NETWORK SCREENING METHODS TO IDENTIFY ROADWAY SITES FOR SAFETY INVESTIGATION: AN EXAMINATION OF SOME CRITICAL ISSUES

by

Brent Gotts, B.Eng., Ryerson, 2002

A thesis

Presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the Program of Civil Engineering

Toronto, Ontario, Canada, 2004

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Author's Declaration

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

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

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

Borrower's Page

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

Name	Address	Signature	Date

Network Screening Methods to Identify Roadway Sites for Safety Investigation: An Examination of Some Critical Issues

Master of Applied Science, 2005

By Brent Gotts

Department of Civil Engineering
Ryerson University

Abstract

Traffic accidents are responsible for about 3,000 deaths and $25 billion in economic losses annually in Canada. One way for transportation authorities to improve safety is to identify potentially hazardous roadway elements through network screening. The process of network screening is a low-cost statistical analysis of highway safety data, which yields a ranked list of sites to be investigated in detail.

Critical issues of two network screening methods are investigated in this thesis. The first method is a peak-searching algorithm for screening roadway segments, with attention focused on threshold values of a key user-selected variable, namely the coefficient of variation. The second method examined is a method of screening for high proportions of specific accident types. For this method, parameter estimation techniques are compared, and the effect of the 'critical proportion,' a key user-selected variable in the method, on site rankings is investigated.

In addition to the two network screening methods, an investigation is carried out into some aspects of safety performance functions calibrated using negative binomial regression. Specific attention is given to how the negative binomial dispersion parameter changes over the range of some independent variables.

Acknowledgements

I would like to thank Dr. Bhagwant Persaud, P.Eng., for his guidance and encouragement in the completion of this thesis. I would also like to thank Mr. Craig Lyon, P.Eng., for his continued support, and for writing some of the computer programs used in this research.

I would further like to thank Mr. Calvin Mollett, P.Eng., of the Regional Municipality of York, and Ms. Karen Richard of the Midwest Research Institute for their technical help.

Finally, I would like to thank my parents for their unwavering support in this endeavour.

Dedication

For my parents.

Table of Contents

Page

Page

List of Tables

Page

Page

List of Figures

List of Acronyms

AADT	Annual Average Daily Traffic
AMF	Accident Modification Factor
APM	Accident Prediction Model
AWSC	All-Way-Stop-Controlled
CV	Coefficient of Variance
DES	Detailed Engineering Study
EB	Empirical Bayes
EPDO	Equivalent Property Damage Only
FHWA	Federal Highway Administration (USA)
FI	Fatal/Injury
GLM	Generalized Linear Model
HSIS	Highway Safety Information System
IHSDM	Interactive Highway Safety Design Module
MBB	“Most Bang for the Buck”
ML	Maximum Likelihood
MLR	Multiple Linear Regression
MM	Method of Moments
NB	Negative Binomial
NFI	Non-Fatal Injury
NSC	National Safety Council (USA)
PDO	Property Damage Only
PSI	Potential for Safety Improvement
RENB	Random-Effect Negative Binomial
RIDE	Reduce Impaired Driving Everywhere
SPF	Safety Performance Function
TWSC	Two-Way-Stop-Controlled
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

1 Introduction

1.1 Road Safety

Canadians are among the most mobile people in the world, and this year 21 million licensed drivers in Canada will travel over 300 billion kilometers. Unfortunately, almost 3,000 people will die in traffic collisions, and over 200,000 will be injured. Transport Canada estimates that the economic cost of collisions is about $25 billion annually (*1*).

By any measure, these numbers are unacceptable. The economic burden as a result traffic accidents is staggering, let alone the less tangible costs of life and limb. Thus, the problem is clear: traffic accidents must be reduced in terms of both frequency and severity.

Unfortunately, there is no single solution to the problem, although progress has been made on many fronts. For example, vehicles now have better safety features, such as airbags; governments have implemented programs such as graduated licensing; police enforcement programs such as Reduce Impaired Driving Everywhere (RIDE) have been used to try to reduce the number of impaired-driving crashes; highway design standards are constantly updated with respect to safety; and many other avenues have been explored, all with the goal of making the transportation network safer.

1.2 Network Screening

A road network is made up of all the transportation facilities (freeways, highways, intersections, ramps, etc.) in a given jurisdiction. The jurisdiction may be a province, state, region, municipality, or any other clearly-defined area of interest. One of the tasks of transportation authorities in a given jurisdiction is to examine the road network for sites (individual road segments, intersections, etc.) that demonstrate a need for improved safety. This task is commonly referred to as *network screening*.

The result of network screening is generally a ranked list of sites, with those sites at the top of the list representing the most 'unsafe' sites. Some or all of these sites are then 'flagged' for a more detailed investigation, sometimes called a detailed engineering study (DES). The aim of a DES is to suggest feasible crash-reduction countermeasures for the flagged sites.

The criteria for what makes a given site unsafe vary by jurisdiction and by the method of network screening. In the most simple forms of network screening, sites are flagged based on observed accident counts, or – by taking into account some measure of exposure (e.g., traffic volume, etc.) – accident rates. These approaches, and their inherent limitations, are discussed in more detail in Chapter 2. More advanced methods employ an empirical Bayes (EB) procedure to estimate long-term expected accident frequencies, or proportions of specific accident types. Two such EB methods are described in Chapters 5 and 6.

In the past two decades, network screening has made tremendous advances; much of this is due, either directly or indirectly, to huge improvements in computing power, and the ubiquitous use of computers. In particular, EB methods offer the advantage of accounting for the random fluctuations inherent in annual accident counts. It is this randomness that renders the simpler screening techniques prone to errors. The consequences of making errors in network screening are that relatively safe sites may receive unnecessary remedial work, while unsafe sites may be ignored. These errors are costly both economically, and in terms of life and limb.

1.3 SafetyAnalyst

The United States' Federal Highway Administration (FHWA) has recognized the need for state-of-the-art highway safety practices to be applied on a wide scale. To that end, the FHWA, along with other agencies and institutions, is developing *SafetyAnalyst*, a set of software tools designed to improve the safety management programs of highway agencies.

SafetyAnalyst is currently in the development stage, but will ultimately provide highway safety practitioners with a number of tools, each representing a different "stage" of a safety management program. The tools include: the Network Screening Tool to identify sites that may have safety deficiencies; the Diagnosis Tool, used for site-specific diagnoses of safety problems; a Countermeasures Selection Tool for identifying specific remedial projects for a given site; an Economic Appraisal Tool for identifying cost-effective countermeasures at a given site; a Priority Ranking Tool, which ranks those sites that have been selected for the application of specific countermeasures based on the cost/benefit analysis performed using the Economic Appraisal Tool; and, finally, the Evaluation Tool, which is used to evaluate the effectiveness of highway safety projects by employing before-after studies.

This thesis is concerned with the first step in the process, the Network Screening Tool. Ultimately, *SafetyAnalyst* will offer several network screening options, all of which will employ EB techniques. Sites flagged in the course of network screening are often referred to as "sites with promise" for safety improvement. Sites with promise may be identified by one of the following criteria:

- sites with higher-than-expected accident frequencies which may indicate the presence of safety problems that are potentially correctable in a cost-effective manner;
- sites whose accident frequencies are not higher than expected, given the traffic volumes and other characteristics present at the site, but which nevertheless experience sufficient numbers of accidents that may potentially be improved in a cost-effective manner;
- sites with high accident severities, and;
- sites with high proportions of specific accident types.

The goal of *SafetyAnalyst* is to provide highway agencies with the means to apply sound statistical methods in the quest for improved highway safety, and thus yield more reliable results. This should ultimately result in both better use of highway safety resources and improved highway safety.

1.4 Objectives

Some of the screening methods that have been proposed for the Network Screening Tool have not been widely used, and there are questions about them that must be answered. Two of the screening methods are studied in detail: the peak-searching algorithm for screening roadway segments, and screening for high proportions of specific accident types.

Some screening methods make use of accident prediction models, or safety performance functions (SPFs), and the reliability of these models has an effect on the results of the screening. Of particular interest is the nature of the negative binomial dispersion parameter. Chapter 4 describes a brief investigation into how SPF parameters change over the range of some independent variables.

Chapter 5 describes the peak-searching algorithm for screening road segments that has been proposed for *SafetyAnalyst*. The focus of attention is how network screening rankings are affected by different values of the coefficient of variance.

In Chapter 6, the method of screening for high proportions of specific accident types is examined in detail. The method was developed by Heydecker and Wu (*2*), and may be included in the Network Screening Tool. Methods of parameter estimation are compared, and the effect of the 'critical proportion' on site rankings is investigated. Screening for high proportions of specific accident types is then compared with more traditional SPF-based screening methods.

Chapter 2 provides the theoretical background for the methods used, and Chapter 3 describes the data used for analysis. Conclusions and recommendations for future work are given in Chapter 7.

2 Background and Literature Review

2.1 What Is Safety?

If one is to improve the safety of a road network, one must first decide how 'safety' shall be defined and measured. Everyone is acquainted with the notion of safety as being freedom from harm or loss; however, safety may be measured both subjectively and objectively. The latter is clearly needed if engineering decisions are to be made with respect to safety.

Hauer (*3*) describes *road safety* (or 'road unsafety') as an objective measure reflected in the prevalence of traffic accidents and their harm. The subjective perception of safety is referred to as the feeling of *security*.

2.2 How Is Safety Measured?

If safety is to be measured quantitatively, a unit of measure must be adopted. While it is easy to decide on a unit of measure for, say, the length of a rod, road safety is less straightforward. First, we must specify the entities being evaluated; these could include road segments, intersections, ramps, or other highway facilities. In general, entities, or 'sites,' are compared with others of the same type. For example, the safety of a 2-lane rural highway is not normally compared with that of an urban freeway owing to the many differences between them. In metaphoric terms, we must compare apples to apples, not apples to oranges.

If the safety of an entity is manifest in accident occurrence, then the number of accidents occurring at a site is clearly of interest. First, we consider accident counts as a measure of safety. Figure 2.1 shows monthly accident counts for a two-lane rural highway in Washington state. The counts vary from month to month for no obvious reason. While it is conceivable that these fluctuations are due to some unmeasured phenomenon, it is far more likely that this is simply evidence of the random nature of accident counts. Indeed, other sites experience similar random behavior. Hauer (*3*) refers to these as random fluctuations that cannot be attributed to causes of interest. Given this randomness, accident counts can be an unreliable estimate of safety. The pitfalls of evaluating safety by the count of accidents are discussed in more detail in the next section.

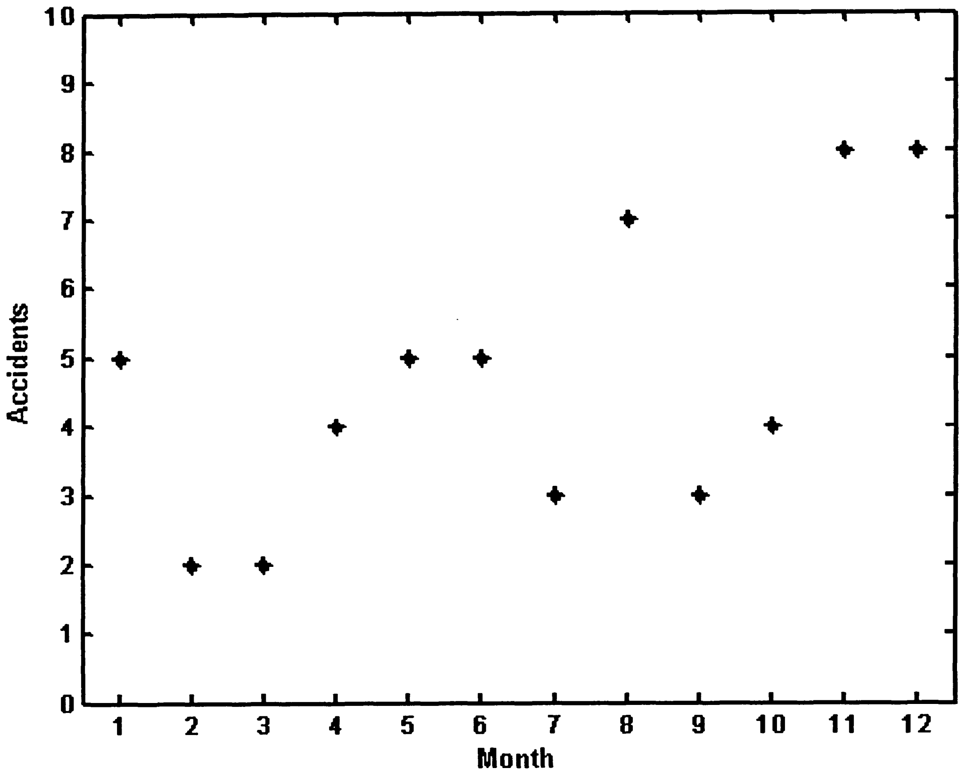


Fig. 2.1: 1996 monthly accident counts for a 2-lane rural highway in Washington State.

Hauer (*3*) and others suggest that safety must be defined as a stable property of an entity. Hauer (*3*) deems this the "underlying stable property that has the nature of a long-term average." This long-term average is unknown to us, but it can be estimated using statistical methods. This estimate yields the number of accidents *expected* to occur on an entity during a specified period.

More commonly, we are interested in an expected *accident frequency.* Accident frequency is simply a number of accidents per unit time, usually one year. Thus, if an intersection had 12 accidents in 3 years, we would say that it had an *observed accident frequency* of 4 accidents per year. We could also examine the *predicted accident frequency* based on mathematical models; these models are more commonly called *accident prediction models* (APMs), or *safety performance functions* (SPFs). The *expected accident frequency* (i.e., the estimated long-term average) is calculated by combining the information from a mathematical model and the observational data. The estimation of expected accident frequency usually involves applying empirical Bayes methods, which will be described later in this Chapter.

Another common measure of safety is *accident rate*. Accident rate is defined as:

$$Accident\ Rate \approx \frac{Accident\ Frequency}{(Exposure / Unit\ of\ Time)} \tag{2.1}$$

Exposure is generally derived from traffic flow, and generally involves estimates of *annual average daily traffic* (AADT). For example, accident rate may be measured as accidents per million vehicle-kilometers. So, for a 1.5 km road segment with an AADT of 25,000 vehicles per day (vpd) and an accident frequency of 5 accidents per year, the accident rate is calculated as follows:

$$Accident\ Rate = \frac{5\,acc / yr}{(25000 vpd)(365 d / yr)(1.5 km)}$$

$$= 0.37 acc / million\ veh \cdot km$$

Other measures of accident rate may also be used. For example, accident rates for intersections are often defined as accidents per million entering vehicles. AADT is normally measured via automatic counters, usually deployed for a few days or weeks at a given site.

Many have criticized the use of accident rates as a measure of safety, as this amounts to assuming a linear relationship between accidents and traffic flow; see, for example, Persaud (*5*). It has been shown that this relationship is generally non-linear, and thus accident rate is not a good measure of safety. Figure 2.2 depicts an accident prediction model of off-the-road accidents on 2-lane highways in Georgia. The model is clearly non-linear, and is of the form:

$$Accidents / mile / yr = \alpha \cdot AADT^{\beta} \tag{2.2}$$

where α=0.0042, and β=0.51 are regression coefficients calibrated from historical data.

At point A, the accident rate is calculated as 0.36 accidents/million veh·km, while at point B, the rate is 0.16 accident/million veh·km. Thus, one may be tempted to conclude that a site corresponding to point B is twice as safe as at point A; however, since twice as many accidents per year are experienced at point B, one may argue that point A is in fact the safer of the two. To conclude the argument, if accident rate is to be selected as the measure of safety, then an accompanying assumption of a linear relationship between traffic and accidents must be made. As this is not generally the case, accident rate should not be used as the primary measure of safety. It should be noted that accident rates should not be entirely ignored. To illustrate, consider a site that experiences a high accident rate but a low accident frequency; if a bona fide safety problem exists, such a site may not be identified by frequency-based screening methods, and it is not fair to expose even a few people to undue risk.

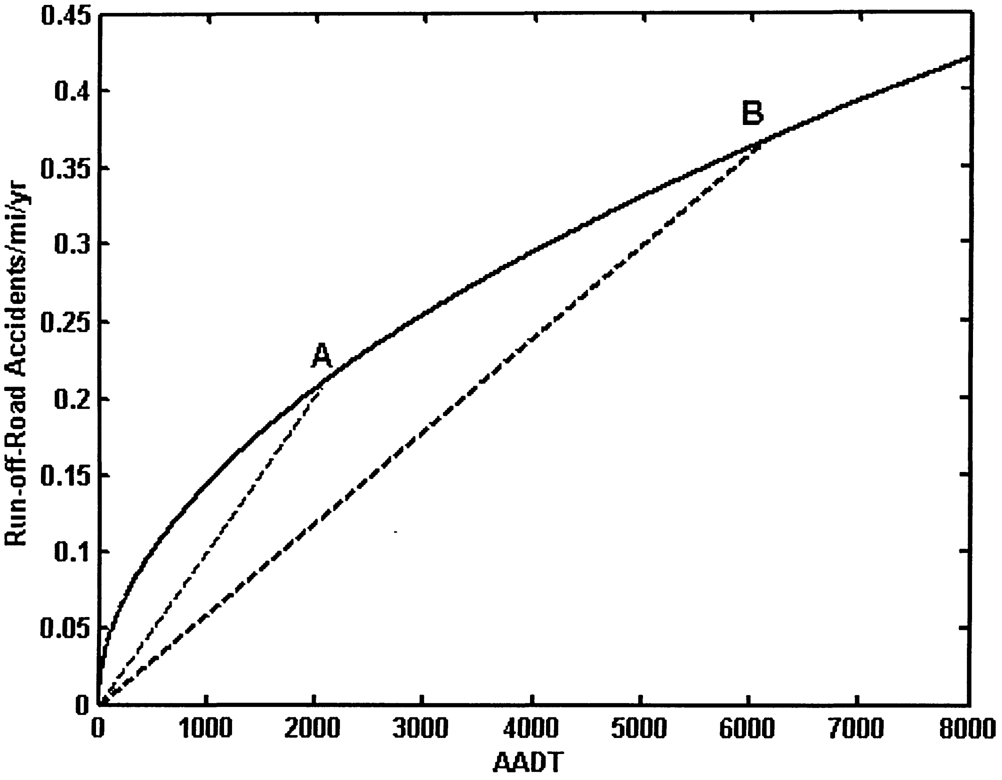


Fig. 2.2: Accident prediction model for off-the-road accidents on 2-lane rural highways in Georgia.

2.3 The Regression-to the-Mean Phenomenon

A common method of identifying sites for remedial safety work is to consider accident counts over a period of time (e.g. 3 years). In practice, sites are grouped according to some set of criteria (e.g. type of site, traffic volume, etc.), and those sites that experienced an accident count greater than some specified limit are 'flagged' for safety investigation. For example, a two-way, stop-controlled (TWSC) intersection might be flagged if there were 12 or more accidents over period of 3 years.

The advantages of a screening process based on accident counts are that it is conceptually very simple, and the data requirements are minimal. The major problem with using accident counts as a measure of safety is that unusually high accident counts are likely to decrease in the future, even if no remedial safety accident is taken. This is a phenomenon know as regression to the mean, and it has been well-documented by Persaud (*5*), for example, and others. Thus, when sites are subjected to a safety treatment based on high accident counts, the safety effect of the improvement is likely to be overstated.

Regression-to-the-mean bias, or 'selection bias,' is a critical flaw of count-based screening methods, and so an improved method of identifying sites with promise is needed.

2.4 Empirical Bayes Analysis

The more recently proposed methods for identifying sites with promise are mostly based on the empirical Bayes (EB) technique. Bayesian statistical methods permit the combination of observations from stochastic processes (i.e.: accident counts) with information from other sources, such as mathematical models for accident prediction. The resulting estimates are a weighted combination of the two sources of information, and the weights are calculated in such a way as to implicitly account for the amount of information in each source. Thus, the EB estimates are as accurate as possible given the two sources of information. See Higle and Witkowski (*4*), Persaud (*5*), Hauer (*3,6)*, and others for details.

In the road safety context, EB methods combine site-specific accident observations with information from other sites of the same kind. The 'information from other sites' is generally in the form of a mathematical model.

The mathematical models, or accident prediction models, are calibrated by an appropriate regression technique. In general, where the parameters of an accident prediction model has a relatively high variance, it will have less influence (i.e., a smaller EB weight) than one with a relatively low variance. At the same time, sites with high accident counts (and thus low variance) will be given a greater weight than sites with low accident counts.

The main feature EB methods is that the random fluctuations in accident counts are 'smoothed out' by defining the safety of a site as its expected long term average, rather than a short-term accident count. Thus EB methods control for regression-to-the mean bias, marking a major improvement over 'traditional' techniques.

Accident prediction models play a central role in the application of the EB procedure, and these models are the subject of much current research. Details of some of the more important accident prediction models are given in the next section.

2.5 Accident Prediction Models

2.5.1 Introduction

One of the important tasks of the road safety practitioner is the development of safety performance functions (SPFs), also called accident prediction models. Generally speaking, SPFs are found by performing statistical regression on accident count data and other relevant information. The result is a mathematical function that returns the predicted accident frequency of an entity given one or more independent variables. Most models have the following general form:

$$\kappa = f(\mathbf{X}, \boldsymbol{\beta}^{T}) \tag{2.3}$$

$$\kappa = f(X_1, X_2, \ldots, X_m, \quad \beta_1, \beta_2, \ldots, \beta_m) \tag{2.4}$$

where κ is the predicted accident frequency, **X** is a vector of characteristic traits of the entity, **β** is a parameter vector, and *f()* is some function.

As discussed above, SPFs are generally developed for different types of entities. For example, if one wished to predict accidents within a given area (e.g., city, county, etc.), separate models would be created for four-legged signalized intersections, three-legged signalized intersections, all-way stop-controlled (AWSC) intersections, urban freeways, 2-lane rural highways, and so on. Because it is impossible to find two sites exactly alike, it is necessary to group them in such a way that they are both logically comparable ("apples to apples") and statistically valid. The latter generally depends on the sample size, which in this case is the number of sites used in the model; the greater the number of sites used, the better the model will be.

At a minimum, traffic volume (AADT) is required to develop a practicable model; however, it is very common to see models including other traits, such as geometric features, area features, driver characteristics, many others. The number of independent variables that may be used is limited only by data availability, and the statistical significance of those variables chosen for the model.

Many different model types have been developed over the past few decades, with varying degrees of success. Below are descriptions of several modelling methods.

2.5.2 Linear Least-Squares Models

Early attempts to predict accidents employed multiple linear regression (MLR) models. MLR techniques use linear least-squares equations to fit a model of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad , \qquad (2.5)$$

where $\hat{y}$ is the dependent (response) variable, x_i are the independent (predictor) variables, m is the number of independent variables, and β_i are regression parameters. In general, traffic volume would be the most important predictor variable. Thus, a simple linear model (i.e.: m=2), with accidents as a function of AADT would be of the form:

$$\hat{\kappa} = \beta_0 + \beta_1 \cdot AADT \quad , \qquad (2.6)$$

where $\hat{\kappa}$ is the predicted accident frequency per mile per year. Here, β_0 is the y-intercept of a straight line, and β_1 is the slope. The addition of segment length allows accidents to be predicted on a per-kilometer or per-mile basis. This equation appears to have a minor logical difficulty: if β_0 is nonzero, then the number of accidents predicted for a site with no traffic will also be nonzero; however, as models only pertain over the range of independent variables used for the calibration, this is not problem.

If we examine a multiple linear model (m>1), we can I mmediately see a flaw. Consider the multiple linear model with AADT and shoulder width as the independent variables:

$$\hat{\kappa} = \beta_0 + \beta_1 \cdot AADT + \beta_2 \cdot (Shoulder\ Width) \ . \qquad (2.7)$$

Hauer (7) uses a similar model to illustrate another problem with MLR models. He describes the above equation as being additive in nature; thus, a change in shoulder width would have the same effect on an entity's accident frequency whether the AADT was 100 vpd or 100,000 vpd. Hauer (*7*) argues that this can surely not be true, thus showing another limitation of MLR models.

Maher and Summersgill (*8*) and Persaud et al. (*9*) have discussed problems with the early application of MLR models, namely the assumptions of normally distributed errors and homoscedasticity. Generally, the "errors" are represented by the distribution of accident counts; however, it accident counts are not normally distributed, given the discrete nature of count data.

A further assumption, the requirement of a linear relationship, also has a serious practical problem: the model predicts that as traffic volume increases, accident rate increases proportionally. In reality, as traffic volumes approach or exceed a highway facility's capacity, operating speeds decrease substantially; thus a lower accident rate is generally observed with high-volume sites as compared with low-volume sites.

Some of the advantages of using linear models are that they are functionally and computationally simple; however, great leaps in computing power, and commercially-available statistics software, such as SAS® and GLIM®, over the last 20 years have made more complex models as easy to calibrate as linear models.

2.5.3 Non-Linear Least-Squares Models

An extension of linear least-squares modelling is non-linear least-squares modelling, which allows a larger and more general class of functions to be used. Without the linearity constraint, there are very few limitations in the way parameters may be used in the functional part of the model.

As with linear least-squares models, homoscedasticity and normally-distributed errors are assumed. Thus, while the problems of a linear function are resolved, the other difficulties remain.

In order to overcome these difficulties, the next major step in improving accident prediction models involved the use of generalized linear models.

2.5.4 Generalized Linear Models

2.5.4.1 Introduction

Generalized linear models (GLMs) were developed in the 1970s as a method of modelling data where the distribution of the response variable (the distribution of accident counts, in this case) is a member of the exponential family (*10*). Exponential distributions include Poisson, binomial, gamma, negative binomial, and others.

This section will describe accident prediction models based on Poisson and negative binomial distributions. The random-effect negative binomial model will also be discussed.

2.5.4.2 Poisson Models

The first attempts to use GLMs involved Poisson regression. The Poisson distribution is a discrete distribution that is well suited for modelling count data, particularly where the data contains a large number of zeros. This applies to road safety modelling, as accident counts are discrete and often have many zeros (i.e.: it is common for an entity to have zero accidents over a given period).

Maher and Summersgill (*8*), Miaou and Lum (*11*), and others showed that by assuming accident counts followed a Poisson distribution, the resulting accident prediction models were a significant improvement over the standard regression models used previously.

One of the assumptions of the Poisson distribution is that the variance is equal to the mean. This implies that the model variance is explained entirely by the chosen independent variables; however, accident prediction models are usually found to have a variance greater than the mean (*8*, and others). This additional variance is termed *overdispersion*. Maher and Summersgill (*8*) suggest several possible causes of overdispersion: there are unobserved explanatory variables which effectively add to the random error; there are errors in the explanatory variables; or, the model may be mis-specified.

Another advantage of Poisson models is that they can be effectively used in empirical Bayes (EB) analysis. EB analysis combines model predictions with observed accident counts to improve prediction. EB techniques were discussed in Section 2.4.

Most recently, Qin et al. (*12*) used zero-inflated Poisson (ZIP) models for predicting accidents on two-lane highway segments in Michigan. They argue that the over-representation of zero-crash observations in the data may suggest overdispersion in the data even though the assumption of a Poisson distribution is actually correct.

The ZIP model accounts for the large probability "spike" at zero by combining two probabilities; the first is the probability that a site will have zero accidents, the second that the site will have a Poisson-distributed accident count. The total probability of observing zero accidents is then found by mixing these probabilities together. Thus, it is assumed that there is more than one underlying process that may be influencing accident frequency.

Qin et al. (*12*) argue that the ZIP model is an improvement over both pure Poisson models and negative binomial models; however, Lord et al. (*13*) refute this claim and give alternative reasons for the preponderance of zeros in crash data. Negative binomial models will be discussed in the next section.

2.5.4.3 Negative Binomial Models

Another way of dealing with overdispersion is to assume that accident counts come from a negative binomial distribution. Negative binomial (NB) accident prediction models have been used extensively in the past few decades; see, for example, Hauer (*3*), Persaud et al. (*9*), Vogt and Bared (*14*), and many others.

The standard negative binomial model assumes that the observed accidents (accident counts), y_i, are distributed about a true mean, $\mu=\lambda T$, where λ is the true accident rate per year, and T is the length of the observation period. A second assumption is that λ is gamma-distributed with a mean of $\hat{\lambda}$, and a shape of α. Here, $\hat{\lambda}$ is the estimate based on known values of the explanatory variables, as shown here:

$$\hat{\lambda} = \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}) \quad , \qquad (2.8)$$

where x is a vector of independent variables (e.g. AADT, shoulder width, etc.) with the first term equal to 1, and $\hat{\boldsymbol{\beta}}$ is a vector of parameters estimated by the fitting process (*8*).

The NB model has the same functional form used in the Poisson model. Negative binomial parameters may be found using weighted least squares if α is known; otherwise, α, $\hat{\boldsymbol{\beta}}$, and the negative binomial dispersion parameter, k, may be estimated by the method of maximum likelihood.

Mayer and Summersgill (*8*) showed that, for a given data set, the NB regression parameters were very similar to those estimated by assuming a Poisson model. Since the functional form of the two models is the same, the accident frequency predicted by the models is almost identical. The difference between the Poisson and NB models is in the calculation of variance; the negative binomial model adds a quadratic term to the variance, such that the variance is calculated as:

$$Var(y) = \mu + k\mu^2 \quad , \tag{2.9}$$

where $E(y)=\mu$.

From Equation 2.9, we can see that the negative binomial distribution will approach the Poisson distribution as $k \rightarrow 0$. Negative binomial models will exhibit greater variance than Poisson models, and the models will thus carry less weight in the context of the EB procedure; however, because the NB model more accurately reflects the observed variance, NB models are generally more appropriate than Poisson models.

Much research is being conducted on negative binomial accident prediction models. Recently, zero-inflated negative binomial (ZINB) models have been used to try to improve prediction; see Lord, et al. (*13*) for details. Chin and Quddus (*15*) have also proposed a random-effect negative binomial (RENB) model.

2.5.5 Nonparametric Models

The accident prediction models described thus far are all parametric; that is, each model is function of the chosen independent variables and the regression parameters. The regression parameters are constant over the range of independent variables; thus, a single process is assumed to govern all sites to which the model is applied. The problem with this assumption is that there may be many different processes at work; for example, sites with very high AADT may 'behave' differently than sites with low AADT, with respect to safety performance.

One way to address this issue is to employ a nonparametric model. Nonparametric methods trace the dependent variable (i.e., accident frequency) as a response to one or more independent variables (i.e., AADT) without specifying a functional relationship in advance. See Kononov and Allery (*16*) for details.

Nonparametric methods are relatively new, and have been facilitated by rapid advances in computing power. While many people prefer the simplicity and relative 'elegance' of mathematical models, nonparametric models show much promise, and should be the subject of further research.

2.6 Potential for Safety Improvement

An alternative to screening based on EB-adjusted expected accident frequency is to consider the 'potential for safety improvement' (PSI), or 'excess accident frequency.' In its simplest form, PSI is defined as the difference between the expected accident frequency, X, and SPF-predicted accident frequency, κ:

$$PSI = X - \kappa \quad , \tag{2.10}$$

where PSI is measured in accidents/mi/yr, or accidents/yr, whichever is appropriate.

The idea was proposed by Persaud et al. (*17*), Persaud et al. (*18*), and others. The idea is based on comparing the expected accident frequency of a site to what is 'normal' for other sites. This measure of safety is attractive, as it aims to quantify the accident frequency that is 'correctable.' To illustrate, suppose an intersection had an EB-adjusted expected accident frequency of 10 accidents/yr, but the model prediction for the site was 4 accidents/yr. The site's PSI would be 10-4=6 accidents/yr. Since the site experiences an 'excess' of 6 accidents/yr over what a 'normal' site would, then it may be reasonable to believe that these 6 accidents can be reduced to normal level of safety.

It can be argued that a similar site experiencing the normal accident frequency of 4 accidents/yr would pose a more difficult challenge for remedial safety work, as this would amount to making the site 'safer than normal.' A negative value of PSI indicates that the site's safety performance is better than normal for similar sites.

3 General Features of the Data Used in the Research

3.1 HSIS Data

All data used in this research was taken from the Highway Safety Information System (HSIS) database. The HSIS database, developed by the FHWA, is a collection of crash, traffic, roadway inventory and geometric data compiled from a select group of states. The data are processed into a common computer format and made available for research purposes.

Three main sets of data were used in this thesis, one from Washington, two from California. The Washington were for 2-lane rural highways (*19*); the California data were for rural, 4-leg two-way-stop-controlled (TWSC) intersections, and rural, 4-leg signalized intersections (*20*). The Washington data were for 4 years, 1993-1996, and the California data were for 5 years, 1997-2001. Relevant statistics for the datasets used for analysis in this thesis are provided in the appropriate sections.

For each dataset, all milepost data were reported to a precision of 0.01mi (16.1m).

The HSIS data were received is SAS format, and that program was used to manipulate all data prior to analysis.

3.2 Accident Data

For the Washington road segment data, intersection and intersection-related accidents were excluded from all analyses. For the California intersection data, only accidents occurring within 75m (0.0466mi) are included for analysis; all others are omitted.

For the Washington data, the reported injury severities were fatal, non-fatal injury (NFI), and property damage only (PDO). Fatal/injury (FI) accidents were computed as the sum of fatal and NFI accidents.

The California accident data were reported using one of five severities: fatal (K), severe injury (A), other visible injury (B), complaint of pain (C), or property damage only (O). The severity levels correspond to the KABCO injury scale (*21*), which will be described in more detail in Chapter 6. FI accidents were calculated as the sum of K, A and B accidents.

3.3 Traffic Data

Traffic volume data were reported as average annual daily traffic (AADT). Where a site had no reported AADT data, that site was excluded from analysis. Where the site had at least one year of AADT data, missing values were dealt with in accordance with the method used in the Interactive Highway Safety Design Module (*22*), which is as follows:

- If AADT data were available for only a single year, that same value was assumed to apply to all years of the analysis period.
- If two or more years of AADT data were available, the AADTs for intervening years were computed by linear interpolation.
- AADTs for years before the first year for which data were available were assumed to be equal to the AADT for that first year.
- AADT for years after the last year for which data were available were assumed to be equal to the last year.

3.4 Units

All HSIS data were presented in U.S. customary units, and all analyses were performed in those units. All methods described in this thesis are equally effective when metric units are specified. The only non-metric units used are miles (mi), which may be converted to kilometres by 1mi=1.61km.

4 An Investigation of the Overdispersion Parameter of Safety Performance Functions Used in Network Screening

4.1 Background

Safety performance functions, also called accident prediction models, are mathematical models that have accident frequency as the dependent variable, and one or more traits of the sites under investigation as the independent variables. The list of independent variables usually includes traffic volume (i.e., AADT) at a minimum, and often includes measured geometric characteristics (e.g., lane width, speed limit, etc.) and other relevant traits. In theory, any measurable characteristic may be included in the model provided that it is shown to have a statistically significant effect on the dependent variable (i.e., accident frequency).

Safety performance functions (SPFs) have two important applications. First, they may be used to predict the safety performance (i.e., accident frequency) of a proposed transportation facility, or to predict the change in safety performance where modifications are proposed to existing facilities. This can help designers to compare the predicted safety performance of alternative designs; the Interactive Highway Safety Design Module (IHSDM) makes use of SPFs for this purpose (*22*).

Second, SPFs are commonly used as part of empirical Bayes (EB) analyses to identify "sites with promise" for safety improvement. In this case, the model predictions from the SPF are the Bayesian prior estimates, which are then combined with site-specific accident data to produce EB-adjusted expected accident frequency estimates. The peak-searching algorithm for screening of roadway segments, described in the next chapter, employs SPF model predictions in this manner; thus, the SPF plays an important role in network screening.

While it is understood that the application of SPFs has much improved the process of network screening, there are also some issues that must be addressed. Not the least of these problems is that the independent variables in the model are assumed to be free of errors. This assumption is perhaps the single greatest weakness of SPFs. To illustrate this point, consider that the most influential (and often the only) dependent variable is traffic volume. The traffic volume of a site is usually measured with an automated traffic counter over a short period of time (days or weeks), and at a particular point on the roadway. The results of these short-term traffic counts are then used to generate annual traffic statistics (i.e.: AADT) for a given length of road (or a given intersection). To suggest that AADT data, or any other data to be used in the model, are error free is, at best, questionable; however, the generalized linear modelling methods used to develop SPFs do not account for variance in the dependent variables. At present, there is no well-accepted method of dealing with these errors, except to exclude the variable from the model. The exclusion of AADT from the model would result in losing what is by far the most significant independent variable in the model. So, when developing SPFs, the lesser of two evils is chosen: a model with error-prone independent variables is preferred to having no viable model at all.

SPFs are now most commonly developed by assuming that traffic counts at individual sites correspond to a negative binomial distribution (see Section 2.5.4.3 for details). For these models, it is generally assumed that the negative binomial dispersion parameter, k, is constant over the range of independent variables; however, it has been suggested by Hauer (*23*), Miaou and Lord (*24*) and others that this is not, in fact, the case.

In this paper, the effects of two independent variables on the negative binomial dispersion parameter are investigated, namely segment length and AADT.

4.2 Overdispersion Issues

In Chapter 2, the reasons for assuming that accident counts follow a Poisson distribution are explained. The variance of the Poisson distribution is assumed to be equal to the mean; in other words, it is assumed that the variance observed in a Poisson model can be explained by the independent variables. In the case of accident prediction models, model variance is usually found to exceed the mean. This 'extra-Poisson' variance (i.e., overdispersion) may indicate that the assumption of a Poisson distribution is inappropriate.

The negative binomial distribution is similar to the Poisson distribution in most respects; however, negative binomial models can accommodate overdispersion explicitly. Overdispersion is represented negative binomial dispersion parameter, which is estimated in the regression process along with other model parameters. It is most commonly denoted k, such that as $k \rightarrow 0$, the negative binomial distribution approaches a Poisson distribution. It is also sometimes given as d, which is the inverse of k; thus, $d=1/k$, and as $d \rightarrow \infty$, the negative binomial distribution approaches the Poisson distribution. In this thesis, the negative binomial dispersion parameter shall be defined as k, above.

Overdispersion may be thought of as the model variance not explained by the independent variables. One should not be surprised to see overdispersion in accident prediction models, as the true number of variables involved in accident occurrence is much larger than the number of modelled variables.

Overdispersion is usually held constant for SPFs; however, this assumption has come under question of late. The next sections describe a brief investigation into how overdispersion estimates vary with segment length and AADT.

4.3 Data Used in This Part of the Research

HSIS data for 2-lane rural highways in Washington state, from 1993 to 1995, were used to calibrate the models. A summary of relevant statistics for the Washington data is shown in Table 4.1.

The data were disaggregated by terrain type – level, rolling and mountainous. SPFs were calibrated for each terrain type; however the rolling terrain data will be the focus of this examination, owing to the larger numbers of sites and collisions. Segment lengths ranged from 0.01 to 28.66mi. Sites were defined as contiguous road segments that were homogeneous with respect to measured traffic and geometric characteristics (AADT, shoulder width, speed limit, etc.).

Table 4.1: Relevant statistics of the Washington HSIS data used for calibrating SPFs.

	Level	Rolling	Mountainous	Total
Total Length (mi)	952.63	3835.44	460.37	5248.44
Total No. Sites	1941	5792	663	8396
Mean Site Length (mi)	0.49	0.66	0.69	0.66
Mean AADT ('93-'95)	4740	4320	1770	3060
Total Collisions	4016	12917	1248	18181
Fatal/Injury Collisions	1906	6200	535	8641

4.4 Negative Binomial Model Calibration

The SPFs were all calibrated by negative binomial regression, which was performed SAS® software, with all parameters being estimated by maximum likelihood using the GENMOD procedure (*25, 26*). A simple 'AADT model' was specified, where the predicted accident frequency, κ, in accidents/mile/year, is estimated as a function of

AADT only. The segment length, *SL*, in miles, and number of years of data being used, *n*, were treated as offset variables. The model form is as follows:

$$\kappa = f(AADT) = \alpha \cdot AADT^{\beta} \qquad \text{(4.1) (same as 2.2)}$$

where α and β are parameters estimated from the data.

While SPFs were developed for each of the three terrain types, sites in rolling terrain only are used for the screening methods presented in this thesis; thus, only results for rolling terrain are reported herein.

For total accidents, α and β were estimated to be 0.0012 and 0.87, respectively, and k was estimated to be 0.49. Figure 4.1 shows a scatter plot of the observed accident counts, and the SPF estimated from those data.

