U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road. Ann Arbor. MI 48106-1346 USA
313/761-4700 800/521-0600

Identification and risk assessment in traffic event system

Lin, Chunming, Ph.D.

The Pennsylvania State University, 1992

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

The Pennsylvania State University

The Graduate School

IDENTIFICATION AND RISK ASSESSMENT IN TRAFFIC EVENT SYSTEM

A Thesis in

Mechanical Engineering

by

Chunming Lin

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 1992

We approve the thesis of Chunming Lin.

Date of Signature

B.T. Kulakowski 12/13/91

Bohdan T. Kulakowski
Professor of Mechanical Engineering
Thesis Advisor, Chair of Committee

Charles E. Antle Dec 9, 1991

Charles E. Antle
Professor of Statistics

Dec 5, 1991

John J. Henry
Professor of Mechanical Engineering

Dennis R. Hiltunen 12/6/91

Dennis R. Hiltunen
Assistant Professor of Civil Engineering

James C Wambold 12/14/91

James C. Wambold
Professor of Mechanical Engineering

12/6/91

Kon-Well Wang
Assistant Professor of Mechanical Engineering

Harold R Jacobs 12/16/91

Harold R. Jacobs
Professor of Mechanical Engineering
Head of the Department of
Mechanical Engineering

ABSTRACT

The enormous number of accidental deaths associated with motor-vehicle accidents each year remains a main issue of highway safety. For the assessment of the accident risk associated with particular highway locations, probabilistic type empirical Bayes methods have been considered a viable approach. However, considerations with regard to the adequate sample size, the random effect of vehicle exposure, the utilization of both accident histories and measurements of roadway and traffic characteristics to identify significant causal factors have not been discussed in detail.

In this thesis, four new methods--a modified Arnold and Antle procedure, two new median estimators for a gamma distribution, a knowledge-based model, and a hierarchical accident index method--were developed to identify significant causal factors and assess traffic accident probability in the highway system. An evaluation of these methods was performed on real data from over 300 sites in Pennsylvania. A comparison of these methods and classical regression methods is also presented.

Based on an absolute error loss function, it was concluded that the modified empirical Bayes procedures, especially the two new median estimators, are superior to the other methods in estimating accident risk when accident statistics and measurement data are available. The knowledge-based model approach proved valuable for predicting accident risk for roadway sections as well as identifying significant causal factors. The hierarchical accident index method, using both accident records and subjective judgement, performs almost as well as the modified empirical Bayes procedures in evaluating accident risk of wet pavement accidents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and thanks to my thesis advisor, Dr. Bohdan T. Kulakowski, for his patience, support, and thoughtful review of the drafts of my thesis.

Thanks are extended to my thesis committee, Drs. John J. Henry, James C. Wambold, Charles E. Antle, Kon-Well Wang, and Dennis R. Hiltunen, for their support and understanding during the process of completing this thesis. Special appreciation goes to Dr. Antle for his guidance and suggestions on many statistical aspects.

I would like to acknowledge Pennsylvania Department of Transportation for their financial support and providing a real data set for model evaluation.

Sincere gratitude is offered to my wife, Kuei-Hwa Kuo, for her understanding, support, and thoughtfulness during my study.

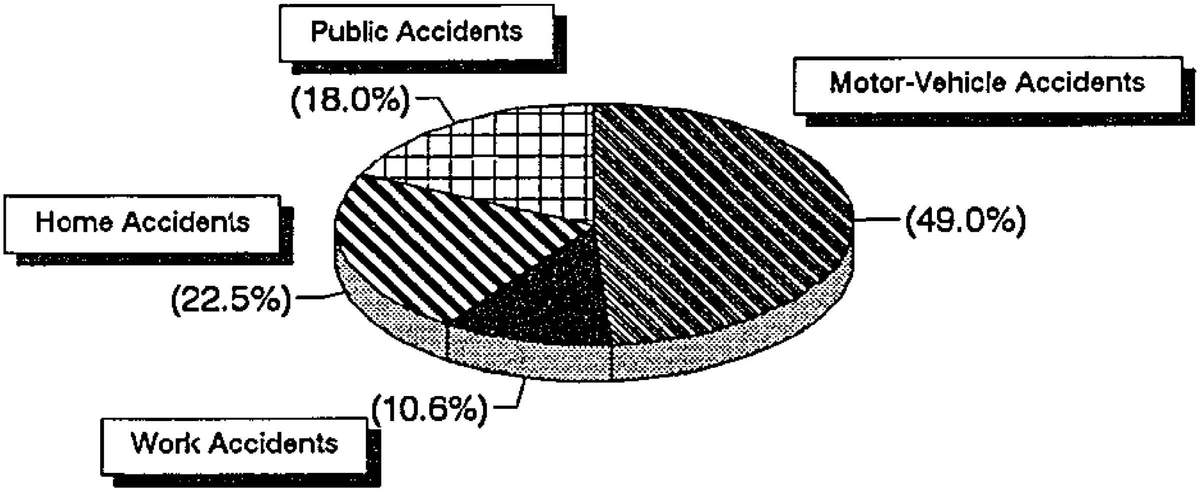
Chapter 1

INTRODUCTION

1.1. Background

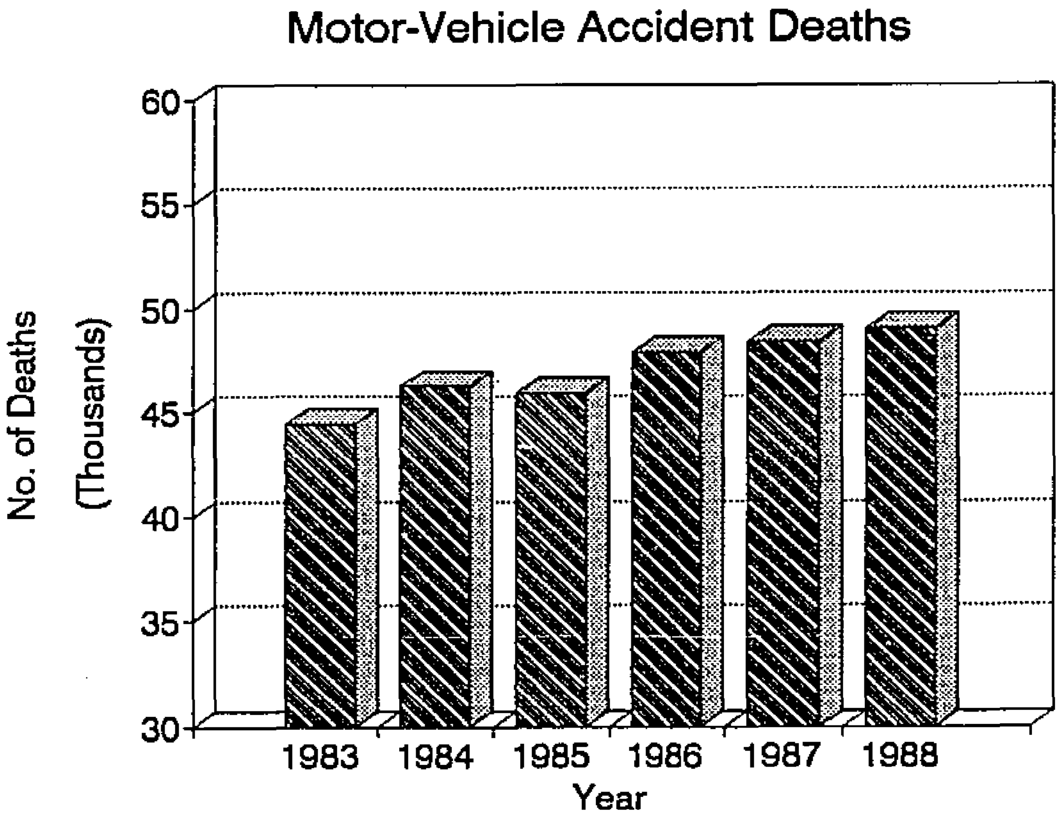
The identification and risk assessment for an event system has been a topic of research for several decades. Interest in this problem originates from the designing of large-scale engineering systems such as nuclear power plants, chemical processing plants, traffic systems, and the like. The identification and risk assessment technique is applied to analyze the causation or assess the risk of a possible undesired event in the system. From a safety point of view, an undesired event could be a leak of radiation, a water pump failure, a fire, a toxic gas leak, or a traffic accident. In this thesis, the research was focused on the traffic event system.

Traffic accidents are the most common and uncommon events in our daily lives. According to the 1989 National Safety Council report, the number of motor vehicle accident deaths represents 49% of all accidental deaths (see figure 1.1). If one looks at the number of motor vehicle accident deaths and the death rates in the consecutive years from 1983 to 1988, it is clear that there exists a serious issue in the prevention of traffic accidents (see figures 1.2 and 1.3). In striving to improve the safety of the traffic event system, numerous methodologies have been



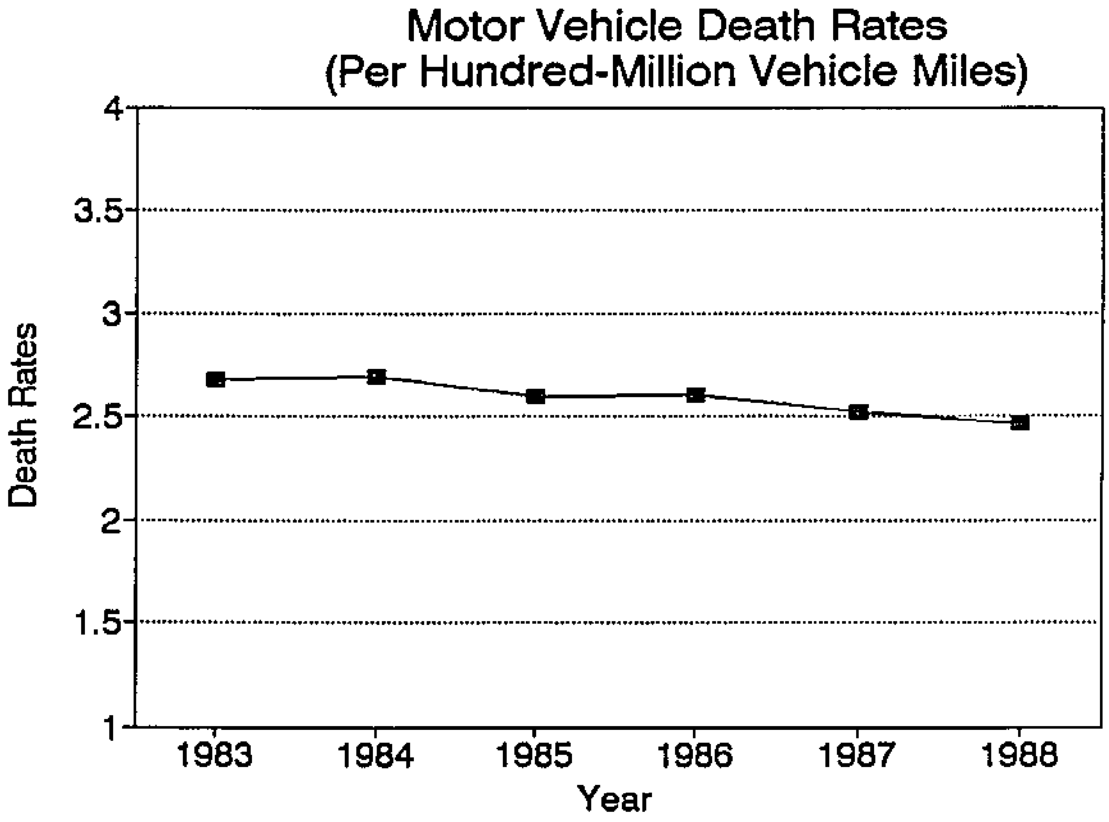
Source: National Safety Council , 1989.

Figure 1.1. Causes of accidental death as reported by the National Safety Council.



Source: National Safety Council, 1989.

Figure 1.2. Accidental deaths associated with motor vehicle accidents (1983-1988).



Source: National Safety Council, 1989.

Figure 1.3. Death rates in motor vehicle accidents (1983-1988).

developed. Most of them concentrate on identifying the problem (hazardous, accident-prone, abnormal, etc.) locations. However, the effectiveness and accuracy of these procedures are limited. Today, the identification of accident-prone locations becomes more crucial due to the increasing demand for system safety under severe resource and budget constraints.

The traffic event system is generally recognized as a human-vehicle-roadway system. The causation of a traffic accident is considered as an outcome of interaction among the human, vehicle, roadway, and traffic factors. The complexity and uncertainty in the system mandate the development of a workable model and thus the identification and risk assessment problem remains a topic worthy of study.

1.2. Review of Literature

Studies of the traffic event system identification can be traced back to the 1950s. Most of the work relied principally on the accident statistics in attempting to identify the so-called problem locations without analyzing the causation of traffic accidents. In highway agencies, the accident frequency method is the most commonly used method. It has the advantage of simplicity and is easy to implement. Using this method, a location is identified as a hazardous location if the accident frequency at the location in a specific time period is higher than a critical value.

A second approach, the accident rate method, involves the concept of risk, which is determined simply by dividing the number of accidents by the vehicle exposure (usually in millions of vehicle miles, MVM, or millions of vehicles, MV) at the location of interest. The assumptions behind the definition are that (1) there exists a linear relationship between the accident frequency and the vehicle exposure (a slope), and (2) the exposure is a value measured without an error. However, a comparison of the accident risk between sites may draw an incorrect conclusion because of neglect of the variation (random) effect of the exposure either between locations or at the location.

A third technique, the quality-control method, was developed by Norden et al. (1956). The method applies the statistical quality-control technique to calculate a critical accident rate for the same "category" of road. The critical accident rate serves as the upper and lower bounds of a control chart. Any road site is identified as a problem location if its accident rate falls outside the control interval. Basically this approach is promising if the sample size of the same "category" of road sites is large enough and the variation of individual vehicle exposure is small. Nevertheless, the problem of neglect of the random effect of exposure found in the accident rate method is also encountered here.

Another approach is called an accident severity method which identifies and/or ranks locations based on the number of severe accidents at each location. Accident severity is defined by the National Safety Council (1976) according to the following categories: (1) Fatal accident, (2) A-type injury (incapacitating) accident, (3) B-type injury (nonincapacitating) accident, (4) C-type injury (probable injury) accident, and (5) PDO (property damage only) accident. Weighting factors are assigned to different categories to obtain an index for identifying or ranking. This method requires a detailed description for each accident at each location. Additionally, it involves subjective assignment of weighting factors.

Other developed methods that concentrate on analyzing the accident causation can be grouped into two categories focusing on the aspects of the roadway and the drivers, respectively (Laughland et al. 1975). Methods in the roadway category include:

1. Skid testing
2. Hazardous indicator reporting
3. Correlation of geometrics with accidents
4. Accident risk factor
5. Formula methods
6. Field observation

For the second category, the methods include:

1. Conflict analysis
2. Speed distortion skew
3. Correlation of speed changes with accident rates
4. Accident rate versus minimum safe headway
5. Physiological response testing

The skid testing method assumes that the friction, as measured by skid resistance, is an index of potential risk of skidding accidents for a given location. A critical value of skid resistance is usually assumed to represent the minimum skid resistance that is considered necessary to provide sufficient traction for vehicles traveling on a particular roadway section. A location is identified to be slippery if its skid resistance is lower than the critical value. Determination of the critical skid resistance for the population of road sites, however, is clouded by the site- and time-specific characteristics of skid resistance (Giles and Sabey 1959; Rice 1977). Thus, the effectiveness of using skid resistance alone to assess accident potential may not be adequate. A remedy procedure that adjusts the measured skid resistance with respect to standard conditions was developed by the Pennsylvania Transportation Institute (Wambold et al. 1988). It is designed to take care of the random effects caused by environmental conditions. An application of this procedure is described in chapter 6.

The hazard indicator reporting method is designed to identify locations or conditions that help to cause or increase the severity of highway accidents. The accuracy of the hazard indicator reporting depends on the knowledge and judgement of highway personnel.

The method of correlating road geometrics to accidents has been used increasingly through a statistical technique known as regression analysis. This analysis method basically assumes that between the accident and highway geometrics exists a cause-and-effect relationship and that a location can be identified to be hazardous through the constructed relationship. The difficulty involved with this method is that various interactions between the components of geometrics need to be taken into account.

Fine (1971) proposed the approach of accident risk factor for identifying hazardous locations. Essentially, the method uses a scheme of rating assignments to deal with vague information in the traffic system and classify them into different levels. The assignment of the ratings is crucial to the identification procedure. Different ratings might result in different lists of hazardous locations.

The formula method is a deterministic approach to the modeling of the traffic accident system. It attempts to relate the suspected casual factors to the accidents. The assumption behind this method is that hazardousness can be computed by using measurable "independent" variables. Unfortunately, the identification of the independent variables is difficult and an improved scheme might be needed.

