurately, the additional five performance measures mentioned in the beginning of this Chapter were examined. The result of these performance assessments, which are presented in Table 7.1.2, shows the following:

- The values of MPB and MAD are less for Model 2 with varying dispersion. This indicates that this model provides less variation and less bias in the prediction.
- The value of MSE and MSPE are also less for Model 2 indicating that this model has less prediction error than Model 1.
- The difference in the Pearson product moment correlation coefficient(r) values between the two models is marginal. This indicates that both models provide the same amount of correlation between the observed and predicted values.

On the basis of the above results Model 2 appears to be the better model.

Table 7.1.2 Performance measures of Model1 (with fixed dispersion) and Model 2(with varying dispersion

Performance Measure	Values	
	Model 1-with fixed dispersion parameter	Model 2-with varying dispersion parameter
$MPB = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)}{n}$	1.89	-0.136
$MAD = \frac{\sum_{i=1}^{n}\left\lvert\hat{Y}_i - Y_i\right\rvert}{n}$	10.22	9.83
$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_1 - p}$	227.17	216.01
$MSPE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_2}$	226.84	215.69
$r = \frac{\sum(Y_{i1} - \overline{Y}_1)(Y_{i2} - \overline{Y}_2)}{\left[\sum(Y_{i1} - Y_1)^2 \sum(Y_{i2} - Y_2)^2\right]^{1/2}}$	0.862	0.861

As described earlier, the dispersion parameter is a key variable in estimating the weight in the EB (Empirical Bayes) estimation process. So it is most likely that a varying dispersion will affect the EB estimates. Figure 7.1.2 plots the EB estimates with fixed and varying dispersion parameter and the observed total accidents for the first twenty segments in the dataset. It is evident that for a given segment all three estimates are different and for shorter segments (0.1km, 0.6km, 0.9km, 1.6km and 2.2km) EB estimates with varying dispersion parameter are close to the observed accidents.

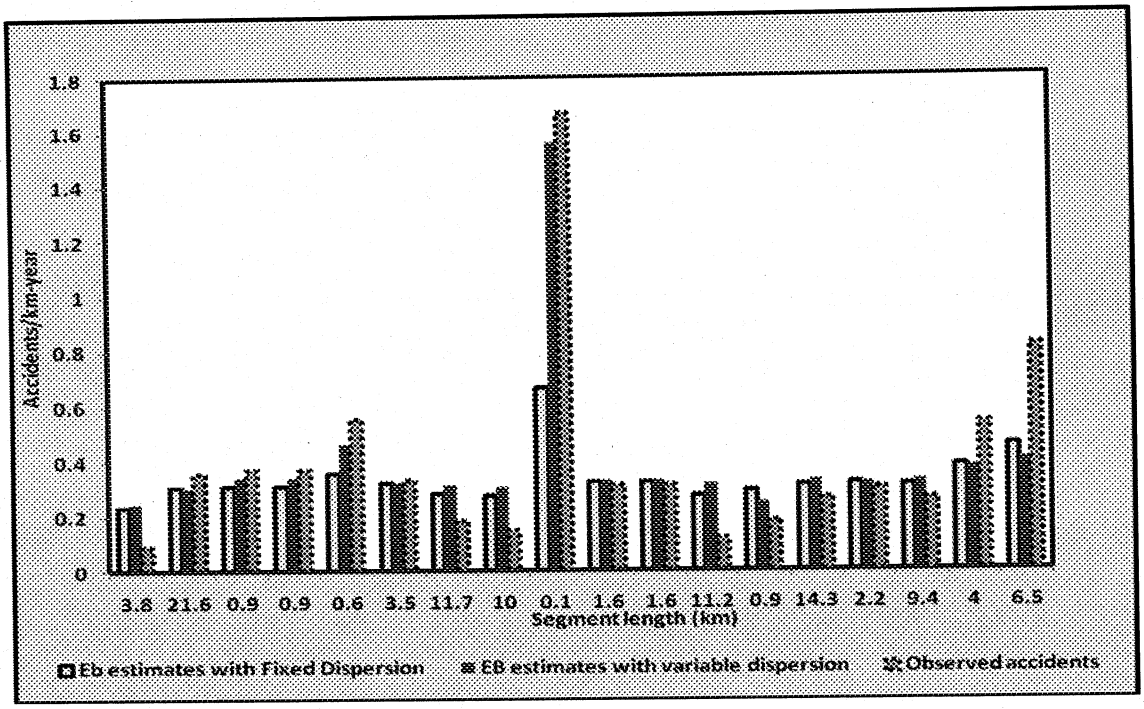


Figure 7.1.2 Plot of the EB estimates with fixed and varying dispersion parameter obtained from "AADT-only" model for Rural 2-Lane Kings Highways and the observed total accidents for a sample of 20 segments

Tables 7.1.3 shows the EB estimates/km with fixed and varying dispersion parameter and thirty top ranked sites based on a network screening with the two estimates. Table 7.1.4 shows a comparison of rankings by these two estimates.

Table 7.1.4 shows that although the EB estimates are different, 28 out of 30 top ranked sites by the EB estimate with fixed dispersion parameter were also in the top 30 ranked sites by EB estimates with a varying dispersion parameter. It is also evident from this table that there are differences in ranking by the twomethods.

Table 7.1.3 EB estimates/km with fixed and varying dispersion parameter (AADT-only model) and thirty top ranked sites on the basis of these estimates