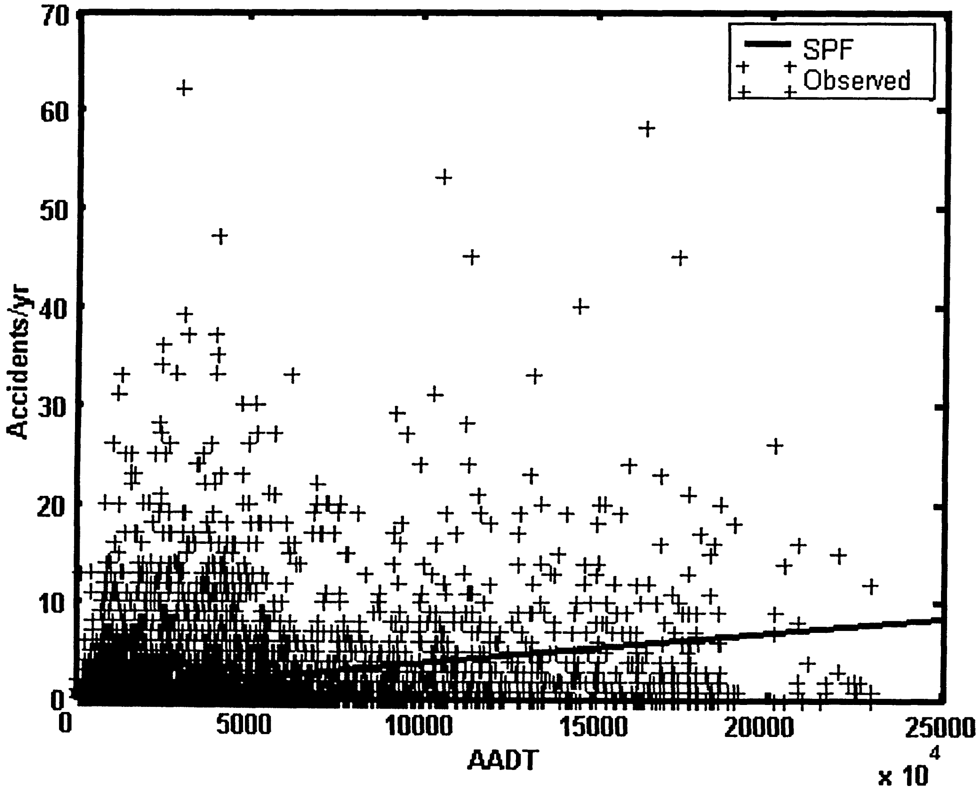


Fig. 4.1: Scatter plot of observed total accident counts and SPF for Washington 2-lane rural highways in rolling terrain.

For fatal/injury (FI) accidents, the estimates of α and β were 0.0006 and 0.87, respectively, and k was estimated to be 0.48. Plots of the SPFs for both total and FI accidents are shown in Figure 4.2.

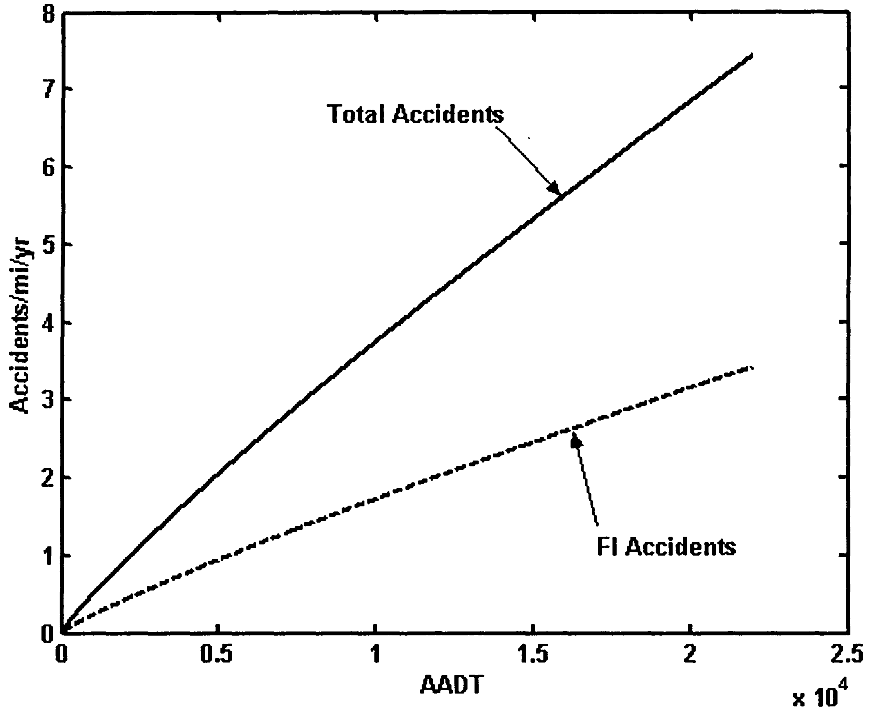


Fig. 4.2: SPFs for total and FI accidents on Washington 2-lane rural highways

4.5 Overdispersion as a Function of Segment Length

The first part of the investigation was to see how SPF parameters are related to the lengths of the segments used in the regression. (Recall that segment length is considered an offset variable, rather than a regressor variable.) To begin, a dataset made up of 2-mile segments was extracted from the original Washington 2-lane rural highway data. This was done by taking every site in the original data that had a length of at least 2 miles and disaggregating these into as many 2-mile segments as possible.

For example, a 4.3-mile site would yield two contiguous 2-mile segments beginning at the same point as the original site; the 0.3-mile remainder would be excluded from the analysis. In the end, the subset of the original data included 831 2-mile segments, some of which were congruent, and others isolated. Each segment was homogeneous along its length, with respect to measured characteristics. Segments were not, in general, homogeneous with respect to one another.

The same negative binomial regression procedure that was used in the previous section was applied to the 2-mile subset. The SPF, based on 1662 total miles of road, and 3289 total accidents, is shown in Figure 4.3, along with the SPF for the 'full' dataset.

The SPF calibrated from the 2-mile segments predicts fewer accidents than does the SPF from the original data. One of the reasons for this may relate to the fact that the original data includes many short segments (the mean site length is 0.66mi); this implies that geometric and/or traffic characteristics are changing over a short distance, and it has been suggested that changes in roadway, traffic, or environmental characteristics may be associated with increased accident risk. A deeper examination of this issue is beyond the scope of this work; see, for example, Anderson et al. (*27*), Ng and Sayed (*28*), and others for details.

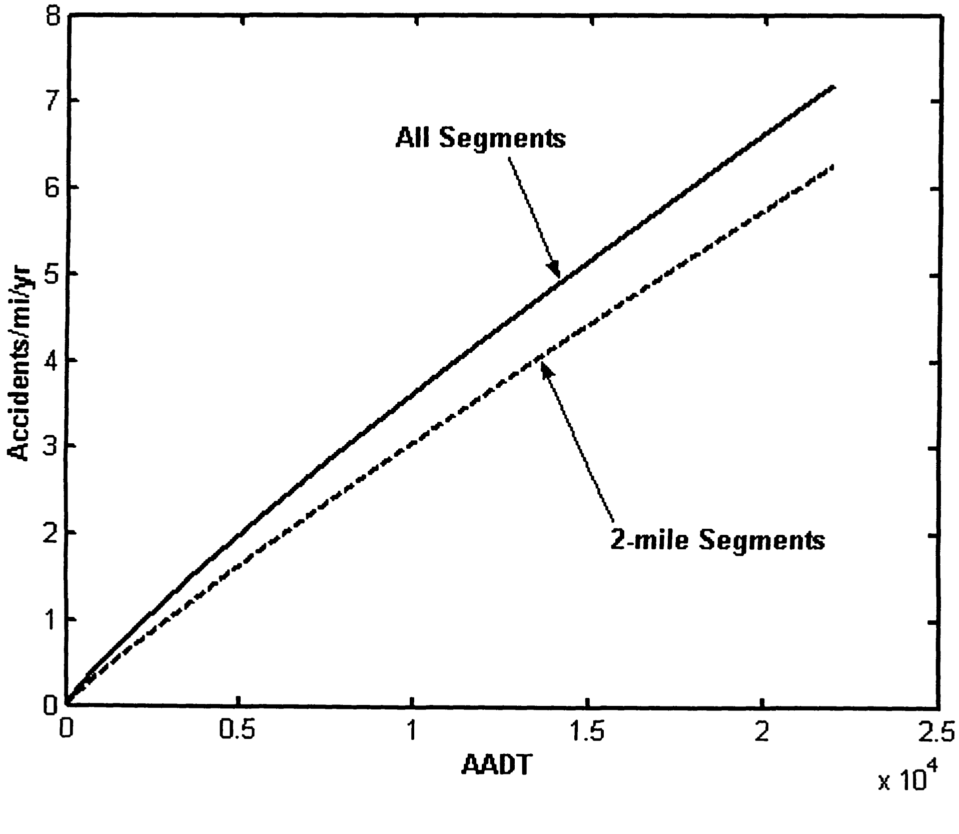


Fig. 4.3: Total-accident SPFs for all 2-lane rural highway segments in the Washington dataset, and a subset of 831 2-mile segments.

Next, the 2-mile segments were split into a new subset of 1-mile segments. The '1-mile dataset' had the same total number of accidents, total length of road, and traffic volumes as did the 2-mile dataset. Negative binomial regression was performed on the 1-mile dataset as before, but there were now twice as many observations used.

The process of dividing segments into shorter, equal-length subsegments was continued until the minimum possible length for analysis, 0.01mi, was reached. Figure 4.4 describes how the datasets were created.

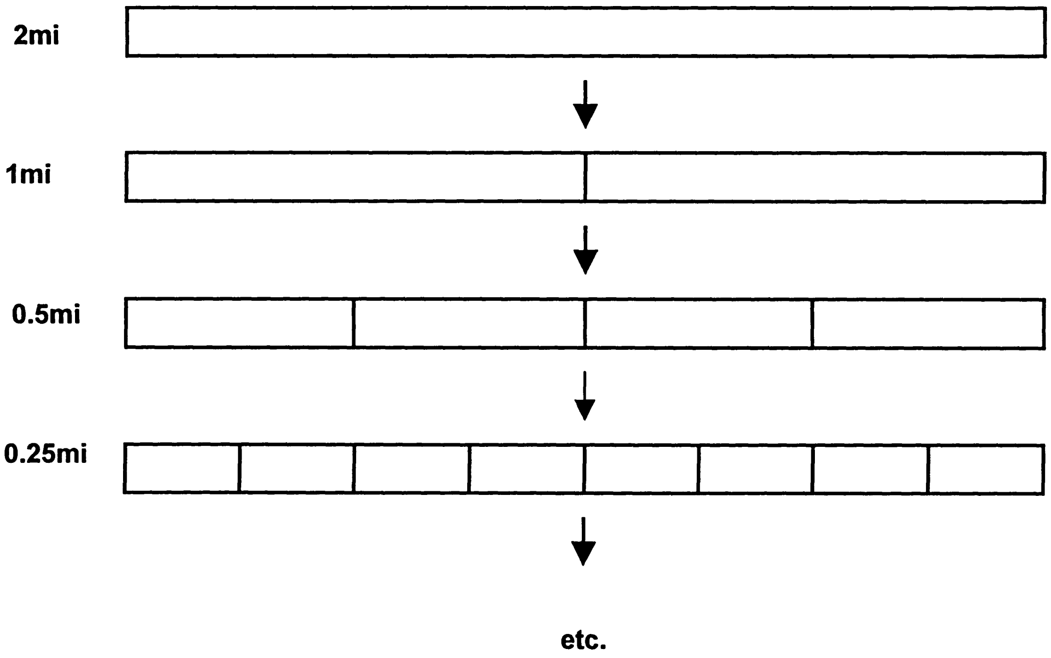


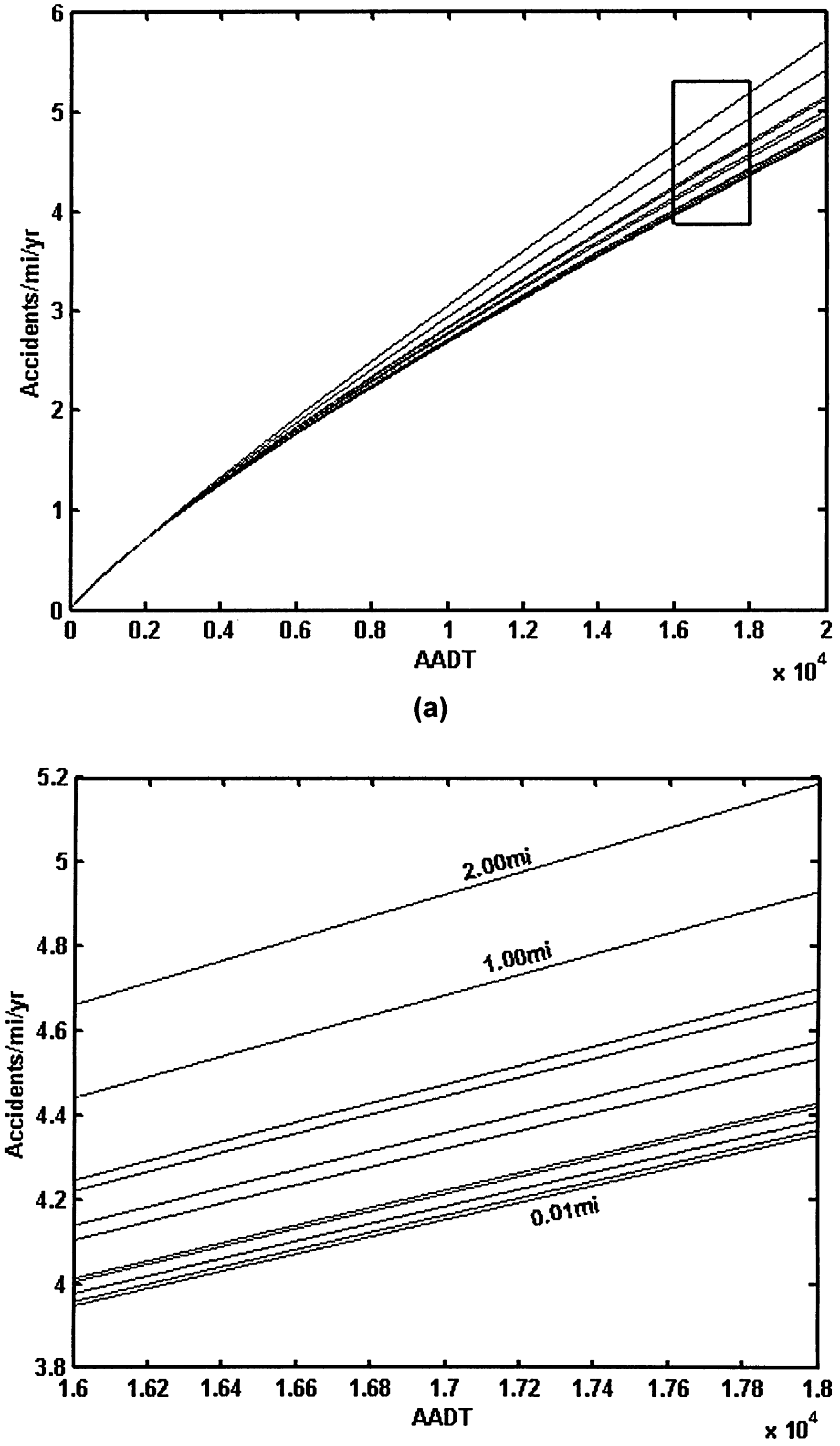
Fig. 4.4: Division of a site.

SPF parameters were calibrated for each segment length used, and the results are shown in Table 4.2. All datasets were the same, except for the length and number of subsegments in each; thus, differences in parameter estimates can be attributed only to the differences in the segment lengths of the datasets.

For each new set of data, a negative binomial regression model was calibrated. The parameter estimates are shown in Table 4.2. The parameter estimates for α and β are not equal across datasets, and this will result in differences in model prediction; this is evident in Figure 4.5, which shows the family of SPFs calibrated from the different datasets. Additional model information, including goodness-of-fit statistics, are given in Appendix A.

Table 4.2: Negative binomial regression parameters for data subsets.

Dataset Statistics				Negative Binomial Regression Parameters		
Segment Length (mi)	**No. of obs.**	**Total Accidents**	**Total Length (mi)**	α	β	k
0.01	166200	3289	1662	0.0013	0.83	6.00
0.02	83100	3289	1662	0.0013	0.83	3.23
0.04	41550	3289	1662		did not converge	
0.05	33240	3289	1662	0.0013	0.83	1.54
0.08	20775	3289	1662	0.0012	0.83	1.17
0.10	16620	3289	1662	0.0012	0.84	1.00
0.20	8310	3289	1662	0.0011	0.85	0.83
0.25	6648	3289	1662	0.0011	0.85	0.79
0.40	4155	3289	1662	0.0010	0.86	0.68
0.50	3324	3289	1662	0.0010	0.86	0.61
1.00	1662	3289	1662	0.0009	0.88	0.59
2.00	831	3289	1662	0.0007	0.91	0.55



(a)

(b)

Fig. 4.5: (a) Family of SPFs calibrated using subsets of 1662 miles of Washington 2-lane rural highway segments with varying segment lengths. Inset is shown in (b).

While the differences in the estimates α and β are relatively small and subtle, the differences between overdispersion estimates are much more pronounced. To visualize how the overdispersion estimates vary with segment length, the k-values were fitted with a shape-preserving interpolant using MATLAB® software; this is shown in Fig. 4.6. A shape-preserving interpolant of d (d=1/k) is shown in Figure 4.7.

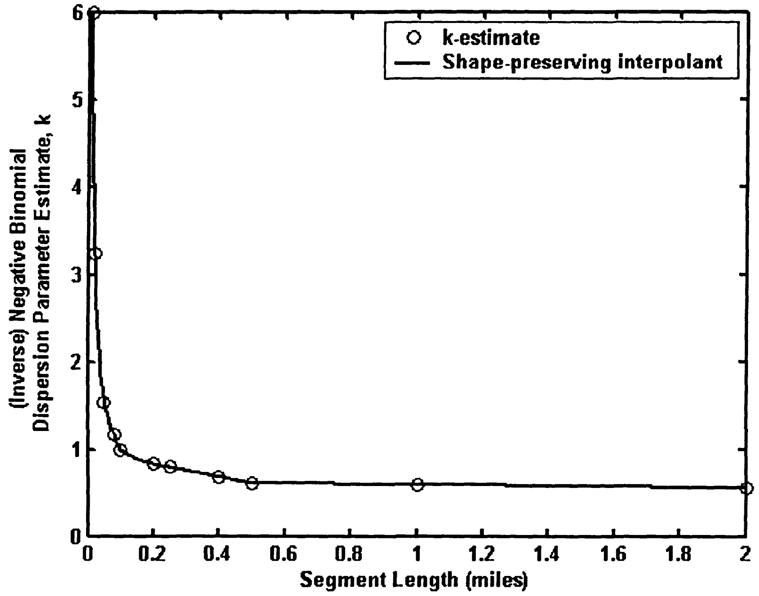


Fig. 4.6: Shape-preserving interpolant of k vs. segment length.

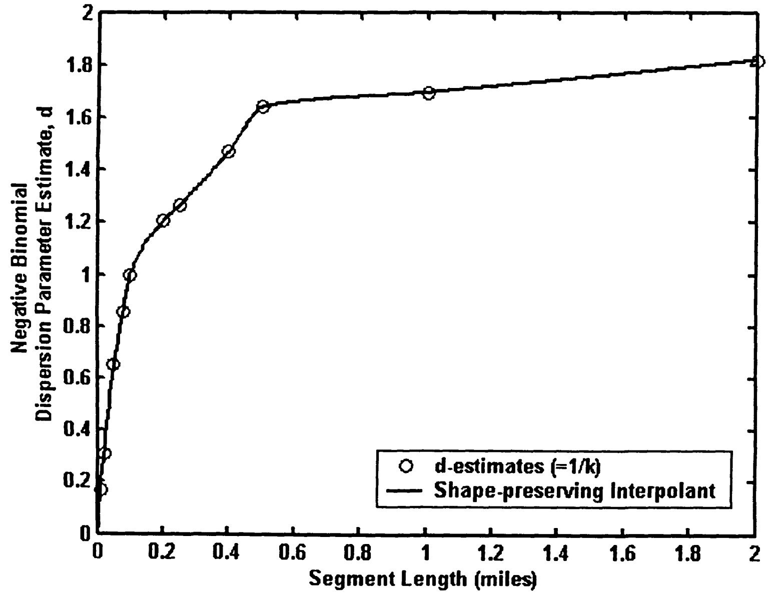


Fig. 4.7: Shape-preserving interpolant of d vs. segment length.

The interpolants shown in Figures 4.6 and 4.7 would suggest that a functional relationship exits between overdispersion and segment length. To determine this relationship, the data points were fitted to several different models. The following non-linear model was found to fit the data very well:

$$k = f(SL) = \frac{\beta_1 + SL}{\beta_2 SL} \quad , \qquad (4.2)$$

where β_1 and β_2 are parameters. This overdispersion model was calibrated using the NLIN procedure in SAS (*26*), and β_1 and β_2 were estimated to be 0.107 and 1.97, respectively. The model implies that as $SL \rightarrow \infty$, $k \rightarrow 1/\beta_2$, and as $SL \rightarrow 0$, $k \rightarrow \infty$; thus, k will approach a minimum value, k_{min}, as segment length increases. Figure 4.8 shows that this model agrees closely with the observed values of k. Figure 4.9 shows a plot of the inverse of Equation 4.2, which is simply d vs. segment length.

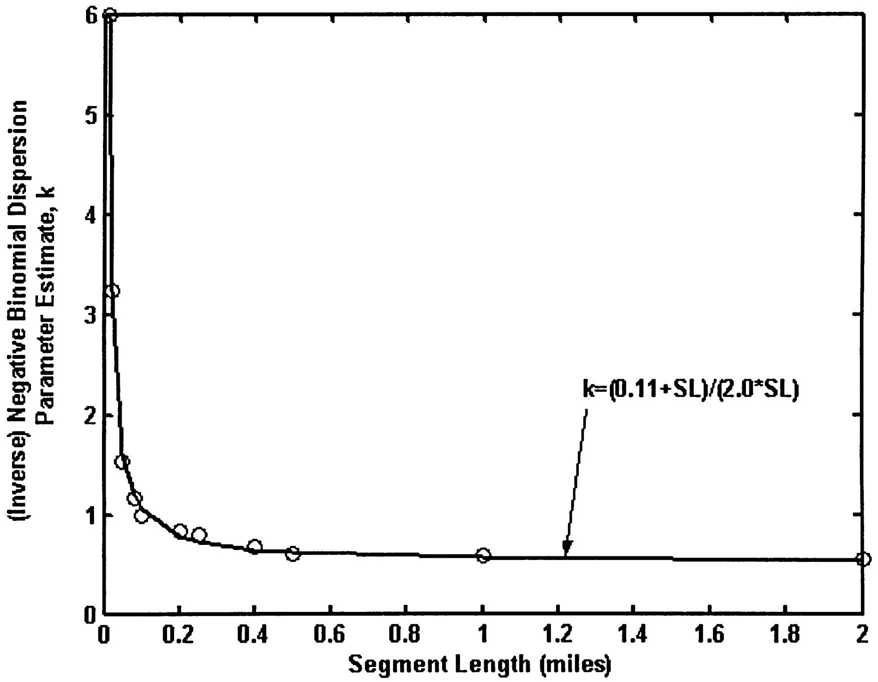


Fig. 4.8: Equation 4.2 fitted to k vs. segment length.

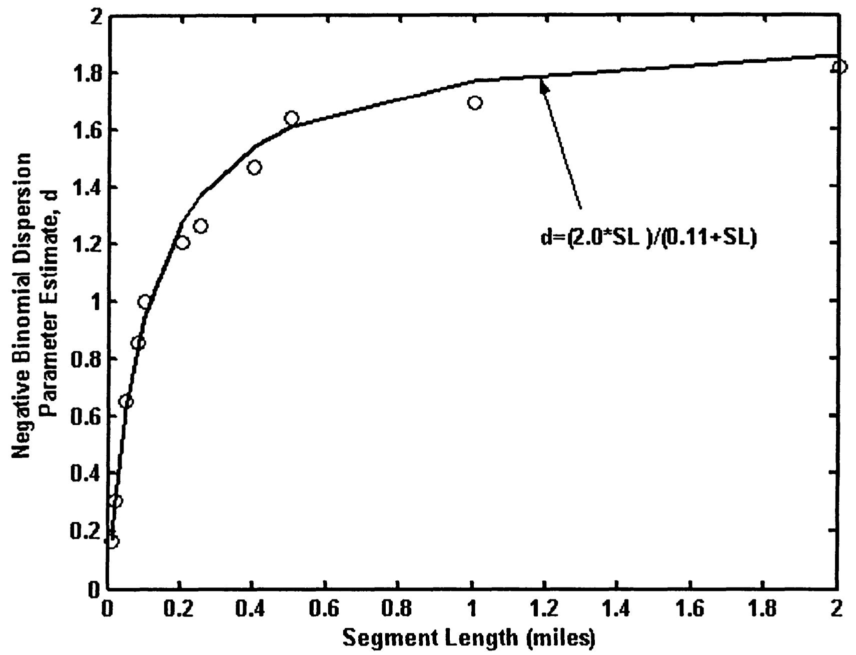


Fig. 4.9: Inverse of Equation 4.2 fitted to d vs. segment length.

Clearly, the value of the value of k is large for small segment lengths, and smaller at longer lengths. Knowing that the variance of the negative binomial distribution is given by:

$$Var(y) = \mu + k\mu^2 \quad , \qquad (4.3) \text{ (same as 2.9)}$$

it can be seen that the variance of the SPF will increase as k increases. Note that the mean of the negative binomial distribution, μ, at any value of AADT would be the same as the model prediction, κ, for any site with that AADT. To illustrate, Table 4.3 shows how the model variance changes with k, for a hypothetical value of the mean, μ=3.

Table 4.3: Variance of model prediction, μ, where μ=3 accidents/yr.

Segment Length (mi)	Dispersion Parameter, *k*	Var(*κ*)
0.01	6.00	57.0
0.02	3.23	32.1
0.05	1.54	16.8
0.1	1.00	12.0
0.25	0.79	10.1
0.5	0.61	8.48
1	0.59	8.28
2	0.55	7.98

The model variance increases dramatically as *k* becomes large, and one of the implications of this is that EB weights will be affected. The EB weight, *w*, is a function of *k* (and hence the variance), and is given by:

$$w = \frac{1}{1 + k\mu} \quad . \tag{4.4}$$

The EB weight is used when model predictions and site-specific observations are combined to estimate the expected accident frequency, *κ*. The larger the value of *w*, the greater the influence of the model prediction, μ, and hence there is less influence on the observed accident counts. Table 4.4 shows the value of *w* calculated for each of the different *k* estimates; again, an SPF prediction of *κ* =3 accidents/mi/yr is assumed.

Table 4.4: EB weights, where μ=3 accidents/yr.

Segment Length (mi)	Overdispersion Parameter, *k*	EB weight, *w*
0.01	6.00	0.053
0.02	3.23	0.093
0.05	1.54	0.18
0.1	1.00	0.25
0.25	0.79	0.30
0.5	0.61	0.35
1	0.59	0.36
2	0.55	0.38

As k approaches zero, the weight given to the model prediction appears to approach zero. Conversely, the weight of the observed accident counts would approach unity.

The reasons why short segments should exhibit greater overdispersion than longer ones are not entirely clear. If one were to imagine a single 10-km-long road segment, and compare it to 100 discontiguous 0.1-km segment segments, it would be reasonable to expect a relatively low overdispersion from the longer site, if only because the same drivers are using the same 10-km of road in the same environmental conditions on each trip. The same cannot be said for the group of shorter segments. To date, very little work has been done on this topic, and further research is needed to gain an understanding of the underlying causes of overdispersion variation.

To see what effect different values of k will have on network screening results, the EB procedure was applied to the 2-mile segments of Washington 2-lane rural highway dataset. The sum of the SPF predictions (i.e., the Bayesian prior estimates), $\Sigma\kappa$, for each value of k, and sum of the EB expected accidents, ΣX, are shown in Table 4.5, and compared with the average annual counts from 1993-1995, adjusted to 1995. The results are shown in Figure 4.10.

Table 4.5: Total observed, predicted, and expected accidents for different values of k, for 2-mile segments of 2-lane rural highway in Washington.

Segment Length (mi)	Observed (93-95)	Predicted (95)	EB Expected (95)
SL	ΣK	Σκ	ΣX
0.01	1096	1127	1120
0.02	1096	1127	1120
0.05	1096	1128	1122
0.1	1096	1130	1122
0.25	1096	1136	1123
0.5	1096	1141	1123
1	1096	1150	1123
2	1096	1162	1123

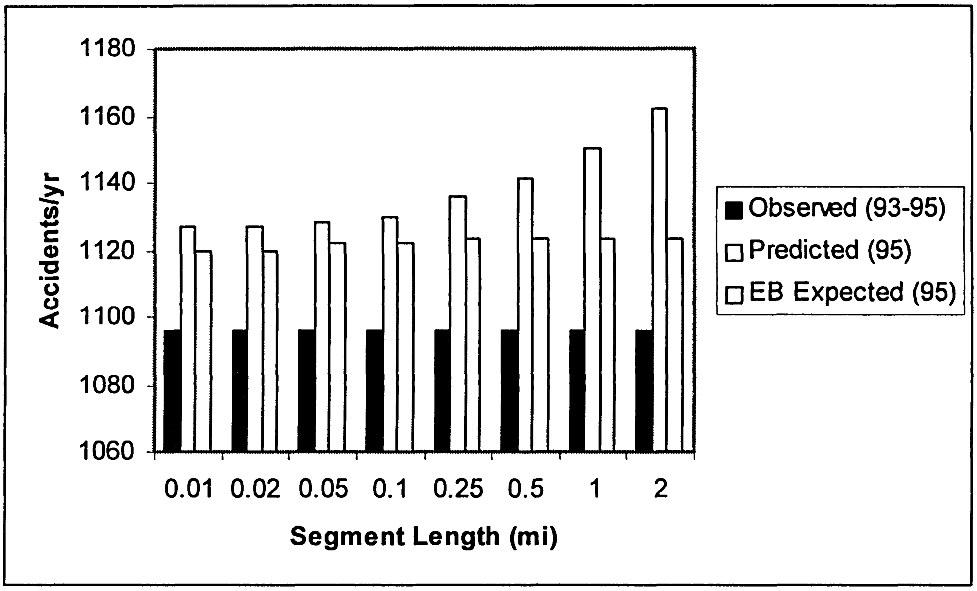


Fig. 4.10: Comparison of observed, predicted, and EB expected accidents for different segment lengths of Washington 2-lane rural highways.

The number of accidents predicted by the model increases as segment length increases; however, the EB estimates of expected accident frequency show a remarkable stability. Thus, it would seem that large differences in the estimated overdispersion do not carry over to the final weighted EB estimates.

4.6 Overdispersion As A Function of AADT

To see if AADT has an effect on estimates of the overdispersion parameter, data for 5,792 2-lane rural highway segments in Washington were selected. Four years of accident and traffic data were available, from 1993-1996. Table 4.6 shows relevant statistics from the dataset.

Sites were grouped into five groups, or 'bins', based on AADT. The choices of bin size were made so that each group had a large sample size; thus the bin ranges are not equal. Table 4.7 shows relevant data for each bin.

Table 4.6: Relevant statistics of Washington 2-lane rural highways used for overdispersion investigation.

Total Sites	5 792
Total Miles of Highway	3 835.44
Total Accidents (1993-1996)	17 634
Minimum AADT	110
Maximum AADT	23 500
Mean AADT	4 360
Mean Site Length (mi)	0.66

Table 4.7: Bin descriptions for AADT data.

Bin	AADT Range	# Sites	Mean AADT	Total Length (mi)	Mean Site Length (mi)
1	0 – 1500	1479	900	1608.80	1.09
2	1500 – 3000	1407	2200	971.36	0.69
3	3000 – 5000	1161	4020	639.64	0.55
4	5000 – 8000	885	6070	345.77	0.39
5	8000+	860	12560	269.87	0.31

For each bin, negative binomial regression was performed using SAS as before. The SPF parameter estimates for each bin are shown in Table 4.8. Parameter estimates for bin 3 were not statistically significant. Further details are provided in Appendix A.

Table 4.8: Results of negative binomial regression

Bin	AADT Range	Parameter Estimates		
		α	β	k
1	0 – 1500	0.00033	1.1	0.60
2	1500 – 3000	0.00024	1.1	0.43
3	3000 – 5000		*parameters not significant	
4	5000 – 8000	0.00016	1.1	0.42
5	8000+	0.00084	0.94	0.40

The results would suggest estimates of *k* are quite stable for AADT values in excess of 1500. For bin 1, which had the lowest range of AADT, the value of *k* was higher than for other bins; however the difference between bin 1 and bin 5 was the only significant difference (i.e., estimates for bins 1-4 were not significantly different, and estimates for bins 2-5 were not statistically different).

Table 4.7 showed that the mean site length decreases substantially as AADT increases. This is because higher volume roads generally have more access points than lower-volume roads; thus, AADT or other site characteristics are likely to change more frequently, and these changes define site boundaries.

Given the results of the previous section, it may be reasonable to expect that overdispersion would rise for high-AADT sites, as the segment lengths are relatively short; however, overdispersion does not appear to increase with AADT, and may even decrease. This could mean that sites with high AADT exhibit less variance than low-AADT sites, and this moderates the effects of segment length; however, it is too early to tell if this is the case, and further study is needed.

4.7 Chapter Summary

This was intended to be a brief excursion into the nature of overdispersion, so it is difficult to draw hard conclusions. The results of Section 4.3 showed that EB-adjusted expected accidents are not seriously affected by differences in overdispersion estimates; however, it is difficult to generalize this result; so the indications that overdispersion is not, in fact, constant over the range of a given independent or offset variable is suggestive that a better understanding of overdispersion could lead to better accident prediction models, and this is an avenue of research that should be pursued.

5 Investigation of Application Issues for the Peak-Searching Algorithm for Screening Roadway Segments

5.1 Background

The peak-searching algorithm examined in this chapter is a network screening approach for use with road segments. The method was developed by Hauer (*29*), and has been suggested for inclusion in the *SafetyAnalyst* Network Screening Tool (*30*).

The peak-searching algorithm has several similarities to the 'sliding window' approach to screening road segments. In the sliding window approach, all sites under investigation are divided into subsegments of some equal length, SL_{SUB}, such as 0.1 or 0.01 km (or mi). A 'window' of length W is 'placed' at the beginning of the road segment, and the EB-adjusted expected accident frequency, X_Y, or the excess accident frequency, $Excess_Y$ is calculated for the given window on a per-mile basis, as shown in Figure 5.1. The subscript Y denotes the year for which the accident frequency is estimated. The accident frequencies may be estimated for total accidents, fatal/injury (FI) accidents, or any other accident type, provided sufficient data exist. Estimates of X_Y and $Excess_Y$ are calculated using empirical Bayes methods; therefore, appropriate SPFs must be calibrated prior to screening.

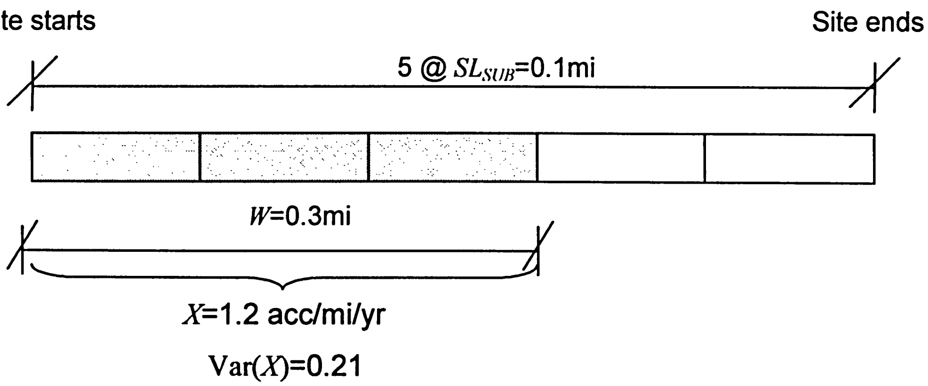


Fig. 5.1: Sliding window concepts: placement of the first window.

Once the calculations for the first window have been made, the window is then moved a length of one increment, L_{INC}, down the roadway, and the calculations are repeated for the new window. The increment size is usually taken to be the same as the subsegment length, SL_{SUB}; however, any value of L_{INC} may be used, provided $L_{INC} \geq SL_{SUB}$. So long as the chosen window size is larger than the increment, as is usually the case, adjacent windows will overlap. This has the effecting of "smoothing" the averages, and is one of the main features of the sliding window approach.

The process continues until the window reaches the end of the site. If the site has a 'remainder' at the end, the last window is taken to be the distance W from the site endpoint. A remainder is a length of road shorter than the subsegment length, SL_{SUB}, and one is created wherever the site length is not a multiple of the subsegment length.

Figure 5.2 illustrates the window locations and accident frequency calculations at a fictitious 0.44-mi site, with W=0.3mi and L_{INC}=0.1mi.

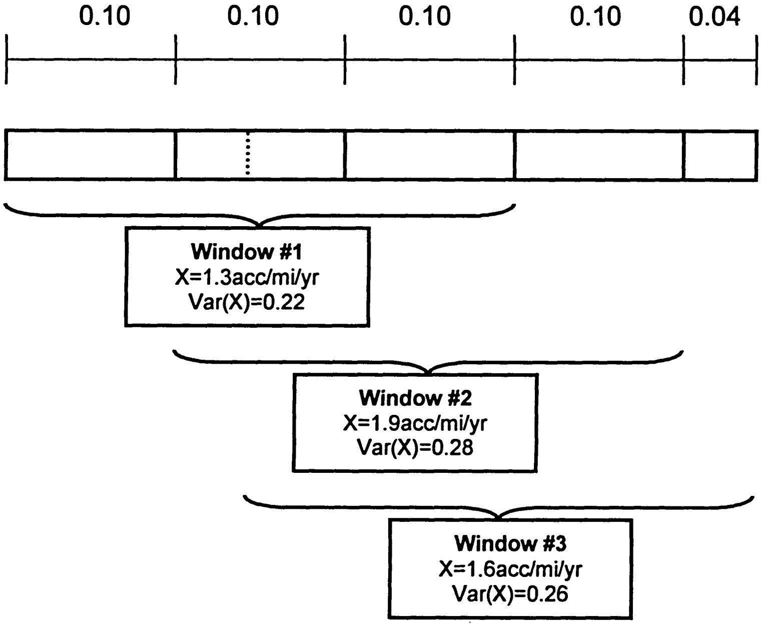


Fig. 5.2: Example of sliding window procedure.

For a given site, the window with the highest expected accident frequency or excess accident frequency, is selected to 'represent' that site. All sites in the network are then ranked based on the given frequency, with those sites having the highest frequencies being ranked at the top. The variance of X_Y or $Excess_Y$ are also estimated for each site; however, they do not influence the rankings.

Another feature of the sliding window approach is that the windows may 'bridge' adjacent sites provided that the sites are contiguous with respect to one another. This feature is illustrated in Figure 5.3.

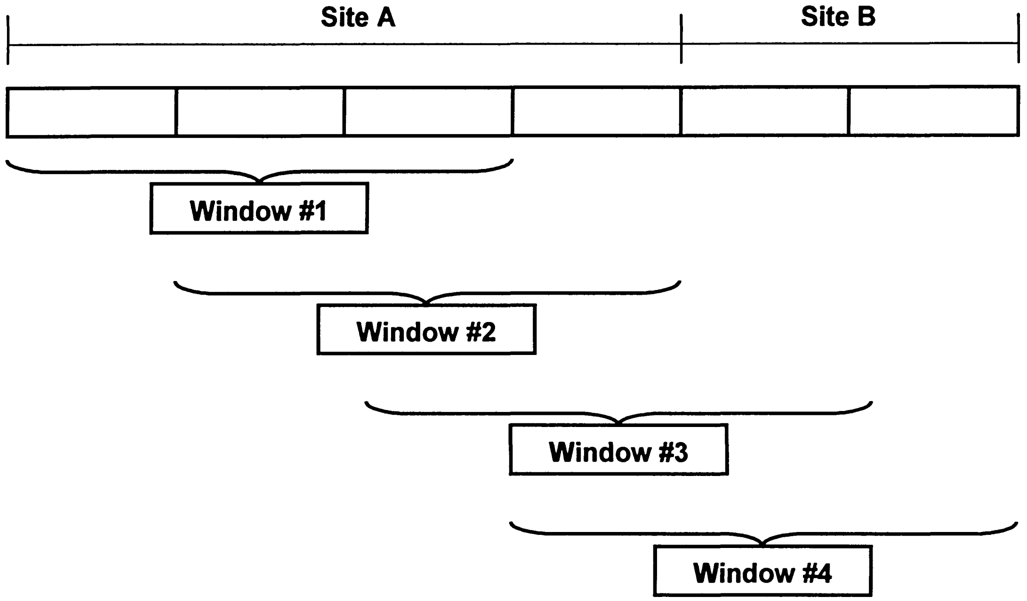


Fig. 5.3: Sliding window 'bridging' adjacent sites.