The field observation approach can give valuable suggestions if the observers are well trained to understand the causation of traffic accidents. Hazards are identified through the observer's judgement during a routine field trip or a specific trip to a location having high accident frequency. It should be noted that the location with identified hazards may not always have high accident frequency since drivers may also perceive the hazards and drive more cautiously.

Perkins and Harris (1968) developed the traffic conflict analysis technique to analyze the accident potential at intersections. An evaluation of this method was then done by Baker (1972). Using this approach, a location is identified as hazardous if it has a high number of traffic conflicts. The most promising use for this technique is to prescribe applicable improvements for the hazardous locations.

The speed distribution skew method was proposed by Caples and Vanstrum (1969). It assumes that the increase of accident frequencies is proportional to the increase of speed difference between vehicles. The existence of a wide speed difference between vehicles, displaying a skewed speed distribution, is identified as an accident potential.

A method that correlates speed changes with accident rates assumes that the number of vehicle speed changes indicates the accident potential of a highway section. Researchers at the North Carolina State University found that an absolute speed change of 4 mi/h per unit of time would be critical (Heimbach et al. 1968).

Rockwell and Treiterer (1968) proposed the accident rate versus the minimum safe headway method to identify the hazardous location. The basic assumption of this method is that the accident potential increases when relative velocity is high and headways are short. If most of the vehicles at a location operate at less than a minimum safe headway, the accident potential of the location will be high.

The physiological response testing method measures the driver's response at the driving task. Special equipment (which is usually not owned by the highway agencies) and techniques are required to perform this testing, and the operating cost is high. Hence, this approach may be more suitable for a research study than for routine testing.

An extensive study of the wet weather accidents using analytical and empirical techniques was conducted by Ivey and Griffin (1977) at the Texas Transportation Institute (TTI). An analytical wet weather index and an empirical wet weather index were formulated and used to identify hazardous locations and predict wet weather accidents. The effectiveness of the analytical wet weather index, as the authors claimed, may not be superior to the empirical wet weather index.

The Bayesian estimation method is a different approach from those methods described previously. It has been applied to a vast area of science and engineering systems emphasizing the characteristic of uncertainty. In 1955, Robbins developed the nonparametric (frequency ratio) empirical Bayes method (EBM) to estimate the posterior mean of a Poisson distribution. The parametric EBM's were then developed by Maritz (1966, 1969, 1970), Rutherford and Krutchkoff (1969), and Lemmon and Krutchkoff (1969). Essentially, the parametric EBM follows the formalism of the Bayes theorem (1763) except that it evaluates its prior information and hyperparameters empirically.

The EBM has been generalized or reformulated by many researchers to deal with different problems such as survival time, risk and reliability estimation, and so on. The estimation of the number of accidents and/or the accident risk is one of the possible applications. The application of the EBM to a traffic event system was realized by Arnold and Antle (1978), who formulated a parametric EBM to estimate the risk of traffic citations for drivers. Abbess et al. (1981) also employed an empirical Bayes procedure to evaluate the effectiveness of the remedial treatment of road surfaces based on the expected number of accidents. Hauer and Persaud (1984) estimated the probabilities of accidents to determine the hazardousness of various locations. A variant EBM was developed by Brüde and Larsson (1988) to deal with problems that arise when using conventional EBM for a small sample size. For evaluating accident risk, Higle and Witkowski (1988) formulated a two-step EBM to identify hazardous locations, using biased estimators of the sample mean and variance. A study of wet pavement accidents using Arnold and Antle's procedure with grouping strategies was conducted by Kulakowski et al. (1990b). Because of the relatively small size in each group, the effect of grouping is not distinct. An extensive study of this wet pavement accident problem, using newly developed methods, was performed in this thesis. In a recent report, Morris et al. (1991) presented a hierarchical empirical Bayes procedure to rank highway sections based on expected accident rate and to evaluate the effectiveness of remedial measures. A reference data set with a large number of road sites is necessary for the method to evaluate its model parameters.

When the objective information (parameter measurements or collected data) is not available and human factors play an important role in the system, a subjective type of approach would be an alternative way to model the system.

The application of Shafer's belief function theory (1976) to the risk assessment problem in traffic event systems would be a new application in this research. A brief introduction to the Bayesian methods and the belief function theory is given in chapter 2.

1.3. Statement of the Problem

Generally speaking, accident reduction problems in the traffic event system can be viewed as a control problem. Two schemes are often applied. One is a direct approach and the other is an indirect approach. The direct approach methods include accident frequency method, accident rate method, severity rate method, EBM, and so on; these methods rely solely on the accident histories to assess the accident proneness for the location of interest without making any effort to identify the system. Conversely, the indirect approach methods--which include the correlation of geometrics method, accident risk factor method, formula method, wet weather index, and so on--attempt to identify the system first using the measurements of suspected attributing factors only or the measurements and the accident histories simultaneously. Appropriate countermeasures are prescribed based on the identified model. Figures 1.4 and 1.5 illustrate the structures of the two approaches.

The common problem to the direct approach methods is that the accident histories suffer from the deficiencies of time delay and insufficient degree of

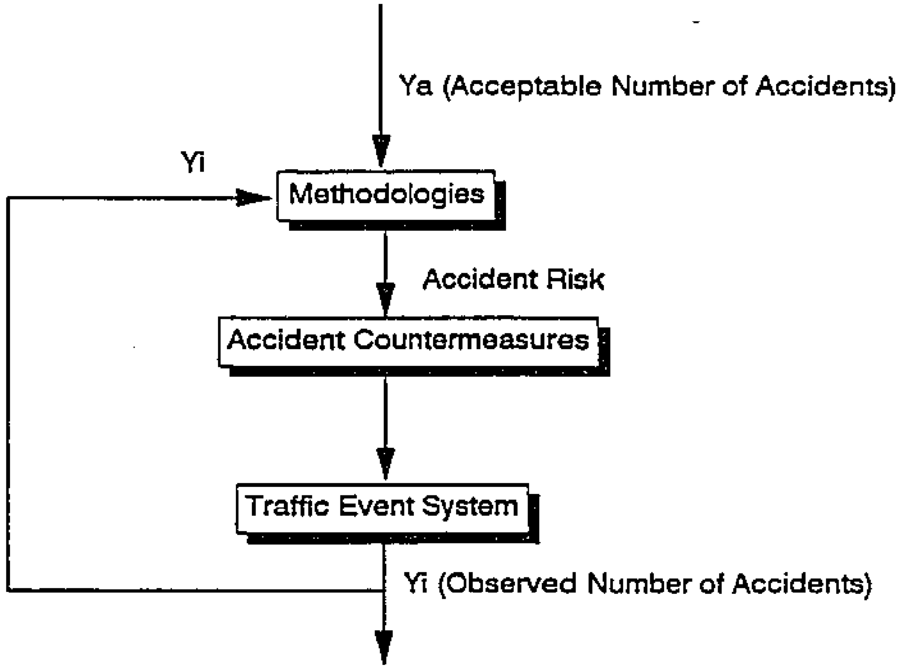


Figure 1.4. The structure of the direct approach.

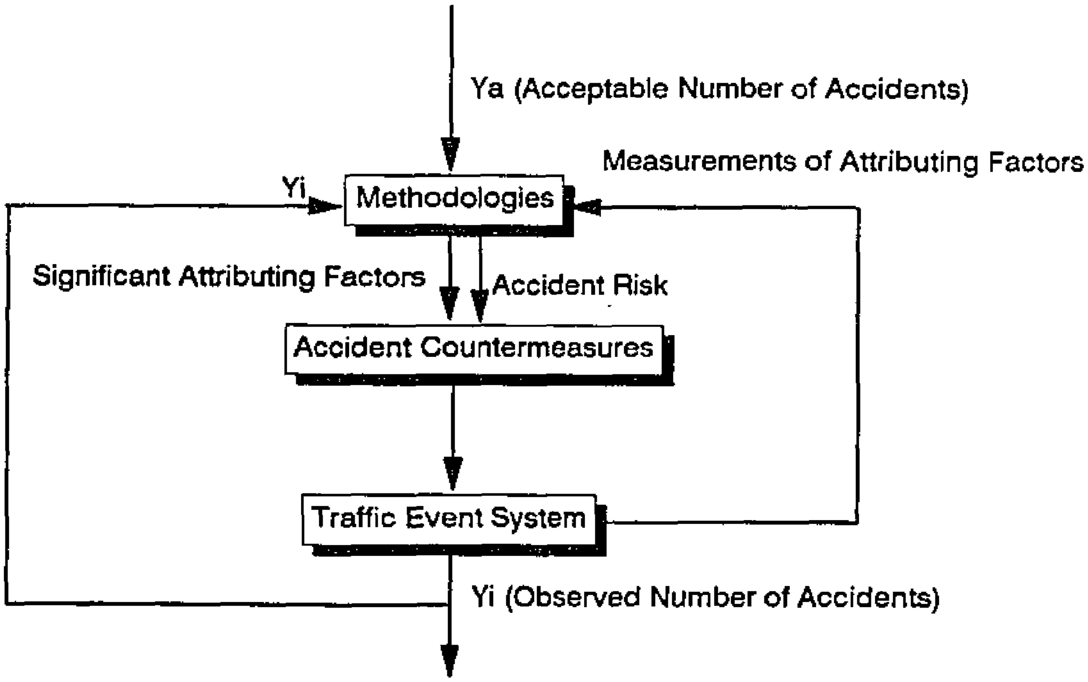

Ya (Acceptable Number of Accidents)
Measurements of Attributing Factors
Yi
Methodologies
Significant Attributing Factors
Accident Risk
Accident Countermeasures
Traffic Event System
Yi (Observed Number of Accidents)

Figure 1.5. The structure of the indirect approach.

accuracy when reporting. Moreover, a bias introduced by the regression-to-mean effect for the road site of interest cannot be neglected. A regression-to-mean effect is described as a phenomenon that a location with a large (small) number of accidents during a "before" period tends to decrease (increase) to a small (large) number of accidents in a similar "after" period *without having implemented any improvement measures.*

The difficulty with the indirect approach methods is that the accuracy of the identified model may be limited. This may be due to the selection of model structure or the system not being identifiable based on available data. In general, a deterministic model is favorable because of its simplicity; however, for the traffic event system, the deterministic model may not be suitable due to the fact that the system is stochastic in nature.

In response to the problems of regression-to-mean effect and the inherent randomness of the system, the EBM has been shown advantageous (Arnold and Antle 1978; Abbess et al. 1981; Hauer and Persaud 1984). However, the following shortcomings still remain:

- The adequacy of sample size for estimating the model parameters has not been determined.
- The random effect of vehicle exposure in the population of road sites has not been taken into account.

- No effort has been made to utilize both the accident histories and the measurements of roadway and traffic characteristics to identify significant causal factors for a traffic accident.

1.4. Research Objective

The objective of the proposed study is to develop a modeling technique, based on the probabilistic-type approaches, for traffic event system identification and prediction. Specifically, the problems of determining the adequate sample size for estimating model parameters, the random effect of vehicle exposure, and the utilization of both accident histories and measurement data will be addressed.

1.5. Thesis Overview

The methodologies presented in this thesis attempt to address the task of traffic event system identification and risk assessment. The work begins with a review of literature, problem statement, and description of research objective in chapter 1. An introduction to the Bayesian methods and belief function theory is presented in chapter 2 to provide a theoretical foundation for the later development work. The development work is divided into two parts--one which is considered as an objective type of approach, based on the Bayesian methods, and the other based on the belief function theory, a subjective type of approach. For the objective approach, a modification (Lin et al. 1991) was recommended in chapter 3 to improve the empirical Bayes procedure developed by Arnold and Antle in 1978. An evaluation of the empirical Bayes, maximum likelihood, and Bayes estimators using the Monte Carlo simulation technique was also performed and is presented in chapter 3. Two new median estimators were developed and are evaluated (Lin et al. 1991) in chapter 4, where an absolute error loss function is considered. In considering the subjective type of approach, a knowledge-based model was developed and presented in chapter 5. After the development work, validation was performed on a real data set provided by the Pennsylvania Department of Transportation using the developed methodologies; the validation is shown in chapter 6. A comparison of these methodologies based on their estimation accuracy was then made to determine the best method or methods to address the problem of risk assessment and identification. Conclusions and recommendations are then described in chapter 7.

Chapter 2

INTRODUCTION TO BAYESIAN METHODS AND BELIEF FUNCTION THEORY

2.1. Bayesian Methods

Consider an estimation problem in which observation y of a discrete random variable Y is available. The probability function is $f(y \mid \eta)$, and the parameter η is estimated in a minimum square error sense. Classically, the $f(y \mid \eta)$ is interpreted as a sampling distribution. The consideration of $f(y \mid \eta)$ as a function of y, with η fixed is called the sampling distribution of Y, given η. If results from a random sample, say y_1, y_2, y_3, ..., y_n, are available, then a likelihood function is defined as:

$$\prod_{i=1}^{i=n} f(y_i \mid \eta) \tag{2.1}$$

It is a function of the parameter η. The utilization of the likelihood function to represent the sample information is based on the likelihood principle, which states that the likelihood function contains all the information from the sample that is relevant for inference making. Following this principle, two parameter estimation schemes, the maximum likelihood estimator (MLE) and the Bayesian estimator are often used. The MLE is designed to find the value of parameter η that maximizes the likelihood function. It can be interpreted as the value of η that makes the observed sample results appear most likely (Winkler 1972). Frequently, it requires a complex iterative procedure to obtain a desired maximum value.

The well known Bayes theorem (1763) provides a simple and useful formalism to incorporate subjective prior knowledge into the analysis of an experiment. It is derived from manipulating the joint, conditional, and marginal probabilities. If A and B are considered to be two events, a mathematical expression of the Bayes theorem will be

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A) \tag{2.2}$$

To the estimation problem, a probability density function for the parameter η, referred to as prior distribution, is required in the Bayesian approach. Bayes theorem gives the posterior density of the parameter η as:

$$f(\eta|y) = \frac{f(y|\eta)f(\eta)}{\int f(y|\eta)f(\eta)d\eta} \tag{2.3}$$

Bayes' estimators will be the optimal estimators when the prior distribution is precisely known. In practice, this is rarely the case; therefore, an estimation procedure for the prior distribution must be developed. Before introducing the estimation procedure for the parameter η in the traffic event system, some fundamental assumptions were made and described in the next subsection. These assumptions are considered to be essential to the probabilistic-type models discussed in later chapters.

2.1.1. Fundamental Assumptions and Notations in Estimating Accident Rates

The occurrence of traffic accidents on a road site is generally assumed as a Poisson random process. This originates from the theory of queues and can be interpreted as below:

> If a random variable t_i is considered as the arrival time of the i^{th} customer (traffic accident) requiring service in a service system (at a road site) and the interarrival sequence associated to t_i $(i \geq 0)$ is represented by T_{i+1} with
>
> $$T_{i+1} = t_{i+1} - t_i \qquad (i \geq 0)$$
>
> then, the sequence T_1, T_2, T_3, ..., T_n, is called a point process over the positive real axis R^+. The point process is called a homogeneous Poisson random process with an intensity H if and only if its associated counting process of the number of customers (the number of traffic accidents) N(t) satisfies that
>
> 1. For every pair of {r,s} and s > r; {N(s)-N(r)} is a Poisson random variable with mean (s-r)H.
> 2. N(t), $t \geq 0$ has independent increments. That is, $\{N(t_2)\text{-}N(t_1)\}$, $\{N(t_3)\text{-}N(t_2)\}$, $\{N(t_4)\text{-}N(t_3)\}$, ..., $\{N(t_n)\text{-}N(t_{n-1})\}$ are independent for every $0 \leq t_1 \leq t_2 \leq t_3 \leq ..... \leq t_n$.

The assumption of a Poisson process implies that for a road site, the probability of y accidents occurring in time interval t given a constant average rate h can be represented by a Poisson distribution with the expression:

$$P(\ Y{=}y\,|\,t,\ H{=}h\ ) = \frac{(ht)^{y}e^{-(ht)}}{y!}\ , \quad with\ y = 0,1,2,3,\cdots \tag{2.4}$$

where Y and H are two random variables and ht (= λ) represents the average (expected) number of accidents in the time interval t. It should be noted that the average number of accidents λ, which is assumed to be the true number of accidents for any specific road site, can never be known and may vary from site to site. Estimating the average accident rate h and/or the expected number of accidents λ will be the main task of parameter estimation in the empirical Bayes procedures. In this research it is assumed that for the $i^{\underline{th}}$ highway section of interest, the number (y_i) of traffic accidents for a time period of interest will be a Poisson random variable with parameter λ_i given by:

$$\lambda_i = M_i h_i \tag{2.5}$$

where M_i is the traveled vehicle miles for the $i^{\underline{th}}$ section, calculated as:

$$M_i = SL_i * ADT_i * DAYS_i \tag{2.6}$$

where

SL	=	section length
ADT	=	average daily traffic
DAYS	=	time duration

The h_i in equation 2.5 is the accident rate for the i^{th} site. M_i will usually be given in millions of vehicle miles. M_i will be called the exposure and h_i the risk for the i^{th} site.