Road segment LHRS Number	Segment length(km)	EB estimate with fixed dispersion	Rank	Road segment LHRS Number	Segment length(km)	EB estimate with varying dispersion	Rank
14946	0.2	32.75	1	14946	0.2	40.84	1
16480	0.5	20.09	2	24200	0.1	24.48	2
40610	0.3	17.07	3	27870	0.1	22.95	3
27870	0.1	16.06	4	16480	0.5	22.35	4
24200	0.1	15.65	5	29715	0.1	21.24	5
29715	0.1	13.89	6	40610	0.3	20.08	6
19320	0.1	13.59	7	19320	0.1	18.10	7
17320	0.1	12.88	8	17320	0.1	16.48	8
33690	0.4	11.55	9	33690	0.4	14.01	9
21020	0.3	11.36	10	21020	0.3	14.00	10
16160	0.1	10.45	11	16160	0.1	13.19	11
38610	0.3	9.83	12	38610	0.3	11.41	12
14946	1.7	8.92	13	48652	0.8	10.00	13
48652	0.8	8.90	14	21032	0.1	9.90	14
48660	2.2	8.19	15	23710	0.1	9.88	15
21032	0.1	8.07	16	27640	0.1	9.84	16
23710	0.1	7.62	17	14946	1.7	9.44	17
19450	0.3	7.24	18	31060	0.1	8.25	18
27640	0.1	7.09	19	48660	2.2	8.24	19
25657	1.2	6.78	20	19450	0.3	8.17	20
31060	0.1	6.67	21	14946	0.2	8.16	21
14946	0.2	6.61	22	28580	0.1	8.13	22
19450	0.9	5.93	23	25657	1.2	7.31	23
14946	1.9	5.57	24	29700	0.2	7.14	24
16150	0.1	5.56	25	16150	0.1	6.61	25
16150	0.2	5.56	26	29110	0.1	6.59	26
29110	0.1	5.23	27	11730	0.1	6.56	27
28580	0.1	5.01	28	27835	0.1	6.56	28
11730	0.1	4.80	29	16150	0.2	6.55	29
27835	0.1	4.79	30	17260	0.1	6.55	30

Table 7.1.4 Comparison of ranking of sites on the basis of EB estimates/km with fixed and varying dispersion parameter

Road segment LHRS Number	Segment length(km)	Ranking by the EB estimates with fixed dispersion	Ranking by the EB estimates with varying dispersion
14946	0.2	1	1
16480	0.5	2	4
40610	0.3	3	6
27870	0.1	4	3
24200	0.1	5	2
29715	0.1	6	5
19320	0.1	7	7
17320	0.1	8	8
33690	0.4	9	9
21020	0.3	10	10
16160	0.1	11	11
38610	0.3	12	12
14946	1.7	13	17
48652	0.8	14	13
48660	2.2	15	19
21032	0.1	16	14
23710	0.1	17	15
19450	0.3	18	20
27640	0.1	19	16
25657	1.2	20	23
31060	0.1	21	18
14946	0.2	22	21
19450	0.9	23	33
14946	1.9	24	34
16150	0.1	25	25
16150	0.2	26	29
29110	0.1	27	26
28580	0.1	28	22
11730	0.1	29	27
27835	0.1	30	28

Figure 7.1.3 plots the EB estimates with fixed and varying dispersion parameter against the observed total accidents. The linear trend line fitted to the data points of the EB estimates with varying dispersion parameter and observed total accidents shows a slope close to 1 than a similar line for the EB estimate on a fixed dispersion parameter. This indicates that the EB estimate with a varying dispersion parameter is more precise than EB estimates with a fixed dispersion parameter.

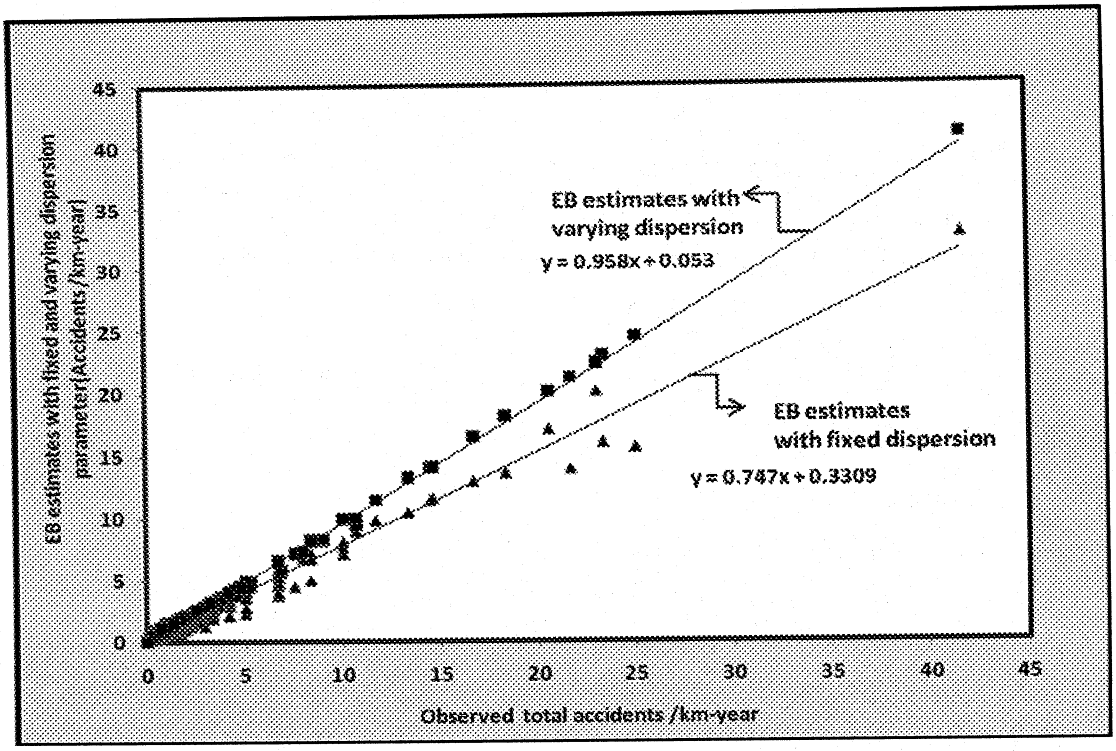


Figure 7.1.3 Plot of EB estimates with fixed and varying dispersion parameter obtained from “AADT-only” model against the observed total accidents

7.2 Investigation of the varying dispersion parameter with the full model of Rural 2-Lane Kings Highways

Recall that the full model had the form:

$$E(m) = aAADT^{b} \times (segment\ length)^{c} \times e^{(d_i NTC_i + e_j SWC_j)}$$

Where ,

$E(m)$ is the predicted no of accidents/km-year

AADT is the annual average daily traffic flow

NTC_i and SWC_j are categorical variables for variable terrain and shoulder width.

a ,b, c, d_i and e_j are regression coefficients

Using the maximum likelihood method the following relationship was obtained between the dispersion parameter and segment length:

$$k = 1/1.9L \qquad (7.10)$$

where *k* is the dispersion parameter and *L* is the segment length.