One of the problems with the sliding window method is the question of what window size should be used. If a very small window size is selected, then, as shown in Chapter 4, the EB estimates will exhibit large variances; thus, the reliability of the screening results would be questionable. If a large window size is selected, any isolated sites with a total length of less than the window size would be excluded from the screening procedure.

To overcome these difficulties, Hauer (*29*) developed a peak-searching algorithm to screen roadway segments. In this method, the window length is not fixed, and the variances of the safety estimates play a central role in the ranking of sites.

5.3 The Peak-Searching Algorithm

5.3.1 General

For the peak-searching algorithm, sites are divided into subsegments in the same way as for the sliding window approach. A minimum window size, W_{min}, is selected such that it represents the shortest segment length that can be evaluated with a reasonable degree of accuracy. The size of W_{min} will be dictated by the precision of the network data under investigation, and reasonable assumptions regarding the accuracy of those data. For example, some jurisdictions report accident locations to 0.001mi (1.6m) precision; however, it is difficult to believe that the size of a 'real' hazardous location could be defined so precisely.

The first window is of size W_{min}, and it is placed at the beginning of the site. The expected accident frequency, X_Y, or the excess accident frequency, $Excess_Y$, is then calculated for the window in the same way as for the sliding window approach, along with the estimate's variance. The estimate of X_Y or $Excess_Y$ is then subjected to a test of statistical precision, which is done by calculating the coefficient of variation (CV) of the estimate. CV is defined as the standard deviation of the estimate divided by the estimate; thus, CV is given by:

$$CV = \frac{\sqrt{Var(Estimate)}}{Estimate} \quad . \tag{5.1}$$

Estimates with relatively large variances will have large values of CV. Thus, the lower the value of CV, the more precise the estimate. A CV of zero would indicate near-perfect precision.

Prior to screening, a limiting value is selected for the coefficient of variation, CV_{lim}. If the first window has a CV of less than or equal to CV_{lim}, the site is 'flagged,' and the site is ranked based on the expected accident frequency or excess accident frequency of the window on a per-mile basis. If the first window is not flagged, then a second window is placed one increment down the roadway in the same manner as with the sliding window approach. The same calculations are made for the new window, and it is subjected to the same statistical precision test. If the second window does not pass the CV test, then the window is moved another increment down the roadway. This process continues until the site end point is reached, or until a window is flagged. Figure 5.4 shows the procedure where W=W_{min}=0.10, and CV_{lim}=0.5.

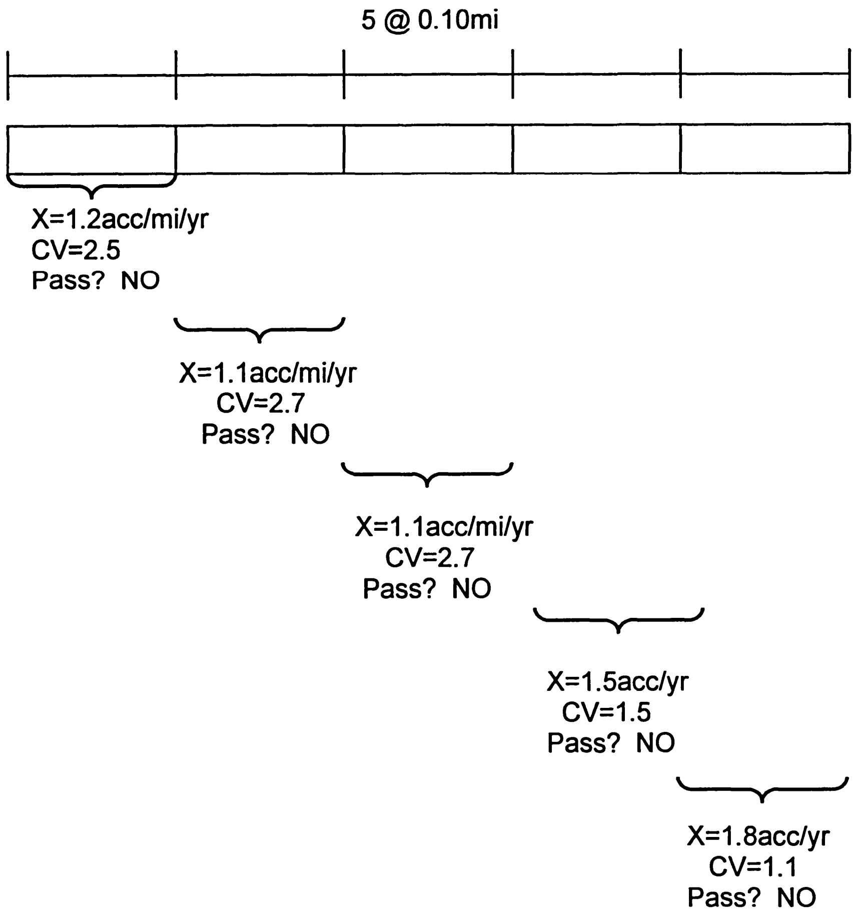


Fig. 5.4: Peak-searching concepts; $W=W_{min}$.

If no windows pass the test, then the window size is increased by one unit, which shall be taken to be the same as the increment size, L_{INC}. Again, the window moves down the roadway, this time with larger, overlapping windows, until the end of the site is reached or a window has an acceptably low CV. This is shown in Figure 5.5 for L_{INC}=0.1; thus $W=W_{min}+L_{INC}$=0.20mi.

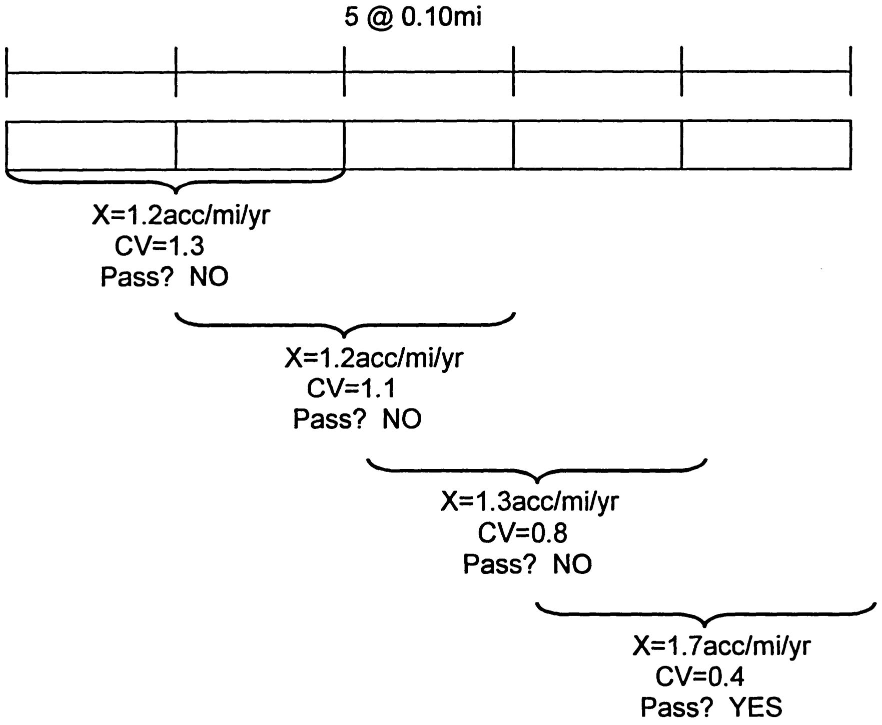


Fig. 5.5: Peak-searching concepts, W=0.20.

If there are still no windows flagged, then the procedure is repeated with successively larger window sizes until either the site has been flagged or the window size is equal to the site length. If all possible window sizes have been tried, and none have passed the test, then the site is not included in the ranked list of sites.

The algorithm is repeated for every site in the network under investigation. The end result is a list of sites which have an expected accident frequency or excess accident frequency estimated to a desired level of precision. Sites are then ranked as in the sliding window method, where sites with the largest X_Y or $Excess_Y$ estimates are ranked highest.

The value of CV_{lim} represents the minimum level of precision the EB estimates must show in order to be included in the ranked list. This has the effect of excluding those sites whose safety estimates are deemed to be too unreliable.

The peak-searching algorithm may be performed for any type of accident for which a sufficient SPF can be developed. In this thesis, screening procedures are demonstrated for total, fatal/injury, property-damage-only, and equivalent-property-damage-only accidents, denoted by the subscripts TOT, FI, PDO, and EPDO, respectively. The network may be evaluated for each accident type, and for either expected accident frequency or excess accident frequency. Step-by-step instructions for each case are given below.

5.3.2 Ranking of Sites Based on Expected Accident Frequency on Road Segments

Step 1: Safety performance functions for both total and FI accidents are calibrated for the network under investigation. Alternatively, SPFs from another source may be transferred from another jurisdiction applied to the current network; however, methods for this are not yet well developed. See Persaud et al. (*9*) for details.

It is planned that future users of the Network Screening Tool will have the option of using default SPFs calibrated using data from other jurisdictions, or other SPF parameters specified by the user. Presumably, the latter SPF would be calibrated using data from the network being investigated.

Step 2: Once the network data have been broken down into subsegments of length SL_{SUB}, the SPFs may be applied. Using the appropriate SPF model parameters, the *predicted* number of accidents per mile, κ_y, for each year, y, y=1,2,...,Y, is calculated for both total and FI accidents as follows:

$$\kappa_{y(TOT)} = \alpha_{TOT} AADT_y^{\beta_{TOT}} \tag{5.2}$$

$$\kappa_{y(FI)} = \alpha_{FI} AADT_y^{\beta_{FI}} \tag{5.3}$$

Step 3: Using the model predictions calculated in Step 2, the yearly correction factor, C_y, was computed for total and FI accidents, and for each year, as follows:

$$C_{y(TOT)} = \frac{\kappa_{y(TOT)}}{\kappa_{1(TOT)}} \tag{5.4}$$

$$C_{y(FI)} = \frac{\kappa_{y(FI)}}{\kappa_{1(FI)}} \tag{5.5}$$

Step 4: Using κ_1,..., κ_Y, and the negative binomial dispersion parameter, k, the EB weight, w, was calculated for total and FI accidents with the following:

$$w_{TOT} = \frac{1}{1 + k_{TOT} \sum_{y=1}^{Y} \kappa_{y(TOT)}} \tag{5.6}$$

$$w_{FI} = \frac{1}{1 + k_{FI} \sum_{y=1}^{Y} \kappa_{y(FI)}} \tag{5.7}$$

Step 5: Next, the base EB-adjusted expected number of accidents, X_1, for total and FI accidents, were calculated for year 1:

$$X_{1(TOT)} = w_{TOT}\kappa_{1(TOT)}SL_{sub} + (1 - w_{TOT})\left(\frac{\sum_{y=1}^{Y} K_{y(TOT)}}{\sum_{y=1}^{Y} C_{y(TOT)}}\right) \tag{5.8}$$

$$X_{1(FI)} = w_{FI}\kappa_{1(FI)}SL_{sub} + (1 - w_{FI})\left(\frac{\sum_{y=1}^{Y} K_{y(FI)}}{\sum_{y=1}^{Y} C_{y(FI)}}\right) \tag{5.9}$$

Step 6: Then X_Y, the EB-adjusted expected number of accidents for $y=Y$, the final year for which data exist for the site. Ultimately, flagged sites will be ranked based on these estimates. For total, FI, and PDO accidents, X_Y is calculated as follows:

$$X_{Y(TOT)} = X_{1(TOT)}C_{Y(TOT)} \tag{5.10}$$

$$X_{Y(FI)} = X_{1(FI)}C_{Y(FI)} \tag{5.11}$$

$$X_{Y(PDO)} = X_{Y(TOT)} - X_{Y(FI)} \tag{5.12}$$

Step 7: The variance of each EB-adjusted expected accident frequency estimate was then calculated, and is the measure of precision for the estimates. Note that the variance for PDO accidents is given by the sum of the total and FI variances. This amounts to a 'worst-case' estimate for the PDO variance by assuming that total and FI accident counts are statistically independent, which is clearly not the case. Were PDO accidents modelled explicitly, the variance of the estimates would likely be smaller than the sum of the total and FI variances. Thus, the PDO variance is overestimated.

$$Var\left(X_{Y(TOT)}\right)=X_{Y(TOT)}\left(1-w_{TOT}\right)\left(\frac{C_{Y(TOT)}}{\sum_{y=1}^{Y}C_{Y(TOT)}}\right) \tag{5.13}$$

$$Var\left(X_{Y(FI)}\right)=X_{Y(FI)}\left(1-w_{FI}\right)\left(\frac{C_{Y(FI)}}{\sum_{y=1}^{Y}C_{Y(FI)}}\right) \tag{5.14}$$

$$Var\left(X_{Y(PDO)}\right)=Var\left(X_{Y(TOT)}\right)+Var\left(X_{Y(FI)}\right) \tag{5.15}$$

The following two steps apply only to screening for EPDO accidents. If screening is not being performed for severity-weighted accidents, Steps 8 and 9 are skipped.

Step 8: The EPDO expected accident frequency was calculated by applying a relative severity weight, SW, for each level of crash severity (e.g., fatal, severe injury, etc.) described in the data. The relative severity weight is the cost of an accident of the given severity level in terms of PDO accidents; thus, PDO accidents always have a weight of 1. Because accident reporting practices vary over jurisdictions, different severity scales and different accident cost estimates may be used. In this Chapter, only fatal (F), nonfatal injury (NFI) and PDO severity types are considered. The number of FI accidents is the sum of F and NFI accidents. To calculate the EPDO expected accident frequency, let RC_{FI} be the relative weight of FI accidents as compared to PDO accidents. RC_{FI} was calculated as follows:

$$RC_{FI}=P_F SW_F+P_{NFI}SW_{NFI} \tag{5.16}$$

where P_F is the proportion of FI accidents that were fatal, and P_{NFI} is the proportion of all FI accidents that were nonfatal. Different injury scales may have a different number of levels and different severity types, but the premise is the same.

The EB-adjusted EPDO expected accident frequency was then be estimated by:

$$X_{Y(EPDO)} = X_{Y(PDO)} + RC_{FI} X_{Y(FI)} \quad . \qquad (5.17)$$

Step 9: Next, the variance of the EPDO estimate was found by:

$$Var\left(X_{Y(EPDO)}\right) = Var\left(X_{Y(TOT)}\right) + \left(RC_{FI} - 1\right)^2 Var\left(X_{Y(FI)}\right) \quad . \qquad (5.18)$$

Whether or not the network is being screened for EPDO accidents, calculations for individual subsegments are now complete. The next steps describe the calculations performed for individual windows.

Step 10: The average expected accident frequency of a given window of length W was calculated by:

$$Avg\left(X_{Y(TOT)}\right) = \frac{\sum_{SUB} X_{Y(TOT)}}{W} \qquad (5.19)$$

$$Avg\left(X_{Y(FI)}\right) = \frac{\sum_{SUB} X_{Y(FI)}}{W} \qquad (5.20)$$

$$Avg\left(X_{Y(PDO)}\right) = \frac{\sum_{SUB} X_{Y(PDO)}}{W} \qquad (5.21)$$

$$Avg\left(X_{Y(EPDO)}\right) = \frac{\sum_{SUB} X_{Y(EPDO)}}{W} \qquad (5.21)$$

Step 11: The variance of each average was calculated by summing the variance of the above statistics for the respective accident severity levels as shown below:

$$Var\left[Avg\left(X_{Y(TOT)}\right)\right] = \frac{\sum_{SUB} Var\left(X_{Y(TOT)}\right)}{W^2} \qquad (5.22)$$

$$Var\left[Avg\left(X_{Y(FI)}\right)\right] = \frac{\sum_{SUB} Var\left(X_{Y(FI)}\right)}{W^2} \qquad (5.23)$$

$$Var\left[Avg\left(X_{Y(PDO)}\right)\right]=\frac{\sum_{SUB} Var\left(X_{Y(PDO)}\right)}{W^2} \quad (5.24)$$

$$Var\left[Avg\left(X_{Y(EPDO)}\right)\right]=\frac{\sum_{SUB} Var\left(X_{Y(EPDO)}\right)}{W^2} \quad (5.25)$$

Step 12: The expected accident frequency or the excess accident frequency for each window of length W_{min} was then subjected to the statistical precision test by calculating the CV of the estimated expected accident frequency, X_Y. The CV of X_Y was calculated as follows for each severity level:

$$CV_{(TOT)}=\frac{\sqrt{Var\left(X_{Y(TOT)}\right)}}{X_{Y(TOT)}} \quad (5.26)$$

$$CV_{(FI)}=\frac{\sqrt{Var\left(X_{Y(FI)}\right)}}{X_{Y(FI)}} \quad (5.27)$$

$$CV_{(PDO)}=\frac{\sqrt{Var\left(X_{Y(PDO)}\right)}}{X_{Y(PDO)}} \quad (5.28)$$

$$CV_{(EPDO)}=\frac{\sqrt{Var\left(X_{Y(EPDO)}\right)}}{X_{Y(EPDO)}} \quad (5.29)$$

The CVs from all of the windows of length W_{min} are then compared to the value of CV_{lim}. When at least one CV is less than CV_{lim}, the entire roadway segment (i.e., site) is flagged. From all windows that have a CV less than CV_{lim}, the window with the largest peak expected accident frequency, $Peak(X_Y)$, is selected. To express X_Y on a per-mile basis, X_Y is multiplied by $1/W$ to account for the window length.

The entire flagged roadway segment is placed on the list of roadway segments to be ranked and the location of the window "passing the test" and the value, on a per-mile basis, of its expected accident frequency is included in the output.

If a roadway segment is not flagged, then the window size is increased by one unit of increment, L_{INC}. The now-larger window is now placed at the beginning of the site, and subsequent windows are moved to the right one increment at a time. The expected accident frequency for the window is determined by calculating the average of the expected accident frequency, $Avg(X_Y)$, across all subsegments contained in the window. The expected accident frequency for each subsegment was calculated in accordance with Steps 2 though 9 above.

Once the algorithm has been completed for each site in the network, a final ranked list of sites can be generated; sites are ranked by expected accident frequency on a per-mile basis. If a shorter list of sites is desired, the value of CV_{lim} may be decreased arbitrarily so that fewer sites will pass the CV test. Increasing the value of CV_{lim} will generate a longer list of sites. In practice, sites at the top of the list would be subjected to a more detailed safety investigation.

5.3.3 Calculation of PSI Based on Excess Accident Frequency on Road Segments

The procedures for screening roadway segments based on excess accident frequency, $Excess_Y$, are very similar to those for screening based on expected accident frequency. Excess accident frequency is defined as the difference between the SPF-predicted accident frequency and the EB-adjusted expected accident frequency. Steps 1 to 7 are identical to those used for screening based on expected accident frequency; thus, the excess expected accident frequency calculation are shown beginning at Step 8'.

Step 8': Calculate the excess accident frequency for all severity levels:

$$Excess_{Y(TOT)} = X_{Y(TOT)} - \kappa_{Y(TOT)} SL_{SUB} \tag{5.30}$$

$$Excess_{Y(FI)} = X_{Y(FI)} - \kappa_{Y(FI)} SL_{SUB} \tag{5.31}$$

$$Excess_{Y(PDO)} = Excess_{Y(TOT)} - Excess_{Y(FI)} \tag{5.32}$$

Step 9': Calculate the variance of the excess accident frequency for all severity levels:

$$Var\left(Excess_{Y(TOT)}\right) = Var\left(X_{Y(TOT)}\right) + \frac{1}{k_{TOT}}\left(\kappa_{Y(TOT)} SL_{SUB}\right)^2 \tag{5.33}$$

$$Var\left(Excess_{Y(FI)}\right)=Var\left(X_{Y(FI)}\right)+\frac{1}{k_{FI}}\left(\kappa_{Y(FI)}SL_{SUB}\right)^{2} \tag{5.34}$$

$$Var\left(Excess_{Y(PDO)}\right)=Var\left(Excess_{Y(TOT)}\right)+Var\left(Excess_{Y(FI)}\right) \tag{5.35}$$

Step 10': The relative weight of a given FI accident, RC_{FI}, is found in the same manner as Step 10 in Section 5.3.1. Then, excess expected EPDO accidents is calculated by:

$$Excess_{Y(EPDO)} = Excess_{Y(PDO)} + RC_{FI}\, Excess_{Y(FI)} \quad . \tag{5.36}$$

Step 11': The variance of the excess expected EPDO accident estimate is given by:

$$Var\left(Excess_{Y(EPDO)}\right)=Var\left(Excess_{Y(TOT)}\right)+\left(RC_{FI}-1\right)^{2}Var\left(Excess_{Y(FI)}\right) \tag{5.37}$$

Step 12': Calculate the average excess accident frequency of the given window:

$$Avg\left(Excess_{Y(TOT)}\right)=\frac{\sum_{SUB} Excess_{Y(TOT)}}{W} \tag{5.38}$$

$$Avg\left(Excess_{Y(FI)}\right)=\frac{\sum_{SUB} Excess_{Y(FI)}}{W} \tag{5.39}$$

$$Avg\left(Excess_{Y(PDO)}\right)=\frac{\sum_{SUB} Excess_{Y(PDO)}}{W} \tag{5.40}$$

$$Avg\left(Excess_{Y(EPDO)}\right)=\frac{\sum_{SUB} Excess_{Y(EPDO)}}{W} \tag{5.41}$$

Step 13': The variance for each window, and for each severity is given by:

$$Var\left[Avg\left(Excess_{Y(TOT)}\right)\right]=\frac{\sum_{SUB} Var\left(Excess_{Y(TOT)}\right)}{W^{2}} \tag{5.42}$$

$$Var\left[Avg\left(Excess_{Y(FI)}\right)\right]=\frac{\sum_{SUB} Var\left(Excess_{Y(FI)}\right)}{W^2}. \quad (5.43)$$

$$Var\left[Avg\left(Excess_{Y(PDO)}\right)\right]=\frac{\sum_{SUB} Var\left(Excess_{Y(PDO)}\right)}{W^2} \quad (5.44)$$

$$Var\left[Avg\left(Excess_{Y(EPDO)}\right)\right]=\frac{\sum_{SUB} Var\left(Excess_{Y(EPDO)}\right)}{W^2} \quad (5.45)$$

As before with $Avg(X_Y)$, the estimate of $Avg(Excess_Y)$ for each window is subjected to the statistical precision test, where the CV is calculated as the ratio of $\sqrt{\mathrm{Var}}$ over Avg of the appropriate statistic. From among all of the windows that pass the test, the window with the largest $Avg(Excess_Y)$ is selected, denoted $Peak(Excess_Y)$, and this is used to rank the entire roadway segment (i.e., site). The boundaries of the respective window are also included in the output.

If statistical significance is not achieved for any window of length $W=W_{MIN}+L_{INC}$, then $W=W_{MIN}+2L_{INC}$ is tried, and so on, until the window length is equal to the entire length of the roadway segment.

If there is still no window flagged, the site is not included in the list of ranked sites. The final result is a list of all sites passing the test ranked in order of the largest excess accident frequency.

5.4 Application to Washington Rural, 2-lane Highway Data

The peak-searching algorithm was performed on a set of HSIS data for 2-lane rural highways in Washington. Three years of accident data, from 1993 to 1995, were used for model prediction and EB estimation.

Sites were defined as contiguous highway segments that were homogeneous with respect to AADT, lane width, shoulder width, and other measured characteristics. Each site was divided into subsegments 0.01mi in length. This was the level of precision for geometric, traffic, and accident data for the Washington HSIS data.

The smallest allowable window size, W_{lim}, was taken to be 0.1mi, as it was perceived that this was a reasonable limit to the accuracy of accident locations. Thus, any site with a total length of less than 0.1mi was not included in the screening procedure. The increment size, L_{INC}, however, was set at 0.01mi. This had the effect of eliminating the need for dealing with the 'remainder' segment at the end of the site, at the expense of increased computing times.

5.5 Development of Safety Performance Functions

Safety performance functions were calibrated for the Washington highway data in the manner described in Chapter 4, using the GENMOD procedure in SAS. Once again, SPF model predictions were calculated using the following functional form:

$$\kappa = \alpha \cdot AADT^{\beta} \quad , \qquad \text{(5.46) (same as 4.1, 2.2)}$$

where κ is the predicted accident frequency in accidents/mi/yr.

The SPF parameter estimates for the current dataset are shown in Table 5.2 for both total and FI accidents.

Table 5.1: SPF parameter estimates for Washington 2-lane rural highways.

	SPF Parameter Estimates		
	α	β	k
Total Accidents	0.0012	0.87	2.0
FI Accidents	0.00058	0.87	2.1

The parameter estimates were used to calculate the predicted accident frequency, κ_y, for each site, for both total and FI accidents, and for each year. The estimates of the dispersion parameter, k, were used to calculate the EB weights used in the peak-searching algorithm.

5.6 Execution of the Peak-Searching Algorithm

The peak-searching algorithm was programmed using MATLAB® software. For simplicity, the value of k was assumed to be constant with respect to segment length and AADT.

The algorithm was used to screen 100 sites from the Washington 2-lane rural highway network for each of the four accident types (total, FI, PDO, and EPDO), and for both expected accident frequency X_Y, and excess accident frequency, $Excess_Y$.

The algorithm was run using different values of CV_{lim} to see what effect this had on the number of sites ranked, and the distribution of segment lengths for ranked sites.

The results of ranking for the expected accident frequency of total accidents is given in Table 5.3. Results for excess accident frequency of total accidents are given in Table 5.4. Results for other accident types are given in Appendix B. An example of the MATLAB peak-searching program is given in Appendix D.

Table 5.3a: Results of peak-searching algorithm for expected accident frequency of total accidents, using CV_{lim}=1.8.

Screening Criterion:	EB-adjusted expected accident frequency
Accident Type:	Total accidents
CVlim	1.8
Number of Sites Ranked:	98/100
Mean length of all ranked sites (mi):	1.78
Mean length of all flagged windows (mi):	0.11

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	14.8	42.1	0.44	1
6	0.25	0.1	14.2	47.6	0.49	2
22	0.89	0.1	13.4	34.6	0.44	3
12	1.18	0.1	12.0	44.8	0.56	4
23	0.13	0.1	11.3	30.2	0.49	5
5	0.64	0.1	10.9	36.8	0.56	6
26	2.4	0.1	10.4	25.6	0.49	7
29	5.37	0.1	10.1	24.4	0.49	8
7	0.41	0.1	9.83	30.1	0.56	9
48	0.24	0.1	9.48	28.1	0.56	10
18	0.8	0.1	9.47	28.0	0.56	11
43	0.25	0.1	8.87	24.6	0.56	12
25	3.34	0.1	8.09	20.4	0.56	13
16	0.84	0.1	7.14	23.1	0.67	14
15	0.13	0.1	7.11	22.9	0.67	15
17	1.32	0.1	7.11	22.9	0.67	16
11	0.72	0.1	6.82	21.1	0.67	17
50	0.57	0.1	6.50	19.2	0.67	18
45	0.18	0.1	6.35	18.3	0.67	19
54	1.2	0.1	6.18	17.3	0.67	20

Table 5.3b: Results of peak-searching algorithm for expected accident frequency of total accidents, using CV_{lim}=1.0.

Screening Criterion:	EB-adjusted expected accident frequency
Accident Type:	Total accidents
CVlim	1.0
Number of Sites Ranked:	90/100
Mean length of all ranked sites (mi):	1.91
Mean length of all flagged windows (mi):	0.14

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	14.8	42.1	0.44	1
6	0.25	0.1	14.2	47.6	0.49	2
22	0.89	0.1	13.4	34.6	0.44	3
12	1.18	0.1	12.0	44.8	0.56	4
23	0.13	0.1	11.3	30.2	0.49	5
5	0.64	0.1	10.9	36.8	0.56	6
26	2.4	0.1	10.4	25.6	0.49	7
29	5.37	0.1	10.1	24.4	0.49	8
7	0.41	0.1	9.83	30.1	0.56	9
48	0.24	0.1	9.48	28.1	0.56	10
18	0.8	0.1	9.47	28.0	0.56	11
43	0.25	0.1	8.87	24.6	0.56	12
25	3.34	0.1	8.09	20.4	0.56	13
16	0.84	0.1	7.14	23.1	0.67	14
15	0.13	0.1	7.11	22.9	0.67	15
17	1.32	0.1	7.11	22.9	0.67	16
11	0.72	0.1	6.82	21.1	0.67	17
50	0.57	0.1	6.50	19.2	0.67	18
45	0.18	0.1	6.35	18.3	0.67	19
54	1.2	0.1	6.18	17.3	0.67	20

Table 5.3c: Results of peak-searching algorithm for expected accident frequency of total accidents, using CV_{lim}=0.5.

Screening Criterion:	EB-adjusted expected accident frequency
Accident Type:	Total accidents
CVlim	0.5
Number of Sites Ranked:	65/100
Mean length of all ranked sites (mi):	2.44
Mean length of all flagged windows (mi):	0.39

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
4	2.8	0.1	16.1	49.6	0.44	1
7	0.41	0.1	16.0	49.0	0.44	2
12	1.18	0.1	15.7	58.7	0.49	3
47	2.7	0.1	14.8	42.1	0.44	4
6	0.25	0.1	14.2	47.6	0.49	5
9	3.27	0.1	13.5	43.5	0.49	6
22	0.89	0.1	13.4	34.6	0.44	7
14	0.82	0.1	13.2	41.3	0.49	8
37	2.07	0.1	13.0	32.6	0.44	9
5	0.64	0.11	13.0	40.1	0.49	10
18	0.8	0.1	12.4	36.7	0.49	11
17	1.32	0.11	12.4	36.3	0.49	12
41	1.56	0.1	12.1	34.7	0.49	13
54	1.2	0.1	11.8	33.1	0.49	14
49	0.3	0.11	11.8	32.7	0.49	15
56	6.2	0.1	11.5	31.6	0.49	16
2	0.62	0.11	11.3	30.4	0.49	17
23	0.13	0.1	11.3	30.2	0.49	18
20	1.76	0.1	11.1	29.2	0.49	19
25	3.34	0.1	10.6	26.8	0.49	20

Table 5.3d: Results of peak-searching algorithm for expected accident frequency of total accidents, using CV_{lim}=0.2.

Screening Criterion:	EB-adjusted expected accident frequency
Accident Type:	Total accidents
CVlim	0.2
Number of Sites Ranked:	13/100
Mean length of all ranked sites (mi):	5.07
Mean length of all flagged windows (mi):	2.20

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.5	14.2	8.09	0.20	1
4	2.8	0.79	10.0	3.91	0.20	2
12	1.18	1.06	8.87	3.13	0.20	3
9	3.27	1.33	6.22	1.50	0.20	4
40	6.87	1.16	5.31	1.11	0.20	5
20	1.76	1.53	4.33	0.745	0.20	6
25	3.34	1.48	4.27	0.728	0.20	7
29	5.37	1.68	3.79	0.544	0.19	8
39	3.93	1.84	3.13	0.380	0.20	9
56	6.2	3.13	2.22	0.195	0.20	10
97	7.52	5.89	0.913	0.0333	0.20	11
86	9.9	6.38	0.619	0.0153	0.20	12
67	10.9	8.34	0.403	0.00648	0.20	13

Table 5.4a: Results of peak-searching algorithm for excess accident frequency of total accidents, using CV_{lim}=4.0.

Screening Criterion:	Excess accident frequency
Accident Type:	Total accidents
CV_{lim}	4.0
Number of Sites Ranked:	77
Mean length of all ranked sites (mi):	2.10
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	$Peak(X_Y)$ (acc/mi/yr)	$Var[Peak(X_Y)]$	$CV[Peak(X_Y)]$	Rank
47	2.7	0.1	10.9	45.2	0.62	1
22	0.89	0.1	10.8	36.0	0.56	2
23	0.13	0.1	8.65	31.6	0.65	3
26	2.4	0.1	8.31	26.5	0.62	4
29	5.37	0.1	8.15	25.2	0.62	5
6	0.25	0.1	7.21	57.5	1.05	6
12	1.18	0.1	6.31	51.3	1.14	7
25	3.34	0.1	6.03	21.3	0.77	8
43	0.25	0.1	5.99	26.2	0.86	9
18	0.8	0.1	5.75	30.8	0.97	10
48	0.24	0.1	5.68	31.0	0.98	11
5	0.64	0.1	3.85	46.8	1.78	12
40	6.87	0.1	3.66	13.6	1.01	13
19	0.72	0.1	3.65	20.7	1.25	14
9	3.27	0.1	3.56	42.4	1.83	15
52	1.06	0.1	3.54	18.7	1.22	16
38	1.88	0.1	3.53	16.0	1.13	17
54	1.2	0.1	3.53	18.8	1.23	18
53	0.92	0.1	3.53	18.8	1.23	19
96	4.5	0.1	3.50	9.74	0.89	20

Table 5.4b: Results of peak-searching algorithm for excess accident frequency of total accidents, using CV_{lim}=2.5.

Screening Criterion:	Excess accident frequency
Accident Type:	Total accidents
CV_{lim}	2.5
Number of Sites Ranked:	74
Mean length of all ranked sites (mi):	2.15
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	10.9	45.2	0.62	1
22	0.89	0.1	10.8	36.0	0.56	2
23	0.13	0.1	8.65	31.6	0.65	3
26	2.4	0.1	8.31	26.5	0.62	4
29	5.37	0.1	8.15	25.2	0.62	5
6	0.25	0.1	7.21	57.5	1.05	6
12	1.18	0.1	6.31	51.3	1.14	7
25	3.34	0.1	6.03	21.3	0.77	8
43	0.25	0.1	5.99	26.2	0.86	9
18	0.8	0.1	5.75	30.8	0.97	10
48	0.24	0.1	5.68	31.0	0.98	11
32	1.03	0.1	5.64	18.5	0.76	12
10	2.87	0.1	4.79	40.4	1.33	13
2	0.62	0.1	4.65	32.6	1.23	14
5	0.64	0.1	3.85	46.8	1.78	15
40	6.87	0.1	3.66	13.6	1.01	16
56	6.2	0.1	3.65	17.7	1.15	17
19	0.72	0.1	3.65	20.7	1.25	18
37	2.07	0.1	3.63	14.5	1.05	19
9	3.27	0.1	3.56	42.4	1.83	20

Table 5.4c: Results of peak-searching algorithm for excess accident frequency of total accidents, using CV_{lim}=1.5.

Screening Criterion:	Excess accident frequency
Accident Type:	Total accidents
CV_{lim}	1.5
Number of Sites Ranked:	62/100
Mean length of all ranked sites (mi):	2.34
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	10.9	45.2	0.62	1
22	0.89	0.1	10.8	36.0	0.56	2
7	0.41	0.1	9.32	58.0	0.82	3
4	2.8	0.1	9.05	59.6	0.85	4
14	0.82	0.1	8.82	45.2	0.76	5
23	0.13	0.1	8.65	31.6	0.65	6
26	2.4	0.1	8.31	26.5	0.62	7
29	5.37	0.1	8.15	25.2	0.62	8
6	0.25	0.1	7.21	57.5	1.05	9
9	3.27	0.1	6.77	52.7	1.07	10
12	1.18	0.1	6.31	51.3	1.14	11
25	3.34	0.1	6.03	21.3	0.77	12
5	0.64	0.11	6.00	49.2	1.17	13
43	0.25	0.1	5.99	26.2	0.86	14
18	0.8	0.1	5.75	30.8	0.97	15
48	0.24	0.1	5.68	31.0	0.98	16
32	1.03	0.1	5.64	18.5	0.76	17
17	1.32	0.1	5.54	38.0	1.11	18
49	0.3	0.1	5.52	33.8	1.05	19
50	0.57	0.1	5.41	31.2	1.03	20

Table 5.4d: Results of peak-searching algorithm for excess accident frequency of total accidents, using CV_{lim}=1.0.

Screening Criterion:	Excess accident frequency
Accident Type:	Total accidents
CV_{lim}	1.0
Number of Sites Ranked:	45
Mean length of all ranked sites (mi):	2.61
Mean length of all flagged windows (mi):	0.11

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
37	2.07	0.1	11.1	33.3	0.52	1
47	2.7	0.1	10.9	45.2	0.62	2
22	0.89	0.1	10.8	36.0	0.56	3
12	1.18	0.1	10.0	65.3	0.80	4
7	0.41	0.1	9.32	58.0	0.82	5
4	2.8	0.1	9.05	59.6	0.85	6
14	0.82	0.1	8.82	45.2	0.76	7
23	0.13	0.1	8.65	31.6	0.65	8
9	3.27	0.11	8.53	53.1	0.85	9
20	1.76	0.1	8.35	30.7	0.66	10
26	2.4	0.1	8.31	26.5	0.62	11
29	5.37	0.1	8.15	25.2	0.62	12
6	0.25	0.12	7.77	49.5	0.91	13
39	3.93	0.1	7.75	21.5	0.60	14
17	1.32	0.11	7.59	40.6	0.84	15
49	0.3	0.11	7.47	36.2	0.81	16
41	1.56	0.1	7.08	27.4	0.74	17
19	0.72	0.1	6.59	29.4	0.82	18
2	0.62	0.11	6.52	34.6	0.90	19
56	6.2	0.1	6.39	25.2	0.79	20

Table 5.4e: Results of peak-searching algorithm for excess accident frequency of total accidents, using CV_{lim}=0.5.

Screening Criterion:	Excess accident frequency
Accident Type:	Total accidents
CV_{lim}	0.5
Number of Sites Ranked:	13
Mean length of all ranked sites (mi):	3.46
Mean length of all flagged windows (mi):	0.14

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
7	0.41	0.1	24.7	105	0.42	1
4	2.8	0.1	21.4	97.7	0.46	2
39	3.93	0.1	16.7	41.5	0.39	3
47	2.7	0.1	16.6	61.4	0.47	4
20	1.76	0.1	16.3	51.5	0.44	5
40	6.87	0.1	13.4	37.2	0.46	6
41	1.56	0.12	12.8	36.6	0.47	7
29	5.37	0.12	10.6	25.9	0.48	8
33	0.91	0.13	9.74	22.1	0.48	9
37	2.07	0.21	6.97	10.9	0.47	10
25	3.34	0.21	6.87	11.1	0.49	11
26	2.4	0.29	5.25	6.52	0.49	12
67	10.9	0.16	3.99	3.78	0.49	13

5.7 Discussion of Peak-Searching Results

When screening for either expected accident frequency or excess accident frequency, and for any severity, fewer sites were flagged when the value of CV_{lim} was lowered, as was expected.

It is clear that, as $\mathrm{CV}_{\mathrm{lim}}$ is lowered, the average length of ranked sites increases; thus, small values of CVlim favour sites with relatively long segment lengths. This is owing to the fact that longer segments exhibit less variance than shorter sites.

Also, the average window length for flagging a given site increases as $\mathrm{CV}_{\mathrm{lim}}$ is decreased. It should be noted that computing time is substantially longer for small values of $\mathrm{CV}_{\mathrm{lim}}$, as compared with larger values.

6 Screening for High Proportions of Specific Accident Types

6.1 Introduction

Bayesian network screening methods most commonly identify sites with promise based on some measure of accident frequency. These methods, including the peak-searching algorithm described in the preceding chapter, are data-intensive in that they require traffic volume data at a minimum; however, these data are sometimes unavailable.

An alternative to screening for high accident frequencies is to screen the network for high proportions of specific accident types. A site with an unusually high proportion of a certain accident type could be a candidate for safety countermeasures specific to that accident type. Accident types may be simple, such as head-on or wet-pavement, or may be compound, such as night-time run-off-road crashes. For example, a road segment with an unusually high number of opposite-direction crashes could be a candidate for the installation of centreline rumble strips.

One of the advantages of this method is that traffic volume data are not required as they are for SPF-based methods; however, this may also be a disadvantage, as traffic volume is an important safety variable.

The method of screening for high proportions of specific accident type, or simply 'screening for proportions,' is described below.