The empirical Bayes procedure developed by Arnold and Antle in 1978 realizes the application of the Bayesian estimation technique to a traffic event system. An introduction to the procedure is given in the next subsection.

2.1.2. Arnold and Antle Procedure

It is commonly assumed for a collection of different road sites that the expected accident rates h_i, $i=1,2,3,...$ are independent and identically distributed as gamma random variables with a density $f(h|\alpha,\beta)$ (Arnold and Antle 1978). The parameters α and β are called the shape parameter and the scale parameter, respectively, of the gamma distribution. The prior probability density for the accident rate at the i^{th} road site, h_i, can then be expressed as:

$$f(h_i|\alpha,\beta) = \frac{h_i^{\alpha-1}\exp^{-\frac{h_i}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}\ ,\quad \textit{with } h_i > 0,\ \alpha > 0,\ \beta > 0 \tag{2.7}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp^{-t}dt\ ,\quad \alpha>0 \tag{2.8}$$

here t is a real variable.

Since the parameters in the prior distribution for the h_i are not usually known, a procedure for estimating them from a given set of observations on highway sections *similar* to the i^{th} highway section must be developed. Following the approach given by Arnold and Antle (1978) (hereafter referred to as the AA procedure), it is assumed that for each of N similar highway sections, the exposure (M_i) and the number of accidents (y_i) for the time period of interest are known. If SY and SSY are the sums of y_i and y_i^2, respectively, and likewise if SM and SSM are the sums of M_i and M_i^2, respectively, it can be shown that when using the method of moments, estimates for α and β are given by:

$$\hat{\beta} = \frac{(SSY)(SM)}{(SY)(SSM)} - \frac{SM}{SSM} - \frac{SY}{SM} \qquad (2.9)$$

and

$$\hat{\alpha} = \frac{SY}{\hat{\beta}(SM)} \qquad (2.10)$$

The expected value of the accident rate $\hat{h}_i$, given the observation y_i, can be expressed as:

$$\hat{h}_i = E(h_i|y_i) = \frac{(y_i+\hat{\alpha})\hat{\beta}}{1.0+\hat{\beta}(M_i)} \qquad (2.11)$$

In this case, the maximum likelihood estimate for h_i is given by y_i / M_i.

In the following chapters, the AA procedure was considered for the problem of risk assessment and identification in a traffic event system. A modification of this procedure was recommended for those problems in which the estimates of the parameters obtained in the prior distribution become unreasonable. Two new approximate median estimators based on an absolute error loss function are also presented.

2.2. Belief Function Theory

The belief function theory (Dempster 1967; Shafer 1976) is intended to provide a mathematical foundation and systematic procedure for combining bodies of evidence of a proposition. It originates from the formalism of Bayesian inference. However, it assigns lower probabilities (Shafer's degree of belief) to propositions rather than simple additive probabilities as the Bayesian does. The theory uses a number between 0 and 1 to indicate the degree of belief (support) for a body of evidence to a proposition. Its combination scheme for the degrees of support to a proposition is called Dempster's rule of combination. An application of this theory to the problem of identification and risk assessment in traffic event systems to construct a knowledge-based model is presented in chapter 5. A brief introduction to this theory is presented in the following subsections to give a general picture. The notations, definitions, and theorems are followed by Shafer's mathematical theory of evidence (1976).

2.2.1. Basic Definitions and Theorems

Consider a parameter θ and the finite set of its possible values Θ; the proposition of interest would be "the true value of θ is A". Here A is a subset of Θ. The set of all subsets of Θ is denoted by 2^{Θ}. As an illustration, if:

$$\Theta = \{A, B\}$$

$$\text{then} \quad 2^{\Theta} = \{\{\varnothing\}, \{A\}, \{B\}, \{A, B\}\}$$

This implies that, by properly choosing the θ and Θ, the 2^{Θ} can be the set that contains all propositions of interest. If we let the Θ be the set of all the different possibilities under consideration, then it is called the *frame of discernment* that discerns a proposition corresponds to a subset of Θ (Shafer 1976).

DEFINITION 2.1. If Θ is a frame of discernment and A is one of its subsets, then a function m: $2^{\Theta} \rightarrow [0,1]$ is called a *basic probability assignment* (bpa) whenever

$$m(\varnothing) = 0 \ \text{ and } \sum_{A \subset \Theta} m(A) = 1.$$

The quantity m(A) is called A's *basic probability number.* It is the measure of the belief that is committed exactly to A, but not the total belief committed to A. The m(A) cannot be further subdivided and does not include portions of belief committed to subsets of A. To measure the total belief committed to A, a belief function is defined. The class of belief function based on the bpa m is defined as:

DEFINITION 2.2. A function Bel: $2^{\Theta} \rightarrow [0,1]$ is called a belief function over Θ if for some basic probability assignment m: $2^{\Theta} \rightarrow [0,1]$

$$Bel(A) = \sum_{B \subset A} m(B)$$

It should be noted that the bpa m produces a given belief function that is unique and can be recovered from the belief function by the following theorem:

THEOREM 2.1. Suppose Bel: $2^{\Theta} \rightarrow [0,1]$ is the belief function given by the basic probability assignment m: $2^{\Theta} \rightarrow [0,1]$, then

$$m(A) = \sum_{B \subset A} (-1)^{|A-B|} Bel(B) \ ; \quad \forall A \subset \Theta \qquad (2.12)$$

where $|A-B|$ is the number of elements in the set of $\{A - B\}$.

In addition to the above definition of the belief function, another characterization of the belief function is shown in the following theorem:

THEOREM 2.2. If Θ is a frame of discernment, then a function Bel: $2^{\Theta} \rightarrow [0,1]$ is a belief function if and only if it satisfies

(1). Bel($\varnothing$) = 0.

(2). Bel(Θ) = 1.

(3). For every positive integer n and every collection $A_1, A_2, \ldots, A_n$ of subsets of Θ,

$$Bel(A_1 \cup \cdots \cup A_n) \geq \sum_{\substack{I \subset \{1, \cdots, n\} \\ I \neq \emptyset}} (-1)^{|I|+1} \, Bel(\bigcap_{i \in I} A_i)$$

2.2.2. Belief Interval

The belief function Bel(A), however, does not completely describe one's belief about proposition A. Since Bel(A) does not reveal to what extent one believes its negation $\bar{A}$, a definition of degree of doubt is necessary. The degree of doubt $Doub(A)$ is defined as:

$$Doub(A) = Bel(\bar{A})$$

If one lets

$$P^*(A) = 1 - Dou(A)$$

then the quantity $P^*(A)$ is called the *upper probability* of A, or the *plausibility* of A. It can be expressed in terms of the basic probability assignment m:

$$P^*(A) = 1 - Bel(\bar{A}) = \sum_{B \subset \Theta} m(B) - \sum_{B \subset \bar{A}} m(B) = \sum_{B \cap A \neq \emptyset} m(B). \qquad (2.13)$$

Comparing equation 2.13 with the expression in definition 2.2, one can conclude that

$$Bel(A) \leq P^*(A) \qquad (2.14)$$

This inequality 2.14 characterizes a belief interval for the proposition A if the Bel(A) is called a *lower probability function* and the $P^{*}(A)$ an *upper probability function* to A, respectively. The interval is then denoted by [Bel(A), 1-Bel(Ā)].

2.2.3. Combination Scheme

After defining the belief function and belief interval in the previous subsections, the combination scheme for pooling evidence can be introduced. Dempster's rule of combination (1967) provides a simple and effective way to compute an orthogonal sum of distinct bodies of evidence and produces a new belief function based on the combined evidence. A weight of conflict was also introduced to deal with the problem of conflicting bodies of evidence.

Suppose m_1 is the bpa over a frame of discernment Θ for a belief function Bel_1 and has the elements $A_1, A_2, \cdots, A_n$; the probability masses of m_1 can be depicted as segments of a line of length 1, as shown in figure 2.1. Similarly, for another bpa m_2 over the Θ of a belief function Bel_2 with elements $B_1, B_2, \cdots, B_m$, the probability masses of m_2 are depicted in figure 2.1. The combination of Bel_1 and Bel_2 was performed by calculating the total intersection areas shown in figure 2.2. These areas represent a joint effect of Bel_1 and Bel_2. This implies that the total probability mass exactly committed to A can be represented by:

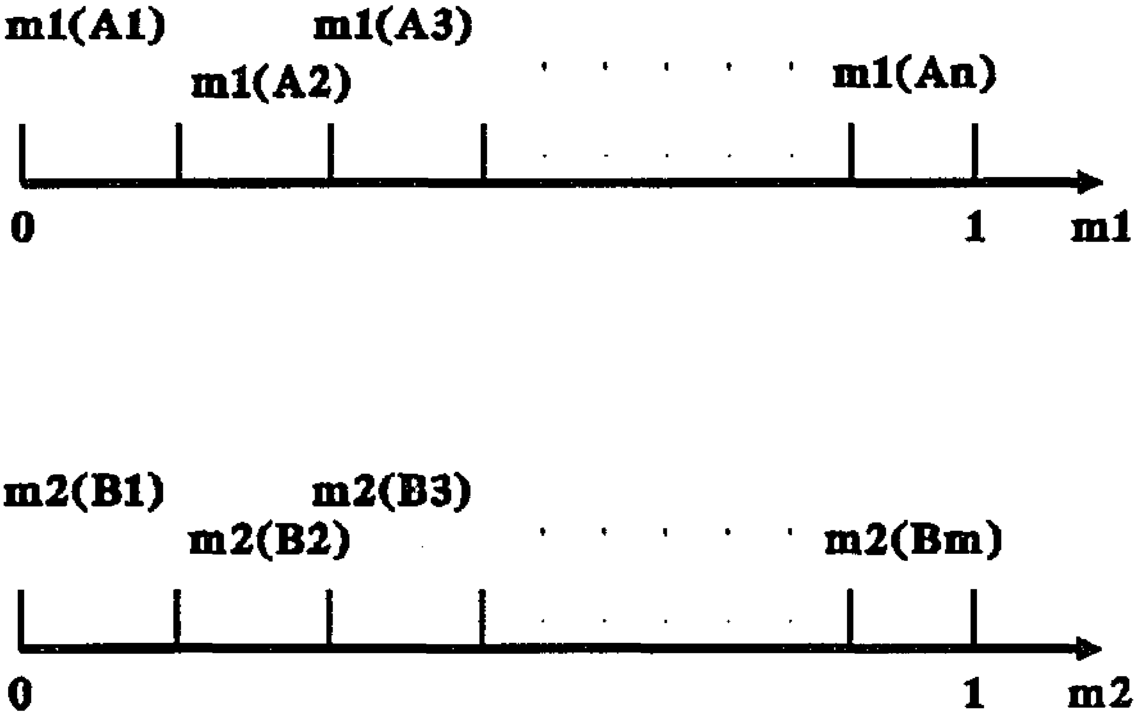


Figure 2.1. Probability mass segments of m_1 and m_2.

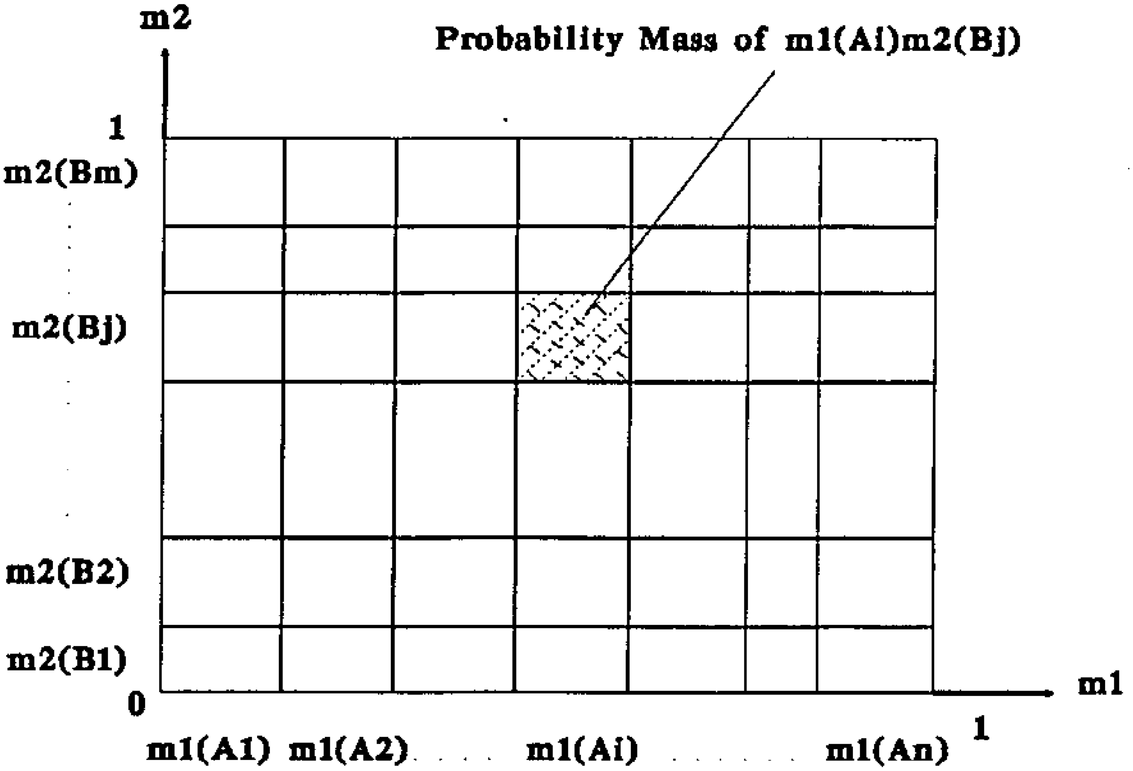

m2
Probability Mass of m1(Ai)m2(Bj)
1
m2(Bm)
m2(Bj)
m2(B2)
m2(B1)
0
m1
m1(A1) m1(A2) . . . . m1(Ai) . . . . m1(An)
1

Figure 2.2. Graphical representation of Dempster's rule of combination.

$$\sum_{\substack{i,j \\ A_i \cap B_j = A}} m_1(A_i) m_2(B_j) \tag{2.15}$$

There exists a problem in calculating the orthogonal sum of m_1 and m_2 denoted by $m_1 \oplus m_2$: some of the intersection areas may commit to the empty set $\{\emptyset\}$; that is, $A_i \cap B_j = \emptyset$. To eliminate this problem, Dempster (1967) discarded those areas committed to the empty set and introduced a weight of conflict--a normalization factor K. The factor K measures the extent of conflict between evidences and is represented by:

$$K = \frac{1}{1 - c} = \frac{1}{1 - \sum\limits_{\substack{i,j \\ A_i \cap B_j = \emptyset}} m_1(A_i) m_2(B_j)} \tag{2.16}$$

If $c < 1$, then the function m: $2^{\Theta} \rightarrow [0,1]$ is a basic probability assignment and is characterized by:

$$m(\emptyset) = 0 \ \ and \ \ m(A) = \frac{\sum\limits_{\substack{i,j \\ A_i \cap B_j = A}} m_1(A_i) m_2(B_j)}{1 - \sum\limits_{\substack{i,j \\ A_i \cap B_j = \emptyset}} m_1(A_i) m_2(B_j)} \tag{2.17}$$

Dempster's rule of combination can be justified by using simple support functions. The simple support functions (Shafer 1976, pp.74-75) are belief functions based on evidence points precisely to a single non-empty subset A of Θ. If S is a simple support function focused on A, then m(A) = S(A), m(Θ) = 1 -

S(A), and m(B) = 0 for all other $B \subset \Theta$. The quantity m(Θ) is a measure of that portion of the total belief that remains unassigned after commitment of belief to various proper subsets of Θ. Suppose $S_1(A) = s_1$ and $S_2(B) = s_2$ are two simple support functions focused on A and B, respectively. If $A \cap B \neq \emptyset$, then both support functions are heterogeneous and the combined effect of the bodies of evidence can be shown by table 2.1. If $A \cap B = \emptyset$, then the two bodies of evidence are conflicting. The weight of conflict K can now be applied to normalize the basic probability numbers m. Table 2.2 shows the combined results.

2.2.4. An Example

Suppose that one wants to diagnose the cause of a bad roadway section and three basic propositions have been proposed--bad pavement condition (PC), bad geometric condition (GC), and bad traffic condition (TC). The frame of discernment Θ can then be represented by a set of {PC, GC, TC}, and its subsets are depicted in figure 2.3 except the empty set $\{\emptyset\}$.