Using Equation 7.10 a unique dispersion parameter was obtained for each site. Two different models were compared as was done in Section 7.1, one considering a fixed dispersion parameter for all the sites and the other a varying dispersion parameter for each site. The estimates of the model parameters obtained from the two approaches are shown in Table 7.2.1

Table 7.2.1 Parameter estimates of the full model with fixed and varying dispersion parameter

Model Form: Accidents/km-year, $E(m) = aAADT^b \times (segment\ length)^c \times e^{(d_i NTC_i + e_j SWC_j)}$

Models	Parameter estimates									Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	Dispersion Parameter
				Terrain indicator variable		Shoulder width indicator variable						
				NTC_1	NTC_2	SWC_1 (<2.5)	SWC_2 (2.5-3m)	SWC_3 (3m)	SWC_4 (>3m)			
	ln(a)	b	c	d_1	d_2	e_1	e_2	e_3	e_4			
Model 1:Model with fixed dispersion	-4.676	0.6371	-0.15	-0.1473	0	-0.188	-0.207	-0.258	0	1.46	1.14	0.1635
p-value	<.0001	<.0001	<.0001	<.0001	-	0.0297	0.017	0.0032	-			
Model 2: Model with varying dispersion	-4.70	0.6322	-0.0991	-0.154	0	-0.207	-0.226	-0.242	0	1.16	1.08	-
p-value	<.0001	<.0001	<.0001	<.0001	-	0.0007	0.0002	0.0001	-			

The values of goodness of fit criteria (Pearson chi-square, scaled deviance and p-value) as shown in Table 7.2.1 indicate that both models are valid. To find the best model that predicts the observed data more accurately, the additional five performance measures as used for the "AADT-only" models were examined. The result of these performance assessments presented in Table 7.2.2 shows the following:

- The values of MPB and MAD are less for Model 2 with varying dispersion. This indicates that this model provides less variation and less bias in the prediction.
- The value of MSE and MSPE are also less for the full model with varying dispersion, indicating that this model has less prediction error than the full model with fixed dispersion.
- The higher value of the Pearson's product moment correlation coefficient (*r*) in case of full model with varying dispersion indicates that this model provides more correlation between the observed and predicted values, as is expected for a better model.

On the basis of the above results Model 2 appears to be the better model.

Table 7.2.2 Performance measures of Full model with fixed and varying dispersion parameter

Performance measures	Values	
	Full model with fixed dispersion parameter	Full model with varying dispersion parameter
$MPB = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)}{n}$	-0.66	-0.24
$MAD = \frac{\sum_{i=1}^{n}\left\lvert\hat{Y}_i - Y_i\right\rvert}{n}$	9.7	9.6
$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_1 - p}$	205.62	202.00
$MSPE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_2}$	205.32	201.70
$r = \frac{\sum (Y_{i1} - \overline{Y}_1)(Y_{i2} - \overline{Y}_2)}{\left[\sum (Y_{i1} - Y_1)^2 \sum (Y_{i2} - Y_2)^2\right]^{1/2}}$	0.868	0.869

Figure 7.2.1 plots the EB estimates with fixed and varying dispersion parameter and the observed total accidents for first twenty segments in the dataset. It is evident that for a given segment all three estimates are different and for shorter segments EB estimates with varying dispersion parameter are close to the observed accidents as was evident with the "AADT-only "models.

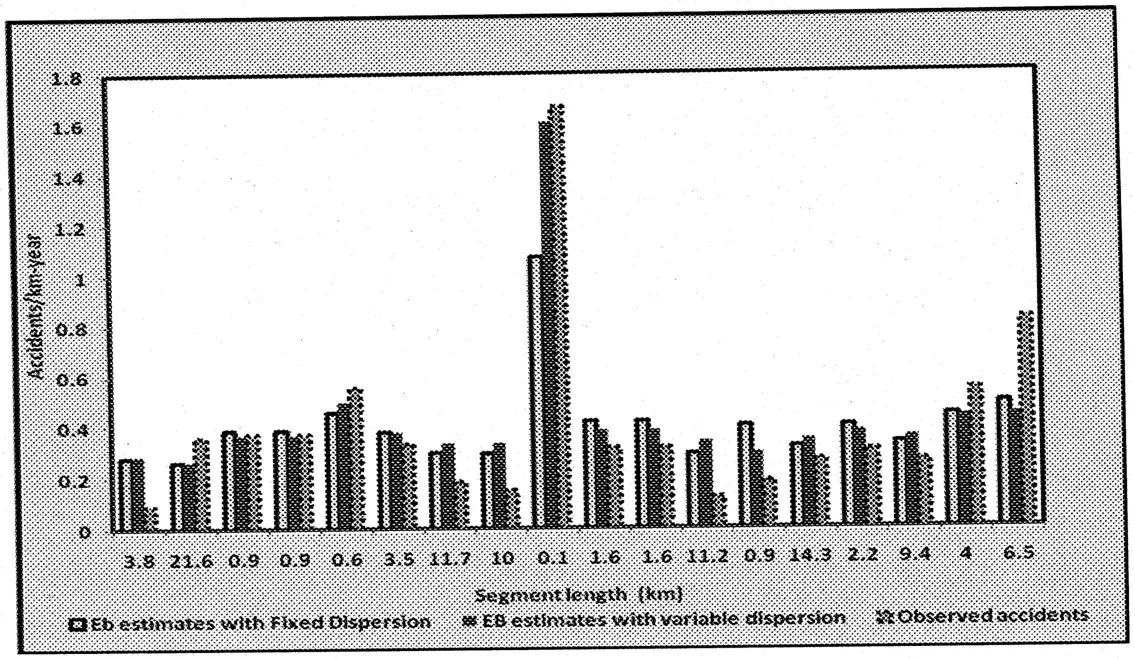


Figure 7.2.1 Plot of the EB estimates with fixed and varying dispersion parameter obtained from full model for Rural 2-Lane Kings Highways and the observed total accidents for a sample of 20 segments

Figure 7.2.2 plots the EB estimates with fixed and varying dispersion parameter against the observed total accidents. The linear trend line fitted to the data points of the EB estimates with varying dispersion parameter and observed total accidents shows a slope much closer to 1 than a similar line for the EB estimate based on a fixed dispersion parameter. This indicates that the EB estimate with a varying dispersion parameter is more precise compared to the EB estimates with a fixed dispersion parameter.