6.2 Empirical Bayes Analysis Using Beta-Binomial Models

6.2.1 Theoretical Framework

Heydecker and Wu (*2*) devised an empirical Bayes (EB) approach to screening for high proportions of specific accident types. It is assumed that whether or not a given accident is of a particular type can be modelled as a Bernoulli trial. Thus, for any site, i, over a given period of time, the count of target accidents, x_i, out of n_i total accidents has a binomial distribution. The binomial distribution, with mean parameter θ, is written as:

$$f(x_i | n_i, \theta) = \binom{n_i}{x_i} \theta^{x_i} (1-\theta)^{n_i - x_i}, \quad 0 \le x_i \le n_i \,, \tag{6.1}$$

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

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad . \tag{6.2}$$

In present context, θ represents the proportion of all accidents that are of a particular type. It is assumed that θ is well defined for each site, but is unknown. Any accident type may be considered, provided that the appropriate data are available. The value of θ is assumed to be fixed for each site, but to vary among sites.

Heydecker and Wu (2) postulated that the distribution of θ among sites could be modelled using a beta distribution. The distribution of θ among sites corresponds to the Bayesian prior distribution; that is, it represents *a priori* knowledge of the process that governs the observed proportions. The beta distribution is written:

$$g_b(\theta \mid \alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, \quad 0 < \theta < 1, \tag{6.3}$$

where α and β are strictly positive parameters, and $\mathrm{B}(\alpha,\beta)$ is the beta function. The subscript b denotes *before*, indicating that the information is *a priori.* The beta function is represented by the upper-case Greek letter beta, and is defined as:

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad , \tag{6.4}$$

where $\Gamma(.)$ is the gamma function, given by:

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt \quad . \tag{6.5}$$

The beta distribution offers several features that make it a good choice of Bayesian prior. First, it is defined on the open interval (0,1); because accidents proportions are defined on the closed interval [0,1], no transformation of variables is necessary.

Second, the beta distribution can have a wide variety of shapes, making it very versatile. Figure 6.1 shows some of the possible shapes. Note that when α=β=1, the uniform distribution results.

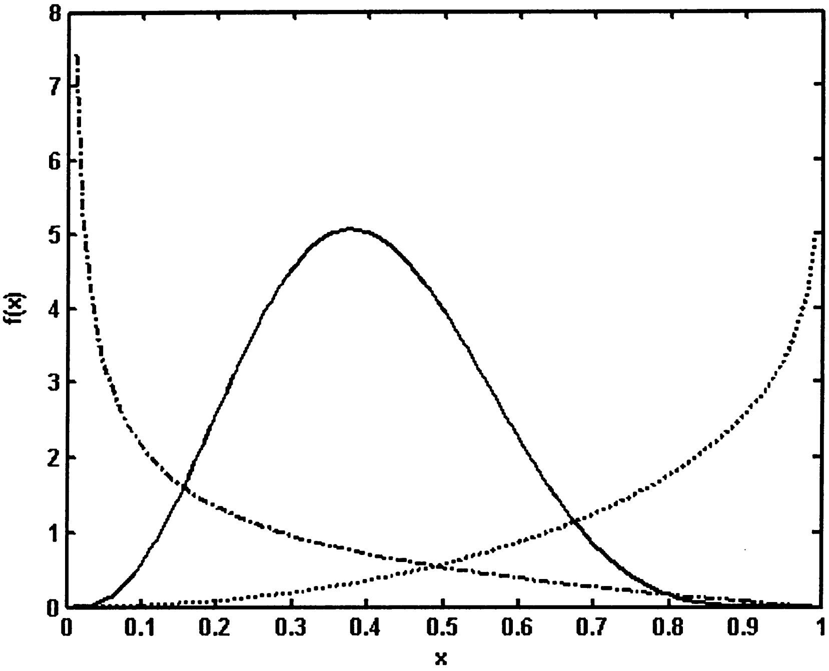


Fig. 6.1: Some of the different shapes of the beta distribution.

Finally, the beta distribution is the natural conjugate prior of the binomial distribution. In theory, any distribution yielding a good fit to the data could be used; however, by using the natural conjugate prior, the EB calculations are much more tractable.

The mean of the prior beta distribution, *E(θ)*, is given by:

$$E(\theta) = \frac{\alpha}{\alpha + \beta} \quad , \tag{6.6}$$

and the variance is:

$$Var(\theta) = \frac{\alpha\beta}{\left[(\alpha + \beta)^2 (\alpha + \beta + 1)\right]} \quad , \tag{6.7}$$

which can also be expressed as:

$$Var(\theta) = \frac{E(\theta)[1 - E(\theta)]}{\alpha + \beta + 1} \quad . \tag{6.8}$$

The binomial distribution of the observed accidents, shown in Eq. 6.1, can be combined with the beta distribution of θ over similar sites that is given in Eq. 6.3. This leads to the unconditional beta-binomial distribution of x_i given n_i, α and β. Note that θ, which is unknown, has been eliminated. The beta-binomial distribution is written:

$$h(x_i \mid n_i, \alpha, \beta) = \binom{n_i}{x_i} \frac{\mathrm{B}(\alpha + x_i, \beta + n_i - x_i)}{\mathrm{B}(\alpha, \beta)} \qquad (6.9)$$

It is now possible to determine the posterior distribution of θ by employing Bayes' theorem. Bayes' theorem can here be written as:

$$g_a(\theta \mid n_i, x_i, \alpha, \beta) = \frac{f(x_i \mid n_i, \theta) \cdot g_b(\theta \mid \alpha, \beta)}{h(x_i \mid n_i, \alpha, \beta)} \qquad (6.10)$$

Here, the subscript a denotes *after*, or *a posteriori*. After substitutions, the posterior distribution becomes:

$$g_a(\theta \mid \alpha + x_i, \beta + n_i - x_i) = \frac{\theta^{\alpha + x_i - 1}(1-\theta)^{\beta + n_i - x_i - 1}}{\mathrm{B}(\alpha + x_i, \beta + n_i - x_i)} \quad , \quad 0 < \theta < 1 \qquad (6.11)$$

The expression for the posterior distribution may be further simplified by introducing the posterior parameters, α' and β', expressed as:

$$\alpha' = \alpha + x_i \quad , \qquad (6.12)$$

$$\beta' = \beta + n_i - x_i \quad . \qquad (6.13)$$

Substituting α' and β' in Eq. 6.11 yields:

$$g_a(\theta \mid \alpha', \beta') = \frac{\theta^{\alpha' - 1}(1-\theta)^{\beta' - 1}}{\mathrm{B}(\alpha', \beta'_i)} \quad , \quad 0 < \theta < 1 \quad , \qquad (6.14)$$

which is clearly a beta distribution of the form of Eq. 6.3.

It is now possible to evaluate the expected value of the posterior, $E(\theta_i)$, for each site, i, as in Eq. 6.6. Thus,

$$E(\theta_i) = \frac{\alpha'}{\alpha' + \beta'} \quad , \qquad (6.15)$$

and, similar to Eq. 6.7, the posterior variance for each site is given as:

$$Var(\theta_i) = \frac{\alpha'\beta'}{\left[(\alpha'+\beta')^2(\alpha'+\beta'+1)\right]} \quad . \tag{6.16}$$

6.2.2 Bayesian Analysis

The posterior distribution represents the state of knowledge of θ after the prior distribution has been 'updated' by the observational distribution. In this case, the parameter θ is the proportion of accidents of a certain type, the prior (Eq. 6.3) is beta-distributed and represents the distribution of θ among similar sites, and the observational distribution is the site-specific binomial distribution (Eq. 6.1), based on observed accident patterns.

For the purposes of network screening, it is necessary to compare individual sites with others of the same type within the area of study. This is accomplished in two steps. First, the critical proportion of the beta prior, θ*, is found by solving the following:

$$\int_0^{\theta=\theta^*} g_b(\theta|\,\alpha,\beta)\,d\theta = \pi \quad , \tag{6.17}$$

where π is defined as the percentage of all sites whose proportion of the accident type in question is less than the value of θ*, and is chosen by the analyst to be some value between 0 and 1. To illustrate, if π is taken to be 0.8, then 80% of the sites in the network would be expected to have a proportion, θ, of that accident type, less than θ*. Heydecker and Wu (*2*), Sayed et al. (*31*), Bolduc and Bonin (*32, 33*), and Mollett (*34*) all used a value of 0.5 for π, which corresponds to the median; however, any value of π may be used.

The beta distribution in Fig. 6.2 has been evaluated at its median using Eq. 6.17; the shaded area represents the value of π. The critical proportion is 0.37; thus, 50% of the sites should experience an accident proportion of 0.37 or less for accident type.

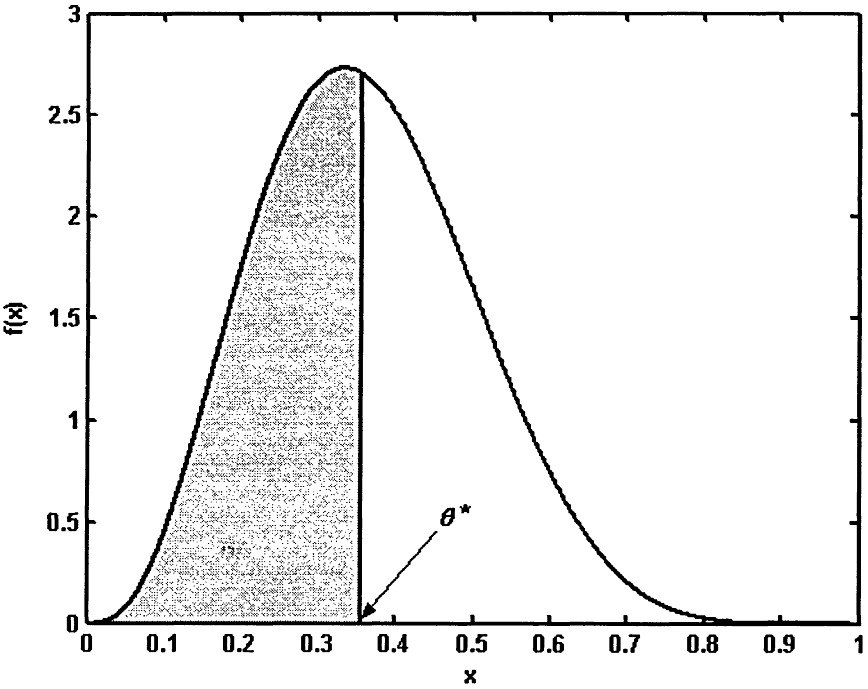


Fig. 6.2: A beta distribution. The shaded area represents the value of π.

Once $\theta*$ is known, the probability that the accident proportion, θ_i, for any site exceeds $\theta*$ can be calculated by the following expression:

$$\Pr(\theta_i > \theta *) = \int_{\theta=\theta*}^{1} g_a(\theta | \alpha', \beta') d\theta \ . \tag{6.18}$$

The integral in Eq. 6.18 represents the area of the posterior beta distribution that is greater than the value of $\theta*$. This value is sometimes called the *pattern score*, and is shown by the shaded area in Fig. 6.3.

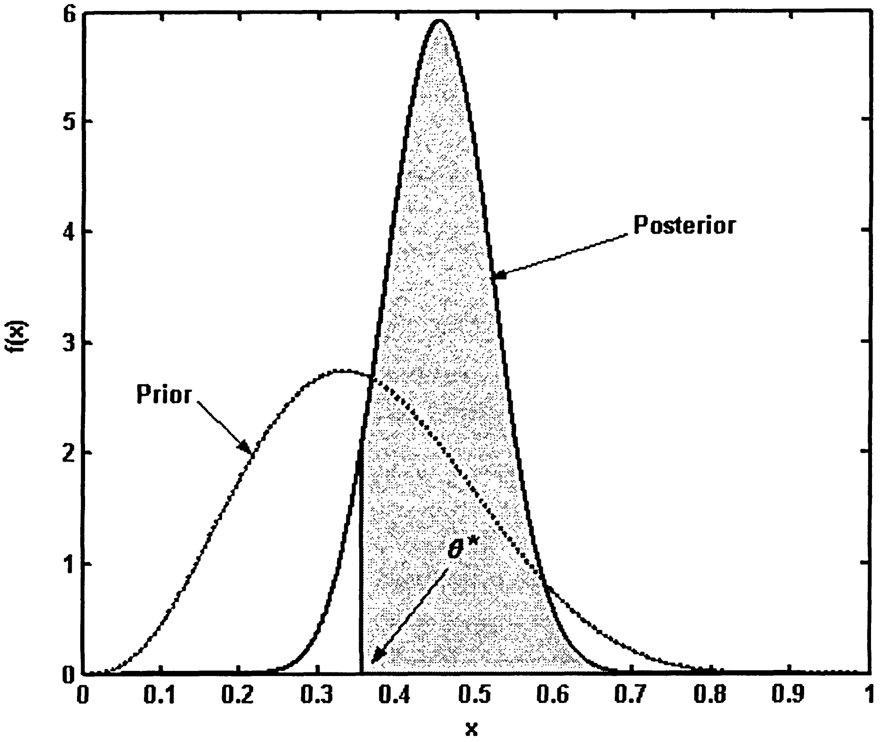


Fig. 6.3: Prior and posterior beta distributions. The shaded area represents the probability that the posterior distribution exceeds θ*.

The pattern score is the degree of belief that a given site experiences a greater proportion of accidents than the one specified by θ*. A pattern score of one indicates that it is almost certain that a site is, in fact, experiencing a relatively high proportion of a given accident type; a pattern score of zero indicates that this is almost certainly not true, with 0.5 being neutral. Thus, for the purposes of network screening, the site with highest pattern score is ranked first.

Only those sites that have a pattern score greater than some critical value, δ, are ranked. Sites with high accident counts are expected to exert more influence over the prior than those with low counts. This is because sites with relatively high counts will have relatively low variances with respect to the observational binomial distribution.

6.3 Parameter Estimation Methodology for Beta Prior Distribution

6.3.1 General

In order to apply Bayes' theorem, the parameters of the beta prior distribution must be estimated by statistical methods. The two most common methods of estimation are the method of moments (MM) and the method of maximum likelihood (ML). Other methods include weighted least-squares regression, and Markov-chain, Monte Carlo simulation.

Here, three methods of parameter estimation are compared: the method of maximum likelihood, and two versions of the method of moments.

6.3.2 Maximum Likelihood Estimates

Maximum likelihood estimation has been used by Heydecker and Wu (*2*); Bolduc and Bonin (*32,33*); and Mollett (*34*) for the calibration of beta priors. In each case, the data were groups of intersections with relatively small sample sizes, and models were calibrated for various accident characteristics.

Maximum likelihood estimates are found by maximizing the likelihood function of the sample data. The likelihood function, L_i, represents the probability of observing the sample data, given the chosen sample distribution, which is in this case the beta distribution. The likelihood function is derived from Eq. 6.9:

$$L_i = \binom{n_i}{x_i} \frac{\mathrm{B}(\alpha + x_i, \beta + n_i - x_i)}{\mathrm{B}(\alpha, \beta)} \quad . \tag{6.19}$$

The ML procedure is usually simplified by taking the logarithm of the likelihood function. This is the *log-likelihood function*, and it is expressed as:

$$\log(L_i) = \log\binom{n_i}{x_i} + \log[\mathrm{B}(\alpha + x_i, \beta + n_i - x_i)] - \log[\mathrm{B}(\alpha, \beta)] \quad . \tag{6.20}$$

The values of α and β that maximize Eq. 6.20 are the maximum likelihood estimates, $\hat{\alpha}$ and $\hat{\beta}$. The caret (^) indicates that the values are estimates, rather than true values.

By applying Eq. 6.4 to the terms in equation Eq. 6.20, the following expressions can be shown:

$$\log[\mathrm{B}(\alpha + x_i, \beta + n_i - x_i)] = \log[\Gamma(\alpha + x_i)] + \log[\Gamma(\beta + n_i - x_i)] - \log[\Gamma(\alpha + \beta + n_i)] \tag{6.21}$$

and

$$\log[\mathrm{B}(\alpha, \beta)] = \log[\Gamma(\alpha)] + \log[\Gamma(\beta)] - \log[\Gamma(\alpha + \beta)] \quad . \tag{6.22}$$

Substituting Equations 6.21 and 6.22 into Eq. 6.20, the log-likelihood function becomes:

$$\begin{aligned}\log(L_i) = \log[\Gamma(\alpha + x_i)] + \log[\Gamma(\beta + n_i - x_i)] - \log[\Gamma(\alpha + \beta + n_i)] \\ - (\log[\Gamma(\alpha)] + \log[\Gamma(\beta)] - \log[\Gamma(\alpha + \beta)])\end{aligned} \tag{6.23}$$

The binomial coefficient is omitted from Eq. 6.23 since it is not a function of the beta parameters, and is thus a constant.

Mollett (*34*) demonstrated a method of maximizing Eq. 6.23 using the Solver® tool in Microsoft Excel®, to yield ML beta prior estimates, $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$.

6.3.3 Method of Moments – Method 1 (MM1)

The first method of moments considered is the simplest of the three approaches described in this thesis. The beta parameters, α and β, can be used to determine the moments (i.e., mean and variance) of the beta distribution, as shown in Equations 6.6 and 6.8. If moments can be estimated from the sample data, these sample moments can be expressed in terms of the parameters; thus, $\hat{\alpha}_{MM1}$ and $\hat{\beta}_{MM1}$ can be estimated by solving the resulting equations.

The first step is to calculate the observed proportion of accidents, θ_i, for each site i, $i=1,2,\ldots,m$, where m is the number of sites. In this case, $\theta_i=x_i/n_i$, $n_i \geq 1$, where x_i and n_i are the count of accidents of the type of interest, and the total number of accidents, respectively, that have been observed at site i within the given study period. The sample mean, $\bar{\theta}$, and the sample variance, s^2, are calculated using the following:

$$\bar{\theta} = \frac{\sum_{i=1}^{m} \theta_i}{m} \quad , \tag{6.24}$$

and

$$s^2{}_{MM1} = \frac{\sum_{i=1}^{m} \left(\theta_i - \bar{\theta}\right)^2}{m-1} \; . \tag{6.25}$$

where the subscript MM1 indicates that the parameters were estimated using the first method of moments.

If the sample size is sufficiently large, $\bar{\theta}$ can be substituted for $E(\theta)$ in Eq. 6.6, while s^2 is substituted for $Var(\theta)$ in Eq. 6.8. The result is two equations with the two unknown parameters, α and β. Solving the equations leads to the following estimates for α and β:

$$\hat{\alpha}_{MM1} = \bar{\theta}\left[\frac{\left(\bar{\theta} - \bar{\theta}^2\right)}{s^2{}_{MM1}} - 1\right] \quad , \tag{6.26}$$

and

$$\hat{\beta}_{MM1} = \left(1 - \bar{\theta}\right)\left[\frac{\left(\bar{\theta} - \bar{\theta}^2\right)}{s^2} - 1\right] \quad . \tag{6.27}$$

The advantage of this technique is that the solutions to Equations 6.25 and 6.26 are straightforward; no iterative routine is required.

6.3.4 Method of Moments – Method 2 (MM2)

A variation of the above method of moments was described by Sayed et al. (*31*), and applied to a sample of intersection crash data. The method uses the same principles as the first method of moments, in that the sample moments are substituted for the 'true' distribution moments.

The difference between the two methods is in the way that the sample variance, s^2, is calculated. For the second method, the sample observations are considered to be a set of paired values, (x_i,n_i), rather than a univariate ratio, θ_i, as in the first method. Thus, the mean is calculated by Eq. 6.24, as in Method 1, and the variance is:

$$s^2_{MM2} = \frac{1}{m-1}\left[\sum_{i=1}^{m}\left(\frac{x_i^2 - x_i}{n_i^2 - n_i}\right) - \frac{1}{m}\left(\sum_{i=1}^{m}\frac{x_i}{n_i}\right)^2\right] \quad , \quad n \geq 2 \quad . \tag{6.28}$$

Substituting the sample mean, $\overline{\theta}$, for $E(\theta)$ in Eq. 6.6, and s^2 for $Var(\theta)$ in Eq. 6.8, the following useful expressions are obtained:

$$s^2_{MM2} = \frac{\frac{\alpha^2}{\overline{\theta}} - \alpha^2}{\left(\frac{\alpha}{\overline{\theta}}\right)^2\left(\frac{\alpha}{\overline{\theta}}+1\right)} \quad , \tag{6.29}$$

$$\beta = \frac{\alpha}{\overline{\theta}} - \alpha \quad . \tag{6.30}$$

Equation 6.29 can be solved for α using the Solver tool in Microsoft Excel. The solution is the parameter estimate $\hat{\alpha}_{MM2}$, which can then be substituted into Eq. 6.30 to give the remaining parameter estimate, $\hat{\beta}_{MM2}$.

Sayed et al. (*31*) argue that Method 2 (MM2) yields beta priors with less variance than those estimated by Method 1 (MM1).

6.4 Application to HSIS Data

The screening for proportions method of network screening was applied to three HSIS databases. The first dataset consisted of 2202 rural, TWSC intersections in California, with accident data from 1997-2002. The second dataset had 108 rural, signalized intersections in California, also with data for 1997-2002. The third dataset was for 831 2-mile segments of 2-lane rural highway in rolling terrain, in Washington state. The Washington crash data were for 3 years, 1993-1995. Table 6.1 gives a brief summary of each dataset.

Table 6.1: Summary of databases.

Set	State	Type of Site	No. Sites	Years	No. Accs.
1	CA	Rural, 4-leg TWSC Intersections	2202	1997-2001	10337
2	CA	Rural, 4-leg Signalized Intersections	108	1997-2001	2565
3	WA	Rural, 2-lane Highway in Rolling Terrain (2-mile segments)	831	1993-1995	3289

For each dataset, beta prior distributions were calibrated for each possible type of accident using each of the three parameter estimation methods described above. Dataset 1 was analyzed with data for the full 5 years of data; dataset 2 was calibrated using both 3 years (1997-1999) and 5 years (1997-2001) for comparison. Dataset 3 was examined for the given 3 years of data.

Table 6.2 shows the accident types that were considered for each dataset, as well as the number and proportion of each accident type.

The results of the parameter estimation techniques are shown in Tables 6.3 to 6.9.

Table 6.2a: Summary of target accident types for California TWSC Intersections

Total Accidents = Σn_i= 10337

Years of Data: 5 (1997-2001)

	Observed Target Accidents	Observed Proportion
Accident Type	**Σx_i**	**$\Sigma x_i/\Sigma n_i$**
Fatal/Injury (FI)	2584	0.25
Head-On	377	0.036
Sideswipe	918	0.089
Rear-End	2274	0.22
Broadside	4544	0.44
Hit Object	1338	0.13
Overturning	260	0.025
Pedestrian	111	0.011
Other/Unknown	515	0.050

Table 6.2b: Summary of target accident types for California signalized intersections.

	5 years (1997-2001)		3 years (1997-1999	
	Total Accidents = Σn_i= 2565		Total Accidents = Σn_i= 1441	
	Observed Target Accidents	Observed Proportion	Observed Target Accidents	Observed Proportion
Accident Type	Σx_i	$\Sigma x_i/\Sigma n_i$	Σx_i	$\Sigma x_i/\Sigma n_i$
Fatal/Injury (FI)	406	0.16	239	0.17
Head-On	128	0.050	73	0.051
Sideswipe	266	0.10	158	0.11
Rear-End	1141	0.44	624	0.43
Broadside	747	0.29	429	0.30
Hit Object	151	0.059	80	0.056
Overturning	28	0.011	16	0.011
Pedestrian	27	0.011	15	0.010
Other/Unknown	77	0.030	46	0.032

Table 6.2c: Summary of target accident types for Washington 2-lane rural highways (2-mile segments).

Total Accidents = Σn_i= 3289

Years of Data: 3 (1993-1995)

	Observed Target Accidents	Observed Proportion
Accident Type	Σx_i	$\Sigma x_i/\Sigma n_i$
Head-On	51	0.016
Angle	1	0.00030
Sideswipe, same direction	47	0.014
Sideswipe, opposite direction	108	0.033
Animal	315	0.096
Bicycle	5	0.0015
Pedestrian	1	0.00030
Parked vehicle	27	0.0082
Overturning	830	0.25
Hit fixed object	1331	0.40
Other multi-vehicle	317	0.096
Other single-vehicle	99	0.030
Other/Unknown	16	0.0049

Table 6.3: Maximum likelihood beta prior estimates for California rural, 4-leg signalized intersections.

Database: HSIS, California
Site Type: Rural, 4-leg, Signalized Intersections
Total Sites: 108
Total Accidents: 2565 (5 years) and 1441 (3 years)
Years of Data: 5 (1997-2001), and 3 (1997-1999)

	5 yrs (1997-2001)					3 yrs (1997-1999)				
Accident Type	**Observed Proportion**	**Maximum Likelihood Estimates**				**Observed Proportion**	**Maximum Likelihood Estimates**			
	(=Σx/Σn)	**α**	**β**	**E(θ)**	**Var(θ)**	**(=Σx/Σn)**	**α**	**β**	**E(θ)**	**Var(θ)**
FI	0.16	10.1	53.5	0.16	0.0021	0.17	15.5	78.4	0.17	0.0015
Head-on	0.050	3.72	71.0	0.050	0.00062	0.051	3.60	67.5	0.051	0.00067
Sideswipe	0.10	3.41	28.9	0.11	0.0028	0.11	3.32	26.6	0.11	0.0032
Rear-end	0.44	4.42	6.30	0.41	0.021	0.043	5.00	7.21	0.41	0.018
Broadside	0.29	3.73	8.38	0.31	0.016	0.30	3.27	7.21	0.31	0.019
Hit Object	0.059	2.86	43.0	0.062	0.0012	0.056	1.83	29.8	0.058	0.0017
Overturning	0.011	1.16	105	0.011	0.00010	0.011	0.381	30.8	0.012	0.00038
Pedestrian	0.011	0.163	11.6	0.014	0.0011	0.010	0.136	11.3	0.012	0.00094
Other/Unk.	0.030	3.73	117	0.031	0.00025	0.032	2.64	80.3	0.032	0.00037

Table 6.4: MM1 beta prior estimates for California rural, 4-leg signalized intersections.

Database: HSIS, California
Site Type: Rural, 4-leg, Signalized Intersections
Total Sites: 108
Total Accidents: 2565 (5 years) and 1441 (3 years)
Years of Data: 5 (1997-2001), and 3 (1997-1999)

Accident Type	5 yrs (1997-2001)					3 yrs (1997-1999)				
	Observed Proportion	Method of Moments Estimates – Method 1				Observed Proportion	Method of Moments Estimates – Method 1			
	(=Σx/Σn)	α	β	E(θ)	Var(θ)	(=Σx/Σn)	α	β	E(θ)	Var(θ)
FI	0.16	1.31	7.11	0.16	00014	0.17	0.710	3.73	0.16	0.025
Head-on	0.050	0.607	11.8	0.049	0.0035	0.051	0.377	7.53	0.048	0.0051
Sideswipe	0.10	0.993	8.38	0.11	0.0091	0.11	0.475	3.67	0.11	0.020
Rear-end	0.44	2.05	3.06	0.40	0.039	0.043	1.41	2.07	0.41	0.054
Broadside	0.29	1.78	3.85	0.32	0.033	0.30	1.38	3.13	0.31	0.039
Hit Object	0.059	0.558	7.46	0.070	0.0072	0.056	2.31	30.5	0.059	0.0019
Overturning	0.011	0.106	9.39	0.011	0.0010	0.011	0.0892	6.51	0.014	0.0018
Pedestrian	0.011	0.0942	5.72	0.016	0.0023	0.010	0.071	5.62	0.012	0.0018
Other/Unk.	0.030	0.425	13.1	0.032	0.0021	0.032	0.258	8.85	0.028	0.0027

Table 6.5: MM2 beta prior estimates for California rural, 4-leg signalized intersections.

Database: HSIS, California
Site Type: Rural, 4-leg, Signalized Intersections
Total Sites: 107
Total Accidents: 2565 (5 years) and 1441 (3 years)
Years of Data: 5 (1997-2001), and 3 (1997-1999)

	5 yrs (1997-2001)					3 yrs (1997-1999)				
Accident Type	Observed Proportion	Method of Moments Estimates – Method 2				Observed Proportion	Method of Moments Estimates – Method 2			
	(=Σx/Σn)	α	β	E(θ)	Var(θ)	(=Σx/Σn)	α	β	E(θ)	Var(θ)
FI	0.16	4.61	24.7	0.16	0.0044	0.17	4.90	27.3	0.16	0.0039
Head-on	0.050	41.8	804	0.049	5.5E-05	0.051	3.11	61.4	0.048	0.00070
Sideswipe	0.10	5.66	47.3	0.11	0.0018	0.11	1.50	11.5	0.12	0.0073
Rear-end	0.44	3.92	5.78	0.40	0.023	0.43	3.51	5.07	0.41	0.025
Broadside	0.29	4.65	10.4	0.31	0.013	0.30	5.09	11.4	0.31	0.012
Hit Object	0.059	2.31	30.5	0.070	0.0019	0.055	2.83	42.2	0.063	0.0013
Overturning	0.011	0.340	30.0	0.011	0.00035	0.011	0.743	56.7	0.014	0.00024
Pedestrian	0.011	0.381	22.9	0.016	0.00066	0.010	0.503	39.4	0.013	0.00030
Other/Unk.	0.030	6.41	195	0.032	0.00015	0.032	3.58	121	0.029	0.00022

Table 6.6: ML beta prior estimates for California rural, 4-leg TWSC intersections.

Database: HSIS, California
Site Type: Rural, TWSC Intersections
Total Sites: 2202
Total Accidents: 10337
Years of Data: 5 (1997-2001)

Accident Type	Observed Prop.	Maximum Likelihood Estimates			
	Σx/Σn	α	β	E(θ)	Var(θ)
FI	0.25	5.47	16.3	0.25	0.0083
Head-on	0.036	3.04	80.0	0.037	0.00042
Sideswipe	0.089	2.26	22.4	0.092	0.0032
Rear-end	0.22	1.48	5.33	0.22	0.022
Broadside	0.44	1.66	2.63	0.39	0.045
Hit Object	0.13	1.18	6.50	0.15	0.015
Overturning	0.025	0.945	34.1	0.027	0.00073
Pedestrian	0.011	0.367	32.1	0.011	0.00033
Other/Unk.	0.050	1.58	27.5	0.054	0.0017

Table 6.7: Method of moments beta prior estimates for California rural, 4-leg TWSC intersections.

Database: HSIS, California
Site Type: Rural, TWSC Intersections
Total Sites: 2202
Total Accidents: 10337
Years of Data: 5 (1997-2001)
Notes: 1692 sites used for Method 1 (≥1 acc.); 1308 sites used for Method 2 (≥2 acc.)

Accident Type	Observed Proportion	Method of Moments Estimates – Method 1				Method of Moments Estimates – Method 2			
	(=Σx/Σn)	α	β	E(θ)	Var(θ)	α	β	E(θ)	Var(θ)
FI	0.25	0.332	0.958	0.26	0.084	7.22	20.8	0.26	0.0066
Head-on	0.036	0.0562	1.58	0.034	0.013	2.23	59.5	0.036	0.00056
Sideswipe	0.089	0.112	1.00	0.10	0.043	1.57	14.8	0.096	0.0050
Rear-end	0.22	0.252	0.971	0.21	0.074	1.95	7.06	0.22	0.017
Broadside	0.44	0.362	0.654	0.36	0.11	1.86	3.13	0.37	0.039
Hit Object	0.13	0.156	0.683	0.19	0.082	1.21	6.02	0.17	0.017
Overturning	0.025	0.0311	0.874	0.034	0.017	0.334	10.5	0.031	0.0025
Pedestrian	0.011	0.0169	1.40	0.012	0.0049	1.13	89.5	0.012	0.00013
Other/Unk.	0.050	0.0744	0.987	0.070	0.032	1.74	24.0	0.068	0.0024

Table 6.8: ML beta prior estimates for 2-mile segments of 2-lane rural highway in Washington.

Database: HSIS, Washington
Site Type: Rural, 2-lane Highways, Rolling Terrain – 2-mile Segments
Total Sites: 831
Total Accidents: 3289
Years of Data: 3 (1997-1999)

Accident Type	Observed Prop.	Maximum Likelihood Estimates			
	Σx/Σn	α	β	E(θ)	Var(θ)
Head on	0.016	dnc	dnc	dnc	dnc
Angle	0.00030	0.00137	2.32	0.00059	0.00018
Sideswipe – same dir	0.014	3.75	260	0.014	5.3E-05
Sideswipe – opp dir	0.033	2.84	84.2	0.033	0.00036
Animal	0.096	1.38	12.7	0.099	0.0059
Cyclist	0.0015	0.00986	3.89	0.0025	0.00051
Pedestrian	0.00030	0.00493	12.8	0.00039	3.2E-05
Parked Veh	0.0082	dnc	dnc	dnc	dnc
Overturn	0.25	2.84	7.66	0.27	0.017
Fixed Object	0.40	5.21	7.91	0.40	0.017
Other Multivehicle	0.096	3.75	35.6	0.095	0.0021
Other Stationary Veh	0.030	0.656	21.6	0.030	0.0012
Unknown	0.0049	0.181	37.6	0.0048	0.00012

Table 6.9: MM1 and MM2 beta prior estimates for 2-mile segments of 2-lane rural highway in Washington.

Database: HSIS, Washington
Site Type: Rural, 2-lane Highways, Rolling Terrain – 2-mile Segments
Total Sites: 831
Total Accidents: 3289
Years of Data: 3 (1997-1999)

Accident Type	Observed Proportion	Method of Moments Estimates – Method 1				Method of Moments Estimates – Method 2			
	(=$\Sigma x/\Sigma n$)	α	β	E(θ)	Var(θ)	α	β	E(θ)	Var(θ)
Head on	0.016	0.0342	3.12	0.011	0.0026	6.66	474	0.014	0.000028
Angle	0.00030	0.00145	14.0	0.00010	6.9E-06	0.0500	377	0.00013	3.5E-07
SSSD	0.014	0.0195	1.45	0.013	0.0053	dnc	dnc	dnc	dnc
SSOD	0.033	0.0490	1.64	0.029	0.011	1.28	39.8	0.031	0.00072
Animal	0.096	0.115	0.986	0.10	0.044	1.36	12.0	0.10	0.0063
Cyclist	0.0015	0.00506	6.31	0.00080	0.00011	dnc	dnc	dnc	dnc
Ped.	0.00030	0.00133	6.00	0.00022	3.2E-05	0.102	359	0.00028	7.9E-07
Parked	0.0082	0.0165	2.99	0.0055	0.0014	dnc	dnc	dnc	dnc
Overturn	0.25	0.289	0.665	0.30	0.11	1.99	5.15	0.28	0.025
Fix Obj	0.40	0.466	0.748	0.38	0.11	4.82	7.40	0.39	0.018
Oth Mult	0.096	0.152	1.79	0.078	0.025	2.49	25.2	0.090	0.0029
Oth Sta	0.030	0.0288	0.864	0.032	0.016	1.12	39.9	0.027	0.00063
Unk.	0.0049	0.00552	0.996	0.0055	0.0027	dnc	dnc	dnc	dnc

6.5 Comparison of Parameter Estimation Methods

6.5.1 General

It is clear that the three parameter estimation methods yielded different results. In a real network screening application, however, it would be impractical to use more than one method; thus, a decision must be made regarding the 'best' alternative.

As a first step in this decision, a set of criteria must be developed in order to compare the methods. For each beta prior calibrated, the estimated expected value of the proportion of that accident type, $\hat{E}(\theta)$, is calculated, as well as the and the estimated variance, $V\hat{a}r(\theta)$. The variance serves as one indicator of the accuracy of the model. A relatively high variance means that the modelled prior distribution will have less influence in the empirical Bayes procedure than would a model with relatively low variance. In other words, a prior model with high variance will carry over less information to the posterior distribution than would a prior model with low variance.

6.5.2 Dataset 1: California Rural 4-Leg TWSC Intersections

Figure 6.4 shows the results of the three methods applied to California TWSC intersections for the following accident types: fatal/injury, rear-end, broadside, hit object, and pedestrian. The histogram shows the number of sites with a given proportion of accidents for each of 15 equally-spaced bins.

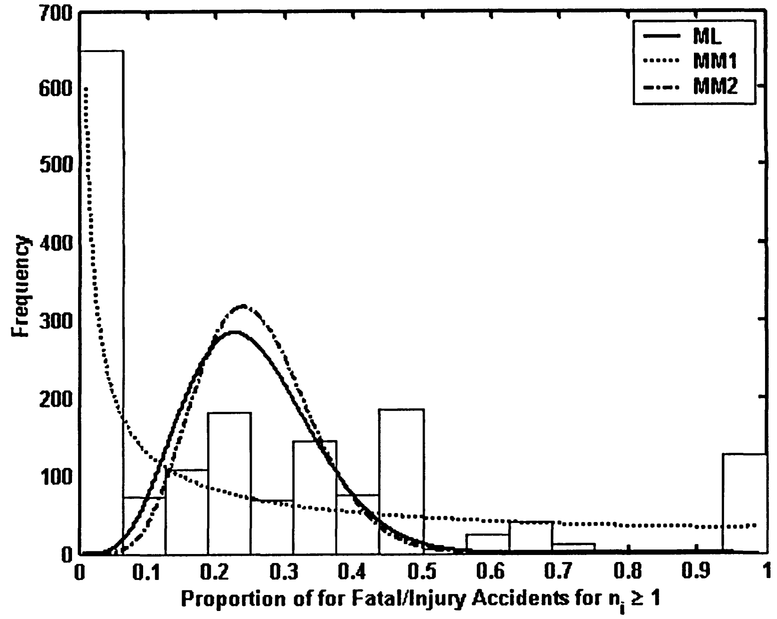


Fig. 6.4a: Beta priors for fatal/injury accidents at California rural, TWSC intersections.

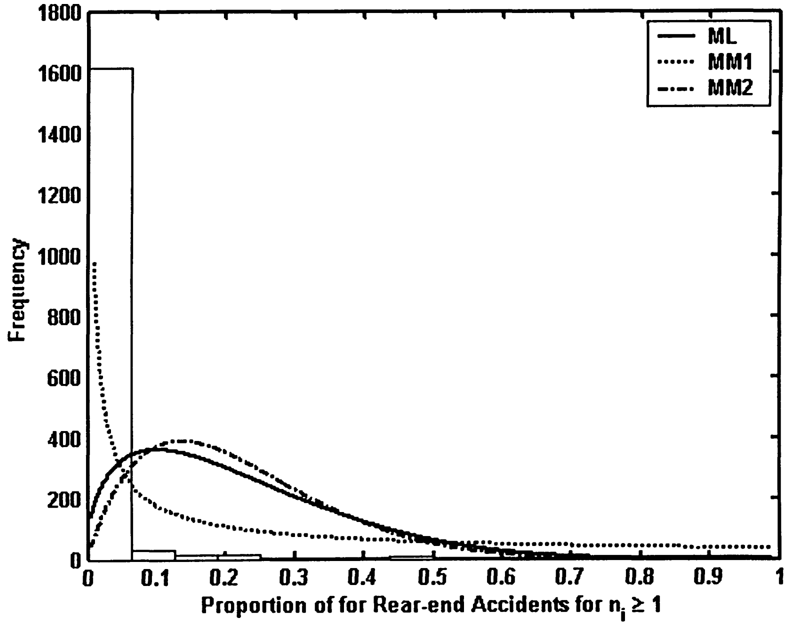


Fig. 6.4b: Beta priors for rear-end accidents at California rural, TWSC intersections.

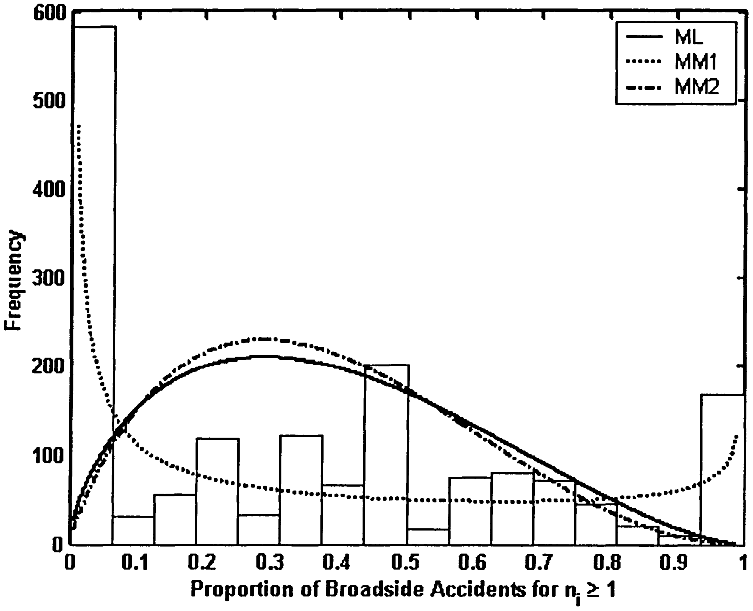


Fig. 6.4c: Beta priors for broadside accidents for California rural, TWSC intersections.