Suppose that a body of evidence confirms the diagnosis of bad pavement condition or bad traffic condition to the degree of 0.6. Then m({PC, TC}) = 0.6, m(Θ) = 0.4, and the value of m for every other subset of Θ is 0. The total belief committed to the subset of {PC, TC} is expressed as:

Table 2.1. Orthogonal sum of heterogeneous evidence for propositions A and B.

Committed to A $m(A) = s_1(1 - s_2)$	Uncommitted $m(\Theta) = (1-s_1)(1-s_2)$
Committed to $A \cap B$ $m(A \cap B) = s_1 s_2$	Committed to B $m(B) = s_2(1 - s_1)$

Table 2.2. Orthogonal sum of conflicting evidence for propositions A and B.

Committed to A $m(A) = \frac{s_1(1 - s_2)}{1 - s_1 s_2}$	Uncommitted $m(\Theta) = \frac{(1-s_1)(1-s_2)}{1 - s_1 s_2}$
Committed to $\emptyset$: $s_1 s_2$	Committed to B $m(B) = \frac{s_2(1 - s_1)}{1 - s_1 s_2}$

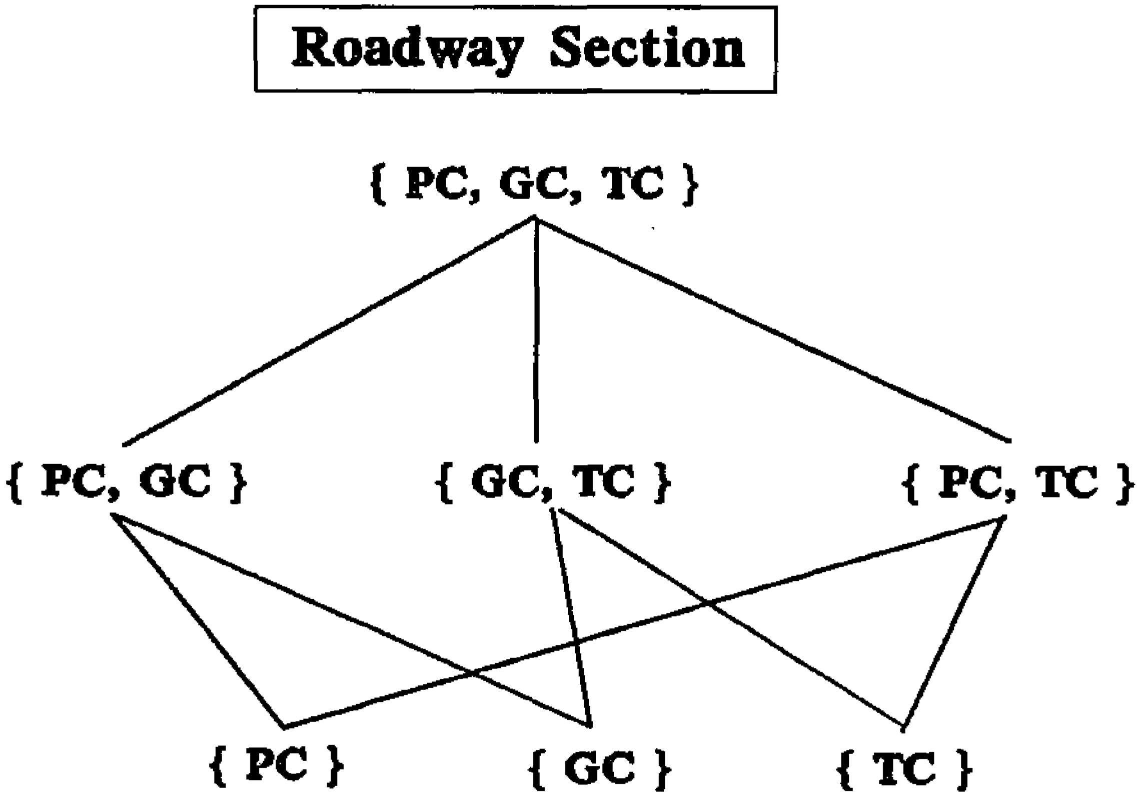


Figure 2.3. The subsets of the set of the roadway section.

$$Bel(\{PC, TC\}) = m(\{PC, TC\}) + m(\{PC\}) + m(\{TC\})$$

Which is different from the amount of belief that committed precisely to A, the m({PC, TC}).

If there are two bodies of evidence, one confirming the proposition to the degree of 0.5 with basic probability assignment m_1 and the other disconfirming the proposition to the degree of 0.3 through the basic probability assignment m_2, then the combined effect on belief is given by $m_1 \oplus m_2$, an orthogonal sum. Following the formulation given in table 2.2, one can obtain the value of normalization factor K = 1 / (1 - 0.15) = 1.176. The combined basic probability numbers are

$$m_1 \oplus m_2(\{PC, TC\}) = 0.35*1.176 = 0.4117$$

$$m_1 \oplus m_2(\{\bar{PC}, \bar{TC}\}) = 0.15*1.176 = 0.1764$$

$$m_1 \oplus m_2(\Theta) = 0.35*1.176 = 0.4117$$

Where $\{\bar{PC}, \bar{TC}\}$ represents the complement subsets of {PC, TC} over the frame of discernment Θ = {PC, TC, GC}. It should be noted that the belief interval of {PC, TC} has been changed from the original interval of [0.5, 0.7] to a new interval of [0.4117, 0.8236] after the combination process. This is intuitively correct since the disconfirmation evidence mildly eroded the degree of support to the proposition of {PC, TC}.

2.3. Summary

In this chapter, the Bayesian methods and belief function theory were introduced. Specifically, the AA procedure was discussed. The procedure will be considered in the following chapters for the problem of accident rate estimation. The brief introduction to the belief function theory provides a foundation for constructing a subjective-type model, which will be discussed in chapter 5.

Chapter 3

A MODIFIED EMPIRICAL BAYES PROCEDURE

While it is a viable method with important strengths, limitations also exist in the AA procedure. First, it cannot secure the requirement of positive parameters for the assumption of a gamma random variable for the accident rate. Second, the sample size problem of the collected data needs to be addressed since it conceivably affects the quality of parameter estimation. Third, the assumption of constant exposure in evaluating the estimation procedure is insufficient since the exposure in a traffic event system is likely to be randomly distributed. A modified rule is proposed in the following sections to improve Arnold and Antle's empirical Bayes procedure.

3.1. A Modified Rule

The phenomenon of nonpositive parameters is often encountered when the sample variance is less than the sample mean for the collected data. These are not allowable in the gamma distribution and must be replaced by some other values. A general remedy measure is to assign a large value for the parameter α (Maritz 1969). This may not be appropriate for the real data because of the large dispersion of the real data. It was observed from several exploratory analyses of simulated data that for the traffic accident problem, the values of α are often between 1 and 6. This provides the basis for considering a large value of shape parameter α (e.g. 10) as unwanted.

To account for the undesired conditions of nonpositive and large values of parameter α, a modified rule for the AA procedure is necessary. Four different rules were examined to determine the best one using the Monte Carlo simulation technique. In essence, they are fixed parameter type and MLE-type rules. The values of 1.5 and 10 were chosen based on several exploratory simulations. Table 3.1 summarizes these rules.

Table 3.1. Remedy rules.

Condition	Remedy Rules			
	Rule 1	Rule 2	Rule 3	Rule 4
If $\alpha \leq 0$	Set $\alpha = 1.5$	Set $\alpha = 1.5$	Use MLE	Use MLE
If $\alpha > 10$	Set $\alpha = 10$	Use MLE	Set $\alpha = 10$	Use MLE

The simulation was made to simulate the Poisson process for traffic accidents. Accident rates h were generated by a gamma distribution with different parameter values. Three sets of parameters, ($\alpha = 1.2$, $\beta = 2.5$), ($\alpha = 3.0$, $\beta = 2.5$), and ($\alpha = 6.0$, $\beta = 2.5$), were selected. For the vehicle exposure, based on the observation of a real data set provided by the Pennsylvania Department of

Transportation, its random nature can be represented by a Weibull random variable with a corresponding high or low exposure. The high exposure Weibull had parameters of $a = 0.15$, $b = 0.25$, and $c = 2.0$, while the low exposure Weibull had parameters of $a = 0.01$, $b = 0.1$, and $c = 2.0$. The parameters a, b, and c represent the location parameter, scale parameter, and shape parameter, respectively. Combining the accident rate h and the vehicle exposure, accidents were generated through a Poisson generator for different sample sizes of road sites. All of the generating processes used the TULSIM (Boswell 1987) software.

The sample size of generated road sites varies from 15 to 260. The criterion for selecting a good rule is based on the mean absolute errors between the actual and the estimated accident rate h for a given sample. An amount of 10% error was arbitrarily chosen. Table 3.2 presents the results of simulation. From the results in table 3.2, the first rule with $\alpha = 1.5$ for $\alpha \leq 0$ and $\alpha = 10$ for $\alpha > 10$ was chosen.

3.2. A Comparison of the Empirical Bayes, Maximum Likelihood, and Bayes Estimators Using Simulated Data

In general, the Bayes procedure and the MLE method will produce error rates that do not depend on the number (N) of highway sections in the group, but the performance of the empirical Bayes procedure will depend upon N. To illustrate the effects of exposure level (high and low) and number of highway sections, several computer simulations were carried out, the results of which are

Table 3.2. Summarized results of Monte Carlo simulation.

Parameter		$\alpha = 1.2$	$\beta = 2.5$	$\alpha = 3.0$	$\beta = 2.5$	$\alpha = 6.0$	$\beta = 2.5$
Vehicle Exposure		weib (2.0, 0.1,0.01)	weib (2.0, 0.25, 0.15)	weib (2.0, 0.1, 0.01)	weib (2.0, 0.25, 0.15)	weib (2.0, 0.1, 0.01)	weib(2.0, 0.25,0.15)
Remedy Rules	Set $\alpha = 1.5$ if $\alpha \leq 0$ Set $\alpha = 10$ if $\alpha > 10$	MAE <10% for N ≥35	MAE <10% for every N	MAE <10% for N ≥45	MAE <10% for N ≥35	MAE <10% for N ≥100	MAE<10 % for every N
	Set $\alpha = 1.5$ if $\alpha \leq 0$ Use MLE if $\alpha > 10$	MAE <10% for N ≥45	MAE <10% for N ≥45	MAE >10% for every N	MAE <10% for N ≥120	MAE >10% for every N	MAE>10 % for every N
	Use MLE if $\alpha \leq 0$ Set $\alpha = 10$ if $\alpha > 10$.	MAE <10% for N ≥140	MAE <10% for N ≥100	MAE <10% for N≥180	MAE <10% for N ≥60	MAE <10% for N ≥220	MAE<10 % for N ≥60
	Use MLE if $\alpha \leq 0$ Use MLE if $\alpha > 10$	MAE <10% for N ≥160	MAE <10% for N ≥60	MAE >10% for every N	MAE <10% for N ≥120	MAE >10% for every N	MAE>10 % for every N

Notes: MAE = mean absolute error.
weib = Weibull distribution.
N = sample size from 15 to 260.

given in figures 3.1 through 3.12. It can be seen from figures 3.1 through 3.6 that under these simulated conditions the modified AA procedure is almost as good as the ideal Bayes procedure whenever there are at least 60 highway sections in the group of interest. It is also clear that the greater the exposure the less the benefit of the Bayes procedure. This is to be expected, since there will be a great deal of information for each site when the exposure at the site is large, and thus less need for using information from other similar sites. For estimating the parameters α and β, the EBM is observed to have a good performance from figure 3.7 to figure 3.12 when sample size is greater than or equal to 100. Hence, a recommended sample size for parameter estimation is 100.

3.3. Summary

A modified rule, $\alpha = 1.5$ for $\alpha \leq 0$ and $\alpha = 10$ for $\alpha > 10$, was proposed in this chapter to deal with the problems of nonpositive and large values of parameter α. A comparison of the modified AA procedure, maximum likelihood, and Bayes estimators was carried out using computer simulation. The simulation results indicate that the modified procedure performs almost as well as any possible rule could perform.

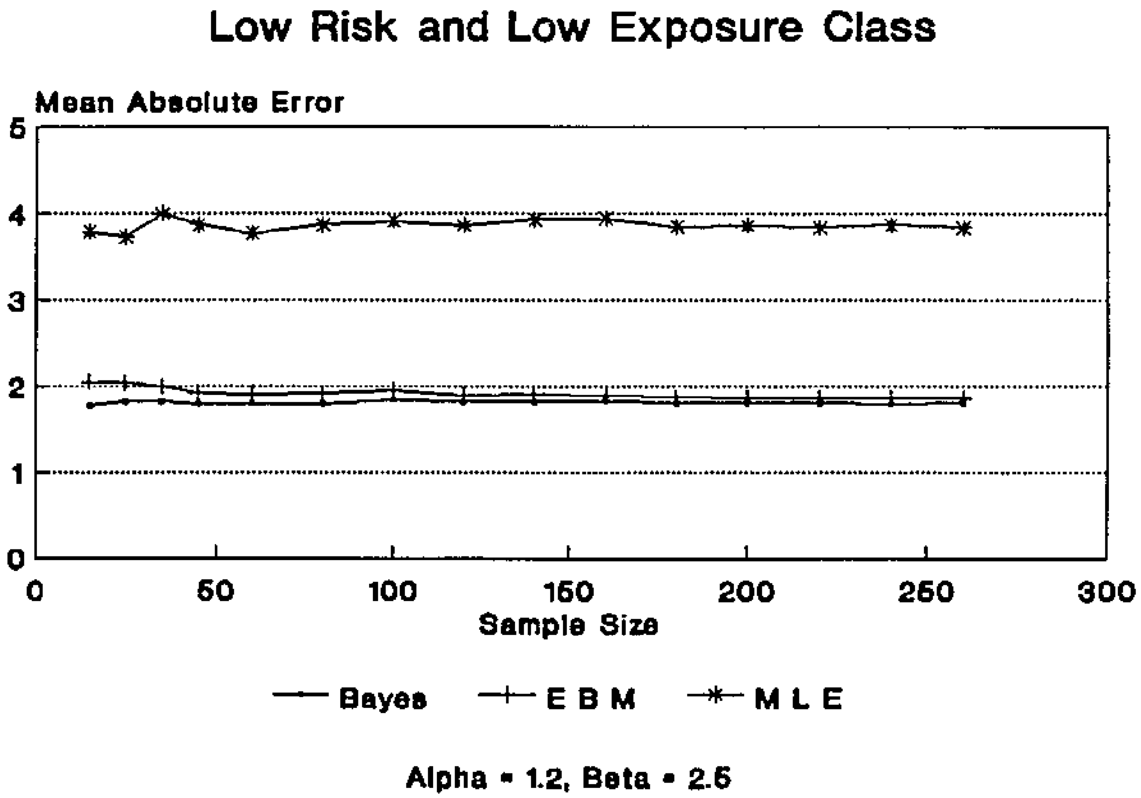


Figure 3.1. Mean absolute errors in estimating the accident risk for a low-risk and low-exposure class of roads.

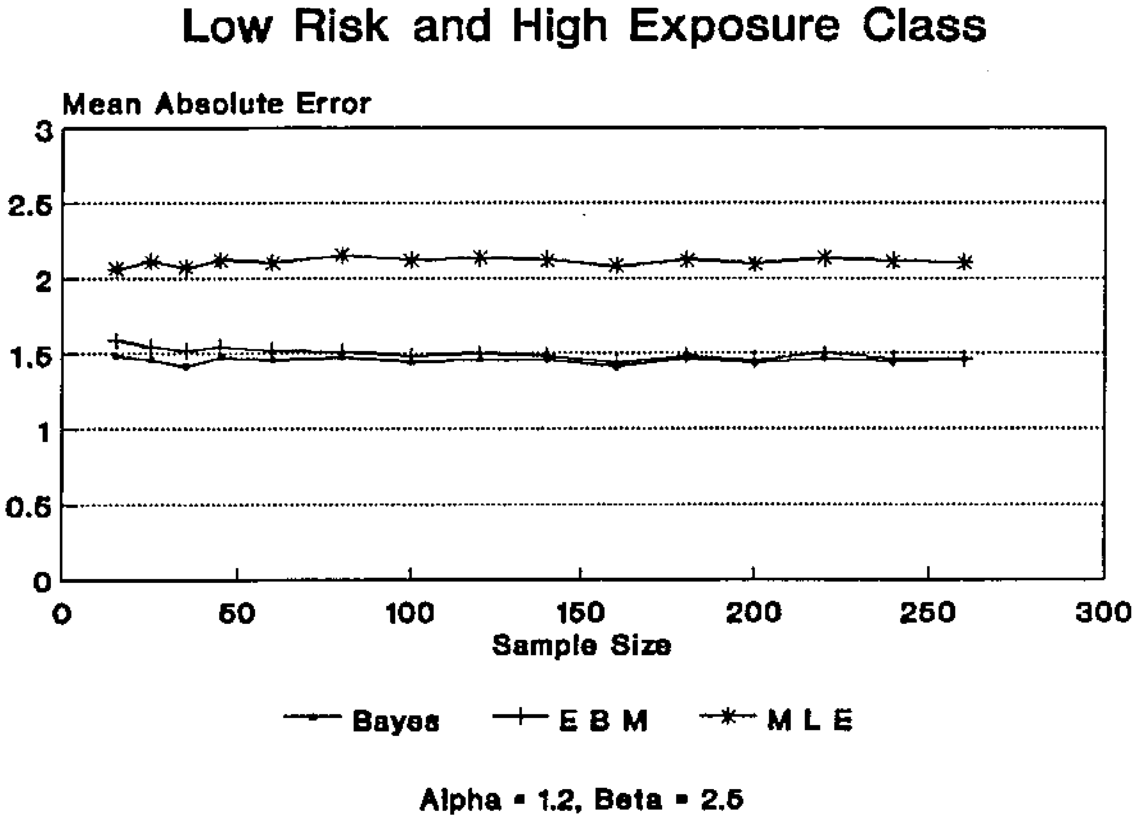


Figure 3.2. Mean absolute errors in estimating the accident risk for a low-risk and high-exposure class of roads.