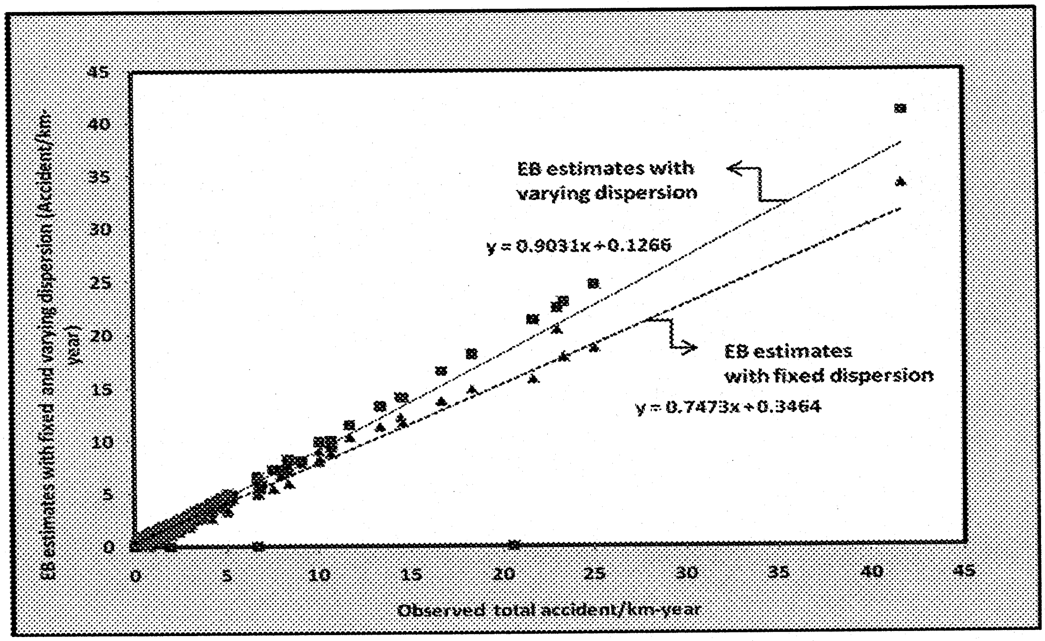


Figure 7.2.2 Plot of the EB estimates with fixed dispersion and varying dispersion parameter obtained from full model against the observed total accidents.

Tables 7.2.3 shows the EB estimates/km with fixed and varying dispersion parameter and thirty top ranked sites based on a network screening by the two estimates. Table 7.2.4 shows a comparison of rankings by these two estimates.

Table 7.2.4 shows that although the EB estimates are different , 29 out of 30 top ranked sites by EB estimate with fixed dispersion parameter were also in the top 30 ranked sites by EB estimates with a varying dispersion parameter. It is also evident from this table that there are differences in ranking by the two methods.

Table 7.2.3 EB estimates/km with fixed and varying dispersion parameter (full model) and ranking of sites on the basis of these estimates

Road segment LHRS Number	Segment length(km)	EB Estimate with fixed dispersion	Rank
14946	0.2	34.09	1
16480	0.5	20.46	2
24200	0.1	18.69	3
27870	0.1	17.88	4
29715	0.1	15.80	5
19320	0.1	14.88	6
17320	0.1	13.75	7
21020	0.3	12.20	8
33690	0.4	11.72	9
16160	0.1	11.34	10
38610	0.3	10.33	11
21032	0.1	8.98	12
48652	0.8	8.80	13
14946	1.7	8.73	14
23710	0.1	8.26	15
27640	0.1	7.98	16
48660	2.2	7.85	17
19450	0.3	7.62	18
31060	0.1	7.19	19
14946	0.2	7.05	20
25657	1.2	6.74	21
16150	0.1	6.17	22
19450	0.9	6.06	23
16150	0.2	6.04	24
28580	0.1	5.94	25
27835	0.1	5.58	26
11730	0.1	5.47	27
29700	0.2	5.45	28
14946	1.9	5.42	29
17260	0.1	5.40	30

Road segment LHRS Number	Segment length(km)	EB Estimate with varying dispersion	Rank
14946	0.2	40.96	1
24200	0.1	24.66	2
27870	0.1	23.04	3
16480	0.5	22.49	4
29715	0.1	21.34	5
19320	0.1	18.15	6
17320	0.1	16.52	7
21020	0.3	14.13	8
33690	0.4	14.06	9
16160	0.1	13.23	10
38610	0.3	11.49	11
48652	0.8	10.01	12
21032	0.1	9.95	13
23710	0.1	9.91	14
27640	0.1	9.89	15
14946	1.7	9.40	16
31060	0.1	8.27	17
19450	0.3	8.23	18
14946	0.2	8.20	19
28580	0.1	8.18	20
48660	2.2	8.13	21
25657	1.2	7.32	22
29700	0.2	7.26	23
16150	0.1	6.64	24
27835	0.1	6.61	25
16150	0.2	6.60	26
17260	0.1	6.60	27
11730	0.1	6.60	28
27925	0.1	6.57	29
34735	0.1	6.55	30

Table 7.2.4 Comparison of ranking of sites on the basis of EB estimates/km with fixed and varying dispersion parameter

Road Segment LHRS Number	Segment length(km)	Ranking by the EB estimates with fixed dispersion	Ranking by the EB estimates with varying dispersion
14946	0.2	1	1
16480	0.5	2	4
24200	0.1	3	2
27870	0.1	4	3
29715	0.1	5	5
19320	0.1	6	6
17320	0.1	7	7
21020	0.3	8	8
33690	0.4	9	9
16160	0.1	10	10
38610	0.3	11	11
21032	0.1	12	13
48652	0.8	13	12
14946	1.7	14	16
23710	0.1	15	14
27640	0.1	16	15
48660	2.2	17	21
19450	0.3	18	18
31060	0.1	19	17
14946	0.2	20	19
25657	1.2	21	22
16150	0.1	22	24
19450	0.9	23	31
16150	0.2	24	26
28580	0.1	25	20
27835	0.1	26	25
11730	0.1	27	28
29700	0.2	28	23
14946	1.9	29	32
17260	0.1	30	27

7.3 Comparison of "AADT-only" model and full model of Rural 2-Lane Kings Highways with fixed and varying dispersion parameter

In the preceding sections it was shown that the model with a varying dispersion parameter is better than the model with fixed dispersion parameter and also that the EB estimates obtained from the model with the varying dispersion is more precise. In this part of thesis a comparison was made between the "AADT-only" model and the full model developed considering fixed and varying dispersion parameters and the ranking of sites for safety improvement on the basis of EB estimates obtained from these models was done.