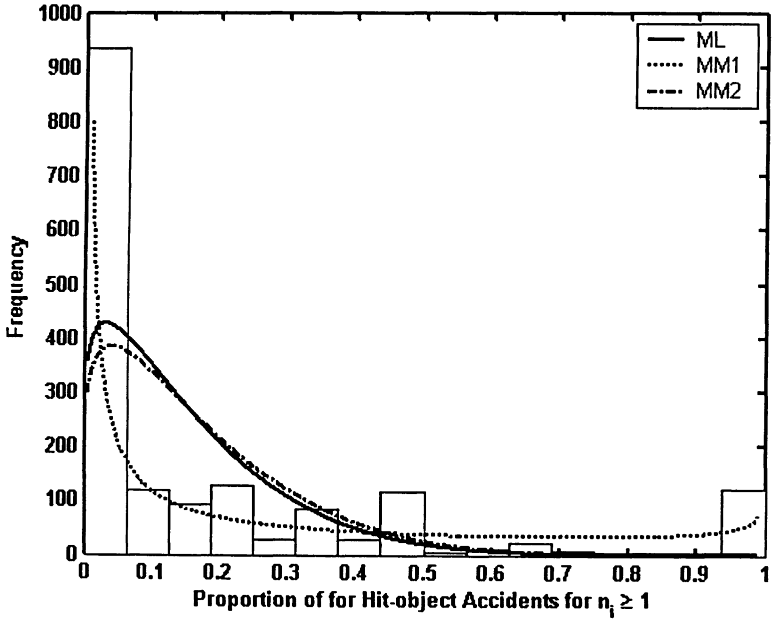


Fig. 6.4d: Beta priors for hit-object accidents for California rural, TWSC intersections.

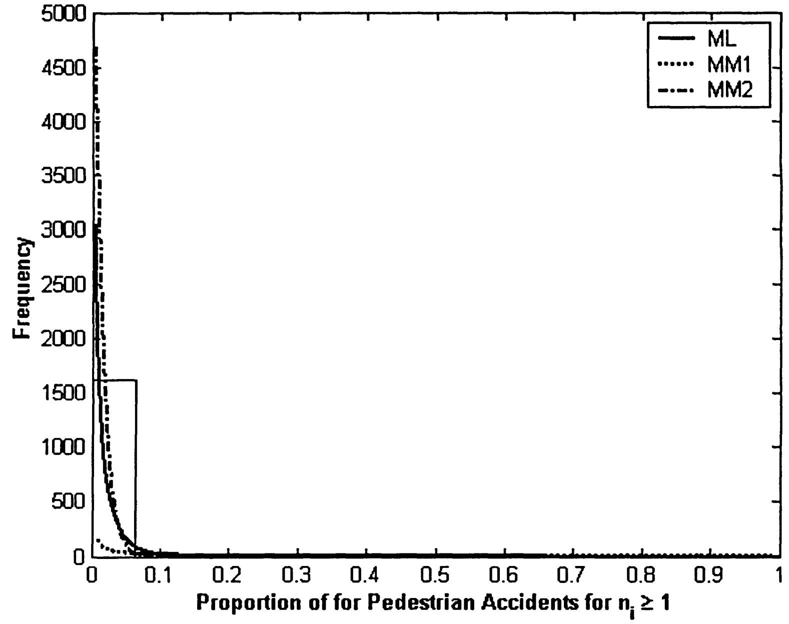


Fig. 6.4e: Beta priors for pedestrian accidents for California rural, TWSC intersections.

The maximum likelihood and MM2 estimates are broadly similar; for each accident type, they have similar shapes, expected values, and variances. The ML and MM2 estimates yielded a bell-shaped curve for all but the two accident types accounting for the lowest proportion of all accidents, overturning (2.5%) and pedestrian (1.1%). A bell-shaped curve results when both α and β are greater than 1. By contrast, the first method of moments (MM1) yielded either L-shaped (α<1,β>1) or U-shaped (α,β<1) curves for every accident type.

The histograms in Fig. 6.4 exhibit large 'spikes' at the ends of the graphs, and, to a lesser extent, at the midpoints of the graphs. This is because a large number of intersections had only one or two total accidents. Thus, for a large number of intersections, it was likely that, for a given accident type, a proportion of 0 (no target accidents) or 1 (all target accidents) would be observed. For intersections with 2 total accidents, many intersections would have a proportion of 0.5, resulting in the high frequency at that level. As the number of total accidents increases, it becomes more likely that a clearly defined pattern emerges; it is this pattern that is of interest, not the randomness inherent in small accident counts.

It is clear from the figures that the MM1 estimates are more heavily influenced by the high frequencies of 'zeros and ones.' This is not surprising, as the sample statistics used for parameter estimation considers only the observed proportion at each site; thus, each site is given equal weight (i.e., each site's observations are assumed to carry the same amount of information), regardless of the number of observations at a given site. The ML and MM2 approaches, however, consider the observed target accidents and total accidents at each site as a pair of values; thus, more weight is given to high-accident sites when estimating the parameters.

In order to see how the estimated prior distributions perform over a range of total accidents, n_i, Figure 6.5 shows the prior distribution for broadside accidents estimated above compared with histograms of observed proportions for $n_i \geq 2$, 5, 10, 15 accidents.

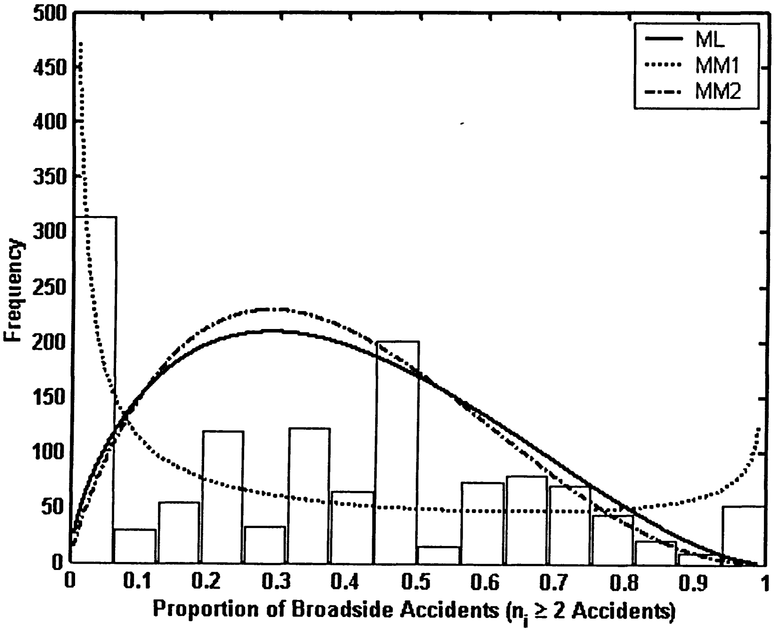


Fig. 6.5a: Proportion of Broadside accidents for $n_i \geq 2$ at Caiifornia TWSC Intersections.

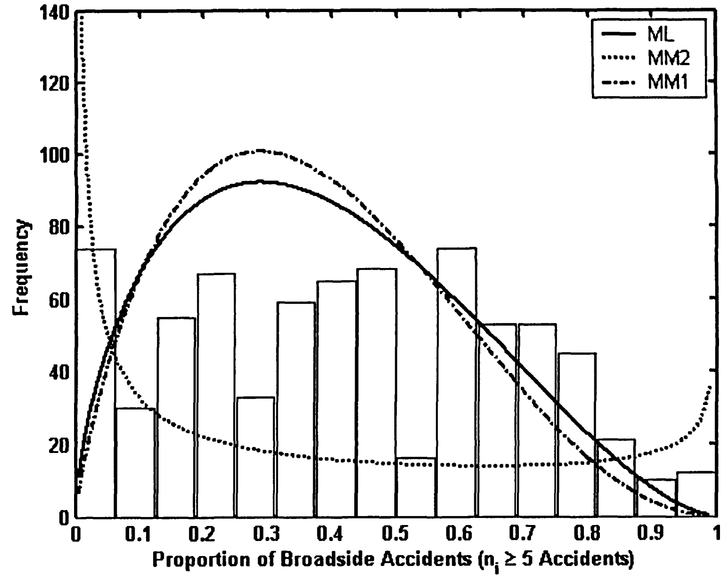


Fig. 6.5b: Proportion of Broadside accidents for n_i≥5 at California TWSC Intersections.

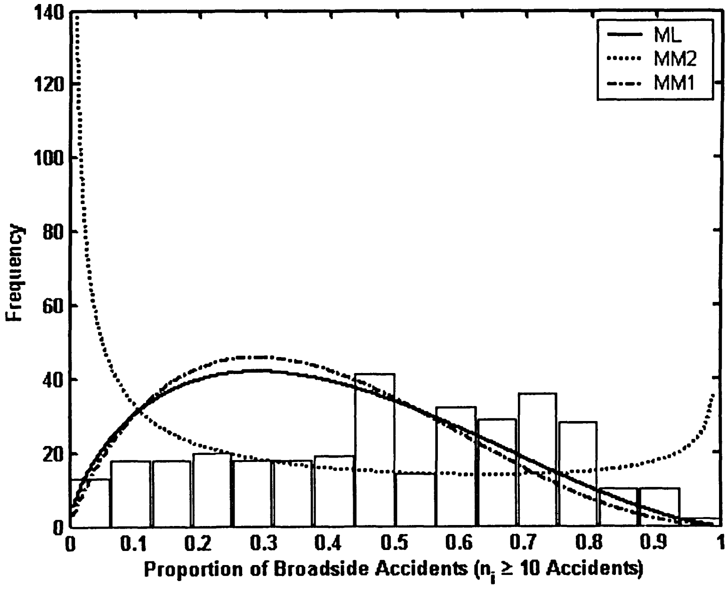


Fig. 6.5c: Proportion of Broadside accidents for n_i≥10 at California TWSC Intersections.

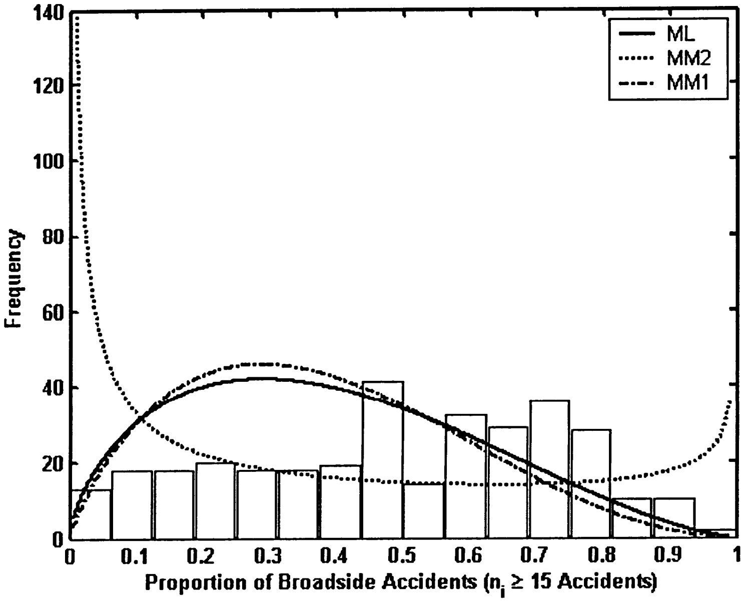


Fig. 6.5d: Proportion of Broadside accidents for $n_i \geq 15$ at California TWSC Intersections.

It is immediately evident that the MM1 model does not fit the data well at higher accident counts. Both the MM2 and ML methods perform better; however, at higher accident counts, the data appear to be negatively skewed (i.e., skewed to the right), whereas the models are positively skewed. This indicates that the predicted mean will differ from the observed mean.

It is important to recall that the MM2 approach considers only sites with $n_i \geq 2$ accidents, while the ML and MM1 approaches considered sites with $n_i \geq 1$ accidents. Thus, the histogram in Fig. 6.5a represents the data from which the MM2 model was calibrated. The fact that the ML estimates agree very closely with the MM2 estimates shows that the ML estimates are relatively robust.

Fig. 6.6 shows a beta prior estimated using only sites with $n_i \geq 10$ accidents.

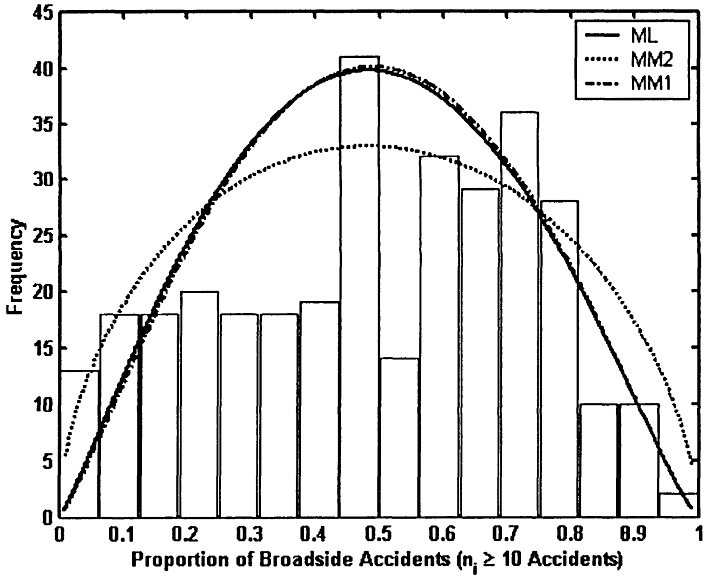


Fig. 6.6: Broadside Beta Priors Calibrated for $n_i \geq 10$ at California TWSC Intersections.

Each new model appears to perform better than its original counterpart; however, the presence of a pattern is not clear, and it is difficult to tell if the assumption of a beta distribution is appropriate.

6.5.3 Dataset 2: California Rural Signalized Intersections

The California TWSC intersections considered above experienced a relatively low number of accidents per site: 0.94 accidents per site, per year on average. The California signalized intersections under investigation experienced an average of 4.8 accidents per site, per year. Thus, it is expected that the models for these data will suffer less from the influence of low-accident sites.

Figure 6.7 shows the beta priors for fatal/injury, rear-end, broadside, hit-object, and pedestrian accidents for California rural, signalized intersections.

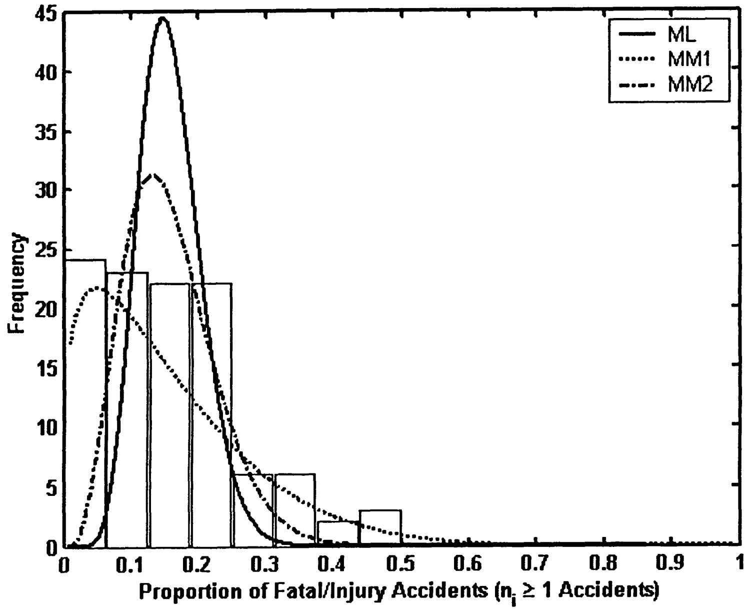


Fig. 6.7a: Beta priors for fatal/injury accidents at California rural, signalized intersections.

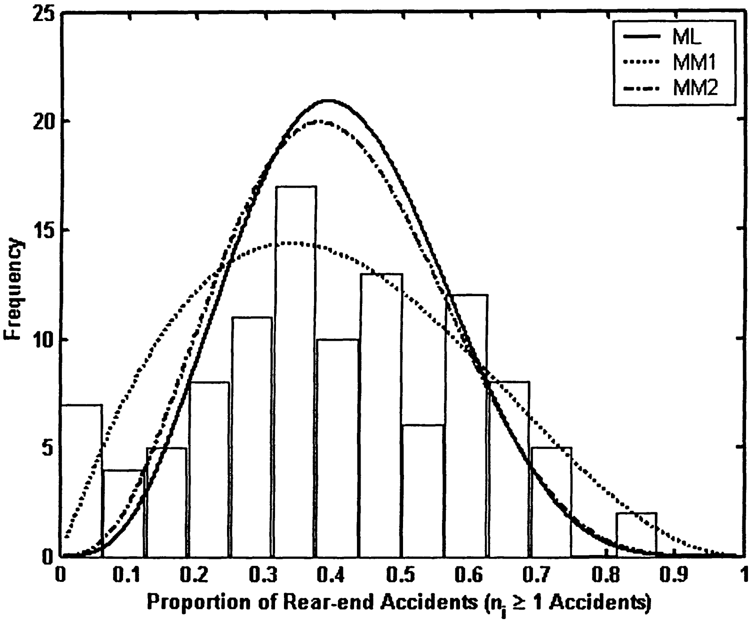


Fig. 6.7b: Beta priors for rear-end accidents at California rural, signalized intersections.

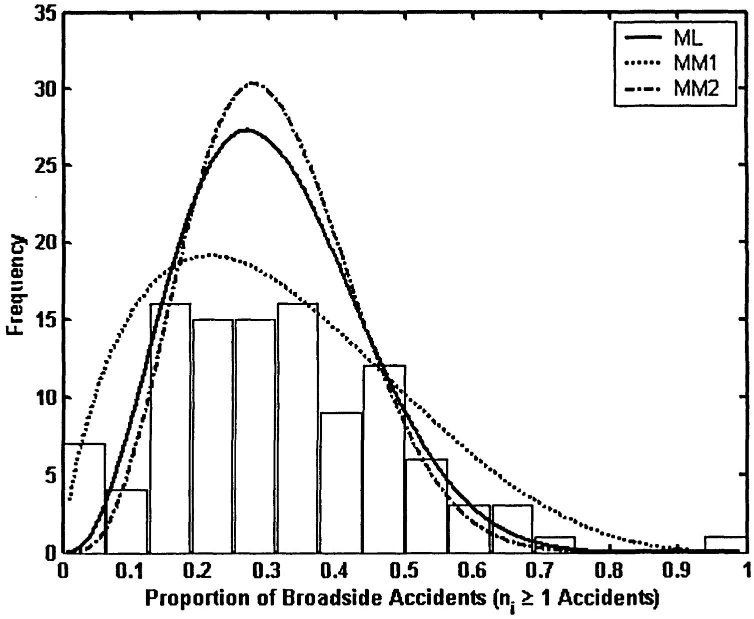


Fig. 6.7c: Beta priors for broadside accidents at California rural, signalized intersections.

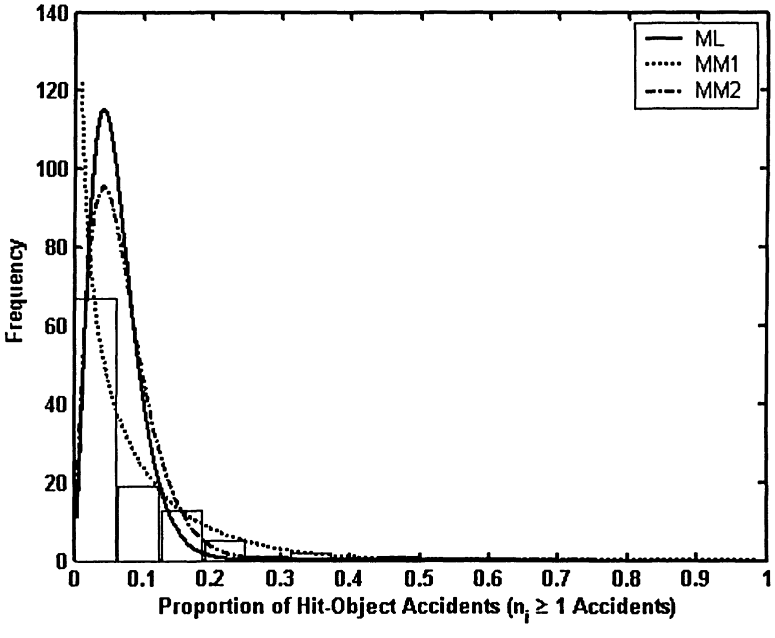


Fig. 6.7d: Beta priors for hit-object accidents at California rural, signalized intersections.

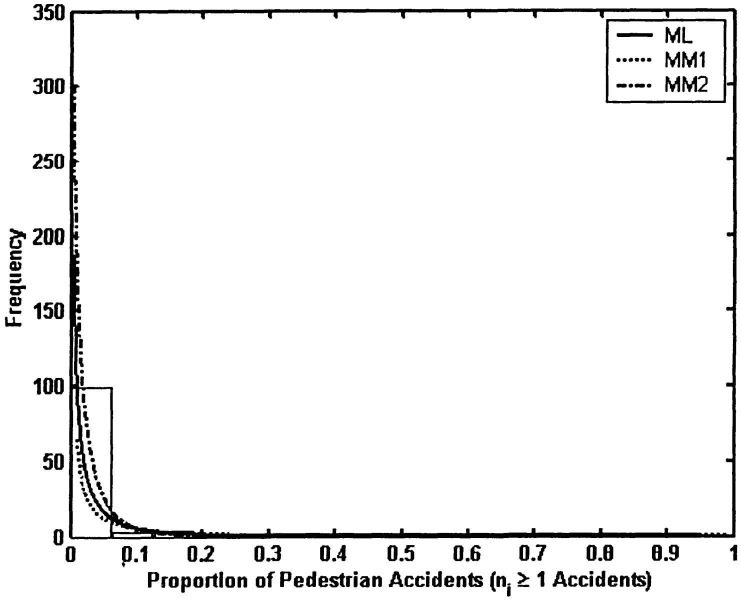


Fig. 6.7e: Beta priors for pedestrian accidents at California rural, signalized intersections.

For the signalized intersections, the ML and MM2 beta priors are similar in shape to those calibrated for the TWSC intersections. The MM1 beta priors are bell-shaped for those accident types representing greater than about 15 percent of all accidents; in this case, only fatal/injury, rear-end, and broadside have proportions in excess of 15 percent. In all other cases, the MM1 distributions were L-shaped.

None of the MM1-fitted distributions had a U-shaped curve, as was the case for TWSC intersections. As discussed above, this was expected, owing to the larger accident counts observed at the signalized intersections.

The ML and MM2 priors were again similar with respect to expected values and variances. In general, the MM1 priors exhibited larger variances than the other two methods.

6.5.4 Dataset 3: 2-Mile Segments of Washington 2-Lane Rural Highways

All of the previous work on screening for high proportions of specific accident types has been applied to intersections. In theory, there is no reason that the method could not be applied to any type of transportation facility. Here, 2-mile segments of Washington 2-lane rural highways in rolling terrain are examined. Traffic volumes and some geometric traits vary between sites.

Figure 6.8 shows the beta prior distributions for head-on, animal, overturning, and fixed object accidents estimated using the ML, MM1, and MM2 approaches. Note that maximum likelihood parameter estimates did not converge for the head-on accident model.

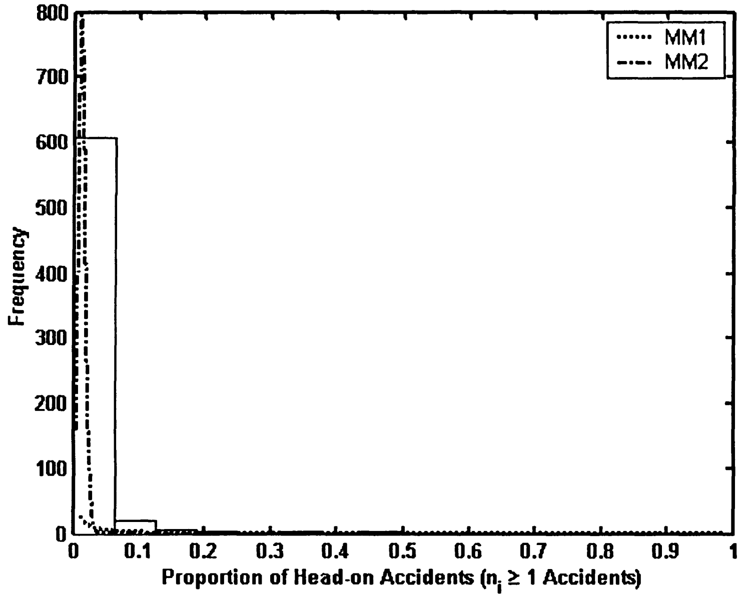


Fig. 6.8.a: Beta priors for head-on accidents on 2-mile segments of Washington 2-lane rural highways.

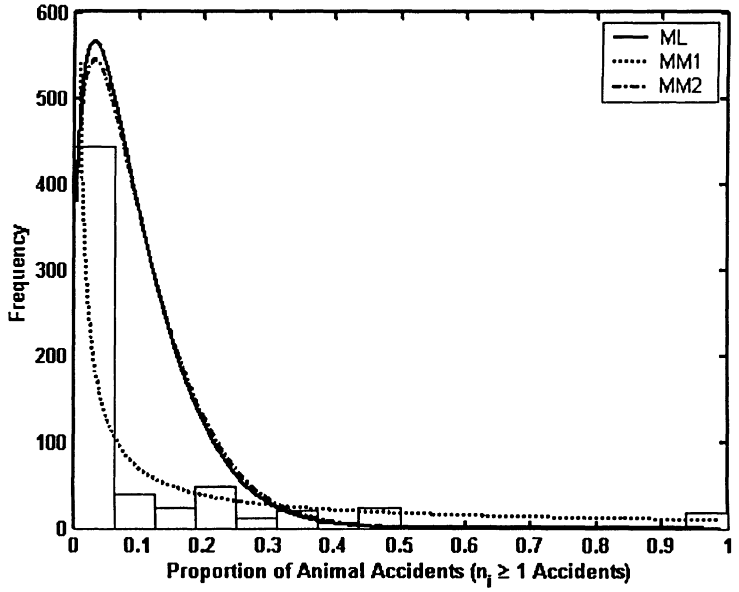


Fig. 6.8.b: Beta priors for animal accidents on 2-mile segments of Washington 2-lane rural highways.

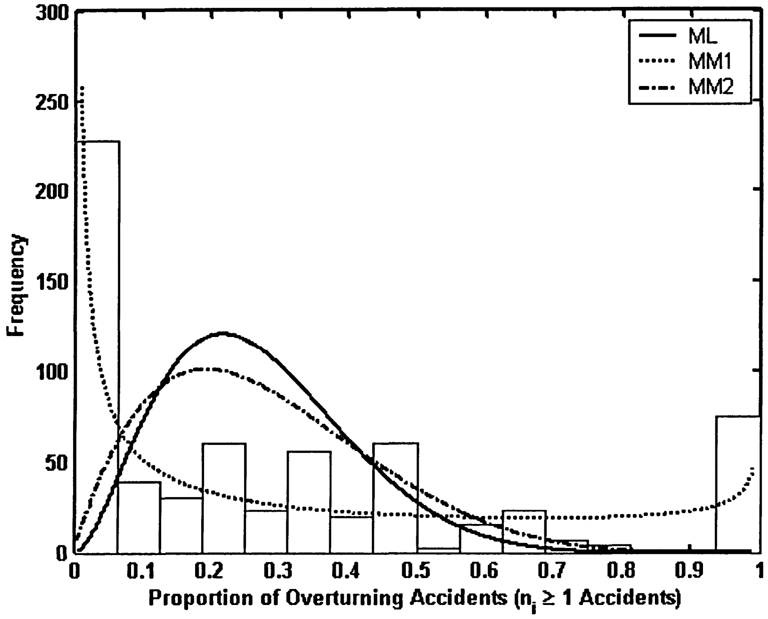


Fig. 6.8.c: Beta priors for overturning accidents on 2-mile segments of Washington 2-lane rural highways.

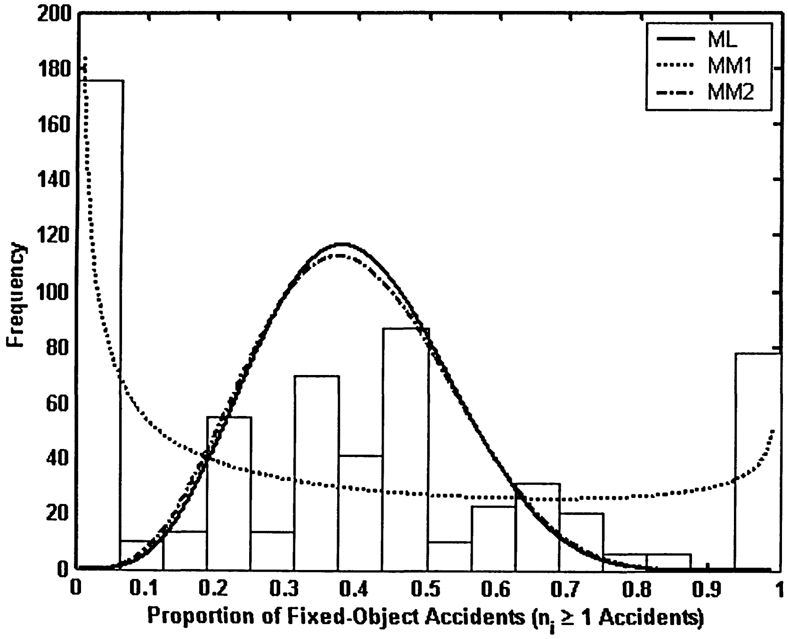


Fig. 6.8.d: Beta priors for fixed-object accidents on 2-mile segments of Washington 2-lane rural highways.

The beta prior distributions calibrated for the Washington highway segments have many similarities with the beta priors estimated for California TWSC intersections. As with the intersection models, the MM1 estimates for highway segments were all either L-shaped or U-shaped. Again, it appears that the MM1 parameter estimates are strongly influenced by the many 'zeros and ones' in the observations.

The ML and MM2 beta priors are again bell-shaped for relatively high-proportion accident types, and are L-shaped for lower-proportion types. Again, the ML and MM2 estimates have similar expected values and variances.

Maximum likelihood estimates did not converge for head-on and parked-vehicle accidents, which represented 1.6 and 0.82 percent of total accidents, respectively. MM2 parameter estimates failed to converge for same-direction sideswipe (1.4%), cyclist (0.15%), parked-vehicle (0.82%), and unknown (0.49%) accident types, where the number in parentheses indicates the percentage of total accidents for the given target type. Models for accident types representing more than 2 percent of total accidents converged in every case.

In order to see the effects of including sites with low accident counts, the beta priors are compared with histograms of observed proportions for $n_i \geq 2$ and $n_i \geq 5$ total accidents. These are shown in Fig. 6.9.

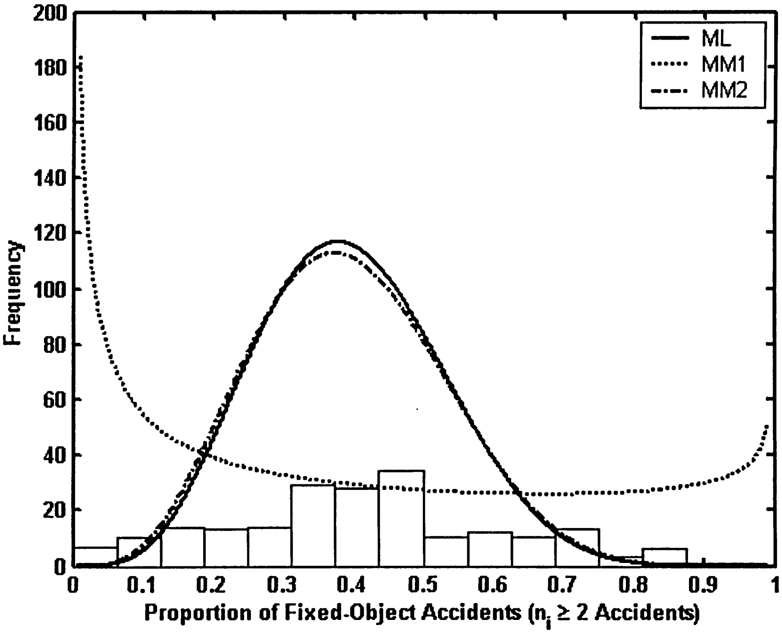


Fig. 6.9a: Proportion of fixed-object accidents for $n_i \geq 2$ on 2-mile segments of Washington rural 2-lane highways.

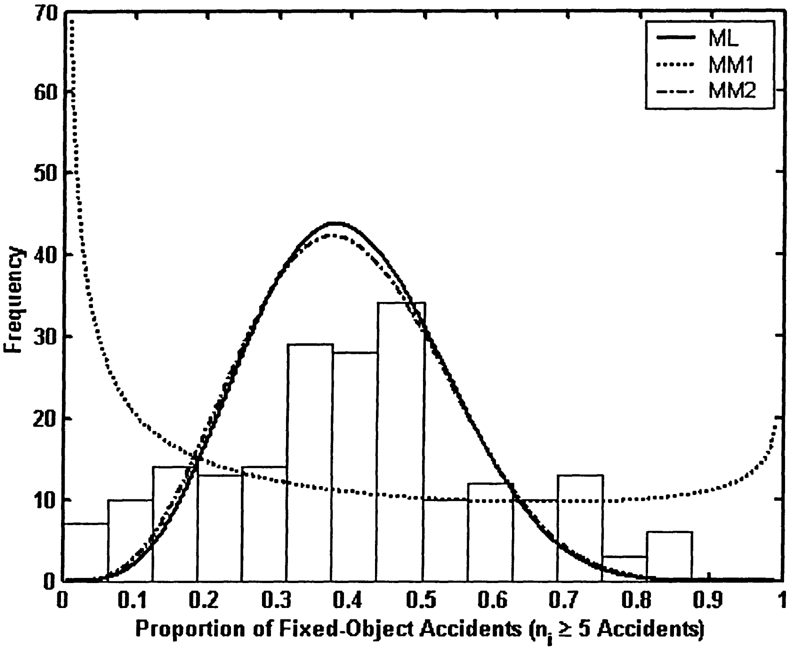


Fig. 6.9b: Proportion of fixed-object accidents for $n_i \geq 5$ on 2-mile segments of Washington rural 2-lane highways.

For both $n_i \geq 2$ and $n_i \geq 5$, the ML and MM2 estimates still appear to fit the data reasonably well; however, it is evident that the MM1 estimates are not appropriate for these data. Fig. 6.10 shows the beta prior distribution estimated using only data for $n_i \geq 5$ sites:

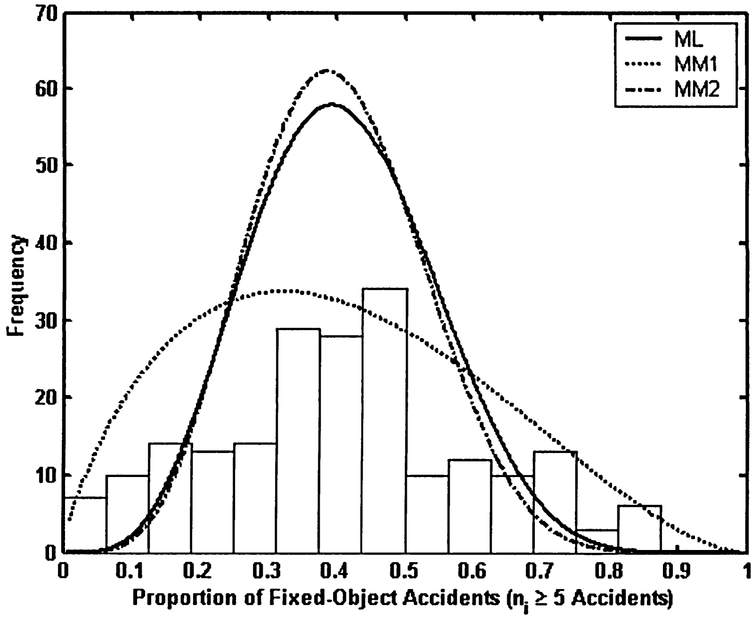


Fig. 6.10: Proportlon of fixed-object accidents calibrated for $n_i \geq 5$ on 2-mile segments of Washington rural 2-lane highways.

The ML and MM2 beta priors show in Fig. 6.10 are similar to those calibrated for the maximum number of sites; however, the MM1 estimates are a much better fit than the MM1 estimates shown in Fig. 6.9. The MM1-fitted prior is no longer influenced by the large numbers of zeros and ones as in the earlier attempts.

6.6 How Many Years of Data Are Needed To Calibrate Beta Priors?

When performing network screening, a decision must be made regarding the number of years of data to include. If many years are used, there are more observations and, thus, better-fitting models should result; however, the characteristics of a given site can change over time, including traffic volume, the roadway environment, vehicles, and drivers, to name a few. If significant changes have occurred at the site over the course of the observation period, then the model parameters gleaned from these data are questionable. If a short observational period is used, there will be fewer data points with which to calibrate a model, and the random fluctuations in accident occurrence will become more influential to the model.

Sometimes, the required accident data is available only for short periods; in this case, all data would normally be included. If a large number of years of data are available, it is common to use 3 years of data. Here, beta priors are estimated using 5 years of accident data are compared to beta priors calibrated from 3 years of data. This comparison is made only with the California signalized rural intersection database, as these sites had much higher average accident counts than did the other two databases.

Figure 6.11 shows the 5-year and 3-year maximum-likelihood-estimated beta priors at California signalized intersections for fatal/injury, rear-end, broadside, hit object, and pedestrian accidents.

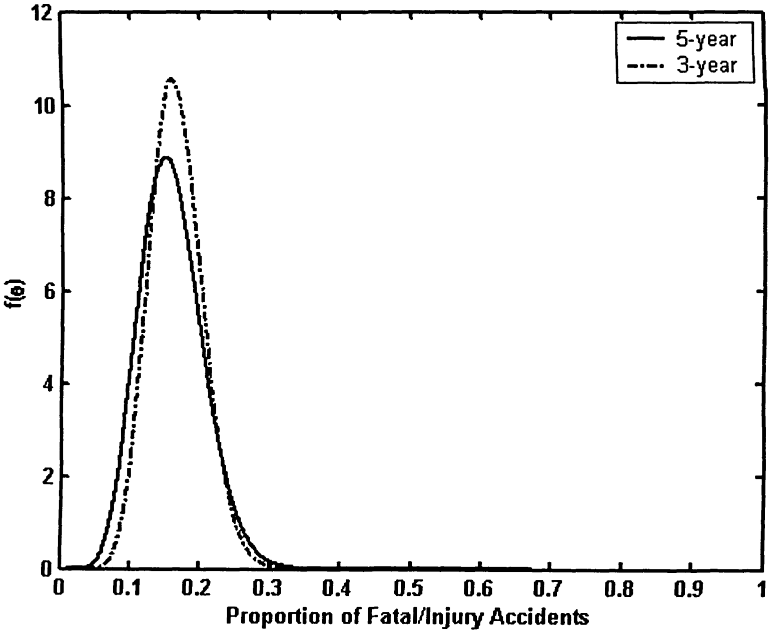


Fig. 6.11a: Comparison of fatal/injury beta priors for 5 and 3 years of accident data, using maximum likelihood estimation. (Washington signalized intersections)

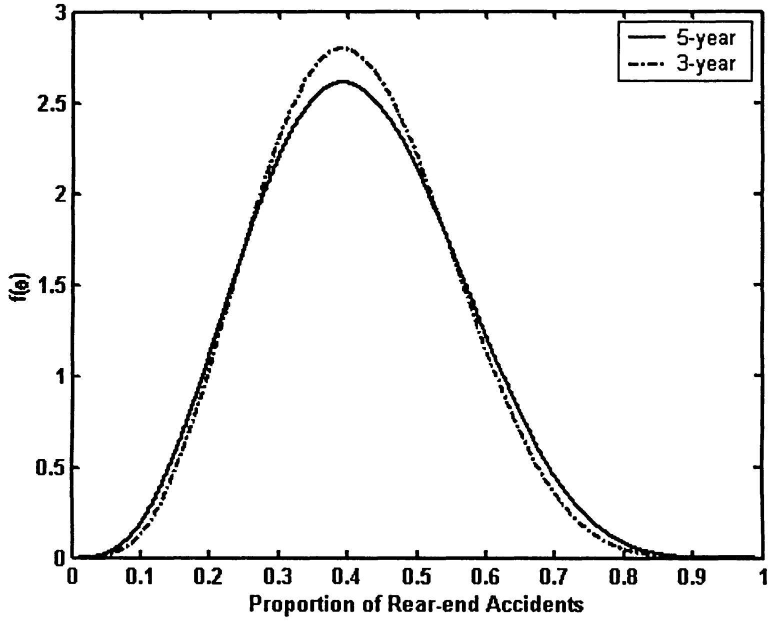


Fig. 6.11b: Comparison of rear-end beta priors for 5 and 3 years of accident data, using maximum likelihood estimation. (Washington signalized intersections)

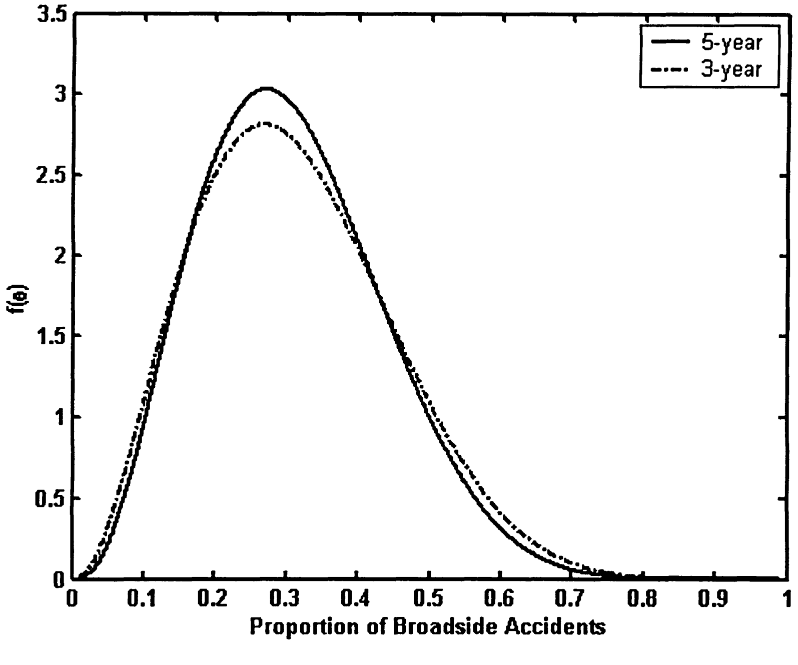


Fig. 6.11c: Comparison of broadside beta priors for 5 and 3 years of accident data, using maximum likelihood estimation.