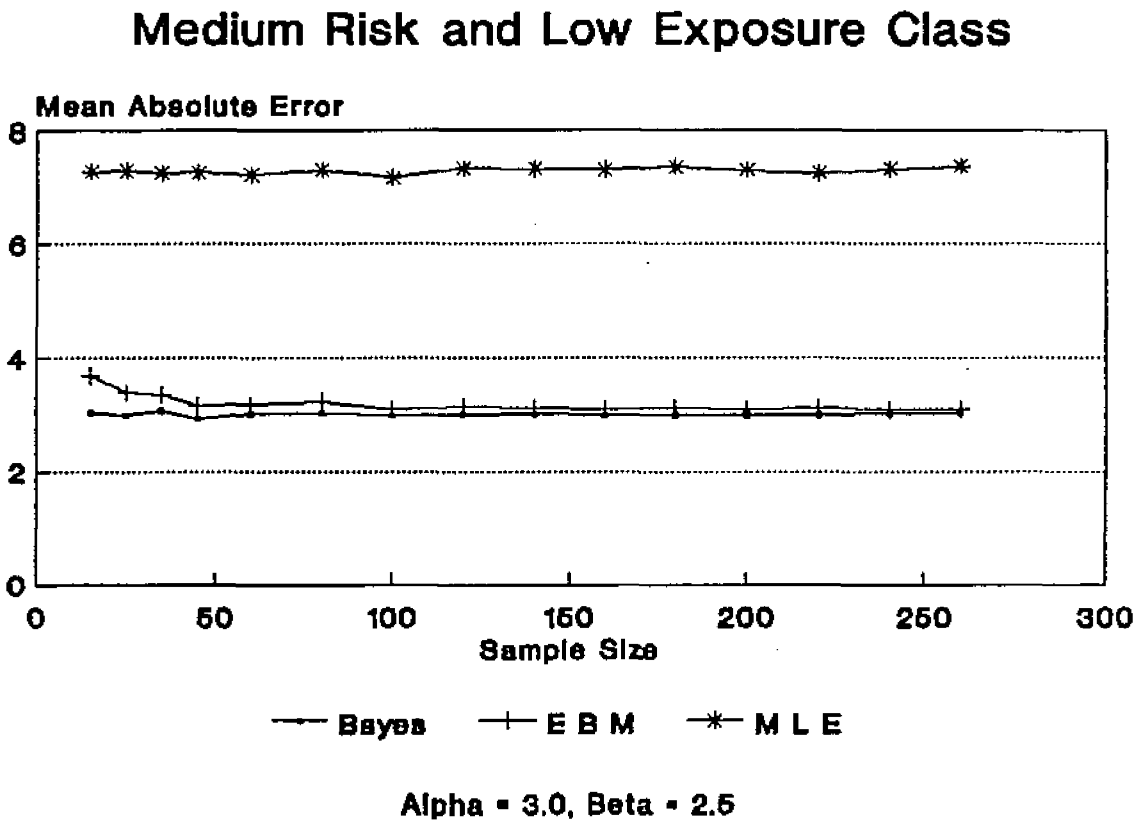


Figure 3.3. Mean absolute errors in estimating the accident risk for a medium-risk and low-exposure class of roads.

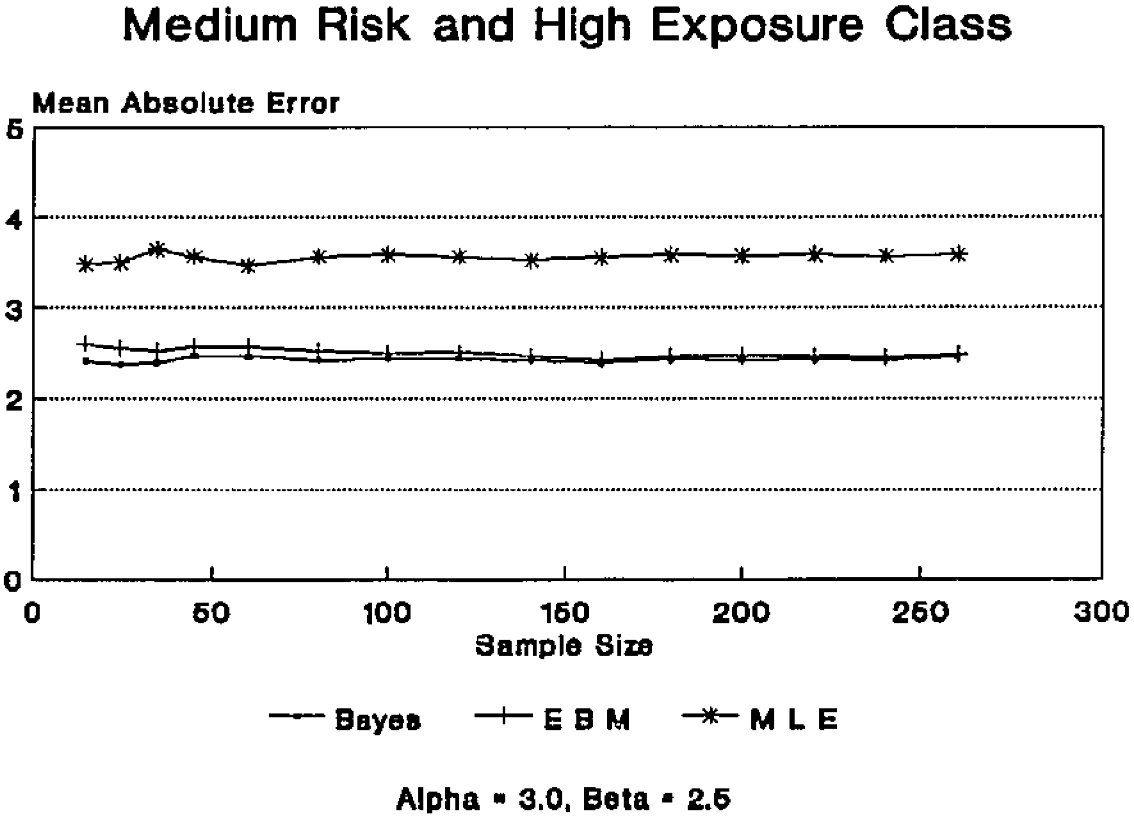


Figure 3.4. Mean absolute errors in estimating the accident risk for a medium-risk and high-exposure class of roads.

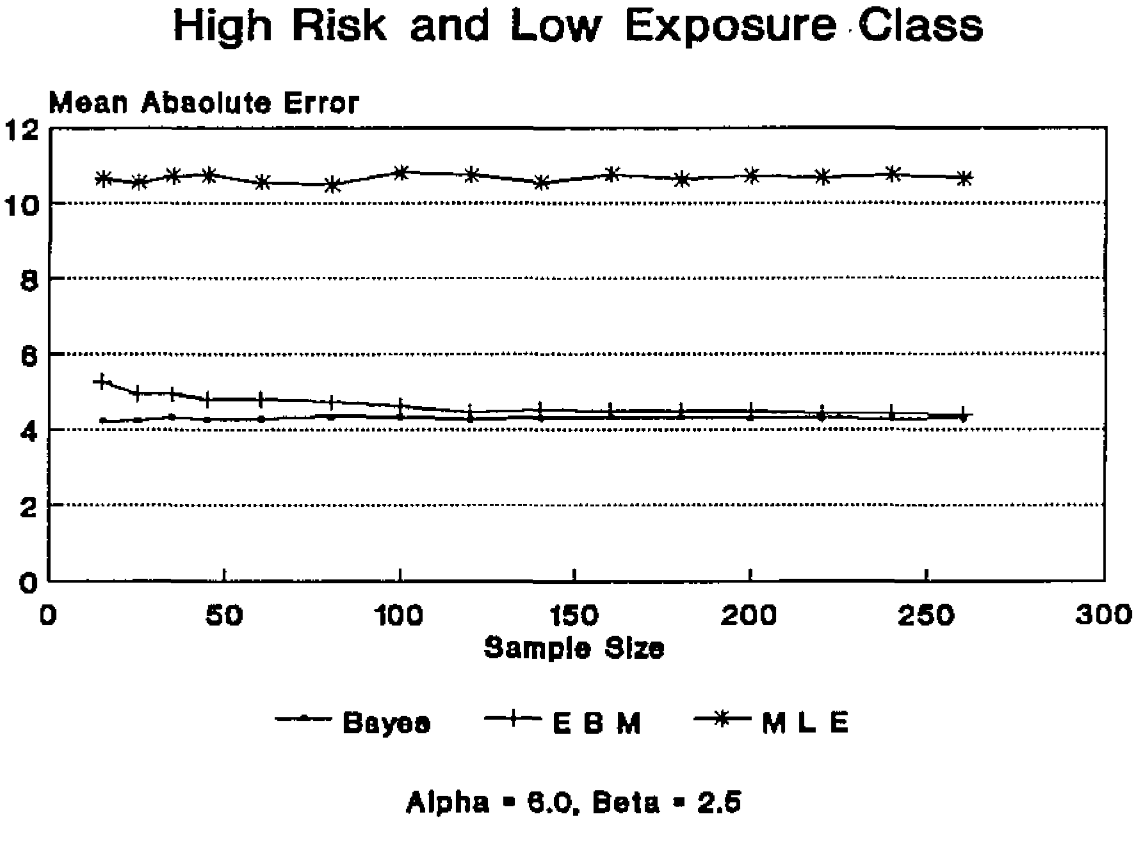


Figure 3.5. Mean absolute errors in estimating the accident risk for a high-risk and low-exposure class of roads.

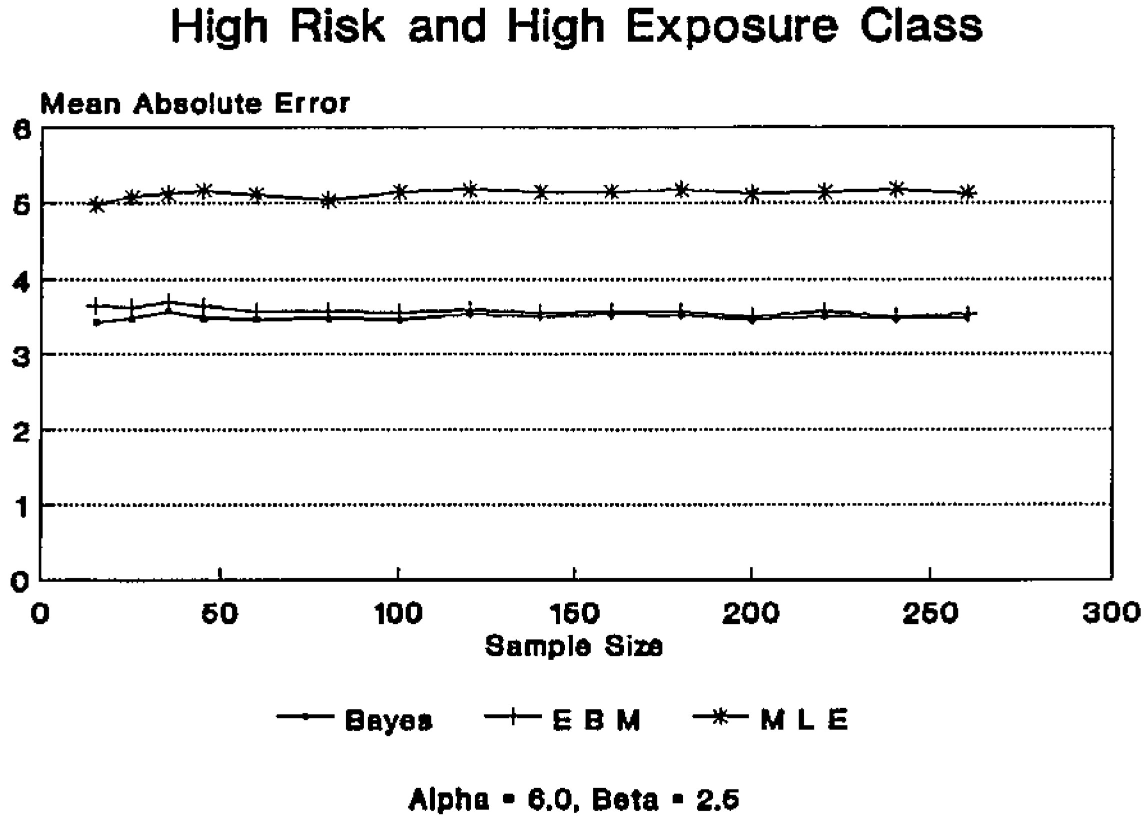


Figure 3.6. Mean absolute errors in estimating the accident risk for a high-risk and high-exposure class of roads.

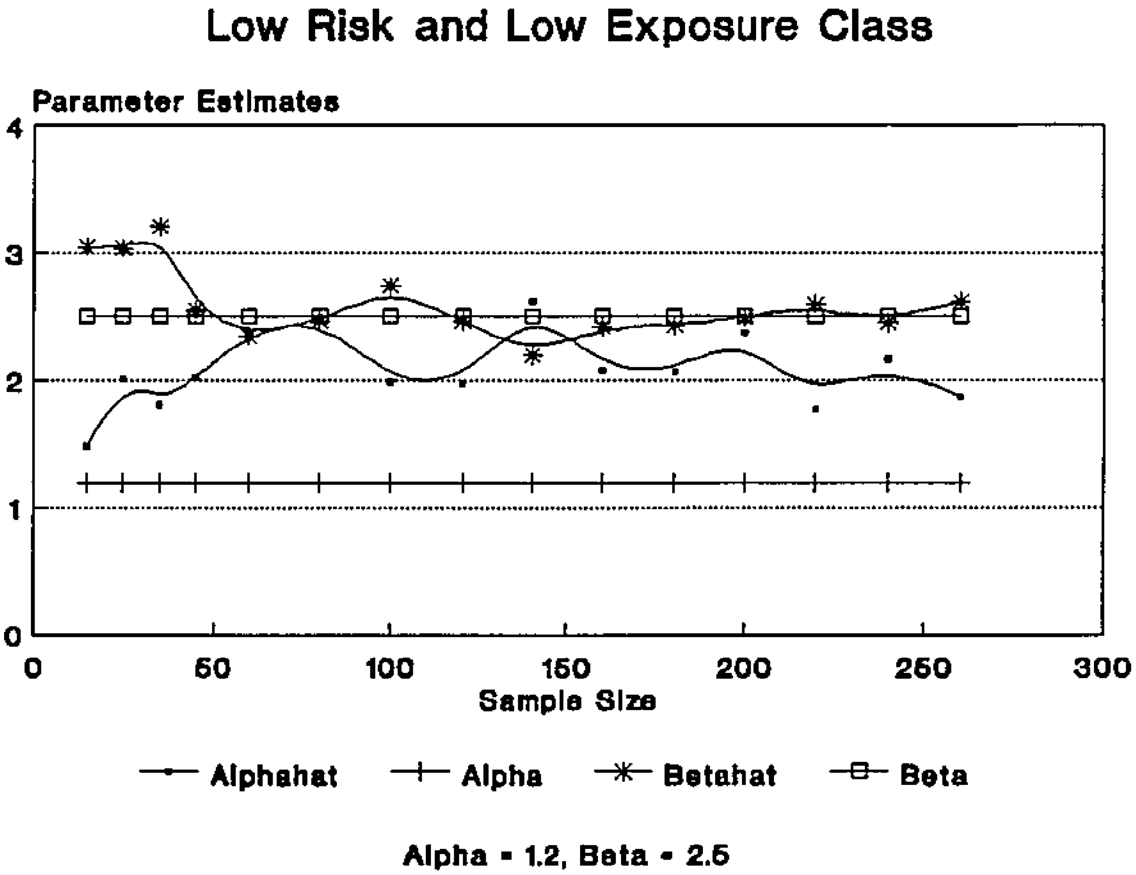


Figure 3.7. Parameter estimates in estimating the accident risk for a low-risk and low-exposure class of roads.