Table 7.3.1 shows the performance measures of the "AADT-only" and full models with fixed and varying dispersion parameters. Except for MPB, the full model with varying dispersion parameter shows better performance. So this model can be considered as the best model among the four. It is expected that the EB estimates calculated using the SPFs obtained from this model is the best estimate to use in ranking the sites for safety improvement.

Table 7.3.1 Performance measures of the "AADT-only" and full models obtained from fixed and varying dispersion parameter

Model Performance Measures	Values			
	"AADT-only" model with fixed dispersion	"AADT-only" model with varying dispersion	Full model with fixed dispersion	Full model with varying dispersion
MPB	1.89	-0.14	-0.66	-0.24
MAD	10.22	9.83	9.7	9.6
MSE	227.17	216.01	205.62	202.00
MSPE	226.84	215.69	205.32	201.70
r	0.862	0.861	0.868	0.869

Table 7.3.2 shows the comparison of the rankings obtained from the four EB estimates. Sites were first ranked on the basis of the best EB estimate (obtained from the full model with varying dispersion parameter) and then the top ranked sites were shown with their corresponding ranks by the other three methods. The average ranking is high for the other three methods compared to the best one, suggesting that it seems worthwhile to undertake the refinements of using a full model and a varying dispersion parameter compared to an "AADT-only" model and a fixed dispersion parameter. The improvement by using the full model seems to be greater than the improvement gained by using a variable dispersion parameter.

Table 7.3.2 Comparison of rankings by best EB estimate (obtained from the best SPF) and by other EB estimates

Road segment LHRS Number	Segment Length(km)	Ranking by the best EB estimates from full model with varying dispersion	Ranking by the EB estimates from full model with fixed dispersion	Ranking by the EB estimates from “AADT-only” model with fixed dispersion	Ranking by the EB estimates from “AADT-only’ model with varying dispersion
14946	0.2	1	1	1	1
24200	0.1	2	3	5	2
27870	0.1	3	4	4	3
16480	0.5	4	2	2	4
29715	0.1	5	5	6	5
19320	0.1	6	6	7	7
17320	0.1	7	7	8	8
21020	0.3	8	8	10	10
33690	0.4	9	9	9	9
16160	0.1	10	10	11	11
38610	0.3	11	11	12	12
48652	0.8	12	13	14	13
21032	0.1	13	12	16	14
23710	0.1	14	15	17	15
27640	0.1	15	16	19	16
14946	1.7	16	14	13	17
31060	0.1	17	19	21	18
19450	0.3	18	18	18	20
14946	0.2	19	20	22	21
28580	0.1	20	25	28	22
48660	2.2	21	17	15	19
25657	1.2	22	21	20	23
29700	0.2	23	28	34	24
16150	0.1	24	22	25	25
27835	0.1	25	26	30	28
Average Ranking		13	13.28	14.68	13.88

7.4 Comparison of the dispersion parameter obtained from two different functional forms

The relationship between dispersion parameter and segment length used in the above two approaches and by many researchers is of the following form:

$k = 1/(d \times L)$

where k is the dispersion parameter,

L is the segment length and

d is estimated by the maximum likelihood method and was found to be 1.75 for the "AADT-only" model and 1.9 for the full model.

In this part of thesis a more general form: $k = 1/(d_1 \times L^d)$ was investigated where d_1 and d were estimated by the maximum likelihood method to obtain the following form for the models with AADT as the only variable:

$k = 1/(2.3 \times L^{0.67})$

Figure 7.4.1 plots the two forms of the relationship between the dispersion parameter and segment length, while the "AADT-only" models calibrated with the varying dispersion parameter obtained from the two forms, along with goodness of fit measures, are presented in Table 7.4.1. The combined evidence from Figure 7.4.1 and Table 7.4.1 suggests that the more general form does not improve materially on the simple form used earlier in this thesis and by other researchers – at least for the "AADT-only" models.

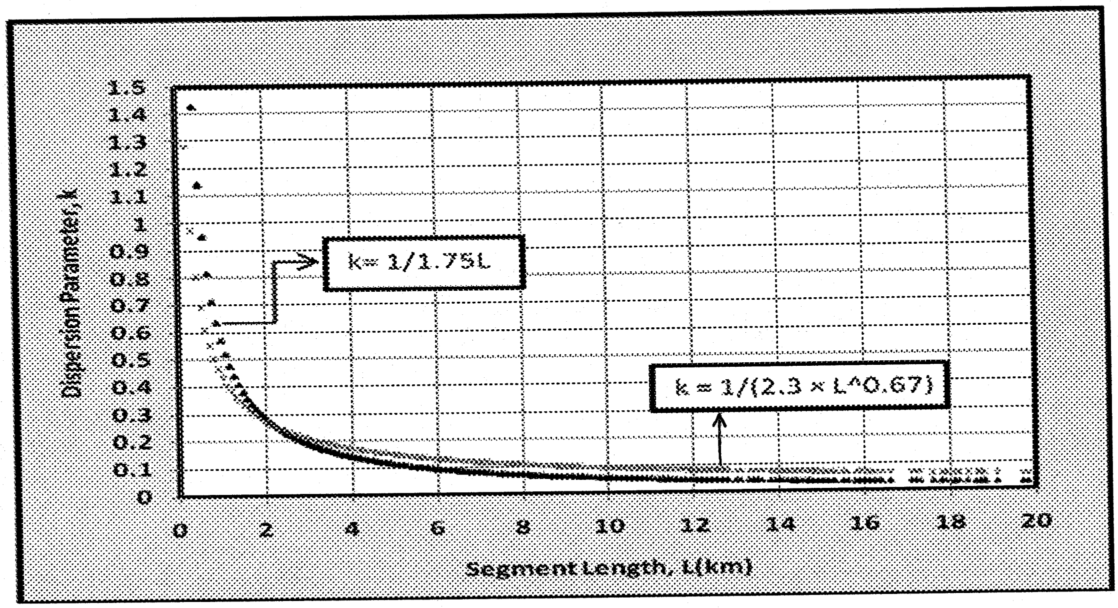


Figure 7.4.1 Plot of two relationship between dispersion paramter and segment length