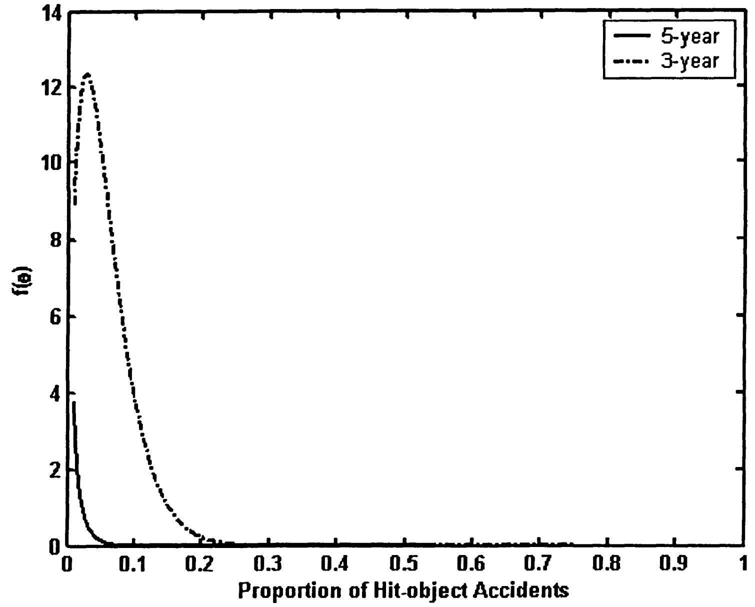


Fig. 6.11d: Comparison of hit-object beta priors for 5 and 3 years of accident data, using maximum likelihood estimation.

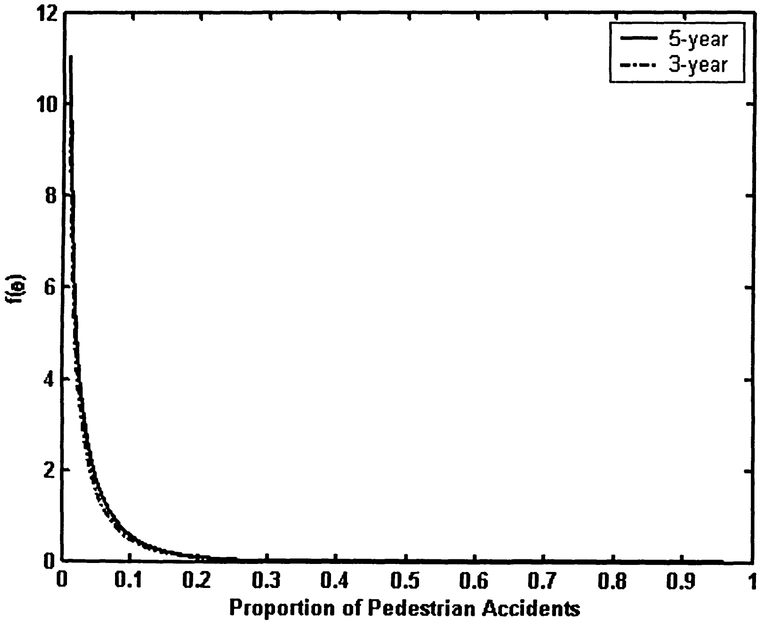


Fig. 6.11e: Comparison of pedestrian beta priors for 5 and 3 years of accident data, using maximum likelihood estimation.

Fig. 6.12 shows both 5-year and 3-year MM1 estimates for rear-end and broadside accidents, and Fig. 6.13 shows the MM2 estimates for the same accident types. Again, all results are from the California signalized intersection database.

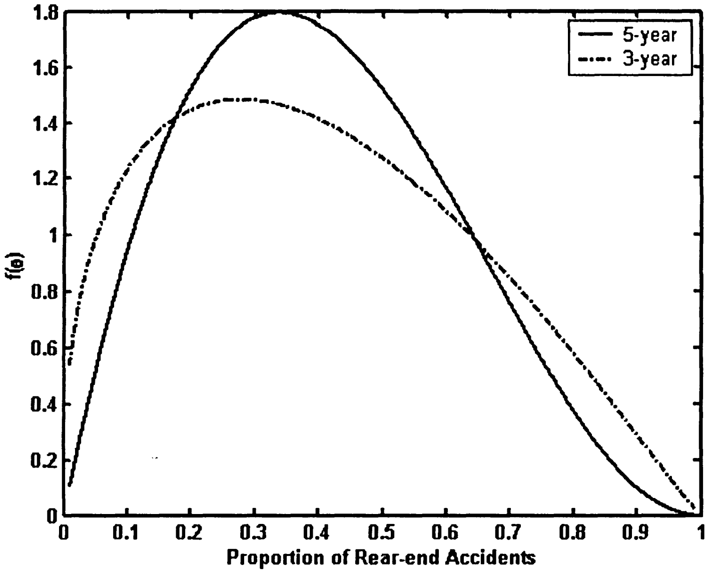


Fig. 6.12a: Comparison of rear-end beta priors for 5 and 3 years of accident data, using MM1 estimation.

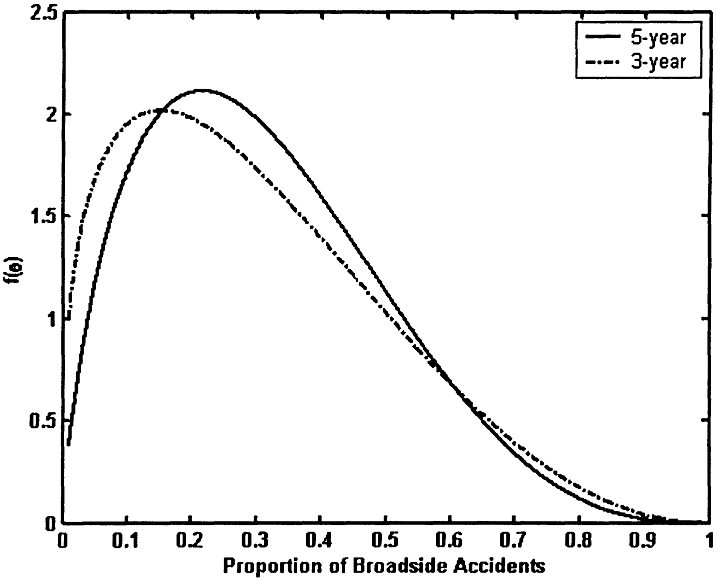


Fig. 6.12b: Comparison of broadside beta priors for 5 and 3 years of accident data, using MM1 estimation.

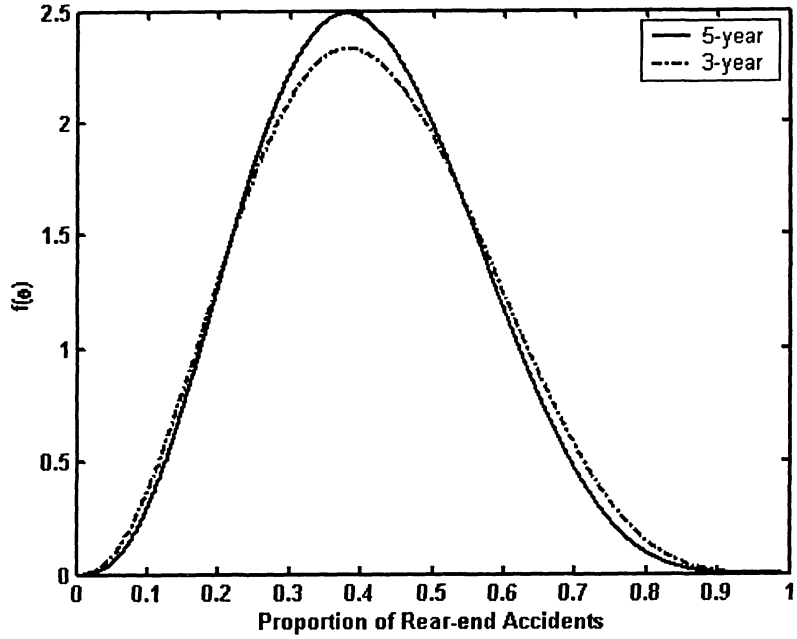


Fig. 6.13a: Comparison of rear-end beta priors for 5 and 3 years of accident data, using MM2 estimation.

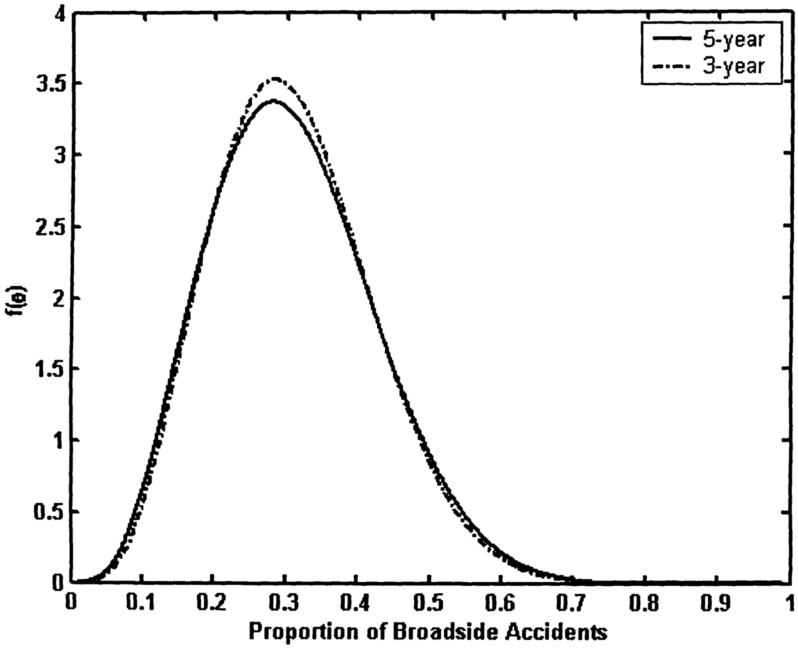


Fig. 6.13b: Comparison of broadside beta priors for 5 and 3 years of accident data, using MM2 estimation.

The 5-year and 3-year beta priors compare very well for both maximum likelihood and MM2 estimates. The MM1 estimates also compare reasonably well; however, the3-year MM1 estimates exhibit a larger variance than the corresponding 5-year estimates.

6.7 Results of Site Ranking

6.7.1 General

The empirical Bayes procedure described here ranks the sites based on the degree of certainty that the proportion of all accidents of a specific accident type, θ_i, is higher than some value, $\theta*$. The value of $\theta*$ has generally been taken to be the median proportion, θ_m, observed at all other sites of the same kind in the network being investigated (*2,31-34*). Heydecker and Wu (*2*) describe using the median as a means of identifying those sites having an accident proportion that is "greater than normal for sites of the same kind."

The 'degree of certainty' is represented by the pattern score, $\Pr(\theta_i > \theta*)$, calculated using Eq. 6.17, with a value of 1 being almost absolute certainty that the statement is true; a value of zero indicates the proposition is almost certainly false.

While a pattern score is calculated for each site, it is usually desirable to limit the list of ranked sites to only those sites with a relatively high probability of having a high proportion of the given accident type. This is accomplished by selecting a limiting value of the pattern score, δ. In theory, any value of δ may be used, but this will have an effect on the size and order of the ranked list. If δ is large (e.g. 0.99), a relatively small number of sites will 'pass the test' and be ranked; however, these sites will have a very high probability of having a high proportion of the given accident type, and thus a relatively low type I error (i.e.: fewer 'false positives' should be observed). As δ decreases, a larger number of sites are ranked, with a corresponding increase in 'false-positive' errors.

The choice of δ can easily be changed to fit the data at hand; if a large number of sites are ranked after setting δ to, say, 0.95, one need only increase δ to reduce the number of sites on the ranked list. Similarly, if too few sites are ranked, δ can be reduced. It would be prudent to select minimum possible value of δ, in order to keep false-positive errors to some 'acceptable' minimum. In this thesis, no site is ranked with a pattern score of less than 0.90. It is important to note that changing the value of δ

does not influence the order the sites are ranked; it merely governs the number of sites making the final list.

A way of modifying the ranked list is by modifying π as described in Section 6.2.2. As already mentioned, a value of 0.5 indicates the median, as 50% of the sites will have a pattern score greater than the median, and 50% will have a lesser score. A value of 0.8, for example, would indicate the probability that the observed proportion at a given site is greater than 80% of similar sites. Several values of π are used in this thesis, and the results are compared. For each of the three datasets, the following π-values were used: 0.5, 0.75, 0.8, 0.9, and 0.95.

All calculations for the screening procedure were performed using Microsoft Excel spreadsheets. Pattern scores are reported to three decimal places; however, rankings are based on scores computed to about 8 decimal places of accuracy. It would be unreasonable to believe that differences of that magnitude are significant, and it could be argued that no differences in ranking should result from them. In practice, however, no decision to conduct road safety remedial work is to be based on a site's ranking; rather, the flagged sites are subjected to a more detailed investigation, the results of which would dictate the allocation of resources.

6.7.2 California Rural 4-leg TWSC Intersections

Rankings of California TWSC intersections were calculated for each accident type, for each of the three parameter estimation method, and for the different values of π described above. Tables 610 to 6.13 show the screening results for rear-end accidents where the beta priors were estimated using maximum likelihood. Rankings are shown for π-values of 0.5, 0.8, 0.9, and 0.95.

When π was taken to be 0.5 (i.e.: the median), 137 sites were flagged. This is likely a far greater number of sites than could be economically studied in detail. Even if a value of 0.99 was selected for δ, there would have been 51 flagged sites, which is still probably too many for practical purposes. It should be noted that the highest-ranked sites are not necessarily the sites with the highest observed proportions.

When π was increased from 0.5 to 0.8, the value of θ^* increased from 0.19 to 0.34. At the same time, the number of flagged sites decreased from 137 to 30. Sites with relatively low accident proportions (with respect to the other ranked sites) were either demoted to lower rankings when the π was increased, or pushed off of the list altogether. Sites with a relatively high proportion were more likely to be promoted.

For π-values of 0.9 and 0.95, the numbers of flagged sites were reduced to 15 and 2, respectively. While the number of flagged sites on these lists could be increased by decreasing , this is done at the peril of increasing the probability of making a false-positive error.

Table 6.10: Screening for proportions rankings of California 4-leg, TWSC rural intersections based on maximum likelihood beta prior estimates and π =0.5.

Dataset:	1 (California rural TWSC intersections)
Target Accidents:	Rear-end
π	0.5
θ^*	0.19
Parameter estimates:	Maximum likelihood
δ	0.90
Number of sites ranked:	137

Site No.	Total Accidents	Rear-end Accidents	Observed Proportion of RE Accidents	Pattern Score	Rank
i	n_i	x_i	θ_i	$\Pr(\theta_i > \theta^*)$	
464	26	19	0.73	1.000	**1**
1003	31	19	0.61	1.000	**2**
1716	23	16	0.70	1.000	**3**
1095	24	16	0.67	1.000	**4**
146	24	16	0.67	1.000	**4**
302	22	15	0.68	1.000	**6**
1585	20	14	0.70	1.000	**7**
645	10	10	1.00	1.000	**8**
1211	16	12	0.75	1.000	**9**
149	16	12	0.75	1.000	**9**
742	11	10	0.91	1.000	**11**
617	19	13	0.68	1.000	**12**
300	28	16	0.57	1.000	**13**
1977	29	16	0.55	1.000	**14**
1011	15	11	0.73	1.000	**15**
152	34	17	0.50	1.000	**16**
177	11	9	0.82	1.000	**17**
1294	11	9	0.82	1.000	**17**
1198	17	11	0.65	1.000	**19**
144	24	13	0.54	1.000	**20**

Table 6.11: Screening for proportions rankings of California 4-leg, TWSC rural intersections based on maximum likelihood beta prior estimates and π =0.8.

Dataset:	1 (California rural TWSC intersections)
Target Accidents:	Rear-end
π	0.8
θ^*	0.34
Parameter estimates:	Maximum likelihood
δ	0.90
Number of sites ranked:	30

Site No.	**Total Accidents**	**Rear-end Accidents**	**Observed Proportion of RE Accidents**	**Pattern Score**	**Rank**
i	n_i	x_i	θ_i	$\Pr(\theta_i > \theta^*)$	
464	26	19	0.73	1.000	**1**
645	10	10	1.00	0.998	**2**
1716	23	16	0.70	0.997	**3**
742	11	10	0.91	0.996	**4**
1095	24	16	0.67	0.995	**5**
146	24	16	0.67	0.995	**5**
302	22	15	0.68	0.995	**7**
1003	31	19	0.61	0.995	**8**
1585	20	14	0.70	0.994	**9**
1211	16	12	0.75	0.993	**10**
149	16	12	0.75	0.993	**10**
617	19	13	0.68	0.989	**12**
1011	15	11	0.73	0.987	**13**
177	11	9	0.82	0.984	**14**
1294	11	9	0.82	0.984	**14**
300	28	16	0.57	0.975	**16**
1198	17	11	0.65	0.967	**17**
863	8	7	0.875	0.967	**18**
867	8	7	0.875	0.967	**18**
1977	29	16	0.55	0.966	**20**

Table 6.12: Screening for proportions rankings of California 4-leg, TWSC rural intersections based on maximum likelihood beta prior estimates and π =0.9.

Dataset:	1 (California rural TWSC intersections)
Target Accidents:	Rear-end
π	0.9
θ^*	0.43
Parameter estimates:	Maximum likelihood
δ	0.90
Number of sites ranked:	15

Site No.	**Total Accidents**	**Rear-end Accidents**	**Observed Proportion of RE Accidents**	**Pattern Score**	**Rank**
i	n_i	x_i	θ_i	$\Pr(\theta_i > \theta^*)$	
464	26	19	0.73	0.988	**1**
645	10	10	1.00	0.984	**2**
742	11	10	0.91	0.969	**3**
1716	23	16	0.70	0.959	**4**
1211	16	12	0.75	0.942	**5**
149	16	12	0.75	0.942	**5**
1585	20	14	0.70	0.941	**7**
1095	24	16	0.67	0.940	**8**
146	24	16	0.67	0.940	**8**
302	22	15	0.68	0.940	**10**
1003	31	19	0.61	0.920	**11**
177	11	9	0.82	0.915	**12**
1294	11	9	0.82	0.915	**12**
617	19	13	0.68	0.913	**14**
1011	15	11	0.73	0.913	**15**

Table 6.13: Screening for proportions rankings of California 4-leg, TWSC rural intersections based on maximum likelihood beta prior estimates and π =0.95.

Dataset:	1 (California rural TWSC intersections)
Target Accidents:	Rear-end
π	0.95
θ^*	0.50
Parameter estimates:	Maximum likelihood
δ	0.90
Number of sites ranked:	2

Site No.	Total Accidents	Rear-end Accidents	Observed Proportion of RE Accidents	Pattern Score	Rank
i	n_i	x_i	θ_i	$\Pr(\theta_i > \theta^*)$	
645	10	10	1.00	0.938	1
464	26	19	0.73	0.922	2

Table 6.14 shows how θ^* changes with the value of π. Table 6.15 shows the top 20 sites ranked by using the median value for θ^* and shows how each ranked using higher values of π. Where the site has been flagged (i.e.: the pattern score was greater than 0.9), the rank is in bold type; where it has not been flagged, the rank is given in plain type.

Table 6.14: Values of θ^* for rear-end accidents at California TWSC intersections, calculated using different π-values.

Percentage	Critical Proportion
π	θ^*
0.5	0.19
0.75	0.31
0.8	0.34
0.9	0.43
0.95	0.50

Table 6.15: Site rankings for rear-end accidents using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Rear-end Accs.	Proportion		Rankings Using Different π-values			
Site No.	n_i	x_i	θ_i	π=0.5	π=0.75	π=0.8	π=0.9	π=0.95
464	26	19	0.73	1	1	1	1	2
1003	31	19	0.61	2	4	8	11	17
1716	23	16	0.70	3	3	3	4	4
1095	24	16	0.67	4	5	5	8	11
146	24	16	0.67	4	5	5	8	11
302	22	15	0.68	6	8	7	10	8
1585	20	14	0.70	7	9	9	7	7
645	10	10	1.00	8	2	2	2	1
1211	16	12	0.75	9	10	10	5	5
149	16	12	0.75	9	10	10	5	5
742	11	10	0.91	11	7	4	3	3
617	19	13	0.68	12	12	12	14	14
300	28	16	0.57	13	16	16	19	25
1977	29	16	0.55	14	17	20	22	26
1011	15	11	0.73	15	13	13	15	13
152	34	17	0.50	16	21	22	34	63
177	11	9	0.82	17	14	14	12	9
1294	11	9	0.82	17	14	14	12	9
1198	17	11	0.65	19	18	17	18	20
144	24	13	0.54	20	22	23	28	41

In order to understand the mechanism by which sites may be 'pushed out,' consider site number 1003. This site was ranked second highest where π was 0.5, but as the π-values increased, it was ranked steadily lower, and was not flagged at all π=0.95. Figure 6.14 depicts the posterior distribution for rear-end accidents at site 1003. The vertical lines represent the values of $\theta*$ for π=0.5 and π=0.95. The area under the posterior curve to the right of the given value of $\theta*$ is the same as the pattern score, and quantifies the degree of certainty that the observed accident proportion is in fact greater than $\theta*$.

From Fig. 6.14, it is clear why the pattern score for any given site decreases as the choice of the π-value increases. Site 1003 experienced a relatively large number of collisions, and for this reason, the posterior distribution exhibits a relatively low variance. It is because of this low variance that the site is ranked so high when the median value of $\theta*$ is used; had the variance been much higher, more of the curve's area would be to the left of $\theta*$.

In general, sites with high accident counts are favoured by lower values of $\theta*$, and sites with high proportions of the given target accident are favoured by higher values. Of course, sites with both high accident counts and high proportions have high rankings regardless of the choice of $\theta*$.

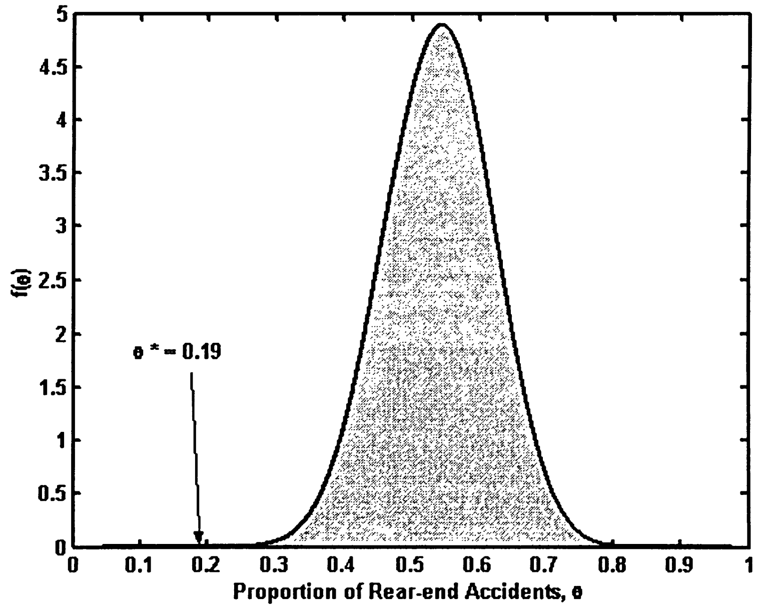


Fig. 6.14a: Posterior distribution of rear-end accidents for site no. 1003. The area of the shaded section is equivalent to the site's pattern score. Here π=0.5.

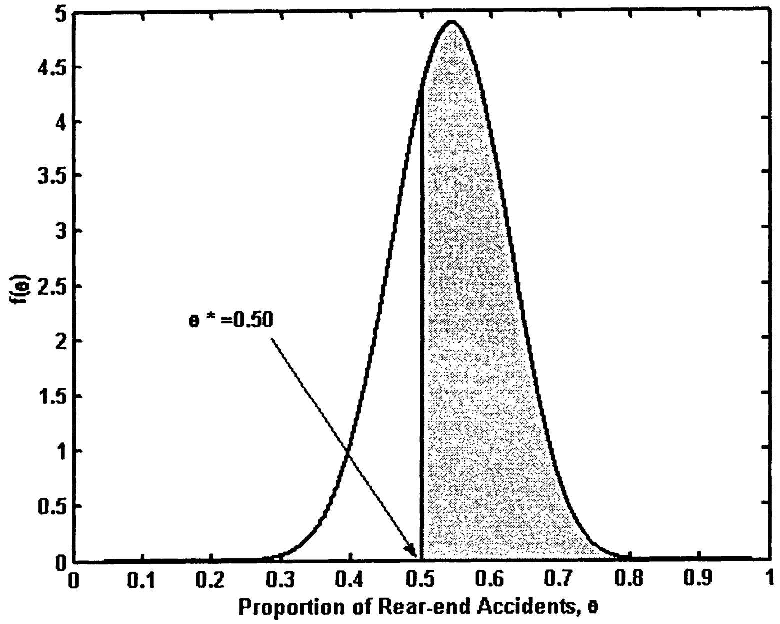


Fig. 6.14b: Posterior distribution of rear-end accidents for site no. 1003. The area of the shaded section is equivalent to the site's pattern score. Here π=0.95.

The ranking 'mechanisms' observed for rear-end accidents are the same for other accident types: as π (or δ) goes up, the number of flagged sites goes down, and sites with lower accident proportions are removed from the list. Table 6.17 shows values of θ^* at different π-values for broadside accidents at California TWSC intersections, and Table 6.16 shows the results of ranking. Results for π=0.95 are omitted, as no sites were flagged.

Table 6.16: Values of θ^* for broadside accidents at California TWSC intersections, calculated using different π-values.

Percentage, π	Critical Proportion, θ^*	No. of flagged sites
0.5	0.37	231
0.75	0.54	67
0.8	0.58	40
0.9	0.69	7
0.95	0.76	0

Table 6.17: Site rankings for broadside accidents using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Broadside Accs.	Proportion	Rankings Using Different π-values			
Site No.	n_i	x_i	θ_i	π=0.5	π=0.75	π=0.8	π=0.9
1006	61	49	0.80	1	1	1	2
153	48	39	0.81	2	3	3	4
1719	40	34	0.85	3	2	2	1
157	50	38	0.76	4	6	9	23
652	37	30	0.81	5	5	5	11
610	45	34	0.76	6	9	12	26
959	31	26	0.84	7	4	4	7
305	26	22	0.85	8	7	7	8
1005	30	24	0.80	9	11	13	18
861	29	23	0.79	10	14	18	24
1498	31	24	0.77	11	19	19	30
464	26	21	0.81	12	17	17	19
1271	16	15	0.94	13	8	6	5
1220	21	18	0.86	14	12	11	12
482	39	28	0.72	15	25	34	66
1501	23	19	0.83	16	18	16	17
318	36	26	0.72	17	26	37	62
1539	44	30	0.68	18	42	50	125
1189	12	12	1.00	19	10	8	3
144	24	19	0.79	20	20	22	31

6.7.3 California Rural 4-Leg Signalized Intersections

The California signalized intersections had, on average, much higher accident frequencies than did the TWSC sites and the Washington highway segments, and far fewer sites. Tables 6.18-6.21 show the screening results for rear-end and broadside accidents at the signalized sites. Where no sites have been flagged for a given π-value, that column is omitted.

For signalized intersections, the results are broadly the same: lower values of π favour sites with high accident frequencies, and higher values favour sites with high proportions. High values of π cannot be used for this database, as no sites are ranked for π>0.8. This is due primarily to the smaller number of sites in the network.

Table 6.18: Values of θ* for rear-end accidents at California signalized intersections, calculated using different π-values.

Percentage, π	Critical Proportion, θ*	No. of flagged sites
0.5	0.41	25
0.75	0.51	5
0.8	0.54	3
0.9	0.61	0
0.95	0.66	0

Table 6.19: Values of θ* for broadside accidents at California signalized intersections, calculated using different π-values.

Percentage, π	Critical Proportion, θ*	No. of flagged sites
0.5	0.30	18
0.75	0.39	4
0.8	0.42	3
0.9	0.48	0
0.95	0.54	0

Table 6.20: Site rankings for rear-end accidents at California signalized intersections using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Rear-end Accs.	Proportion	Rankings Using Different π-values		
Site No.	n_i	x_i	θ_i	π=0.5	π=0.75	π=0.8
28	69	46	0.67	1	2	2
50	55	38	0.69	2	1	1
5	113	66	0.58	3	5	7
67	58	36	0.62	4	6	5
54	35	24	0.69	5	4	4
38	82	47	0.57	6	11	11
29	72	42	0.58	7	8	10
10	17	14	0.82	8	3	3
27	61	36	0.59	9	9	9
69	41	26	0.63	10	7	6
16	54	30	0.56	11	18	22
35	17	12	0.71	12	10	8
6	49	27	0.55	13	22	24
101	22	14	0.64	14	12	12
36	37	21	0.57	15	19	19
75	37	21	0.57	15	19	19
20	23	14	0.61	17	17	17
11	15	10	0.67	18	14	14
79	15	10	0.67	18	14	14
108	13	9	0.69	20	13	13

Table 6.21: Site rankings for broadside accidents at California signaiized intersections using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Rear-end Accs.	Proportion	Rankings Using Different π-values		
Site No.	n_i	x_i	θ_i	π=0.5	π=0.75	π=0.8
39	52	29	0.56	1	2	3
62	28	19	0.68	2	1	1
30	22	15	0.68	3	3	2
77	28	17	0.61	4	4	4
87	23	13	0.57	5	5	5
97	22	12	0.55	6	7	7
95	18	10	0.56	7	8	9
52	32	15	0.47	8	11	12
22	8	6	0.75	9	6	6
44	61	25	0.41	10	23	28
43	16	9	0.56	11	10	10
103	11	7	0.64	12	9	8
86	20	10	0.50	13	12	11
17	26	12	0.46	14	15	17
47	26	12	0.46	14	15	17
21	29	13	0.45	16	17	20
23	21	10	0.48	17	14	14
84	16	8	0.50	18	13	13

6.7.4 Washington Highway Segments

Screening for high proportions of specific target accidents on road segments uses the same procedure as for the intersection screening described above. The only differences one would expect to see would be the target accidents examined, owing to the differences in nature of intersection and highway segment accidents and their associated countermeasures.

Tables 6.22-6.25 show the results of screening on 831 2-mile segments of 2-lane rural highways in Washington. The accidents types shown are overturning and fixed-object.

In general, a smaller portion of sites were ranked than for the intersection databases; use of the median value for θ^* would appear to be the only option, as otherwise too few sites would be flagged.

Table 6.22: Values of θ^* for overturning accidents on 2-mile segments of 2-lane rural highways in Washington, using different π-values.

Percentage, π	Critical Proportion, θ^*	No. of flagged sites
0.5	0.26	23
0.75	0.35	1
0.8	0.38	0
0.9	0.45	0
0.95	0.51	0

Table 6.23: Values of θ^* for fixed-object accidents on 2-mile segments of 2-lane rural highways in Washington, using different π-values.

Percentage, π	Critical Proportion, θ^*	No. of flagged sites
0.5	0.39	20
0.75	0.49	5
0.8	0.51	4
0.9	0.57	0
0.95	0.62	0

Table 6.24: Site rankings for overturning accidents on 2-mile segments of 2-lane rural highways in Washington using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Rear-end Accs.	Proportion	Rankings Using Different π-values	
Site No.	n_i	x_i	θ_i	π=0.5	π=0.75
697	9	7	0.78	1	1
343	14	8	0.57	2	3
72	12	7	0.58	3	4
664	4	4	1.00	4	2
510	10	6	0.60	5	5
173	5	4	0.80	6	6
285	5	4	0.80	6	6
462	5	4	0.80	6	6
67	14	7	0.50	9	21
790	14	7	0.50	9	21
751	8	5	0.63	11	9
468	9	5	0.56	12	23
703	9	5	0.56	12	23
112	3	3	1.00	14	10
288	3	3	1.00	14	10
289	3	3	1.00	14	10
291	3	3	1.00	14	10
388	3	3	1.00	14	10
602	3	3	1.00	14	10
700	3	3	1.00	14	10

Table 6.25: Site rankings for fixed-object accidents on 2-mile segments of 2-lane rural highways in Washington using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Rear-end Accs.	Proportion	Rankings Using Different π-values		
Site No.	n_i	x_i	θ_i	π=0.5	π=0.75	π=0.8
149	43	30	0.70	1	1	1
532	15	13	0.87	2	2	2
782	21	16	0.76	3	3	3
514	23	17	0.74	4	4	4
748	21	15	0.71	5	6	6
779	19	14	0.74	6	5	5
792	12	9	0.75	7	8	8
743	12	9	0.75	7	8	8
784	12	9	0.75	7	8	8
521	8	7	0.88	10	7	7
783	19	12	0.63	11	18	18
750	17	11	0.65	12	17	17
808	13	9	0.69	13	16	16
88	11	8	0.73	14	15	15
815	9	7	0.78	15	13	13
503	9	7	0.78	15	13	13
698	7	6	0.86	17	11	11
742	7	6	0.86	17	11	11
150	18	11	0.61	19	23	23
482	12	8	0.67	20	22	22

6.8 Accounting for Accident Severity in Screening for Proportions

6.8.1 EPDO Accidents

As described in the peak-searching algorithm, equivalent property damage only (EPDO) accidents are used to account for the very large costs to society that are incurred by injury accidents, as compared to property damage only (PDO) accidents.

To incorporate EPDO accidents into the framework of screening for proportions, the observed accident counts at each site, n_i and x_i, are modified by multiplying the number of observed injury accidents, $x_{i(FI)}$, by a relative weight (or relative cost factor), RC_{FI}.

To calculate the relative weight of injury accidents, a cost must be ascribed to each level of accident severity as recorded in the dataset. In the case of the California intersection datasets, accident severity was classified as either fatal (K), severe (A), visible (B), possible (C) injuries, or property damage only (O). These classifications correspond to the KABCO injury scale (*21*), which describes the costs associated with injury accidents in the United States. The KABCO scale is shown in Table 6.26.

The severity weight, SW_j, where j=K,A,B,C,O, is then calculated by dividing the cost of each severity level by the PDO cost. In this way, the economic costs of each level of severity can be described as a multiple of the cost of a PDO accident (hence, equivalent PDO), with PDO accidents having a severity weight of 1. Table 6.27 shows the severity weight for each severity level, and the proportion of all accidents that were of the given severity, P_j, for the California signalized intersections in dataset 2.

Table 6.26: KABCO Scale, from National Safety Council (*22*).

Code	Severity	Cost (2000, US $)
K	Fatal	$3,214,290
A	Incapacitating Injury	$159,449
B	Visible Injury	$41,027
C	Possible Injury	$19,528
O	Property Damage Only	$1,861

Table 6.27: Severity weights and accident proportions of injury accidents for California rural signalized intersections.

Severity	Cost (NSC)	Severity Weight, SW_j	Proportion, P_j
Fatal (K)	$3,214,290	1727	0.01
Incapacitating (A)	$159,449	86	0.03
Visible (B)	$41,027	22	0.12
Possible (C)	$19,528	10	0.23
PDO (O)	$1,861	1	0.61

The relative weight for a given injury accident is given by the following:

$$RC_{FI} = P_K SW_K + P_A SW_A + P_B SW_B + P_C SW_C \quad . \tag{6.30}$$

For the 108 signalized intersections in dataset 2, RC_{FI} was found to be 23.6. Thus, on average, an FI accident was over 20 times more costly to society than did the average PDO accident.

The observed accident counts were then modified as shown in the following:

$$n_{i(EPDO)} = n_{i(PDO)} + RC_{FI} n_{i(FI)} \quad , \tag{6.31}$$

and

$$x_{i(EPDO)} = x_{i(PDO)} + RC_{FI} x_{i(FI)} \quad , \tag{6.32}$$

where the subscripts PDO and FI represent the counts of PDO and injury accidents, respectively, at site *i*.

Once the observed accident counts have been modified to EPDO accidents, parameters for the beta prior distribution can be estimated as before. Table 6.28 shows maximum likelihood parameter estimates for prior beta distributions for California signalized intersection dataset. Tables 6.29 and 6.30 show the results of screening for EPDO broadside accidents at California signalized intersections.

Table 6.28: Maximum likelihood parameter estimates for EPDO accidents at California signalized intersections.

	Total Target Accidents	EPDO Target Accidents	Maximum Likelihood Parameter Estimates	
Accident Type	Σx_i	$\Sigma x_{i(EPDO)}$	α	β
Head-on	128	1048	0.280	3.92
Rear-end	1141	3382	4.42	6.30
Broadside	747	5065	0.849	1.29
Hit Object	151	906	2.86	43.0
Overturning	28	193	1.16	105
Pedestrian	27	404	0.163	11.6
Other/Unknown	77	478	0.199	4.57

Table 6.29: Values of θ^* for broadside accidents at California signalized intersections, using different π-values.

Percentage, π	Critical Proportion, θ^*	No. of flagged sites
0.5	0.36	42
0.75	0.62	18
0.8	0.67	14
0.9	0.81	5
0.95	0.89	4

Table 6.30: Site rankings for broadside EPDO accidents at California signalized intersections using different π-values. Where the site has not been flagged, the ranking is shaded.

	Total Accs.	Broadside Accs.	Broadside FI Accs.	Total EPDO Accs.	Broadside EPDO Accs.	Rankings Using Different π-values			
Site No.	n_i	x_i	$x_{i(FI)}$	$n_{i(EPDO)}$	$x_{i(EPDO)}$	π=0.5	π=0.8	π=0.9	π=0.95
84	22	15	9	234	227	1	1	1	1
98	28	19	7	193	184	1	1	2	2
57	18	10	5	136	128	1	3	3	3
87	9	5	3	80	76	1	4	4	4
35	13	6	3	84	77	1	5	5	5
28	26	12	6	191	154	1	6	12	17
56	23	10	5	165	128	1	8	15	27
83	32	15	5	197	133	1	18	38	46
77	61	25	6	250	167	1	20	43	46
24	52	29	8	335	218	1	25	52	46
97	38	11	3	109	82	11	13	17	26
53	54	20	7	314	185	11	46	57	46
74	10	2	2	57	49	13	7	8	8
34	27	10	6	263	152	14	49	57	46
67	32	13	5	221	131	15	43	54	46
103	23	5	2	70	52	16	16	16	18
66	7	4	1	31	28	17	10	7	7
91	3	1	1	27	25	18	9	6	6
31	28	17	3	146	88	19	34	48	46
65	22	12	4	187	106	20	45	56	46

6.8.2 Joint Probabilities of Target Accidents and FI Accidents

Another way to incorporate accident severity into the screening for proportions methodology is to calculate the joint probability that a given site experiences both a greater than normal proportion of target accidents and a greater than normal proportion of injury accidents. The joint probability is given by the product of the marginal distributions of target accidents and FI accidents, and is written as:

$$\Pr(\theta_{i(type)} > \theta^{*}_{(type)}, \theta_{i(FI)} > \theta^{*}_{FI}) = \Pr(\theta_{i(type)} > \theta^{*}_{(type)})\ \Pr(\theta_{i(FI)} > \theta^{*}_{FI}) \ , \qquad (6.33)$$

where the subscript (*type*) indicates the target accident type. The marginal probabilities in Eq. 6.33 are simply the pattern scores for the target accidents and FI accidents, as calculated in Section 6.7.2. It is very important to note that use of Equation 6.33 assumes independence between injury accidents and the given target accident type, and this is questionable.