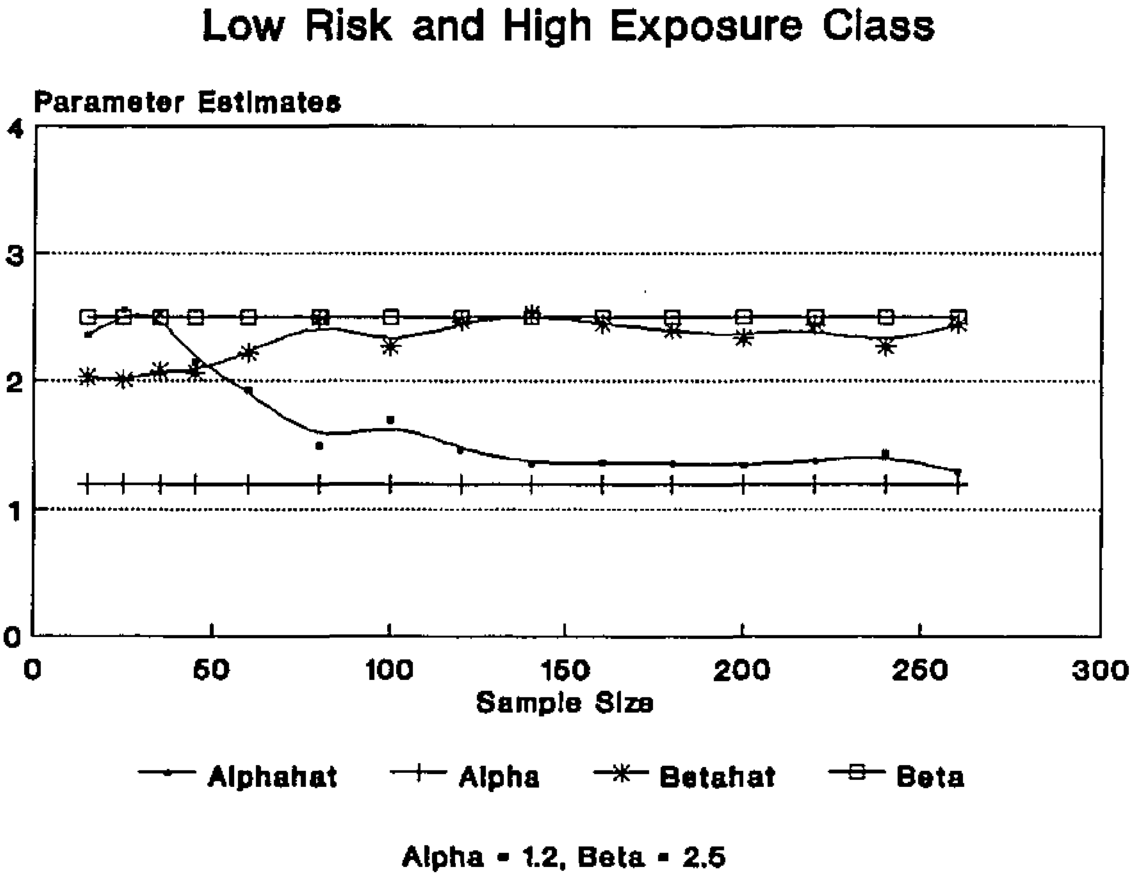


Figure 3.8. Parameter estimates in estimating the accident risk for a low-risk and high-exposure class of roads.

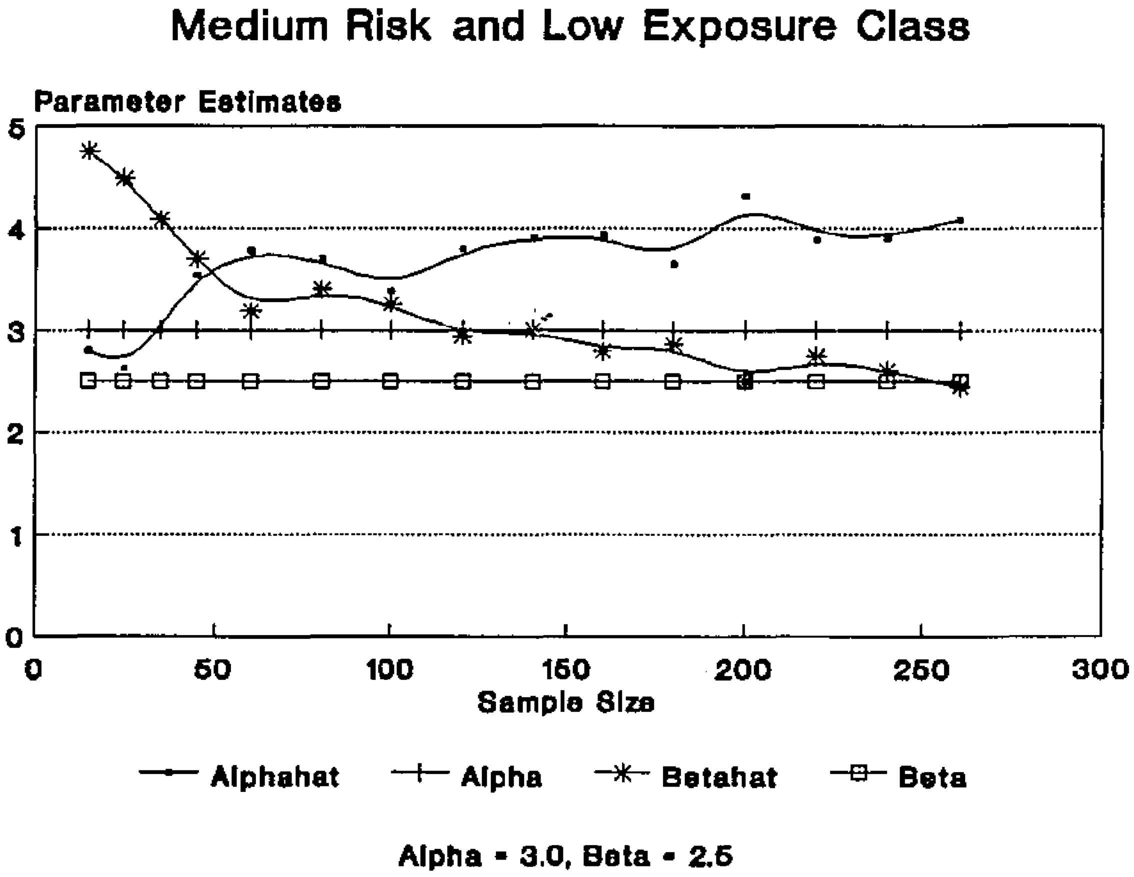


Figure 3.9. Parameter estimates in estimating the accident risk for a medium-risk and low-exposure class of roads.

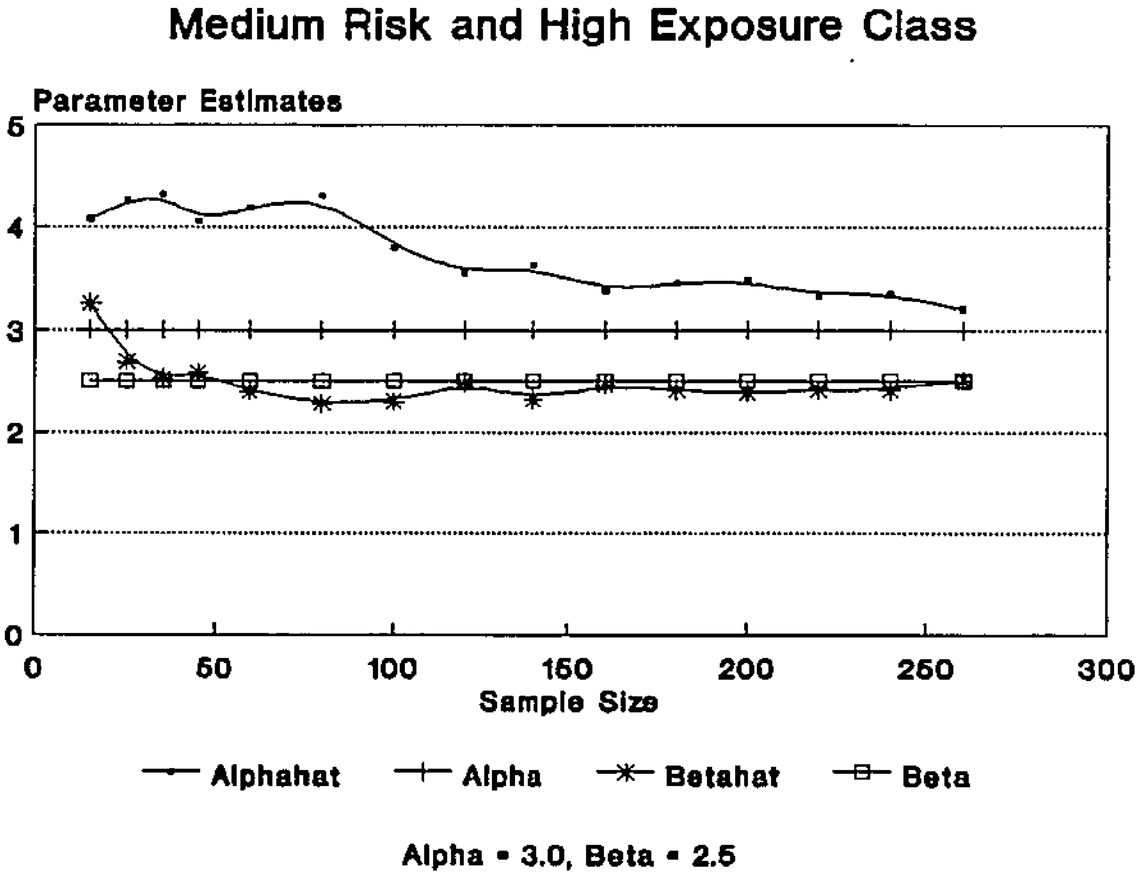


Figure 3.10. Parameter estimates in estimating the accident risk for a medium-risk and high-exposure class of roads.

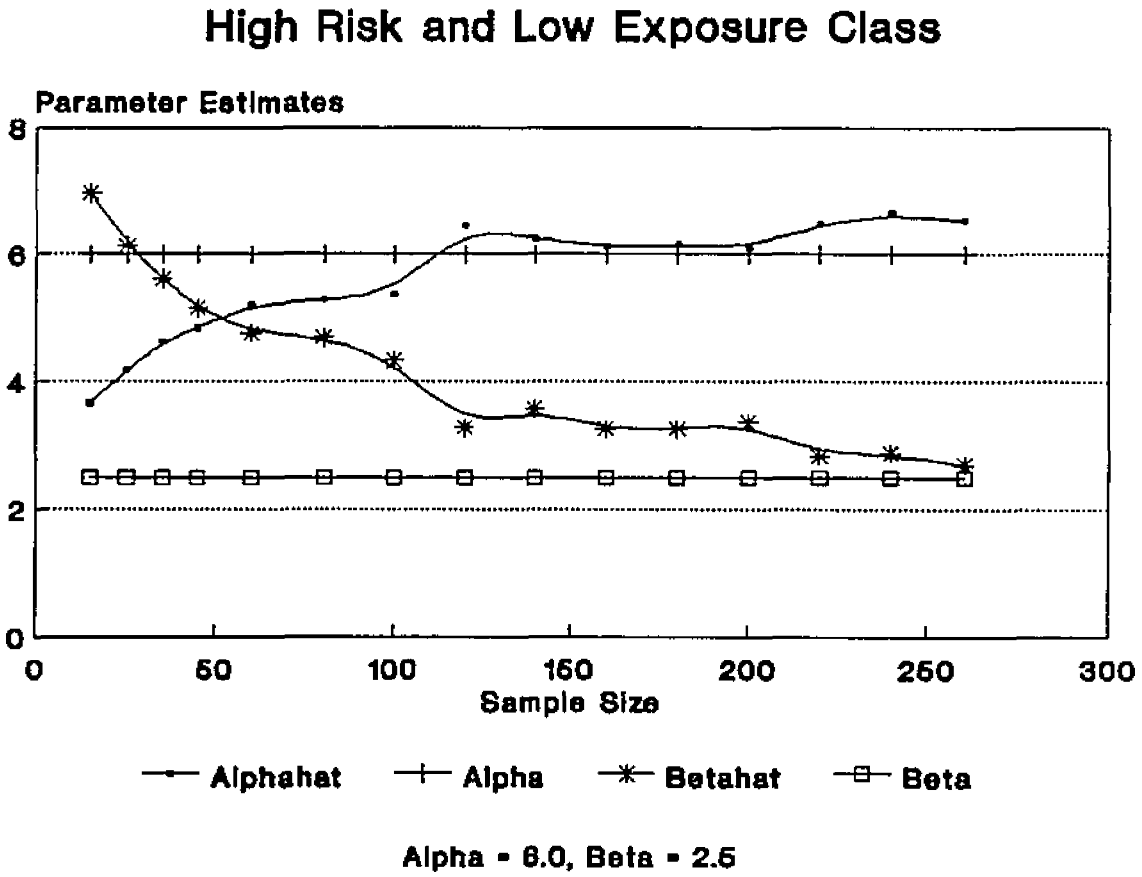


Figure 3.11. Parameter estimates in estimating the accident risk for a high-risk and low-exposure class of roads.

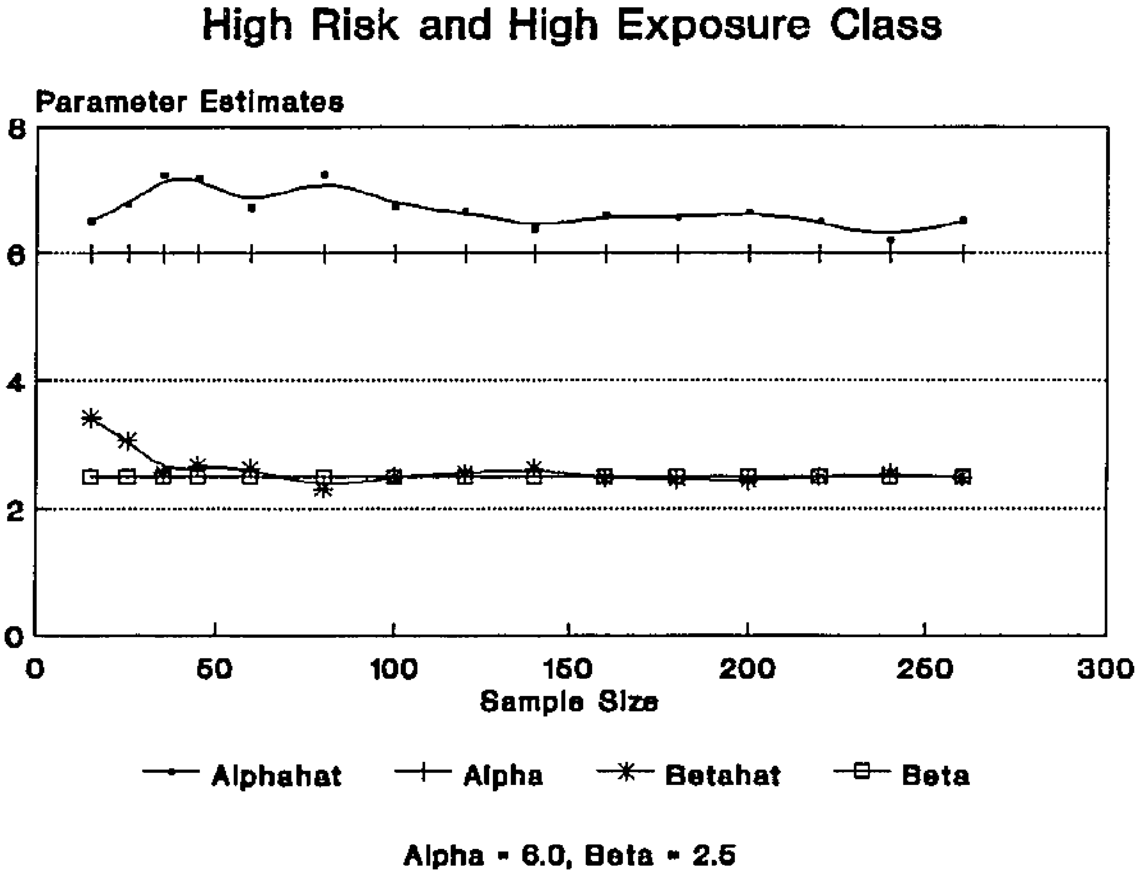


Figure 3.12. Parameter estimates in estimating the accident risk for a high-risk and high-exposure class of roads.