Table 7.4.1 Parameter estimates of the “AADT-only” models with two functional forms of varying dispersion parameter

Model Form: *Accidents/km-Year = a ×AADT*^{b}

Models	Parameter Estimates		Pearson Chi-Square/Degree of freedom	Scaled Deviance/Degree of freedom	
	ln(a)	b			
Model with $k = 1/(2.3 \times L^{0.67})$	-5.4236	0.6636	1.35	0.98	
p-value of parameter	<.0001	<.0001			
Model with $k = 1/(1.75 \times L)$	-5.3681	0.6549	1.31	1.1	
p-value of parameter	<.0001	<.0001			

8 APPLICATION OF GENERALIZED ESTIMATING EQUATIONS (GEE) PROCEDURE TO CALIBRATE MODELS WITH TREND

In this part of study the Rural 2-Lane Kings Highways dataset was used. This contained 6 individual years AADT and crash count data. The procedure followed here is the one proposed by Lord and Persaud (2000).

To examine the year to year variation or trend in accident counts it is required to have data for as many years as possible and separated such that each year can be treated as a separate observation. However, disaggregating data by time period creates temporal correlation within the data set. Regression models are very useful to examine the relationship between collisions and a series of covariates. However, these models cannot take care of temporal correlation that occurs due to repeated observations on the same entity. The Generalized Estimating Equation (GEE) procedure is particularly useful for handling this correlation and calibrating models that incorporate trend, enabling the development of proper and unbiased models (Lord and Persaud, 2000). GEE procedure uses the temporal correlation directly in the estimation of the Model coefficients. The procedure can be used if the extent and the type of correlation are not known, since it uses a working correlation matrix to converge to the exact solution.

Three different models (Model 1, 2 and 3) were calibrated from the Rural 2-Lane Kings Highways dataset for total accidents. The following general model form was used:

$$E(Y) = L \times a \times AADT^{b} \tag{8.1}$$

Where $E(Y)$ is the expected number of accidents/year

L is the segment length and used as offset.

AADT is the annual average daily traffic volume

a and b are the coefficients to be estimated

For model 1 $E(Y)$ is the expected number of accidents between year 2000 to 2005 and AADT is the average traffic flow between the years 2000 to 2005.This model, and which does not include accident trend, was calibrated from aggregated data as before, using the regular Generalized Linear Modelling procedure in SAS.

For Models 2 and 3, trend was included by estimating models of the form:

$$E(Y) = \sum_{j=1}^{6} E(Y_j) = \sum_{j=1}^{6} L \times a_j \times AADT_j^{b} \tag{8.2}$$

Where $E(Y)$ is the expected number of accidents between years 2000 to 2005 and $E(Y_j)$ is the expected number of accidents in year j and $AADT_j$ is the traffic flow for year j. These models were calibrated from data disaggregated by year. Table 8.1 shows the statistics of these disaggregated data.

Table 8.1 Statistics of disaggregated data for calibrating GLM and GEE model

Year	Variables	Values		
		Minimum	Maximum	Sum
2000	Total accident	0	32	6518
	AADT	570	29800	-
2001	Total accident	0	49	6501
	AADT	560	30500	-
2002	Total accident	0	43	7285
	AADT	540	31200	-
2003	Total accident	0	35	7543
	AADT	520	31800	-
2004	Total accident	0	37	7433
	AADT	500	32800	-
2005	Total accident	0	36	7307
	AADT	480	33200	-

The regular generalized linear modelling approach (SAS PROC GENMOD) was used to calibrate Model 2, which includes trend but does not account for temporal correlation in the yearly data. The built-in GEE function in SAS was used to calibrate Model 3 to incorporate both accident trend and temporal correlation.

The coefficient a_j in Models 2 and 3 were estimated by defining each year as a categorical variable in the SAS GENMOD procedure.

The estimated model coefficients are presented in Table 8.2.

Table 8.2 Estimated coefficients of GLM and GEE Models (Models 1, 2 and 3)

		MODEL 1-GLM without trend		MODEL 2-GLM with trend but no temporal correlation		MODEL 3-GEE with trend & temporal correlation incorporated	
Year	Coefficients	Estimates	Standard Error	Estimates	Standard Error	Estimates	Standard Error
1	Ln(a)			-5.42	0.11	-5.42	0.198
2	Ln(a)			-5.43	0.11	-5.43	0.198
3	Ln(a)			-5.34	0.109	-5.34	0.187
4	Ln(a)			5.38	0.109	5.38	0.186
5	Ln(a)			-5.36	0.109	-5.36	0.186
6	Ln(a)			-5.53	0.08	-5.53	0.177
	Ln(a)	-5.677	0.149				
	b	0.6992	0.0177	0.6798	0.0098	0.6798	0.0215
	Dispersion Parameter	0.1831	0.0097	0.171	0.0065	0.171	0.0065
	Model Form	$E(Y) = L \times a \times AADT^b$		$E(Y) = \sum_{j=1}^{6} E(Y_j) = \sum_{j=1}^{6} L \times a_j \times AADT_j^b$		$E(Y) = \sum_{j=1}^{6} E(Y_j) = \sum_{j=1}^{6} L \times a_j \times AADT_j^b$	

Table 8.2 shows the following:

- The estimated coefficients and dispersion parameter are same for Model 2 and 3.
- The standard errors for the AADT exponent in Model 2 is much lower than in Model 1, and the dispersion parameter is also less for model 2. Both models are GLM models but when trend is incorporated, it is expected that a better model would be obtained, reflecting the reality that the relationship between accidents and AADT likely changes over time.
- When temporal correlation is incorporated the standard error for the AADT coefficient increased, as evident in Model 3 and which is close to those obtained in case of Model 1

The above findings indicate that the model with trend is perhaps better than the model without trend and that it seems necessary to account for temporal correlation in estimating models with trend.