This approach was used by Bolduc and Bonin (*32*) to calculated the joint probability that sites experienced both a high proportions of Monday-Tuesday-Wednesday accidents and Thursday-Friday accidents. They compared the results of calculations based on joint probabilities with those based on an assumption of a Dirichlet-multinomial model rather the univariate beta-binomial model used here. The Dirichlet-multinomial approach is not appropriate in the present case, as FI accidents and the given target accidents are not mutually exclusive events; thus, the marginal probabilities must be used.

To investigate the effectiveness of this approach, the marginal probability for FI accidents is combined with rear-end and broadside marginal probabilities using the California signalized intersection dataset. Only maximum likelihood estimates were used, and the median only was used for all values of $\theta*$ (i.e.: π=0.5). For δ=0.9, only one site was flagged for high proportions of both broadside and injury accidents, and no sites were flagged for rear-end and injury accidents. It would, of course, be possible to flag more sites by changing δ to some lower value, but this is at the expense of a larger type I error.

It is clear by the small number of sites flagged that this method would be of little practical use.

6.9 Comparison of Screening for Proportions Approach with an SPF-Based Approach to Screening for High Frequency of Target Accidents

6.9.1 General

One of the main goals of screening for high proportions of specific accident types is to identify those sites that may benefit from the application of a given road safety countermeasure. For example, if a jurisdiction was considering the installation of shoulder rumble strips, it would be desirable to find those sites with a high proportion of run-off-road accidents.

It is usually assumed that the application of a given countermeasure will alter the expected accident frequency by a certain proportion; this proportion is known as an accident modification factor, or AMF. Keeping with the rumble strips example, suppose that the AMF for run-off-road accidents after installation is 0.8; thus, for a site with an expected run-off-road accident frequency of 10 accidents per year, one would expect to observe 10x0.8=8 run-off-road accidents per year after installation.

If one can put faith in the validity of AMFs, then the sites that would benefit the most from a given countermeasure would be those sites with the highest frequency of the appropriate target accident(s). This would be in keeping with what Hauer et al. (*35*) deems the most-bang-for-the buck (MBB) principle. In simple terms, it is more important to reduce the frequency of a given accident than the proportion of that accident.

The basic principles of screening for expected accident frequency and excess expected accident frequency have been discussed in Chapter 5, with applications to road segments. In this section, the screening is applied to intersections, which requires some modification of the road segment screening method. The approach for intersections is described in the next subsection.

6.9.2 Development of Safety Performance Functions

As a first step in the empirical Bayes estimation of expected accident frequency, a safety performance function must be developed. The safety performance functions are again developed by assuming that accident counts at a given site follow a negative binomial distribution.

The SPFs for intersections were calibrated using the GENMOD procedure in SAS as before. The key difference for the intersection models was the use of two independent variables, $AADT_1$ and $AADT_2$, representing the AADTs of the major and minor approaches, respectively. The SPFs for the two California intersection datasets had the following general form:

$$SP = \alpha\, AADT_1^{\beta_1}\, AADT_2^{\beta_2} \quad , \tag{6.34}$$

where SP is the predicted accident frequency in accidents/year, and α, β_1, β_2 are parameters.

Tables 6.31 and 6.32 show the SPF parameter estimates for California rural TWSC and signalized intersections, respectively. Recall that the dispersion parameter k, is such that as $k \to \infty$, the negative binomial distribution approaches a Poisson distribution.

Table 6.31: SPF parameters at California rural TWSC intersections.

Accident Type	Parameter Estimates			Overdispersion
	ln(α)	β_1	β_2	k
Total	-8.3	0.67	0.40	0.64
FI	-8.6	0.57	0.38	0.86
Head-on	-10	0.50	0.44	1.1
Sideswipe	-10	0.62	0.34	0.93
Rear-end	-12	1.0	0.30	1.1
Broadside	-9.6	0.58	0.59	1.3
Hit Object	-8.2	0.55	0.19	0.69
Overturning	dnc	dnc	dnc	dnc
Pedestrian	-16	1.1	0.24	2.5
Other/Unknown	-9.4	0.60	0.15	0.39

Table 6.32: SPF parameters at California rural signalized intersections.

Accident Type	Parameter Estimates			Overdispersion
	ln(α)	β_1	β_2	k
Total	-6.2	0.64	0.20	0.32
FI	dnc	dnc	dnc	dnc
Head-on	dnc	dnc	dnc	dnc
Sideswipe	-7.9	0.50	0.29	0.48
Rear-end	-9.5	0.83	0.27	0.54
Broadside	dnc	dnc	dnc	dnc
Hit Object	dnc	dnc	dnc	dnc
Overturning	dnc	dnc	dnc	dnc
Pedestrian	dnc	dnc	dnc	dnc
Other/Unknown	dnc	dnc	dnc	dnc

For the TWSC intersections, the negative binomial models converged in all but one case; for the signalized intersection data, however, the models converged for only total, sideswipe, and rear-end accidents. The explanation for this difference lies with the difference in sample size: with only 108 sites and a wide range of AADT data, it is more difficult to extract statistically significant parameter estimates.

6.9.3 Screening Based on Expected Accident Frequency and Excess Accident Frequency of Specific Target Accidents for Intersections

The procedure for calculating EB-adjusted expected accident frequency at intersections is similar to the one used in the peak-searching algorithm. As in that approach to screening, a model prediction is combined with site-specific observations using empirical Bayes methods.

For the purposes of this thesis, the only total accidents of a given type were considered; FI, PDO, and EPDO calculations were not performed, although these would be done in a similar fashion to those described in the peak-searching algorithm. A detailed description of the method for intersections is given below.

First, SPF model predictions, κ_y, y=1,2,...,Y were calculated using the SPF from equation 6.34 and the estimated parameters from Section 6.9.1. Thus,

$$\kappa_{y(type)} = \alpha AADT_1^{\beta_1} AADT_2^{\beta_2} \qquad (6.35)$$

where κ_y, is measured in accidents/mi/yr, and the subscript $type$ denotes the accident type (e.g., head-on, broadside, etc.).

Next, the yearly correction factor, C_y, was computed for each year:

$$C_y = \frac{\kappa_{y(type)}}{\kappa_{1(type)}} \quad , \tag{6.36}$$

where κ_1, is the model prediction for year 1.

The EB weight was then found by:

$$w_{(type)} = \frac{1}{1 + k_{(type)} \sum_{y=1}^{Y} \kappa_{y(type)}} \quad . \tag{6.37}$$

The EB-adjusted expected number of accidents, X_1, for total accident of a given type for year 1 were calculated via the following:

$$X_{1(type)} = w_{(type)} \kappa_{1(type)} + \left(1 - w_{(type)}\right) \left(\frac{\sum_{y=1}^{Y} K_{y(type)}}{\sum_{y=1}^{Y} C_{y(type)}} \right) \quad , \tag{6.38}$$

where K_y is the observed accident count in year y. Note that Equation 6.38 is similar to Equation 5.8 from the peak-searching algorithm. The only difference is that the segment length variable, SL, has been omitted for intersections (i.e.: SL is assumed to be 1).

The expected accident frequency is calculated for year Y by:

$$X_{Y(type)} = X_{1(type)} C_{Y(type)} \quad . \tag{6.39}$$

and the variance of X_Y is given by:

$$Var\left(X_{Y(type)}\right) = X_{Y(type)}\left(1 - w_{(type)}\right)\left(\frac{C_{Y(type)}}{\sum_{y=1}^{Y} C_{y(type)}}\right) \quad . \tag{6.40}$$

Where sites were ranked based on the expected accident frequency, $PSI=X_{Y(type)}$. Sites were then ranked based on PSI. While the variance of the estimates is included in the output, it does not affect the rankings.

If sites were to be ranked based on the excess accident frequency of a specific accident type, $Excess_{Y(type)}$, two more equations were needed; estimates of excess accident frequency and the variance of the estimate are given by:

$$Excess_{Y(type)} = X_{Y(type)} - \kappa_{Y(type)} \tag{6.41}$$

and

$$Var\left(Excess_{Y(type)}\right) = Var\left(X_{Y(type)}\right) + \frac{\left(\kappa_{Y(type)}\right)^2}{k_{(type)}} \quad . \tag{6.42}$$

6.9.4 Comparison of Results of Screening for EB-Expected Accident Frequency Screening for High Proportion of Accidents for Specific Accident Types

The method of screening for high frequency of a specific accident was applied to broadside and rear-end accidents at California rural TWSC intersections. The results of this screening, and the results of screening for high proportions for comparison, are shown in Tables 6.33 and 6.34.

For both rear-end and broadside accidents, most of the top 10 sites ranked based on proportion were included in the top 20 sites ranked based on expected accident frequency; thus these sites would probably be flagged for detailed investigation regardless of which of the two screening methods were used.

There were, however, some discrepancies between the frequency-ranked lists and the proportion-ranked lists. For example, from Table 6.33, site No. 1272 is ranked 14th with respect to frequency screening, but it is ranked 2096th with respect to proportion screening. The expected accident frequency screening method will generally favour sites with high accident counts – indeed, this is the basis of the method – and not necessarily sites with high proportion. Keeping with Site No. 1272 as an example, that location experienced 10 rear-end accidents out of a total of 71. Thus, the target accident count is fairly high (10), but the proportion of rear-end accidents is low (0.14).

Table 6.33: Results of screening for high frequency of rear-end accidents at California TWSC intersections. Ranking from screening for high proportions method is shown for comparison.

	Total Accidents	Count of Rear-end Accidents	Expected Rear-end Accident Frequency (acc/yr)		
Site No.	**ni**	**xi**	**X**	**Rank (freq)**	**Rank (prop)**
1716	23	16	3.23	1	3
1003	31	19	3.17	2	2
464	26	19	3.05	3	1
1977	29	16	3.00	4	14
152	34	17	2.71	5	16
1205	31	12	2.64	6	48
176	34	14	2.59	7	34
300	28	16	2.55	8	13
302	22	15	2.54	9	6
1195	33	15	2.54	10	23
146	24	16	2.41	11	4
1095	24	16	2.10	12	4
881	20	10	2.05	13	33
1272	71	10	2.03	14	2096
652	37	15	2.01	15	32
1504	27	12	1.96	16	35
303	20	11	1.91	17	24
149	16	12	1.89	18	9
806	20	11	1.88	19	24
1585	20	14	1.85	20	7

Table 6.34: Results of screening for high frequency of broadside accidents at California TWSC intersections. Ranking from screening for high proportions method is shown for comparison.

	Total Accidents	Count of Broadside Accidents	Expected Broadside Accident Frequency (acc/yr)		
Site No.	ni	xi	X	Rank (freq)	Rank (prop)
1006	61	49	8.87	1	1
1719	40	34	6.78	2	3
1272	71	34	6.75	3	137
153	48	39	6.57	4	2
157	50	38	6.46	5	4
610	45	34	6.43	6	6
482	39	28	5.84	7	15
1539	44	30	5.65	8	18
652	37	30	5.65	9	5
2103	43	21	4.97	10	174
919	37	25	4.79	11	30
318	36	26	4.59	12	17
1270	33	23	4.53	13	29
1012	32	23	4.42	14	24
861	29	23	4.41	15	10
1005	30	24	4.34	16	9
499	38	26	4.34	17	25
1188	34	22	4.12	18	47
305	26	22	4.09	19	8
464	26	21	4.04	20	12

6.9.5 Comparison of Results of Screening for Excess Frequency to Screening for High Proportion of Specific Accident Types

Tables 6.35 and 6.36 show the results of screening for excess accident frequency a specific accident types compared to screening for proportions of specific accident types, for both rear-end and broadside accidents. The two methods agree more closely than the expected-frequency-based ranks did with the proportion-based ranks. For broadside accidents, the top six excess-ranked sites also ranked in the top six proportion-ranked sites.

It should not be surprising that these lists have similarities; if a site is experiencing a higher frequency of target accidents than is expected, as compared with other sites, it is also probable that that site will also be experiencing a higher proportion of those target accidents.

Table 6.35: Results of screening for expected excess frequency of rear-end accidents at California TWSC intersections. Ranking from screening for high proportions method and screening for expected frequency are shown for comparison.

	Total Accidents	Count of Rear-end Accidents	Expected Excess Rear-end Accident Frequency (acc/yr)	Excess Freq.	Freq.	Prop
Site No.	ni	xi	Excess(X)	Rank	Rank	Rank
1003	31	19	2.61	1	2	2
464	26	19	2.50	2	3	1
152	34	17	2.16	3	5	16
1716	23	16	2.16	4	1	3
1977	29	16	2.11	5	4	14
146	24	16	1.99	6	11	4
300	28	16	1.97	7	8	13
1095	24	16	1.79	8	12	4
302	22	15	1.79	9	9	6
1195	33	15	1.74	10	10	23
652	37	15	1.71	11	15	32
1585	20	14	1.54	12	20	7
144	24	13	1.50	13	21	20
272	27	13	1.42	14	24	26
149	16	12	1.41	15	18	9
1504	27	12	1.39	16	16	35
617	19	13	1.38	17	28	12
806	20	11	1.28	18	19	24
1198	17	11	1.24	19	26	19
1424	25	11	1.24	20	22	41

Table 6.36: Results of screening for expected excess frequency of broadside accidents at California TWSC intersections. Ranking from screening for high proportions method and screening for expected frequency are shown for comparison.

	Total Accidents	Count of Broadside Accidents	Expected Excess Broadside Accident Frequency (acc/yr)	Excess Freq.	Freq.	Prop.
Site No.	ni	xi	Excess(X)	Rank	Rank	Rank
1006	61	49	7.96	1	1	1
157	50	38	5.69	2	5	4
153	48	39	5.33	3	4	2
1719	40	34	5.08	4	2	3
610	45	34	5.05	5	6	6
652	37	30	4.23	6	9	5
1539	44	30	4.13	7	8	18
482	39	28	3.93	8	7	15
499	38	26	3.87	9	17	25
318	36	26	3.70	10	12	17
919	37	25	3.61	11	11	30
959	31	26	3.48	12	23	7
1005	30	24	3.34	13	16	9
1498	31	24	3.31	14	24	11
861	29	23	3.14	15	15	10
1188	34	22	2.84	16	18	47
1639	28	20	2.74	17	31	31
2103	43	21	2.70	18	10	174
305	26	22	2.70	19	19	8
144	24	19	2.59	20	38	20

6.10 Chapter Conclusions

The method of screening for high proportions of specific accident types is relatively new, and has not yet been widely applied. SPF-based screening methods have been widely used throughout the world in the past few years; however, using SPF methods to screen for specific accident types has not been commonly used as compared with screening for total, FI, or EPDO accidents.

Possibly the greatest advantage of using the proportion-based screening is that the method is not data-intensive; the data requirements are accident counts by type (and severity, if desired) and enough basic information to classify the site as a particular type (e.g., 4-leg signalized urban intersections).

Of the three parameter estimation techniques, maximum likelihood and MM2 had the best performance. The MM1 technique was the only one to yield estimates in every case; this was because the solution was computed directly, rather than using an iterative method as was the case with the ML and MM2 methods. MM1 estimates, however, yielded the largest variance in every case, and were heavily influenced by observations with very low accident counts. The MM1 estimates for all sites that experienced accidents did not perform well when compared to data containing only sites with higher numbers of accidents. The MM1 estimates also did not perform as well as ML and MM2 when modelling data for only 3 years as opposed to 5. The ML and MM2 estimates, in contrast, agreed closely for both 3 and 5 years of data. For these reasons, the MM1 method is not recommended for calibration of the beta prior distributions.

The maximum likelihood and ML2 estimates agreed very closely in almost every case examined. Both methods yielded similar estimates for 3 and 5 years of data, indicating a reasonable degree of stability. Both ML and MM2 estimates failed to converge in some cases; however, estimates converged for every accident type representing greater than 2% of all accidents.

The value of π, which dictates the size of the critical theta value, θ^*, may be set to any value from zero to 1. The larger the value of π, the larger the value of θ^*, and fewer sites are ranked; however, those sites that are ranked will have the highest proportions of all sites in the network, provided a sufficient number of accidents have occurred at the site.

When π is set to a relatively low value, such as 0.5 as for the median, many sites are ranked. In this case, ranked sites may not have observed proportions that are much higher than median; a site with many accidents may have an accident proportion only

20% greater than the median, but still have a high ranking because of the low variance of the observations.

A small value of π favours sites with high accident counts, while a larger value favours those sites with a high proportion of the given accident type.

It was shown in Section 6.5 that the observed accident proportions at some accident types increase as sites with low accident counts are dropped from the analysis. Using the beta-binomial screening method, there is no obvious way to account for these variations other than to group sites by total accidents, which will result in much-reduced sample sizes, and less-reliable models.

The variation in accident proportion can be easily explained by SPF-based methods: for a given AADT, the expected proportion of specific accident type can be defined as the ratio of the expected values of the SPFs for total accidents, and the target accidents as follows:

$$E(\theta)=\frac{SPF_{(type)}}{SPF_{TOT}} \qquad (6.43)$$

Because the SPFs are continuous functions, there is no requirement for θ to be constant over the range of independent variables. In this case, the value of θ would only be constant when the SPF 'slope' parameters, β_{TOT} and $\beta_{(type)}$ are equal.

The SPF-based methods offer the advantage of being able explicitly model one or several independent variables, most importantly AADT. It is clear from the SPF calibration results that accident frequency is a function of AADT, and it would be reasonable to believe that as traffic volume varies, accident characteristics change as well.

In Section 6.9.4, it was shown that SPF-based screening for excess frequency of specific accident types compares reasonably well with screening form proportions; however, the SPF method accounts for other variables (i.e., AADT) that influence accident frequency, and there is no assumption of a constant proportion for a given accident type.

Screening for excess accident frequency has the appeal of quantifying the level of 'unsafety' by showing how much 'worse' (or better) a given site is than other sites of the same kind. It also suggests that the excess should be correctable, and that reducing the accident frequency, to at least the mean frequency for all sites, is an achievable goal.

A fundamental flaw with the screening for proportions method is that a site may appear to experience a higher than normal proportion of some accident types, simply because the site experiences an unusually low proportion of another accident type or types. Thus, if a site is 'safe' with respect to one accident type, it necessarily appear to be less safe with respect to other accident types. This represents a major weakness of the screening method.

Because of the advantages inherent in using SPF-based screening methods described above, and the fundamental weakness of the screening for proportions method, the method of screening for excess frequency of specific target accidents is preferred over the screening for proportions method. The screening for proportions approach appears to be somewhat viable, and could be used where reliable SPFs are not available.

7 Conclusions and Recommendations for Future Work

The screening methods described in this thesis are state-of-the art procedures that may see wide-scale use if they are included in the *SafetyAnalyst* Network Screening Toolbox. All of the methods applied here employ empirical Bayes methods to account for regression-to-the-mean effects, which makes them more reliable than accident count or accident rate screening methods.

The examination of the negative binomial dispersion parameter showed that the assumption of constant overdispersion in SPF models may not be valid. Overdispersion varies as a function of segment length, and possibly as a function of traffic volume. Empirical Bayes estimates of expected accident frequency, on average, do not seem seriously affected by different values of the dispersion parameter, but this may not be the case for individual sites. Further study of this issue is needed.

The peak-searching algorithm was shown to overcome some of the limitations of the sliding window approach, and performed well when applied to real-world data. It was found that low values of CV_{lim} resulted in the ranked list of sites favouring longer segments, and having fewer sites. Larger values of resulted in many more sites being ranked, and tended to favour shorter segments.

The method of screening for high proportions of specific accident types received the most detailed investigation. It was found that the method of maximum likelihood and a modified method of moments (MM2) were preferred for making parameter estimates for the beta prior distributions.

It was shown that by varying the value of the critical proportion, $\theta*$, the number of ranked sites could be varied; the higher the value, the fewer sites were ranked. Also, high values of $\theta*$ favoured sites with high proportions of the target accident type, while lower values favoured sites with high accident counts.

It was also found that the proportion of a given accident type was not constant between high- and low-accident sites.

The screening for proportions method was compared with SPF-based methods for screening for high expected frequency and excess frequency of specific target accidents. The SPF method for screening for excess accident frequency compared well with the screening for proportions method; however, the SPF methods do not make the assumption that the mean proportion of a given target accident is constant. This makes the SPF-based method a more powerful screening tool.

The main advantage of the screening for proportions methodology is that no AADT or geometric data are required; thus, where reliable SPFs are unavailable, the screening for proportions method would be a good alternative.

References

1. “Road Safety in Canada: An Overview,” Report ISBN 0-662-36440-6. Transport Canada and Health Canada. Online version [Accessed June, 2004], http://www.tc.gc.ca/roadsafety/stats/overview/2004/menu.htm.

2. Heydecker, B., and Wu, J. Using the information in road accident records. *Proceedings*, 19th PTRC Summer Annual Meeting, London, September 1991.

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

4. Higle, J.L., and Witkowski, J.M. Bayesian identification of hazardous locations. *Transportation Research Record 1185*, TRB, National Research Council, Washington, D.C., 1988, pp. 22-36.

5. Persaud, B. “Statistical Methods in Highway Safety Analysis,” *NCHRP Synthesis 295.* Transportation Research Board, Washington, D.C., 2001.

6. Hauer, E. Identification of sites with promise. *Transportation Research Record 1542*, TRB, National Research Council, Washington, D.C., 1996, pp. 54-60.

7. Hauer, E. Statistical road safety modeling. Presented at TRB Annual Conference, Washington, D.C, 2004.

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

9. Persaud, B., Lord, D., and Palmisano, J. Calibration and transferability of accident prediction models for urban intersections. *Transportation Research Record 1784*, TRB, National Research Council, Washington, D.C., 2002, pp. 57-64.

10. Myers, R.H., and Montgomery, D.C. A tutorial on generalized linear models. *Journal of Quality Technology*, Vol. 29, No. 3, 1997, pp. 274-291.

11. Miaou, S.-P., and Lum, H. Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis and Prevention*, Vol. 25, No. 6, 1993, pp. 689-709.

12. Qin, X., Ivan, J.N., Ravishanker, N., and Liu, J. A Hierarchical Bayesian Estimation of Non-Linear Safety Performance Functions for Two-Lane Highways Using MCMC Modeling. Submitted for presentation at TRB Annual Meeting, Washington D.C., 2003.

13. Lord, D., Washington, S.P., and Ivan, J.N. Statistical challenges with modeling motor vehicle crashes: understanding the implications of alternative approaches. Presented at TRB Annual Conference, Washington D.C., 2004.

14. Vogt, A., and Bared, J.G. *Accident Models for Two-Lane Rural Roads: Segments and Intersections.* Report FHWA-RD-98-133. FHWA, U.S. Department of Transportation, 1998.

15. Chin, H.C., and Quddus, M.A. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis and Prevention*, Vol. 35, No. 2, 2003, pp. 253-259.

16. Kononov, J., and Allery, B. Level of service of safety: a conceptual blueprint and the analytical framework. Presented at TRB Annual Conference, Washington D.C., 2003.

17. Persaud, B., Cook, W., and Kazakov, A. Demonstration of new approaches for identifying hazardous locations and prioritizing safety treatment. Presented at 7th International Conference: Traffic Safety On Two Continents, Lisbon, September 1997.

18. Persaud, B., Lyon, C., and Nguyen, T. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety Improvement. *Transportation Research Record 1665*, TRB, National Research Council, Washington, D.C., 1999, pp. 7-12.

19. Council, F.M., Williams, C.D., and Mohamedshah, Y.M. *Highway Safety Information System: Guidebook for the Washington Data Files, Volume 1.* FHWA, Washington, D.C., 1995.

20. Council, F.M., Williams, C.D., Chen, L.-w., and Mohamedshah, Y.M. *Highway Safety Information System: Guidebook for the California Data Files, Volume 1.* FHWA, Washington, D.C., 2000.

21. National Safety Council. Estimating the Costs of Unintentional Injuries. http://www.nsc.org/lrs/statinfo/estcost0.htm (accessed May 2, 2004). 2000.

22. Interactive Highway Design Safety Module: Crash Prediction Module Engineer's Manual. FHWA, Washington, D.C., CPM Version 1.01, January, 2004.

23. Hauer, E. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis and Prevention*, Vol. 33, No. 6, 2001, pp. 799-808.

24. Miaou, S.P., and Lord, D. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record 1840*, TRB, National Research Council, Washington, D.C., 2003, pp. 31-40.

25. SAS Institute Inc. *SAS/STAT User's Guide, Version 8.* SAS Institute Inc., Cary N.C., 1999.

26. Der, G., and Everitt, B.S. *A Handbook of Statistical Analyses Using SAS*, 2nd ed. Chapman and Hall, New York, 2002.

27. Anderson, I., Bauer, K.M., Collins, J.M., Fitzpatrick, K., Green, P., Harwood, D.W., Koppa, R., Krammes, R.A., Parma, K.D., Poggioli, B., Tsimhoni, O., and Wooldridge, M.D. *Alternative Design Consistency Rating Methods for Two-Lane Rural Highways*, . Report FHWA-RD-99-172. FHWA, U.S. Department of Transportation, 1999.

28. Ng, J.C.W., and Sayed, T. Effect of geometric design consistency on road safety. *Canadian Journal of Civil Engineering*, Vol. 31, No. 2, 2004, pp. 218-227.

29. Hauer, E. Unpublished Working Paper 10. Screening: Methods and Software. *Draft*. 2000.

30. SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites – Task M: Functional Specification for Module 1 – Network Screening. Final Draft. Task No. DTFH61-01-F-00096. FHWA, 2003.

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

32. Bolduc, D., and Bonin, S. Bayesian analysis of road accidents: a general framework for the multinomial case. *Cahiers de recherche*, 9802, Université Laval, Département d'économique, 1998.

33. Bolduc, D., and Bonin, S. Bayesian analysis of road accidents: an application of the multinomial setting. *Proceedings*, Canadian Multidisciplinary Road Safety Conference XI, Halifax, May 1999.

34. Mollett, C.J. Developing a traffic safety improvement program: a review and comparison of different network screening approaches. *Proceedings*, Canadian Multidisciplinary Road Safety Conference XIV, Ottawa, June 2004.

35. Hauer, E., Kononov, J., Allery, B., and Griffith, M. Screening the road network for sites with promise. *Transportation Research Record 1784*, TRB, National Research Council, Washington, D.C., 2002, pp. 27-32.

Appendix A

Regression Results

Table A.1: Negative binomial regression parameter estimates and goodness-of-fit statistics for 2-miie segments of 2-lane rural highway in Washington. Segments are progressively broken down into smaller segments. Crash data for 1993-1995.

Seg. Length	No. Obs.	Parameter Estimates										Goodness-of-Fit Statistics	
(mi)	n	ln(α)	Std. Error	X^2	Pr(<X^2)	β	Std. Error	X^2	Pr(<X^2)	k	Std. Error	Deg. Of Freedom	Pearson X^2
2	831	-7.22	0.31	554	<.0001	0.91	0.041	487	<.0001	0.55	0.05	829	0.95
1	1662	-7.06	0.26	759	<.0001	0.88	0.034	679	<.0001	0.59	0.04	1660	0.99
0.5	3324	-6.90	0.22	978	<.0001	0.86	0.029	892	<.0001	0.61	0.05	3322	1.01
0.25	6648	-6.81	0.20	1117	<.0001	0.85	0.027	1031	<.0001	0.79	0.06	6646	1.00
0.1	16620	-6.70	0.19	1301	<.0001	0.84	0.024	1219	<.0001	1.00	0.08	16618	0.98
0.05	33240	-6.67	0.18	1363	<.0001	0.83	0.023	1283	<.0001	1.54	0.12	33238	0.99
0.02	83100	-6.65	0.18	1399	<.0001	0.83	0.023	1321	<.0001	3.23	0.24	83098	0.98
0.01	166200	-6.64	0.18	1413	<.0001	0.83	0.828	1336	<.0001	6.00	0.43	166198	0.98

Table A.2: Negative binomial regression parameter estimates and goodness-of-fit statistics for 5 bins of different AADT ranges, for 2-lane rural highways in Washington. Crash data for 1993-1995.

Bin	No. Obs.	Parameter Estimates										Goodness-of-Fit Statistics	
	n	ln(α)	Std. Error	X^2	Pr(<X^2)	β	Std. Error	X^2	Pr(<X^2)	k	Std. Error	Deg. Of Freedom	Pearson X^2
1	1479	-8.00	0.53	226	<.0001	1.1	0.079	198	<.0001	0.60	0.059	1477	2.08
2	1407	-8.34	1.24	45	<.0001	1.1	0.16	49	<.0001	0.43	0.040	1405	1.73
3	1161	-2.04	1.77	1.32	0.25	0.35	0.21	2.71	0.10	0.44	0.038	1159	1.12
4	885	-8.76	2.10	17.4	<.0001	1.1	0.24	22.4	<.0001	0.42	0.045	883	1.47
5	860	-7.09	1.13	39.5	<.0001	0.94	0.12	61.7	<.0001	0.40	0.038	858	1.35

Table A.3: Regression results for k vs. segment length, using PROC NLIN in SAS®.

Model:	$k=(\beta_1+SL)/(\beta_2*SL)$	
Regression Method:	Gauss-Newton	
Parameters	Estimate	Approx Std. Error
β_1	0.107	0.0036
β_2	1.96	0.056
Mean-squared residual:	0.0017	
Fit OK?	YES	

Table A.4: Negative binomial regression parameter estimates and goodness-of-fit statistics for total and FI accident, for 2-lane rural highways in Washington, used in the peak-searching algorithm. Crash data for 1993-1995.

Acc Type	No. Obs.	Parameter Estimates										Goodness-of-Fit Statistics	
	n	ln(α)	Std. Error	X^2	Pr(<X^2)	β	Std. Error	X^2	Pr(<X^2)	k	Std. Error	Deg. Of Freedom	Pearson X^2
Total	5792	-6.69	0.14	2185	<.0001	0.87	0.017	2503	<.0001	0.49	0.023	5790	1.50
FI	5792	-7.47	0.17	1902	<.0001	0.87	0.021	1760	<.0001	0.48	0.031	5790	1.42

Table A.5a: Negative binomial regression goodness-of-fit statistics for rural, 4-leg, signalized intersections in California. Crash data for 5 years, 1997-2001.

Accident Type	Parameter Estimates													
	ln(α)	Std. Error	X^2	Pr(<X^2)	β_1	Std. Error	X^2	Pr(<X^2)	β_2	Std. Error	X^2	Pr(<X^2)	k	Std. Error
Total	-6.25	1.28	24	<.0001	0.64	0.13	24	<.0001	0.20	0.043	22	<.0001	0.32	0.049
FI	-7.45	2.1	13	0.0003	0.64	0.21	9.1	0.0026	0.12	0.066	3.4	0.064	0.58	0.13
Head-on	-7.01	2.9	5.8	0.016	0.47	0.30	2.5	0.12	0.13	0.09	2.1	0.14	0.82	0.26
Sideswipe	-7.90	2.1	14	0.0002	0.50	0.22	5.1	0.0243	0.29	0.07	16	<.0001	0.48	0.14
Rear-end	-9.51	1.6	34	<.0001	0.83	0.17	25	<.0001	0.27	0.06	23	<.0001	0.54	0.09
Broadside	-5.00	1.8	7.8	0.0053	0.47	0.19	6.3	0.0118	0.10	0.056	3.4	0.0671	0.46	0.09
Hit Object	-5.69	2.5	5.2	0.0221	0.37	0.26	2.1	0.1471	0.45	0.07	1.9	0.1633	0.45	0.17
Overturn.	-9.63	5.2	3.5	0.0618	0.57	0.54	1.1	0.2863	0.14	0.16	0.83	0.3612	1.5	1.0
Pedestrian	-4.30	6.7	0.42	0.5185	0.34	0.70	0.23	0.6316	0.25	0.20	1.6	0.2137	4.0	2.0
Other/Unk.	-11.4	3.0	14	0.0002	0.86	0.32	7.4	0.0067	0.15	0.08	3.2	0.0737	0.26	0.24

Table A.5b: Negative binomial regression parameter estimates for rural, 4-leg, signalized intersections in California. Crash data for 1997-2001.

Accident Type	Goodness-of-Fit Statistics		
	Degrees Of Freedom	Pearson X^2	Model Valid?
Total	105	1.0	Yes
FI	105	0.87	No – parameter(s) not significant.
Head-on	105	0.94	No – parameter(s) not significant.
Sideswipe	105	1.0	Yes
Rear-end	105	1.0	Yes
Broadside	105	1.1	No – parameter(s) not significant.
Hit Object	105	1.0	No – parameter(s) not significant.
Overturning	105	0.95	No – parameter(s) not significant.
Pedestrian	105	1.1	No – parameter(s) not significant.
Other/Unk.	105	1.1	No – parameter(s) not significant.

Table A.6a: Negative binomial regression goodness-of-fit statistics for rural, 4-leg, TWSC intersections in California. Crash data for 5 years, 1997-2001.

Accident Type	Parameter Estimates													
	ln(α)	Std. Error	X^2	Pr(<X^2)	β_1	Std. Error	X^2	Pr(<X^2)	β_2	Std. Error	X^2	Pr(<X^2)	k	Std. Error
Total	-8.26	0.24	1190	<.0001	0.67	0.03	634	<.0001	0.40	0.01	733	<.0001	0.64	0.03
FI	-8.63	0.35	624	<.0001	0.57	0.04	228	<.0001	0.38	0.02	298	<.0001	0.86	0.06
Head-on	-10.4	0.75	194	<.0001	0.50	0.08	38	<.0001	0.44	0.05	93	<.0001	1.1	0.29
Sideswipe	-10.1	0.52	371	<.0001	0.62	0.06	118	<.0001	0.34	0.03	112	<.0001	0.93	0.14
Rear-end	-12.4	0.46	729	<.0001	1.0	0.05	397	<.0001	0.30	0.03	146	<.0001	1.1	0.09
Broadside	-9.58	0.38	642	<.0001	0.58	0.04	196	<.0001	0.59	0.02	612	<.0001	1.3	0.08
Hit Object	-8.19	0.40	420	<.0001	0.55	0.04	152	<.0001	0.19	0.03	57	<.0001	0.69	0.10
Overturn.	-8.91	0.78	130	<.0001	0.54	0.09	38	<.0001	0.056	0.05	1.3	0.2623	1.3	0.47
Pedestrian	-15.8	1.5	110	<.0001	1.1	0.16	47	<.0001	0.24	0.08	8.8	0.0030	2.5	1.0
Other/Unk.	-9.37	0.59	257	<.0001	0.60	0.06	86	<.0001	0.15	0.04	18	<.0001	0.39	0.17

Table A.6b: Negative binomial regression parameter estimates for rural, 4-leg, TWSC intersections in California. Crash data for 1997-2001.

Accident Type	Goodness-of-Fit Statistics		
	Degrees Of Freedom	Pearson X^2	Model Valid?
Total	2199	1.2	Yes
FI	2199	1.0	Yes
Head-on	2199	0.97	Yes
Sideswipe	2199	1.1	Yes
Rear-end	2199	1.0	Yes
Broadside	2199	1.5	Yes
Hit Object	2199	1.0	Yes
Overturning	2199	1.0	No - parameter(s) not significant.
Pedestrian	2199	0.88	Yes
Other/Unk.	2199	1.1	Yes

Appendix B

Peak-Searching Results

Table B.1: Results of peak-searching algorithm for expected accident frequency of FI accidents, using CV_{lim}=1.5.

Screening Criterion:	Expected accident frequency
Accident Type:	FI accidents
CV_{lim}	1.5
Number of Sites Ranked:	93/100
Mean length of all ranked sites (mi):	1.87
Mean length of all flagged windows (mi):	0.13

Site No.	Site Length (mi)	Length of Flagged Window (mi)	$Peak(X_Y)$ (acc/mi/yr)	$Var[Peak(X_Y)]$	$CV[Peak(X_Y)]$	Rank
22	0.89	0.1	8.68	17.91	0.49	1
12	1.18	0.1	7.08	22.68	0.67	2
5	0.64	0.1	6.67	20.16	0.67	3
6	0.25	0.1	6.63	19.87	0.67	4
26	2.4	0.1	6.08	11.51	0.56	5
15	0.13	0.1	6.07	16.68	0.67	6
7	0.41	0.1	6.07	16.68	0.67	7
48	0.24	0.1	5.46	13.51	0.67	8
52	1.06	0.1	4.88	10.76	0.67	9
23	0.13	0.1	4.70	10.00	0.67	10
24	0.34	0.1	4.69	9.97	0.67	11
9	3.27	0.1	3.47	9.96	0.91	12
8	0.13	0.1	3.46	9.88	0.91	13
10	2.87	0.1	3.44	9.76	0.91	14
4	2.8	0.1	3.36	9.34	0.91	15
17	1.32	0.1	3.34	9.20	0.91	16
11	0.72	0.1	3.33	9.17	0.91	17
16	0.84	0.1	3.33	9.16	0.91	18
1	1.73	0.1	3.22	8.57	0.91	19
14	0.82	0.1	3.21	8.51	0.91	20

Table B.2: Results of peak-searching algorithm for expected accident frequency of FI accidents, using CV_{lim}=1.0.

Screening Criterion:	Expected accident frequency
Accident Type:	FI accidents
CV_{lim}	1.0
Number of Sites Ranked:	86/100
Mean length of all ranked sites (mi):	1.99
Mean length of all flagged windows (mi):	0.18

Site No.	Site Length (mi)	Length of Flagged Window (mi)	$Peak(X_Y)$ (acc/mi/yr)	$Var[Peak(X_Y)]$	$CV[Peak(X_Y)]$	Rank
22	0.89	0.1	8.68	17.91	0.49	1
12	1.18	0.1	7.08	22.68	0.67	2
5	0.64	0.1	6.67	20.16	0.67	3
6	0.25	0.1	6.63	19.87	0.67	4
26	2.4	0.1	6.08	11.51	0.56	5
15	0.13	0.1	6.07	16.68	0.67	6
7	0.41	0.1	6.07	16.68	0.67	7
48	0.24	0.1	5.46	13.51	0.67	8
52	1.06	0.1	4.88	10.76	0.67	9
23	0.13	0.1	4.70	10.00	0.67	10
24	0.34	0.1	4.69	9.97	0.67	11
9	3.27	0.1	3.47	9.96	0.91	12
8	0.13	0.1	3.46	9.88	0.91	13
10	2.87	0.1	3.44	9.76	0.91	14
4	2.8	0.1	3.36	9.34	0.91	15
17	1.32	0.1	3.34	9.20	0.91	16
11	0.72	0.1	3.33	9.17	0.91	17
16	0.84	0.1	3.33	9.16	0.91	18
1	1.73	0.1	3.22	8.57	0.91	19
14	0.82	0.1	3.21	8.51	0.91	20

Table B.3: Results of peak-searching algorithm for expected accident frequency of FI accidents, using CV_{lim}=0.5.

Screening Criterion:	Expected accident frequency
Accident Type:	FI accidents
CV_{lim}	0.5
Number of Sites Ranked:	54/100
Mean length of all ranked sites (mi):	2.79
Mean length of all flagged windows (mi):	0.60

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
4	2.8	0.1	11.69	32.48	0.49	1
5	0.64	0.11	11.62	31.89	0.49	2
7	0.41	0.1	11.56	31.77	0.49	3
9	3.27	0.11	11.04	28.78	0.49	4
47	2.7	0.1	10.12	24.31	0.49	5
20	1.76	0.1	8.91	18.86	0.49	6
22	0.89	0.1	8.68	17.91	0.49	7
39	3.93	0.1	7.00	11.64	0.49	8
40	6.87	0.11	6.86	11.12	0.49	9
54	1.2	0.14	6.79	10.72	0.48	10
12	1.18	0.21	6.77	10.34	0.47	11
10	2.87	0.19	6.58	9.83	0.48	12
56	6.2	0.14	6.52	9.89	0.48	13
26	2.4	0.15	5.45	6.88	0.48	14
29	5.37	0.15	5.31	6.53	0.48	15
48	0.24	0.21	5.23	6.16	0.47	16
36	0.76	0.17	4.69	5.05	0.48	17
33	0.91	0.17	4.67	5.02	0.48	18
37	2.07	0.2	4.14	3.88	0.48	19
41	1.56	0.26	3.75	3.09	0.47	20

Table B.4: Results of peak-searching algorithm for expected accident frequency of FI accidents, using CV_{lim}=0.2.