Chapter 4

DEVELOPMENT OF TWO NEW MEDIAN ESTIMATORS

4.1. Two New Median Estimators for a Gamma Distribution

For a traffic event system, it is usually more useful to calculate the absolute errors between the estimated number and the actual number of accidents than to calculate the squared errors between the actual and estimated number of accidents. Also, it is well known that with an absolute error loss function, the Bayes estimator will be the median of the quantity of interest. Thus, a median estimator for the accident rate based on an absolute error loss function should be considered.

Without loss of generality, the median, *med,* of a gamma distribution with the scale parameter $\beta = 1$, can be obtained from the following expression:

$$\frac{1}{2}\int_0^{\infty} \frac{x^{\alpha-1}\mathrm{exp}^{-x}}{\Gamma(\alpha)}dx = \int_0^{med} \frac{x^{\alpha-1}\mathrm{exp}^{-x}}{\Gamma(\alpha)}dx \qquad (4.1)$$

Solving equation 4.1 numerically for several values of parameter α results in the estimated median values shown in table 4.1. When the data in table 4.1 are plotted in figure 4.1, it is clear that a straight line provides a very good fit. Two simple regression models, one without a constant and one with a constant, were

Table 4.1. Median estimates obtained using equation 4.1.

α	*med*	α	*med*
0.5	0.225	5.5	5.17
1	0.695	6	5.668
1.5	1.183	6.5	6.1698
2	1.678	7	6.671
2.5	2.175	7.5	7.182
3	2.674	8	7.67
3.5	3.173	8.5	8.169
4	3.672	9	8.661
4.5	4.171	9.5	9.169
5	4.6709	10	9.6568

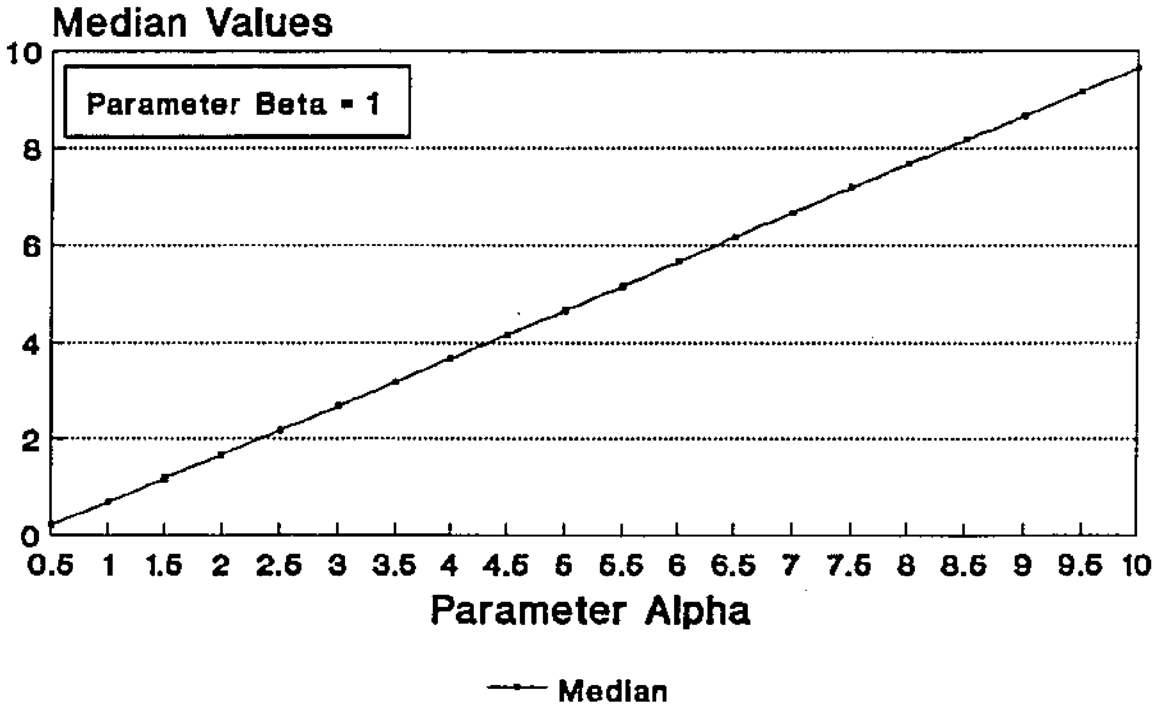


Figure 4.1. The median values versus parameter alpha plot.

developed. The results in equations 4.2 and 4.3 provide two approximations for the relationship between the median values and the shape parameter α.

$$med = 0.9967\alpha - 0.3074, \quad with \;\; R^2 \doteq 1.0; \tag{4.2}$$

$$med = 0.952\alpha, \quad with \; R^2 \doteq 0.99 \tag{4.3}$$

These results provide the basis for considering two new estimators for the accident rate. These estimators are defined by the following equations:

$$\hat{h}_i = \hat{\beta}_p (\hat{\alpha}_p - k_c) \tag{4.4}$$

$$\hat{h}_i = k_b \hat{\beta}_p \hat{\alpha}_p \tag{4.5}$$

where

$\hat{\beta}_p$ = posterior scale parameter $[\hat{\beta} / (1.0 + \hat{\beta}(M_i))]$
$\hat{\alpha}_p$ = posterior shape parameter $(Y_i + \hat{\alpha})$

The k_c and k_b are two constants used for defining the two new estimators. The values selected for k_c and k_b were based on a computer simulation. A sample size of 100 and varying vehicle exposures represented by Weibull distributions were used in the evaluation of possible values for these constants. Summarized results are shown in tables 4.2 and 4.3. The optimal values of k_c and k_b represent the k values at a minimum sum of absolute error or sum of square error between the estimated accident rate and the actual accident rate, respectively.

Table 4.2. The optimal values of k_c and k_b for sample size = 100 and high vehicle exposure level.

Param.	k_c Values				k_b Values			
	EBM_{abs}	EBM_{sse}	$Bayes_{abs}$	$Bayes_{sse}$	EBM_{abs}	EBM_{sse}	$Bayes_{abs}$	$Bayes_{sse}$
$\alpha = 2$ $\beta = 1$	0.29	0.04	0.35	0.02	0.87	0.96	0.88	0.99
$\alpha = 2$ $\beta = 5$	0.28	0.0	0.30	0.0	0.95	0.99	0.95	1.0
$\alpha = 2$ $\beta = 10$	0.34	0.0	0.34	0.0	0.97	1.0	0.97	1.0
$\alpha = 5$ $\beta = 1$	0.23	0.04	0.30	0.0	0.94	0.97	0.95	1.0
$\alpha = 7$ $\beta = 1$	0.18	0.0	0.32	0.0	0.97	1.0	0.97	1.0
$\alpha = 10$ $\beta = 1$	0.21	0.08	0.32	0.01	0.96	0.98	0.98	1.0

Notes: Param. = Parameter.

EBM_{abs} = The condition of using the empirical Bayes procedure to calculate the sum of absolute error.

EBM_{sse} = The condition of using the empirical Bayes procedure to calculate the sum of square error.

$Bayes_{abs}$ = The condition of using the ideal Bayes procedure to calculate the sum of absolute error.

$Bayes_{sse}$ = The condition of using the ideal Bayes procedure to calculate the sum of square error.

Table 4.3. The optimal values of k_c and k_b for sample size =100 and low vehicle exposure level.

Param.	k_c Values				k_b Values			
	EBM_{abs}	EBM_{sse}	$Bayes_{abs}$	$Bayes_{sse}$	EBM_{abs}	EBM_{sse}	$Bayes_{abs}$	$Bayes_{sse}$
$\alpha = 2$ $\beta = 1$	0.18	0.06	0.30	0.01	0.81	0.90	0.86	0.99
$\alpha = 2$ $\beta = 5$	0.30	0.03	0.34	0.02	0.88	0.98	0.88	0.99
$\alpha = 2$ $\beta = 10$	0.34	0.06	0.33	0.04	0.92	0.99	0.92	1.0
$\alpha = 5$ $\beta = 1$	0.10	0.04	0.32	0.0	0.93	0.93	0.94	1.0
$\alpha = 7$ $\beta = 1$	0.07	0.05	0.29	0.0	0.92	0.93	0.97	1.0
$\alpha = 10$ $\beta = 1$	0.02	0.03	0.30	0.0	0.94	0.94	0.97	1.0

Notes: Param. = Parameter.

EBM_{abs} = The condition of using the empirical Bayes procedure to calculate the sum of absolute error.

EBM_{sse} = The condition of using the empirical Bayes procedure to calculate the sum of square error.

$Bayes_{abs}$ = The condition of using the ideal Bayes procedure to calculate the sum of absolute error.

$Bayes_{sse}$ = The condition of using the ideal Bayes procedure to calculate the sum of square error.

It is expected that the k_c value will be 0.0 and k_b value will be 1.0 for square error loss function. This can be verified easily from the columns of EBM_{se} and Bayes_{se} in tables 4.2 and 4.3. This implies that the simulation is on the right track. With the absolute error loss function, it is noted that from equations 4.2 and 4.3 ideal values would be around 0.3 and 0.95 for k_c and k_b, respectively, under the fixed β condition. As expected, the ideal Bayes procedure does possess this feature as shown in the tables 4.2 and 4.3.

For the empirical Bayes procedure, the k_c values decrease as parameter α increases for the fixed β situation, whereas the k_b values increase as parameter α increases. This implies that the estimate of accident rate h using the square error loss function or the absolute error loss function varies little when the parameter α is large. Another observed phenomenon is that the k_c value is sensitive to the variation of vehicle exposure while the k_b values are relatively insensitive. The average values of k_b and k_c for the small and large vehicle exposures are shown in table 4.4.

The small vehicle exposure level, M_i~weib(2.0,0.1,0.01), which indicates that less information is available for the accident rate estimation process, is generally the most difficult situation for obtaining good estimates. Based on the above simulation, the second median estimator, using the k_b value, seems to be more promising than the first median estimator using the k_c.

The final chosen values for k_c and k_b are 0.21 and 0.92, respectively.

Table 4.4. Average values of k_b and k_c.

Vehicle Exposure	k_c Values		k_b Values	
	EBM_{abs}	$Bayes_{abs}$	EBM_{abs}	$Bayes_{abs}$
$M_i \sim weib(2.0,0.25,0.15)$	0.26	0.32	0.94	0.95
$M_i \sim weib(2.0,0.1,0.01)$	0.17	0.31	0.90	0.923

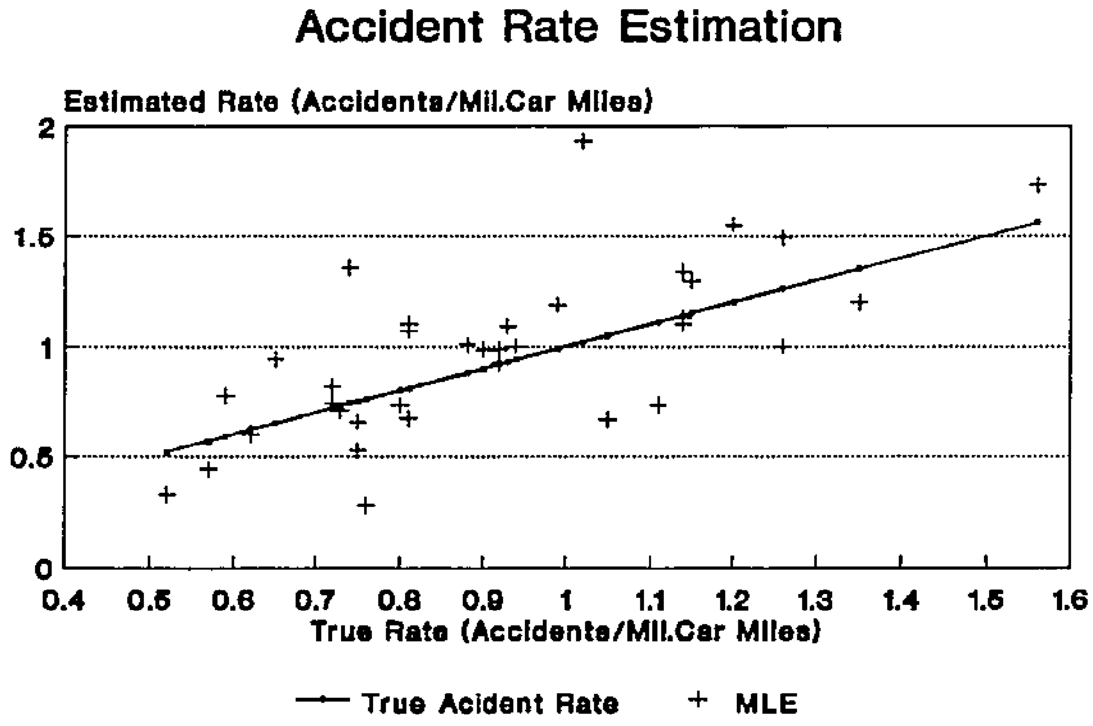
An evaluation of the two median estimators, performed on the same data set reported by Morris et al. (1991), is given in the next section.

4.2. An Evaluation of the Two New Median Estimators

Based on the obtained values of k_c and k_b, an evaluation for the two new estimators was conducted on the simulated data set reported by Morris et al. (1991) and reproduced in table 4.5. This data set consists of simulated values of events (z_i) and exposures (e_i) for 35 sites. Four empirical Bayes procedures were applied to the data set, and the results for these are also presented in table 4.5. They are the Morris hierarchical Bayes (MH), the Arnold and Antle empirical Bayes, and the two new estimators defined by equations 4.4 and 4.5 (L1 and L2). The simple MLE is also given in table 4.5. The results of these methods are shown in figures 4.1 through 4 4 and summarized in table 4.6, where it is seen that the two new estimators have, for this set of data, provided better estimates for the accident risk than the other methods.

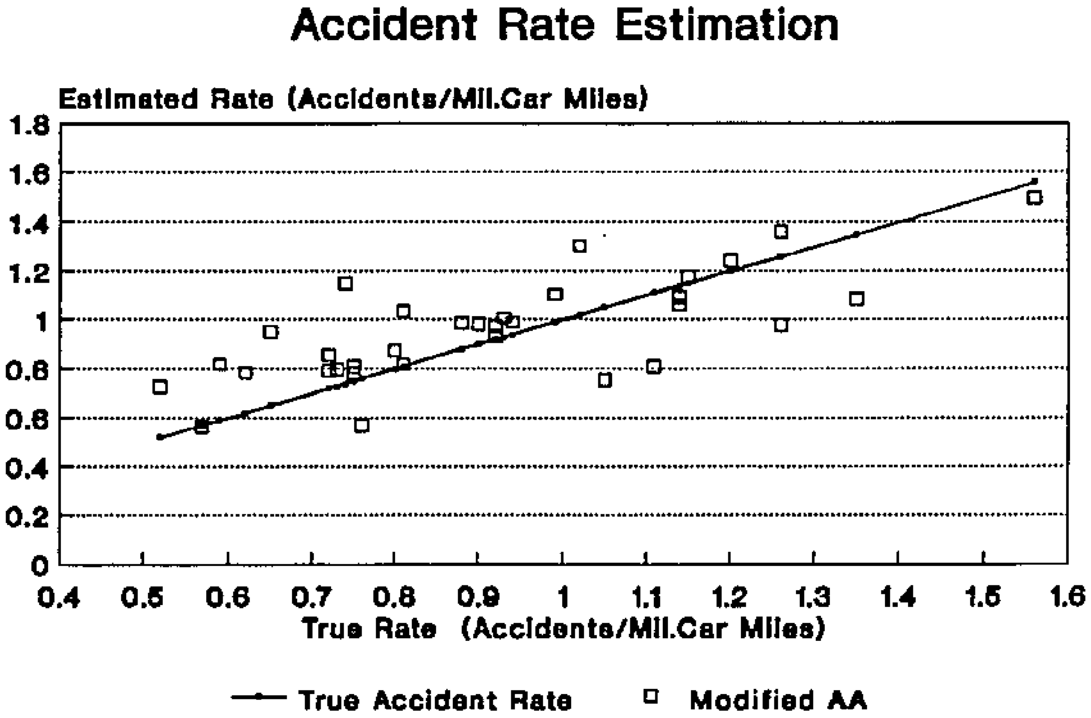
Table 4.5. The simulated data set given by Morris et al. (1991).