Models 1 and 3 were further investigated to emphasize the importance of incorporating trend. First, the models were plotted in Figure 8.1. It is clear that there is difference in prediction between the two models. Model 1, the GLM model, predicts more accidents. This is also evident in Table 8.3, which shows that there is a great discrepancy between the sums of the observed and predicted accidents in the case of Model 1.

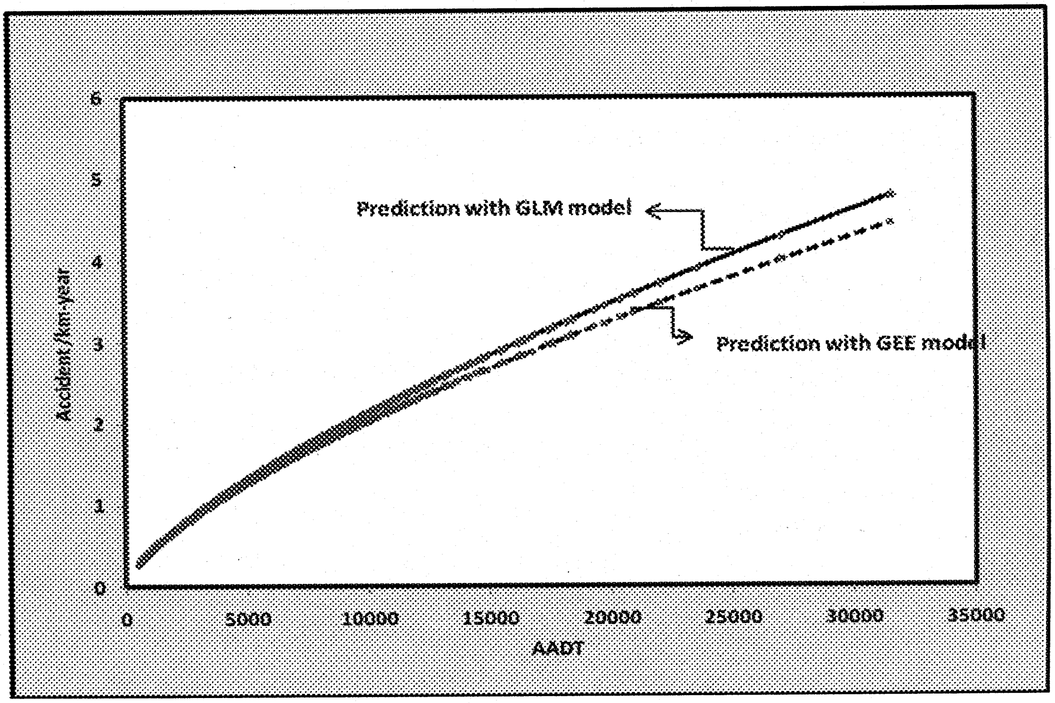


Figure 8.1 Predictions from SPFs calibrated with GLM (Model 1) and GEE (Model 3) models

Table 8.3 Accident statistics of the GLM and GEE models

	MODEL 1-GLM model	MODEL 2-GLM model with trend	MODEL 3-GEE model with trend
Sum of observed accidents	42585	42585	42585
Sum of fitted values	45153	43534	43534
Difference between Observed and fitted values	2568	949	949
Dispersion	0.1831	0.171	0.171

Next, Models 1 and 3 were further examined with the following performance measures as explained in the previous Chapter:

- Pearson's Product Moment Correlation Coefficients between Observed and Predicted Crash Frequencies
- Mean Prediction Bias (MPB) and
- Mean Absolute Deviation (MAD)
- Mean Squared Prediction Error(MSPE)
- Mean Squared Error (MSE)

For Model 1, the SAS gave the output as total of six years prediction but for Model 3 SAS gave the output of predicted values for each year for each segment. In order to find the six years predicted value for Model 3 in order to compare it to Model 1, the yearly values were added.

The values of the above measures which are presented in table 8.4 show the following results:

- Model 3 has lower values for MPB, MAD, MSE, MSPE compared to model 1 in particular for the MPB (Mean prediction bias). This suggests that Model 3 is better at predicting the observed data.
- The correlation coefficient of both the model is same showing same amount of correlation among observed and predicted values. This is not surprising since for Model 1, which uses aggregated data, temporal correlation is not an issue, while for Model 3, the GEE procedure accounts for temporal correlation.

Table 8.4 Performance measures of GEE Model and GLM Model

Performance Measures	Values	
	MODEL 1-GLM without trend	MODEL 3-GEE with trend & temporal correlation incorporated
$MPB = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)}{n}$	1.89	0.70
$MAD = \frac{\sum_{i=1}^{n}\lvert \hat{Y}_i - Y_i \rvert}{n}$	10.22	9.56
$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_1 - p}$	227.17	218.34
$MSPE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n_2}$	226.84	218.02
$r = \frac{\sum(Y_{i1} - \overline{Y}_1)(Y_{i2} - \overline{Y}_2)}{\left[\sum(Y_{i1} - Y_1)^2 \sum(Y_{i2} - Y_2)^2\right]^{1/2}}$	0.862	0.862

Based on the above results Model 3- the GEE Model with time trend that accounts for temporal correlation appears to be the best model.

9 CONCLUSIONS AND RECOMMENDATIONS

Ontario road segment data were used to investigate several issues regarding negative binomial accident prediction models. The conclusions below relate to these five issues.

1. *The difference between models calibrated for fixed 100m segments and models based on aggregated segments of variable lengths*

   It was observed that the dispersion parameter of the negative binomial distribution is large for 100m segments and small for longer segments. The estimates of the model coefficients were also found to vary with the use of 100m segment and longer segment in calibrating the models. Longer segments provide a smaller negative binomial dispersion parameter and so are preffered over shorter segments for calibrating the models.

2. *The effect of AADT on accident occurance for various road classes in ontario*

   The plot of AADT only accident prediction models for Rural 2-Lane Kings Highway showed a decrease in the slope of regression line as the flow increases but this was in contrast to the SPF plots for Urban 2-Lane Kings Highways, which showed increasing slopes. It is possible that the SPF plot for Urban 2-Lane Kings Highways is reflecting the increasing possibility of risk with higher AADT levels.