Screening Criterion:	Expected accident frequency
Accident Type:	FI accidents
CV_{lim}	0.2
Number of Sites Ranked:	9/100
Mean length of all ranked sites (mi):	4.49
Mean length of all flagged windows (mi):	4.19

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
9	3.27	1.89	3.94	0.60	0.20	1
4	2.8	1.99	3.51	0.49	0.20	2
47	2.7	1.91	3.15	0.40	0.20	3
25	3.34	2.86	1.68	0.11	0.20	4
40	6.87	2.95	1.52	0.09	0.20	5
29	5.37	3.68	1.29	0.06	0.20	6
56	6.2	4.77	1.11	0.05	0.20	7
86	9.9	8.58	0.27	0.00	0.20	8
67	10.9	9.05	0.22	0.00	0.20	9

Table B.5: Results of peak-searching algorithm for excess frequency of FI accidents, using CV_{lim}=4.0.

Screening Criterion:	Excess accident frequency
Accident Type:	FI accidents
CV_{lim}	4.0
Number of Sites Ranked:	66/100
Mean length of all ranked sites (mi):	2.28
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
22	0.89	0.1	7.48	18.21	0.57	1
26	2.4	0.1	5.13	11.70	0.67	2
12	1.18	0.1	4.48	24.11	1.10	3
15	0.13	0.1	3.92	17.65	1.07	4
48	0.24	0.1	3.72	14.15	1.01	5
52	1.06	0.1	3.67	11.07	0.91	6
23	0.13	0.1	3.50	10.31	0.92	7
24	0.34	0.1	3.46	10.29	0.93	8
5	0.64	0.1	3.45	22.34	1.37	9
6	0.25	0.1	3.44	22.01	1.37	10
1	1.73	0.1	3.31	17.05	1.25	11
9	3.27	0.1	3.24	20.21	1.39	12
11	0.72	0.1	3.20	18.48	1.34	13
7	0.41	0.1	3.02	18.64	1.43	14
4	2.8	0.1	2.92	19.23	1.50	15
96	4.5	0.1	2.59	4.36	0.81	16
3	1.16	0.1	2.46	10.89	1.34	17
41	1.56	0.1	1.61	5.76	1.49	18
19	0.72	0.1	1.52	6.94	1.74	19
56	6.2	0.1	1.47	5.72	1.62	20

Table B.6: Results of peak-searching algorithm for excess frequency of FI accidents, using CV_{lim}=2.5.

Screening Criterion:	Excess accident frequency
Accident Type:	FI accidents
CV_{lim}	2.5
Number of Sites Ranked:	66/100
Mean length of all ranked sites (mi):	2.28
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
22	0.89	0.1	7.48	18.21	0.57	1
26	2.4	0.1	5.13	11.70	0.67	2
12	1.18	0.1	4.48	24.11	1.10	3
15	0.13	0.1	3.92	17.65	1.07	4
17	1.32	0.1	3.89	17.82	1.09	5
14	0.82	0.1	3.86	16.39	1.05	6
48	0.24	0.1	3.72	14.15	1.01	7
10	2.87	0.1	3.68	19.24	1.19	8
52	1.06	0.1	3.67	11.07	0.91	9
13	0.44	0.1	3.61	15.48	1.09	10
47	2.7	0.1	3.51	13.44	1.05	11
23	0.13	0.1	3.50	10.31	0.92	12
24	0.34	0.1	3.46	10.29	0.93	13
5	0.64	0.1	3.45	22.34	1.37	14
6	0.25	0.1	3.44	22.01	1.37	15
1	1.73	0.1	3.31	17.05	1.25	16
49	0.3	0.11	3.29	13.15	1.10	17
9	3.27	0.1	3.24	20.21	1.39	18
11	0.72	0.1	3.20	18.48	1.34	19
7	0.41	0.1	3.02	18.64	1.43	20

Table B.7: Results of peak-searching algorithm for excess frequency of FI accidents, using CV_{lim}=1.0.

Screening Criterion:	Excess accident frequency
Accident Type:	FI accidents
CV_{lim}	1.0
Number of Sites Ranked:	35/100
Mean length of all ranked sites (mi):	2.59
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	$Peak(X_Y)$ (acc/mi/yr)	$Var[Peak(X_Y)]$	$CV[Peak(X_Y)]$	Rank
7	0.41	0.1	8.51	33.73	0.68	1
22	0.89	0.1	7.48	18.21	0.57	2
12	1.18	0.11	6.81	28.70	0.79	3
5	0.64	0.1	6.47	31.46	0.87	4
9	3.27	0.1	6.11	28.44	0.87	5
47	2.7	0.1	5.91	19.22	0.74	6
10	2.87	0.11	5.75	22.82	0.83	7
4	2.8	0.1	5.69	26.95	0.91	8
37	2.07	0.1	5.15	11.42	0.66	9
26	2.4	0.1	5.13	11.70	0.67	10
36	0.76	0.1	5.06	10.88	0.65	11
33	0.91	0.1	4.99	10.84	0.66	12
6	0.25	0.12	4.94	22.09	0.95	13
17	1.32	0.13	4.74	15.51	0.83	14
19	0.72	0.1	3.85	12.38	0.91	15
41	1.56	0.1	3.76	10.35	0.86	16
48	0.24	0.15	3.72	9.44	0.83	17
54	1.2	0.1	3.67	11.13	0.91	18
52	1.06	0.1	3.67	11.07	0.91	19
56	6.2	0.1	3.60	10.23	0.89	20

Table B.8: Results of peak-searching algorithm for excess frequency of FI accidents, using CV_{lim}=0.5.

Screening Criterion:	Excess accident frequency
Accident Type:	FI accidents
CV_{lim}	0.5
Number of Sites Ranked:	9/100
Mean length of all ranked sites (mi):	3.12
Mean length of all flagged windows (mi):	0.49

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
4	2.8	0.1	16.80	57.81	0.45	1
7	0.41	0.1	14.00	48.82	0.50	2
20	1.76	0.1	9.78	23.67	0.50	3
39	3.93	0.11	7.15	12.06	0.49	4
40	6.87	0.31	3.05	2.24	0.49	5
22	0.89	0.71	2.14	1.01	0.47	6
47	2.7	1.27	1.73	0.72	0.49	7
29	5.37	0.78	1.60	0.62	0.49	8
25	3.34	0.97	1.44	0.49	0.49	9

Table B.9: Results of peak-searching algorithm for expected frequency of PDO accidents, using CV_{lim}=2.5.

Screening Criterion:	Expected accident frequency
Accident Type:	PDO accidents
CV_{lim}	2.5
Number of Sites Ranked:	80/100
Mean length of all ranked sites (mi):	2.09
Mean length of all flagged windows (mi):	0.13

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	14.29	43.29	0.46	1
29	5.37	0.1	9.75	25.16	0.51	2
18	0.8	0.1	8.95	29.25	0.60	3
6	0.25	0.1	7.52	67.51	1.09	4
23	0.13	0.1	6.56	40.18	0.97	5
17	1.32	0.1	6.53	24.54	0.76	6
43	0.25	0.1	6.17	30.55	0.90	7
50	0.57	0.1	5.98	20.47	0.76	8
45	0.18	0.1	5.84	19.51	0.76	9
25	3.34	0.1	5.76	24.91	0.87	10
38	1.88	0.1	5.32	15.82	0.75	11
12	1.18	0.1	4.90	67.45	1.68	12
22	0.89	0.1	4.73	52.48	1.53	13
26	2.4	0.1	4.29	37.10	1.42	14
5	0.64	0.1	4.18	56.95	1.80	15
48	0.24	0.1	4.02	41.58	1.61	16
16	0.84	0.1	3.81	32.29	1.49	17
7	0.41	0.1	3.76	46.82	1.82	18
49	0.3	0.1	3.61	28.81	1.49	19
32	1.03	0.1	3.58	24.88	1.39	20

Table B.10: Results of peak-searching algorithm for expected frequency of PDO accidents, using CV_{lim}=1.5.

Screening Criterion:	Expected accident frequency
Accident Type:	PDO accidents
CV_{lim}	1.5
Number of Sites Ranked:	76/100
Mean length of all ranked sites (mi):	2.15
Mean length of all flagged windows (mi):	0.15

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	14.29	43.29	0.46	1
29	5.37	0.1	9.75	25.16	0.51	2
18	0.8	0.1	8.95	29.25	0.60	3
12	1.18	0.1	8.63	81.43	1.05	4
6	0.25	0.1	7.52	67.51	1.09	5
5	0.64	0.1	7.20	47.83	0.96	6
23	0.13	0.1	6.56	40.18	0.97	7
17	1.32	0.1	6.53	24.54	0.76	8
2	0.62	0.1	6.36	35.76	0.94	9
7	0.41	0.1	6.18	22.32	0.76	10
43	0.25	0.1	6.17	30.55	0.90	11
50	0.57	0.1	5.98	20.47	0.76	12
45	0.18	0.1	5.84	19.51	0.76	13
25	3.34	0.1	5.76	24.91	0.87	14
38	1.88	0.1	5.32	15.82	0.75	15
26	2.4	0.1	4.29	37.10	1.42	16
16	0.84	0.1	3.81	32.29	1.49	17
49	0.3	0.1	3.61	28.81	1.49	18
32	1.03	0.1	3.58	24.88	1.39	19
54	1.2	0.1	3.50	23.25	1.38	20

Table B.11: Results of peak-searching algorithm for expected frequency of PDO accidents, using CV_{lim}=1.0.

Screening Criterion:	Expected accident frequency
Accident Type:	PDO accidents
CV_{lim}	1.0
Number of Sites Ranked:	59/100
Mean length of all ranked sites (mi):	2.46
Mean length of all flagged windows (mi):	0.19

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	14.29	43.29	0.46	1
20	1.76	0.1	10.63	30.12	0.52	2
29	5.37	0.1	9.75	25.16	0.51	3
18	0.8	0.1	8.95	29.25	0.60	4
12	1.18	0.1	7.57	32.95	0.76	5
39	3.93	0.1	7.38	24.33	0.67	6
5	0.64	0.1	7.20	47.83	0.96	7
6	0.25	0.1	6.79	26.86	0.76	8
26	2.4	0.1	6.76	43.18	0.97	9
49	0.3	0.1	6.67	38.19	0.93	10
41	1.56	0.1	6.61	32.02	0.86	11
10	2.87	0.1	6.58	25.04	0.76	12
23	0.13	0.1	6.56	40.18	0.97	13
16	0.84	0.1	6.56	24.72	0.76	14
17	1.32	0.1	6.53	24.54	0.76	15
9	3.27	0.1	6.48	24.51	0.76	16
2	0.62	0.1	6.36	35.76	0.94	17
14	0.82	0.11	6.22	33.35	0.93	18
4	2.8	0.1	6.22	22.62	0.76	19
7	0.41	0.1	6.18	22.32	0.76	20

Table B.12: Results of peak-searching algorithm for expected frequency of PDO accidents, using CV_{lim}=0.5.

Screening Criterion:	Expected accident frequency
Accident Type:	PDO accidents
CV_{lim}	0.5
Number of Sites Ranked:	18/100
Mean length of all ranked sites (mi):	4.01
Mean length of all flagged windows (mi):	1.52

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	14.29	43.29	0.46	1
7	0.41	0.16	13.87	41.16	0.46	2
40	6.87	0.1	12.91	40.45	0.49	3
39	3.93	0.21	8.64	15.72	0.46	4
17	1.32	0.2	8.14	14.87	0.47	5
29	5.37	0.22	7.45	13.78	0.50	6
12	1.18	0.72	4.54	4.72	0.48	7
25	3.34	0.51	3.20	2.27	0.47	8
20	1.76	0.5	2.73	1.86	0.50	9
37	2.07	0.5	2.62	1.65	0.49	10
32	1.03	0.63	2.43	1.35	0.48	11
41	1.56	1.11	2.41	1.44	0.50	12
4	2.8	2.16	2.20	1.21	0.50	13
9	3.27	2.5	2.00	0.96	0.49	14
56	6.2	1.94	1.47	0.51	0.49	15
97	7.52	4.15	0.54	0.06	0.47	16
86	9.9	4.37	0.36	0.03	0.50	17
67	10.9	7.27	0.20	0.01	0.50	18

Table B.13: Results of peak-searching algorithm for excess frequency of PDO accidents, using CV_{lim}=4.0.

Screening Criterion:	Excess accident frequency
Accident Type:	PDO accidents
CV_{lim}	4.0
Number of Sites Ranked:	73/100
Mean length of all ranked sites (mi):	2.15
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	12.17	47.11	0.56	1
29	5.37	0.1	8.68	26.14	0.59	2
18	0.8	0.1	6.94	32.69	0.82	3
12	1.18	0.1	5.57	89.40	1.70	4
23	0.13	0.1	5.15	41.88	1.26	5
14	0.82	0.1	4.96	61.59	1.58	6
25	3.34	0.1	4.64	25.96	1.10	7
43	0.25	0.1	4.62	32.62	1.24	8
49	0.3	0.1	4.35	42.76	1.50	9
38	1.88	0.1	4.13	17.04	1.00	10
16	0.84	0.1	4.07	30.02	1.35	11
17	1.32	0.1	3.94	30.25	1.40	12
50	0.57	0.1	3.79	24.53	1.31	13
6	0.25	0.1	3.77	79.48	2.36	14
2	0.62	0.1	3.77	41.49	1.71	15
45	0.18	0.1	3.72	23.32	1.30	16
46	0.2	0.1	3.67	26.52	1.40	17
10	2.87	0.1	3.53	32.94	1.63	18
5	0.64	0.1	3.41	60.05	2.27	19
22	0.89	0.1	3.32	54.17	2.22	20

Table B.14: Results of peak-searching algorithm for excess frequency of PDO accidents, using CV_{lim}=2.5.

Screening Criterion:	Excess accident frequency
Accident Type:	PDO accidents
CV_{lim}	2.5
Number of Sites Ranked:	69/100
Mean length of all ranked sites (mi):	2.22
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	12.17	47.11	0.56	1
20	1.76	0.1	9.16	31.96	0.62	2
29	5.37	0.1	8.68	26.14	0.59	3
18	0.8	0.1	6.94	32.69	0.82	4
12	1.18	0.1	5.57	89.40	1.70	5
23	0.13	0.1	5.15	41.88	1.26	6
14	0.82	0.1	4.96	61.59	1.58	7
25	3.34	0.1	4.64	25.96	1.10	8
43	0.25	0.1	4.62	32.62	1.24	9
49	0.3	0.1	4.35	42.76	1.50	10
38	1.88	0.1	4.13	17.04	1.00	11
16	0.84	0.1	4.07	30.02	1.35	12
17	1.32	0.1	3.94	30.25	1.40	13
50	0.57	0.1	3.79	24.53	1.31	14
6	0.25	0.1	3.77	79.48	2.36	15
2	0.62	0.1	3.77	41.49	1.71	16
45	0.18	0.1	3.72	23.32	1.30	17
46	0.2	0.1	3.67	26.52	1.40	18
10	2.87	0.1	3.53	32.94	1.63	19
5	0.64	0.1	3.41	60.05	2.27	20

Table B.15: Results of peak-searching algorithm for excess frequency of PDO accidents, using CV_{lim}=1.0.

Screening Criterion:	Excess accident frequency
Accident Type:	PDO accidents
CV_{lim}	1.0
Number of Sites Ranked:	26/100
Mean length of all ranked sites (mi):	2.78
Mean length of all flagged windows (mi):	0.11

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
7	0.41	0.1	12.44	95.43	0.79	1
47	2.7	0.1	12.17	47.11	0.56	2
20	1.76	0.1	9.16	31.96	0.62	3
29	5.37	0.1	8.68	26.14	0.59	4
12	1.18	0.1	8.24	54.90	0.90	5
41	1.56	0.11	7.49	34.53	0.78	6
17	1.32	0.1	7.17	40.66	0.89	7
23	0.13	0.11	7.00	40.88	0.91	8
18	0.8	0.1	6.94	32.69	0.82	9
46	0.2	0.1	6.69	35.62	0.89	10
45	0.18	0.1	6.60	31.61	0.85	11
39	3.93	0.1	6.49	25.00	0.77	12
53	0.92	0.15	4.28	13.42	0.86	13
38	1.88	0.1	4.13	17.04	1.00	14
37	2.07	0.1	4.10	15.42	0.96	15
35	2.42	0.1	4.06	14.40	0.93	16
40	6.87	0.1	4.06	14.39	0.93	17
25	3.34	0.1	4.05	15.88	0.99	18
99	0.73	0.1	4.02	15.87	0.99	19
33	0.91	0.1	3.91	14.41	0.97	20

Table B.16: Results of peak-searching algorithm for excess frequency of PDO accidents, using CV_{lim}=0.5.

Screening Criterion:	Excess accident frequency
Accident Type:	PDO accidents
CV_{lim}	0.5
Number of Sites Ranked:	3/100
Mean length of all ranked sites (mi):	4.50
Mean length of all flagged windows (mi):	0.15

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
47	2.7	0.1	15.02	55.19	0.49	1
40	6.87	0.1	14.43	47.07	0.48	2
39	3.93	0.26	6.97	11.37	0.48	3

Table B.17: Results of peak-searching algorithm for expected frequency of EPDO accidents, using CV_{lim}=2.5.

Screening Criterion:	Expected accident frequency
Accident Type:	EPDO accidents
CV_{lim}	2.5
Number of Sites Ranked:	100/100
Mean length of all ranked sites (mi):	1.75
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
22	0.89	0.1	163.81	5407.63	0.45	1
12	1.18	0.1	134.60	6848.96	0.61	2
6	0.25	0.1	128.93	6008.87	0.60	3
5	0.64	0.1	126.45	6083.02	0.62	4
26	2.4	0.1	115.66	3478.85	0.51	5
7	0.41	0.1	114.97	5032.48	0.62	6
15	0.13	0.1	112.28	5027.25	0.63	7
48	0.24	0.1	104.13	4081.87	0.61	8
23	0.13	0.1	92.70	3030.94	0.59	9
52	1.06	0.1	90.65	3246.35	0.63	10
16	0.84	0.1	64.81	2771.33	0.81	11
11	0.72	0.1	64.50	2770.59	0.82	12
9	3.27	0.1	64.00	3000.51	0.86	13
8	0.13	0.1	63.74	2976.10	0.86	14
49	0.3	0.1	58.06	2454.19	0.85	15
2	0.62	0.1	57.05	2376.22	0.85	16
54	1.2	0.1	52.54	1793.71	0.81	17
19	0.72	0.1	52.43	1985.51	0.85	18
25	3.34	0.1	48.47	1368.43	0.76	19
3	1.16	0.1	47.15	1627.46	0.86	20

Table B.18: Results of peak-searching algorithm for expected frequency of EPDO accidents, using CV_{lim}=1.0.

Screening Criterion:	Expected accident frequency
Accident Type:	EPDO accidents
CV_{lim}	1.0
Number of Sites Ranked:	86/100
Mean length of all ranked sites (mi):	1.99
Mean length of all flagged windows (mi):	0.16

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
22	0.89	0.1	163.81	5407.63	0.45	1
12	1.18	0.1	134.60	6848.96	0.61	2
6	0.25	0.1	128.93	6008.87	0.60	3
5	0.64	0.1	126.45	6083.02	0.62	4
26	2.4	0.1	115.66	3478.85	0.51	5
7	0.41	0.1	114.97	5032.48	0.62	6
15	0.13	0.1	112.28	5027.25	0.63	7
48	0.24	0.1	104.13	4081.87	0.61	8
23	0.13	0.1	92.70	3030.94	0.59	9
52	1.06	0.1	90.65	3246.35	0.63	10
24	0.34	0.1	87.17	3007.44	0.63	11
16	0.84	0.1	64.81	2771.33	0.81	12
11	0.72	0.1	64.50	2770.59	0.82	13
9	3.27	0.1	64.00	3000.51	0.86	14
18	0.8	0.1	64.00	2235.19	0.74	15
8	0.13	0.1	63.74	2976.10	0.86	16
10	2.87	0.1	63.44	2940.66	0.85	17
14	0.82	0.1	62.50	2575.62	0.81	18
4	2.8	0.1	61.93	2812.20	0.86	19
17	1.32	0.1	61.67	2772.19	0.85	20

Table B.19: Results of peak-searching algorithm for expected frequency of EPDO accidents, using CV_{lim}=0.5.

Screening Criterion:	Expected accident frequency
Accident Type:	EPDO accidents
CV_{lim}	0.5
Number of Sites Ranked:	58/100
Mean length of all ranked sites (mi):	2.65
Mean length of all flagged windows (mi):	0.46

Site No.	Site Length (mi)	Length of Flagged Window (mi)	$Peak(X_Y)$ (acc/mi/yr)	$Var[Peak(X_Y)]$	$CV[Peak(X_Y)]$	Rank
7	0.41	0.1	231.60	9625	0.42	1
4	2.8	0.1	218.61	9794	0.45	2
5	0.64	0.11	214.20	9607	0.46	3
9	3.27	0.11	203.49	8671	0.46	4
20	1.76	0.1	173.32	5708	0.44	5
22	0.89	0.1	163.81	5408	0.45	6
47	2.7	0.1	151.23	5610	0.50	7
12	1.18	0.14	144.67	5212	0.50	8
6	0.25	0.15	126.69	4001	0.50	9
54	1.2	0.11	125.86	3954	0.50	10
10	2.87	0.19	123.12	2964	0.44	11
41	1.56	0.11	122.67	3714	0.50	12
26	2.4	0.1	118.12	3485	0.50	13
37	2.07	0.1	117.14	3410	0.50	14
40	6.87	0.1	114.20	3098	0.49	15
39	3.93	0.1	110.78	2703	0.47	16
29	5.37	0.11	105.30	2751	0.50	17
56	6.2	0.13	103.76	2646	0.50	18
33	0.91	0.13	89.31	1951	0.49	19
25	3.34	0.14	88.06	1889	0.49	20

Table B.20: Results of peak-searching algorithm for expected frequency of EPDO accidents, using CV_{lim}=0.2.

Screening Criterion:	Expected accident frequency
Accident Type:	EPDO accidents
CV_{lim}	0.2
Number of Sites Ranked:	11/100
Mean length of all ranked sites (mi):	5.71
Mean length of all flagged windows (mi):	3.57

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
9	3.27	1.51	79.16	241.44	0.20	1
4	2.8	1.49	74.45	221.58	0.20	2
47	2.7	1.4	68.18	178.01	0.20	3
40	6.87	1.88	36.88	54.36	0.20	4
25	3.34	2.32	32.26	41.53	0.20	5
29	5.37	2.47	29.63	34.12	0.20	6
56	6.2	3.99	21.79	18.33	0.20	7
39	3.93	3.24	20.13	15.82	0.20	8
97	7.52	7.28	7.93	2.51	0.20	9
86	9.9	6.31	5.86	1.37	0.20	10
67	10.9	7.37	4.19	0.70	0.20	11

Table B.21: Results of peak-searching algorithm for excess frequency of EPDO accidents, using CV_{lim}=4.0.

Screening Criterion:	Excess accident frequency
Accident Type:	EPDO accidents
CV_{lim}	4.0
Number of Sites Ranked:	66/100
Mean length of all ranked sites (mi):	2.28
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
22	0.89	0.1	140.41	5500	0.53	1
26	2.4	0.1	97.19	3536	0.61	2
12	1.18	0.1	83.83	7283	1.02	3
15	0.13	0.1	70.38	5323	1.04	4
48	0.24	0.1	70.06	4277	0.93	5
23	0.13	0.1	69.29	3123	0.81	6
52	1.06	0.1	67.05	3340	0.86	7
6	0.25	0.1	66.72	6660	1.22	8
10	2.87	0.1	65.33	5801	1.17	9
5	0.64	0.1	63.61	6748	1.29	10
24	0.34	0.1	63.07	3105	0.88	11
1	1.73	0.1	58.39	5140	1.23	12
9	3.27	0.1	56.47	6095	1.38	13
11	0.72	0.1	56.02	5574	1.33	14
7	0.41	0.1	55.40	5630	1.35	15
4	2.8	0.1	50.28	5801	1.51	16
96	4.5	0.1	48.37	1319	0.75	17
3	1.16	0.1	45.46	3290	1.26	18
18	0.8	0.1	30.64	2422	1.61	19
25	3.34	0.1	30.01	1426	1.26	20

Table B.22: Results of peak-searching algorithm for excess frequency of EPDO accidents, using CV_{lim}=2.5.

Screening Criterion:	Excess accident frequency
Accident Type:	EPDO accidents
CV_{lim}	2.5
Number of Sites Ranked:	66/100
Mean length of all ranked sites (mi):	2.28
Mean length of all flagged windows (mi):	0.10

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
22	0.89	0.1	140.41	5500	0.53	1
26	2.4	0.1	97.19	3536	0.61	2
12	1.18	0.1	83.83	7283	1.02	3
15	0.13	0.1	70.38	5323	1.04	4
48	0.24	0.1	70.06	4277	0.93	5
17	1.32	0.1	69.68	5374	1.05	6
23	0.13	0.1	69.29	3123	0.81	7
13	0.44	0.1	67.75	4677	1.01	8
52	1.06	0.1	67.05	3340	0.86	9
6	0.25	0.1	66.72	6660	1.22	10
50	0.57	0.1	65.59	4349	1.01	11
10	2.87	0.1	65.33	5801	1.17	12
5	0.64	0.1	63.61	6748	1.29	13
24	0.34	0.1	63.07	3105	0.88	14
1	1.73	0.1	58.39	5140	1.23	15
2	0.62	0.11	57.96	3919	1.08	16
9	3.27	0.1	56.47	6095	1.38	17
11	0.72	0.1	56.02	5574	1.33	18
7	0.41	0.1	55.40	5630	1.35	19
4	2.8	0.1	50.28	5801	1.51	20

Table B.23: Results of peak-searching algorithm for excess frequency of EPDO accidents, using CV_{lim}=1.0.

Screening Criterion:	Excess accident frequency
Accident Type:	EPDO accidents
CV_{lim}	1.0
Number of Sites Ranked:	41/100
Mean length of all ranked sites (mi):	2.59
Mean length of all flagged windows (mi):	0.11

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
7	0.41	0.1	172.03	10222	0.59	1
22	0.89	0.1	140.41	5500	0.53	2
5	0.64	0.1	115.91	9484	0.84	3
9	3.27	0.1	109.38	8575	0.85	4
10	2.87	0.11	103.47	6881	0.80	5
4	2.8	0.1	101.48	8125	0.89	6
37	2.07	0.1	100.27	3458	0.59	7
26	2.4	0.1	97.19	3536	0.61	8
36	0.76	0.1	93.88	3285	0.61	9
6	0.25	0.12	93.25	6676	0.88	10
33	0.91	0.1	92.19	3270	0.62	11
12	1.18	0.1	87.57	7297	0.98	12
17	1.32	0.13	85.36	4676	0.80	13
14	0.82	0.1	75.65	4963	0.93	14
19	0.72	0.1	70.32	3736	0.87	15
48	0.24	0.1	70.06	4277	0.93	16
23	0.13	0.1	69.29	3123	0.81	17
41	1.56	0.1	69.27	3124	0.81	18
47	2.7	0.1	68.79	4069	0.93	19
50	0.57	0.1	68.54	4358	0.96	20

Table B.24: Results of peak-searching algorithm for excess frequency of EPDO accidents, using CV_{lim}=0.5.

Screening Criterion:	Excess accident frequency
Accident Type:	EPDO accidents
CV_{lim}	0.5
Number of Sites Ranked:	13/100
Mean length of all ranked sites (mi):	2.62
Mean length of all flagged windows (mi):	0.23

Site No.	Site Length (mi)	Length of Flagged Window (mi)	Peak(X_Y) (acc/mi/yr)	Var[Peak(X_Y)]	CV[Peak(X_Y)]	Rank
7	0.41	0.1	273.32	14768	0.44	1
4	2.8	0.1	258.16	15107	0.48	2
20	1.76	0.1	188.22	7159	0.45	3
22	0.89	0.12	144.21	4756	0.48	4
39	3.93	0.1	127.07	3574	0.47	5
47	2.7	0.16	123.79	3776	0.50	6
40	6.87	0.12	108.90	2878	0.49	7
54	1.2	0.29	70.92	1194	0.49	8
26	2.4	0.29	63.59	871	0.46	9
33	0.91	0.25	62.05	910	0.49	10
41	1.56	0.36	58.33	741	0.47	11
25	3.34	0.53	37.91	335	0.48	12
29	5.37	0.51	37.05	329	0.49	13

Appendix C

Goodness of Fit Statistics For Beta Models

Table C.1: Goodness of fit statistics for beta models calibrated using the method of maximum likelihood from 5 years of California signalized intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	-	-	Could not compute χ^2 value.
Head-on	425	<0.0001	Good fit.
Sideswipe	64.8	>0.995	Questionable fit.
Rear-end	168	0.99	Questionable fit.
Broadside	-	-	Could not compute χ^2 value.
Hit Object	-	-	Could not compute χ^2 value.
Overturning	-	-	Could not compute χ^2 value.
Pedestrian	5.68	>0.995	Questionable fit.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.2: Goodness of fit statistics for beta models calibrated using the method of maximum likelihood from 3 years of California signalized intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	4380	<0.0001	Good fit.
Head-on	-	-	Could not compute χ^2 value.
Sideswipe	-	-	Could not compute χ^2 value.
Rear-end	1270	<0.0001	Good fit.
Broadside	-	-	Could not compute χ^2 value.
Hit Object	9.91	>0.995	Questionable fit.
Overturning	149	<0.0001	Good fit.
Pedestrian	37.6	>0.995	Questionable fit.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.3: Goodness of fit statistics for beta models calibrated using the first method of moments (MM1) from 5 years of California signalized intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	5.03	>0.995	Questionable fit.
Head-on	2.09	>0.995	Questionable fit.
Sideswipe	3.12	>0.995	Questionable fit.
Rear-end	15.2	>0.995	Questionable fit.
Broadside	50.1	>0.995	Questionable fit.
Hit Object	0.760	>0.995	Questionable fit.
Overturning	0.578	>0.995	Questionable fit.
Pedestrian	2.77	>0.995	Questionable fit.
Other/Unknown	0.516	>0.995	Questionable fit.

Table C.4: Goodness of fit statistics for beta models calibrated using the first method of moments (MM1) from 3 years of California signalized intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	52.1	>0.995	Questionable fit.
Head-on	179	<0.0001	Good fit.
Sideswipe	6120	<0.0001	Good fit.
Rear-end	28.2	>0.995	Questionable fit.
Broadside	370	<0.0001	Good fit.
Hit Object	20.9	>0.995	Questionable fit.
Overturning	2.75	>0.995	Questionable fit.
Pedestrian	34.0	<0.0001	Good fit.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.5: Goodness of fit statistics for beta models calibrated using the second method of moments (MM2) from 5 years of California signalized intersection data.

Accident Type	$\chi^2_{(0.10)}$	$Pr(<\chi^2_{(0.10)})$	Comments
FI	200	<0.0001	Good fit.
Head-on	-	-	Could not compute χ^2 value.
Sideswipe	1590	<0.0001	Good fit.
Rear-end	87.3	0.50	Questionable fit
Broadside	-	-	Could not compute χ^2 value.
Hit Object	451	>0.995	Questionable fit.
Overturning	33.6	>0.995	Questionable fit.
Pedestrian	25.6	>0.995	Questionable fit.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.6: Goodness of fit statistics for beta models calibrated using the second method of moments (MM2) from 3 years of California signalized intersection data.

Accident Type	$\chi^2_{(0.10)}$	$Pr(<\chi^2_{(0.10)})$	Comments
FI	288	<0.0001	Good fit.
Head-on	-	-	Could not compute χ^2 value.
Sideswipe	-	-	Could not compute χ^2 value.
Rear-end	284	<0.0001	Good fit.
Broadside	-	-	Could not compute χ^2 value.
Hit Object	14.0	>0.995	Questionable fit.
Overturning	7730	<0.0001	Good fit.
Pedestrian	708	<0.0001	Good fit.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.7: Goodness of fit statistics for beta models calibrated using the method of maximum likelihood from 5 years of California TWSC intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	-	-	Could not compute χ^2 value.
Head-on	-	-	Could not compute χ^2 value.
Sideswipe	-	-	Could not compute χ^2 value.
Rear-end	-	-	Could not compute χ^2 value.
Broadside	1350	<0.0001	Good fit.
Hit Object	-	-	Could not compute χ^2 value.
Overturning	-	-	Could not compute χ^2 value.
Pedestrian	-	-	Could not compute χ^2 value.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.8: Goodness of fit statistics for beta models calibrated using the first method of moments (MM1) from 3 years of California TWSC intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	661	<0.0001	Good fit.
Head-on	110	0.25	Questionable fit.
Sideswipe	329	<0.0001	Good fit.
Rear-end	463	<0.0001	Good fit.
Broadside	611	<0.0001	Good fit.
Hit Object	465	<0.0001	Good fit.
Overturning	122	0.10	Good fit.
Pedestrian	38.0	>0.995	Questionable fit.
Other/Unknown	212	<0.0001	Good fit.

Table C.9: Goodness of fit statistics for beta models calibrated using the second method of moments (MM2) from 3 years of California TWSC intersection data.

Accident Type	$\chi^2_{(0.10)}$	$\Pr(<\chi^2_{(0.10)})$	Comments
FI	-	-	Could not compute χ^2 value.
Head-on	-	-	Could not compute χ^2 value.
Sideswipe	-	-	Could not compute χ^2 value.
Rear-end	-	-	Could not compute χ^2 value.
Broadside	2530	<0.0001	Good fit.
Hit Object	-	-	Could not compute χ^2 value.
Overturning	-	-	Could not compute χ^2 value.
Pedestrian	-	-	Could not compute χ^2 value.
Other/Unknown	-	-	Could not compute χ^2 value.

Table C.10: Goodness of fit statistics for beta models calibrated using the method of maximum likelihood from 3 years of 2-mile segments of Washington 2-lane rural highway.

Accident Type	$\chi^2_{(0.10)}$	Pr($<\chi^2_{(0.10)}$)	Comments
Head-on	-	-	Could not compute χ^2 value.
Angle	1.17	>0.995	Questionable fit.
SSSD	-	-	Could not compute χ^2 value.
SSOD	-	-	Could not compute χ^2 value.
Animal	-	-	Could not compute χ^2 value.
Cyclist	1.56	>0.995	Questionable fit.
Pedestrian	0.781	>0.995	Questionable fit.
Parked Vehicle	-	-	Could not compute χ^2 value.
Overturning	-	-	Could not compute χ^2 value.
Fixed Object	-	-	Could not compute χ^2 value.
OMV	-	-	Could not compute χ^2 value.
OSV	-	-	Could not compute χ^2 value.
Unknown	-	-	Could not compute χ^2 value.

Table C.11: Goodness of fit statistics for beta models calibrated using the first method of moments (MM1) from 3 years of 2-mile segments of Washington 2-lane rural highway.

Accident Type	$\chi^2_{(0.10)}$	Pr($<\chi^2_{(0.10)}$)	Comments
Head-on	10.9	>0.995	Questionable fit.
Angle	0.123	>0.995	Questionable fit.
SSSD	29.5	>0.995	Questionable fit.
SSOD	50.1	>0.995	Questionable fit.
Animal	120	0.92	Good fit.
Cyclist	2.20	>0.995	Questionable fit.
Pedestrian	1.83	>0.995	Questionable fit.
Parked Vehicle	7.33	>0.995	Questionable fit.
Overturning	180	<0.0001	Good fit.
Fixed Object	219	<0.0001	Good fit.
OMV	118	0.90	Good fit.
OSV	39.2	>0.995	Questionable fit.
Unknown	12.5	>0.995	Questionable fit.

Table C.12: Goodness of fit statistics for beta models calibrated using the second method of moments (MM2) from 3 years of 2-mile segments of Washington 2-lane rural highway.

Accident Type	$\chi^2_{(0.10)}$	Pr(<$\chi^2_{(0.10)}$)	Comments
Head-on	-	-	Could not compute χ^2 value.
Angle	0.00001	>0.995	Questionable fit.
SSSD	-	-	Could not compute χ^2 value.
SSOD	-	-	Could not compute χ^2 value.
Animal	-	-	Could not compute χ^2 value.
Cyclist	-	-	Could not compute χ^2 value.
Pedestrian	-	-	Could not compute χ^2 value.
Parked Vehicle	-	-	Could not compute χ^2 value.
Overturning	-	-	Could not compute χ^2 value.
Fixed Object	-	-	Could not compute χ^2 value.
OMV	-	-	Could not compute χ^2 value.
OSV	-	-	Could not compute χ^2 value.
Unknown	-	-	Could not compute χ^2 value.

Appendix D

MATLAB Code for Peak-Searching Algorithm

MATLAB Code for Peak-Searching Algorithm

```
%
%       Peak-Searching Algorithm for screening for
%       EB-adjusted expected accident frequency
%
%       by Brent Gotts,  July, 2004
%



no_sites=max(site); %number of sites;
temp=size(site);
no_subs=temp(:,1);  %number of subsegments (total);


CVlim=1.5;
inc=1;
no_subs=0;
run_count=0;

'Beginning algorithm...'

for d=1:1:no_sites;
    d;
    count1=round(1+run_count);
    no_subs=round(site_lng(count1)*100);  %number of subsegs in site i,
assuming all subsegs are 0.01mi;
    size=9;
    flag=0;
    max_win=no_subs-9;  %maximum number of different window sizes;

    for j=1:1:max_win;
        j;
        size=size+1;
        no_wins=round(no_subs-size+1);  %number of windows using a
given window size;
        count2=count1;

        for k=1:1:no_wins;
            k;
            count2;
            winbp=sub_bp(count2);
            winep=sub_ep(count2+size-1);
            winlng=winep-winbp;
            winX=sum(sub_tot(count2:count2+size-1))/winlng;
            winvarX=sum(sub_var_tot(count2:count2+size-1))/winlng^2;
            CV=sqrt(winvarX)/winX;

            if CV<=CVlim;
                flag=1;
                site_out(d)=site(count2);
                sitebp_out(d)=site_bp(count2);
                siteep_out(d)=site_ep(count2);
                site_lng_out(d)=site_lng(count2);
                winbp_out(d)=winbp;
                winep_out(d)=winep;
                winlng_out(d)=winlng;
                winX_out(d)=winX;
                winvarX_out(d)=winvarX;
                CV_out(d)=CV;
                flag_out(d)=1;
                break;
            else
                count2=count2+inc;
            end;

        end

        if flag>0;
            break;  %Quits this loop if CVlim has already been reached;
        end;

    end

    if flag<0.5;
        site_out(d)=site(count2);
        sitebp_out(d)=site_bp(count2);
        siteep_out(d)=site_ep(count2);
        site_lng_out(d)=site_lng(count2);
        winbp_out(d)=0;
        winep_out(d)=0;
        winlng_out(d)=0;
        winX_out(d)=winX;
        winvarX_out(d)=winvarX;
        CV_out(d)=CV;
        flag_out(d)=0;
    end;

    run_count=run_count+no_subs;

end

'Algorithm complete.'

for m=1:1:no_sites;
    output(m,:)=[site_out(m) sitebp_out(m) siteep_out(m)
site_lng_out(m) winbp_out(m) winep_out(m) winlng_out(m) winX_out(m)
winvarX_out(m) CV_out(m) flag_out(m)];
end;

count3=1;
for n=1:1:no_sites;
    if output(n,11)>0.5;
        passed(count3,:)=output(n,:);
        count3=count3+1;
    end;
end

if count3<2;
    'No sites were ranked.'
else
```

```
    sortrows(passed,8);
    % disp(passed);

    header='PS Output';
    colnames={'Site No.', 'Site bp', 'Site ep', 'Site Lng', 'Window
bp', 'Window ep', 'Window Lng', 'Exp. Acc Freq', 'VAR (EAF)',
'CV','Flag'};
    xlswrite(passed,header,colnames);

    'Finished.'
end;
```

3 BL-106-124