3. *Difference in predictions between models with AADT as the only variable and models with additional explanatory variables*

   For the models calibrated for Rural 2-Lane Kings Highways, the variables that had significant effects on accident occurrence, were the terrain, shoulder width and segment length. The availability of variables for this study was limited, and some important variables such as horizontal and vertical alignment, % heavy vehicle, number of access points, driveway density etc were not available for this study. So, further study can be done with these variables, to find their association with the accident frequency on this roadway class in Ontario.

4. *The best form for the dispersion parameter of accident prediction models in empirical Bayes estimation*

   The examination of the dispersion parameter showed that the assumption of constant dispersion parameter in SPF models may not be valid. Models calibrated with separate dispersion parameter for each site depending on the segment length performed better than the models calibrated considering fixed dispersion parameter for all sites. Empirical Bayes (EB) estimates of expected accident frequency were also found to be affected by whether a varying or fixed dispersion parameter is used.

It was observed that EB estimates with a varying dispersion parameter are more precise compared to the estimates with fixed dispersion parameter and the ranking of sites was also different. Both "AADT-only" model and full model of Rural 2-Lane Kings Highways were used in this investigation and four models were developed considering fixed and varying dispersion parameter. Among the four models, the full model with varying dispersion parameter was found to be the best model for prediction. In addition, ranking of sites for safety improvement based on the EB estimates obtained from the best model and the other three models were materially different. Further study on the dispersion parameter issue can be done considering it as a function of other covariates such as traffic volume, geometric characteristics etc.

5. *The importance of estimating models to reflect time trends in accident occurrence and accounting for temporal correlation in data in estimating those models*

For Rural 2-Lane Kings Highways, a model was calibrated with trend considering each year as a separate observation. The GEE (Generalized Estimating Equation) procedure was used to develop these models since it incorporates the temporal correlation that exist in repeated measurements. The model with trend that is calibrated considering each year of the study period as a separate observation was found to be better than the model without trend that was fitted to data aggregated over all years.

References:

Accident Information system MS Access Query User Guide, Ministry of Transportation of Ontario, Published by Ministry of Transportation Traffic Office and Systems Development office, February 2007.

El-Basyouny, K., and Sayed, T, *Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models,* Transportation Research Record 1950, Transportation Research Board 2006.

Geedipally, Srinivas, and Lord, D., *Effects of Varying Dispersion Parameter of Poisson-Gamma Models on Estimation of Confidence Intervals of Crash Prediction Models,* paper no.08-1563, presented at the 87th annual meeting of Transportation Research Board, Washington DC, 2008.

Hauer, E, Harwood D. W., and Council F., *Estimating Safety by the Empirical Bayes Method: A Tutorial*, Transportation Research Record 1784, Transportation Research Board 2002.

Hauer, E and Bamfo, J., *Two tools for finding what function links the dependent variable to the explanatory variables,* Published in Proceedings of ICTCT 97 Conference, November 5-7 1997, Lund, Sweden.

Hauer, E., *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety,* Oxford, England: Pergamon Press, Elsevier Science Ltd, 1997.

Hauer, E, Council, F. and Mohammedshah, Y., *Safety Models for Urban Four-Lane Undivided Road Segments*, Transportation Research Record 1897, Transportation Research Board 2004.

Lord, D. and Persaud B. N, *Accident prediction models with and without trend: Application of the generalized estimating equations procedure*, Transportation Research Record 1717, Transportation Research Board, 2000.

Lord, D., and J. A. Bonneson, *Calibration of Predictive Models for Estimating the Safety of Ramp Design Configurations*, Transportation Research Record 1908, pp. 88-95, Transportation Research Board 2005.

Miaou, S. P., and Lord, D., *Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes*, Transportation Research Record 1840, pp. 31-40, Transportation Research Board 2003.

Persaud B.N., Retting, R., Garder, P. and Lord, D. *Observational before-after study of U.S. roundabout conversions using the empirical Bayes method*, Transportation Research Record 1751, pp. 1-8, Transportation Research Board 2001.

Persaud B. N., *Estimating Accident Potential of Ontario Road Sections*, Transportation Research Record 1327, Transportation Research Board 1991.

Persaud B.N. *Statistical Methods in Highway Safety Analysis*, National Cooperative Highway Research Program Synthesis 294, Transportation Research Board, 2001.

Persaud, B.N. and Dzbik, Leszek, *Accident prediction models for Freeways,* Transportation Research Record 1401, Transportation Research Board 1993.

Persaud, B.N. and Mucsi, K., *Microscopic accident prediction models for two-lane rural roads,* Transportation Research Record 1485, pp 134-140, Transportation Research Board 1995.

Persaud, B.N., Lord, D., and Palminaso, J., *Issues of Calibration and Transferability in Developing Accident Prediction Models for Urban Intersections*, Transportation Research Record 1784, pp. 57-64, Transportation Research Board 2002.

SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites Task K White Paper for Module 1—Network Screening, for Federal Highway Administration GSA Contract No. GS-23F-0379K Task No. DTFH61-01 F-00096 December 2002.

SAS Institute Inc., SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999.

SAS Institute Inc., Version 9 of the SAS System for Windows. SAS Institute Inc., Cary, NC. 2002.

S. Washington, Persaud B., Lyon C., and Oh J., *Validation of Accident Models for Intersections*, Final report, Federal Highway Administration Contract DTFH61-00-C-00073, October 2003. Publication No. FHWA-RD-03-037.

Sawalha, Ziad and Sayed, T, *Statistical Issues in Traffic Accident Modeling*, 82nd annual Meeting CD-ROM. Transportation Research Board 2003.

Transport Canada 2005 annual report TP 13347 E (10-2006) Available from http://www.tc.gc.ca/roadsafety/vision/2005/pdf/rsv2005se.pdf [accessed in March 2008].

Vogt, Andrew and Bared, Joe, *Accident models for two-lane rural segments and intersections*, Transportation Research Record 1635, Transportation Research Board 1998.

Zhang, Y., Ye, Z. and Lord, D. *Estimating the Dispersion Parameter of the Negative Binomial Distribution for Analyzing Crash Data Using a Bootstrapped Maximum Likelihood Method*, Transportation Research Record 1635, Transportation Research Board 1998.

③ 12-36-67