Site No.	Event zi	Expo. ei	MLE hi	MH hi	AA hi	L1 hi	L2 hi	True hi
1	6	5.5	1.0909	1.00	1.0058	0.9926	0.9253	0.93
2	3	5.7	0.5263	0.85	0.8071	0.794	0.7425	0.75
3	11	5.7	1.9298	1.22	1.3037	1.2907	1.1994	1.02
4	8	6.0	1.3333	1.07	1.097	1.0842	1.0093	1.14
5	2	6.1	0.3279	0.79	0.7269	0.7142	0.6688	0.52
6	5	6.8	0.7353	0.9	0.8717	0.8595	0.802	0.8
7	7	7.0	1.000	0.98	0.9766	0.9645	0.8984	1.26
8	8	8.1	0.9877	0.97	0.9726	0.9612	0.8947	0.92
9	13	9.6	1.3542	1.11	1.1495	1.139	1.0576	0.74
10	15	9.7	1.5464	1.19	1.2433	1.2328	1.1438	1.2
11	6	10.0	0.600	0.83	0.784	0.7737	0.7213	0.62
12	7	10.4	0.6731	0.85	0.817	0.8069	0.7516	0.81
13	14	11.7	1.1966	1.06	1.0856	1.0761	0.9987	1.35
14	13	11.8	1.1017	1.02	1.0357	1.0262	0.9528	0.81
15	13	12.9	1.0078	0.99	0.9868	0.9778	0.9078	0.88
16	4	14.2	0.2817	0.64	0.5689	0.5604	0.5234	0.76
17	10	15.2	0.6579	0.82	0.781	0.7728	0.7185	0.75
18	15	15.9	0.9434	0.96	0.9503	0.9423	0.8743	0.65
19	21	17.7	1.1864	1.08	1.1029	1.0954	1.0147	0.99
20	13	18.4	0.7065	0.83	0.7984	0.7911	0.7345	0.73
21	25	19.3	1.2953	1.15	1.1781	1.1711	1.0839	1.15
22	21	19.6	1.0714	1.02	1.0331	1.0261	0.9504	0.81
23	15	20.5	0.7317	0.83	0.8089	0.8021	0.7441	1.11
24	41	23.7	1.73	1.42	1.4952	1.4891	1.3756	1.56
25	16	24	0.6667	0.79	0.7556	0.7495	0.6952	1.05
26	27	24.5	1.102	1.05	1.0599	1.0539	0.9751	1.14
27	25	25.3	0.9881	0.98	0.9802	0.9743	0.9018	0.9
28	28	28.0	1.00	0.99	0.9894	0.9839	0.9102	0.94
29	24	29.4	0.8163	0.87	0.8541	0.8488	0.7858	0.72
30	28	30.3	0.9241	0.94	0.9335	0.9283	0.8588	0.92
31	37	32.7	1.1315	1.08	1.0903	1.0854	1.0031	1.14
32	49	32.9	1.4894	1.32	1.3623	1.3575	1.2533	1.26
33	25	33.8	0.7396	0.81	0.7917	0.787	0.7284	0.72
34	15	33.9	0.4425	0.61	0.5642	0.5595	0.5191	0.57
35	28	36.1	0.7756	0.83	0.8171	0.8125	0.7517	0.59



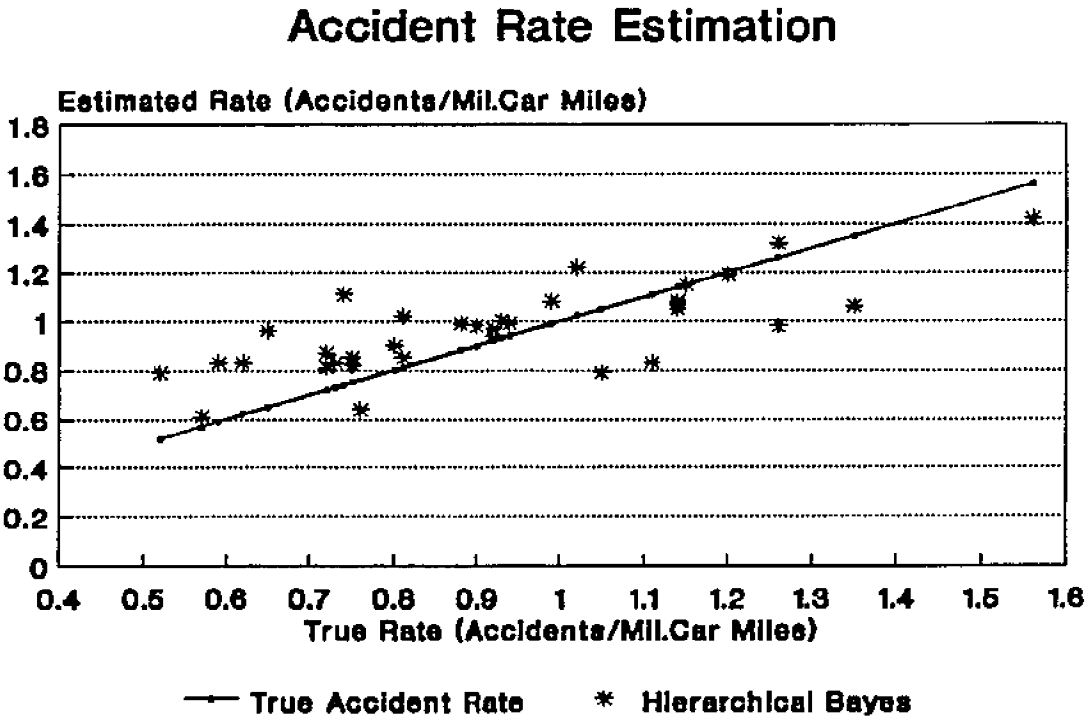
Source: Morris et al. (1991).

Figure 4.2. Accident rate estimation using the maximum likelihood method.



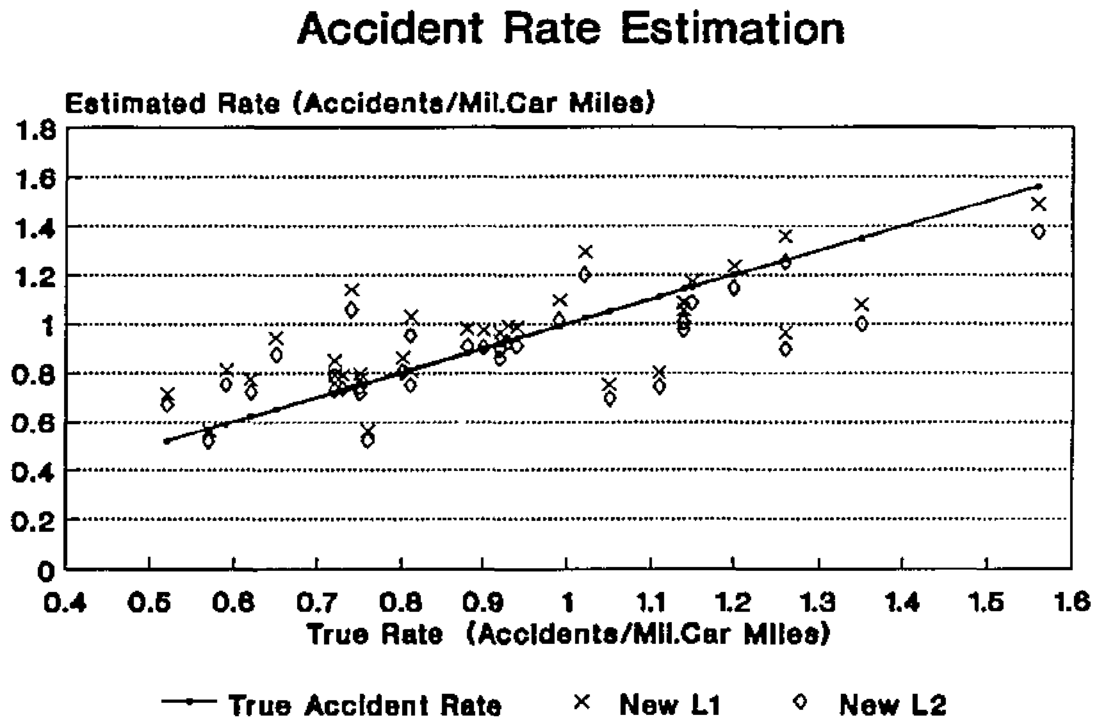
Source: Morris et al. (1991).

Figure 4.3. Accident rate estimation using the modified AA procedure.



Source: Morris et al. (1991).

Figure 4.4. Accident rate estimation using the Morris' hierarchical Bayes method.



Source: Morris et al. (1991).

Figure 4.5. Accident rate estimation using the two new median estimators L1 and L2.

Table 4.6. Total absolute errors for estimating accident risk.

Estimators	Total Absolute Errors
MLE: Maximum Likelihood	7.04
AA: Modified Arnold Antle	4.72
MH: Morris' Hierarchical	4.84
L1: New Estimator $\hat{h} = (\hat{\alpha}_p - 0.21) \hat{\beta}_p$	4.59
L2: New Estimator $\hat{h} = 0.92(\hat{\alpha}_p \hat{\beta}_p)$	4.24

4.3. Summary

This chapter presents an improved empirical Bayes procedure using two new approximate median estimators when considering an absolute error loss function. Monte Carlo simulations were carried out to determine the two constants k_c and k_b of the median estimators. The modified AA procedure and the two median estimators were then evaluated using a simulated data set reported by Morris et al. (1991). Results show that the modified AA procedure and the two new median estimators are very promising.

It should be noted that if, the median estimators and the modified rule presented in chapter 3 are combined, a new rule for the AA procedure can be represented by setting $\hat{\alpha} = 1.5$ if $\hat{\alpha} < 0.3$ and estimating β by the equation $\hat{\beta} = SY / (1.5*SM)$; if $\hat{\alpha}$ is greater than 10, then $\hat{\alpha}$ would be set equal to 10 and, accordingly, β would be estimated by $\hat{\beta} = SY / (10*SM)$. Also, the sample size for estimating the parameters is recommended to be greater than or equal to 100.

Chapter 5

DEVELOPMENT OF A KNOWLEDGE-BASED MODEL

The needs for developing a knowledge-based model that can effectively identify most significant causal factors and assess the accident risk for a traffic event system are threefold. First, the estimation of accident rate through an empirical Bayes procedure using accident data and vehicle exposure cannot clearly show the effect of a suspected causal factor, since all of the factor effects are represented by only one parameter, the mean accident rate (the Poisson mean). An extraction of the factor effects from the mean accident rate is usually difficult and inefficient. Second, the large and good quality data set necessary to produce reasonably accurate estimates of parameters α and β may not be easily obtained. Third, since human factors play an important role in the traffic event system, it is natural to adopt a subjective type of approach to the risk assessment problem.

In this chapter, the procedure for developing a knowledge-based model is discussed. The whole process begins with determining the model structure, collecting and combining expert knowledge, then finalize the model. A strategy based on the belief function theories (Dempster 1967; Shafer 1976) is used to combine expert knowledge. An interpretation of the model is also given. An application of this model to the problem of estimating accident risk and identifying significant causal factors for wet pavement accidents is presented in chapter 6.

5.1. Model Structure

In order to identify the cause and effect relationship between a traffic accident and its suspected causal factors, a diagnostic type of structure was selected. The structure of the model is constructed as a hierarchical event tree. Figures 5.1 and 5.2 depict the model structure. The model has a maximum number of five levels. Its highest level is the top event--a traffic accident. The second level is represented by main events, such as bad driver/vehicle condition and bad roadway section. The third level consists of potential initiating events, such as bad pavement condition, bad geometric condition, and so on. The fourth level (shown in figure 5.2) includes a variety of single events, such as low skid resistance (SN), high surface rutting (RUT), high surface roughness (IPM), high pavement surface age (AGE), horizontal curvature (HC), vertical alignment (VA), driving difficulty (DD), high average daily traffic (ADT), and high percentage of time when a road surface is wet (TW); the lowest level comprises those categorized levels for the factors of the proposition of bad roadway section in level four. It should be noted that the hierarchy of the proposition of bad driver/vehicle condition has only three levels, in recognition of the attainability of driver/vehicle factors in practice. The ratings and definitions of the three quantities HC, VA, and DD are given in Kulakowski et al. (1990b).

The model was realized using a collected expert knowledge base. The collection of expert knowledge is described in the next section.

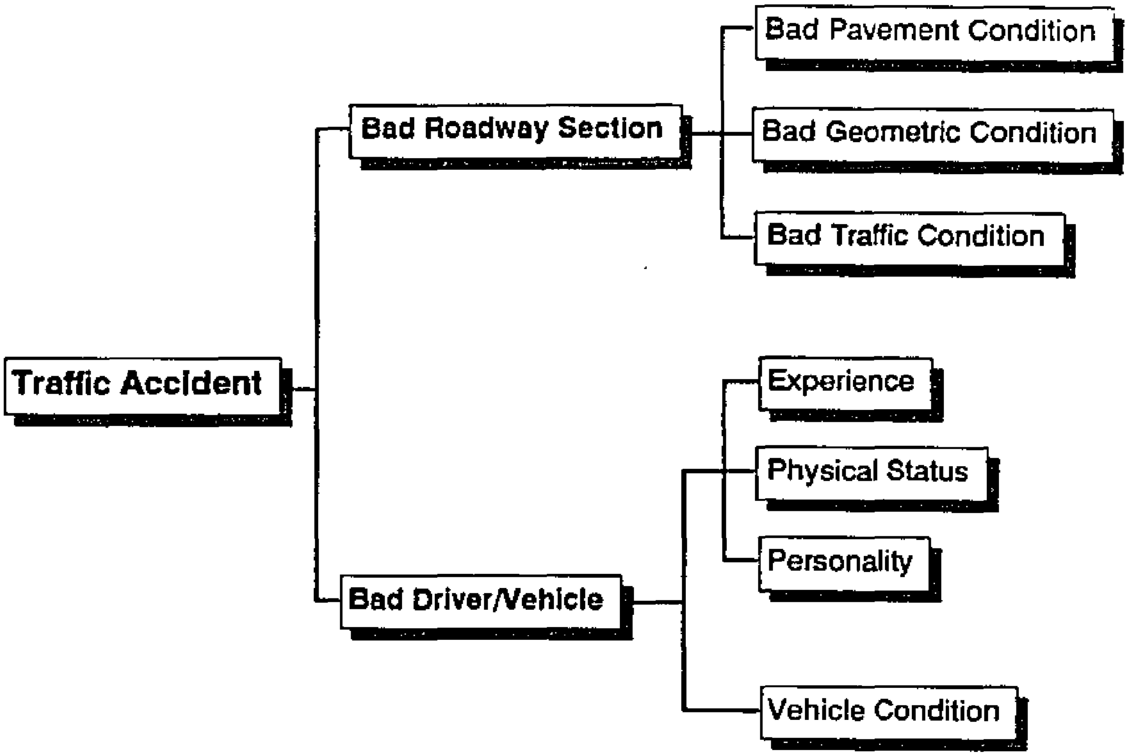


Figure 5.1. The schematic diagram of the knowledge-based model, part I.

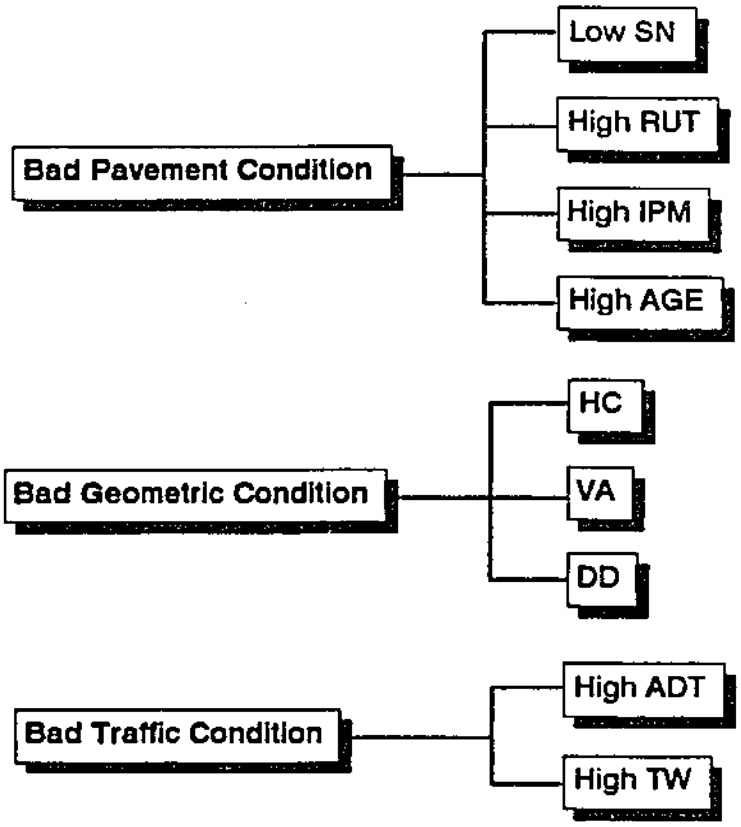


Figure 5.2. The schematic diagram of the knowledge-based model, part II.

5.2. Collection of Expert Knowledge

An expert is generally characterized by expertise in a specific subject or area. The expert's knowledge in this subject or area is usually viewed as a valuable asset. For the problem of identification and risk assessment of traffic accidents, those researchers, highway engineers, and research engineers in the Pennsylvania Department of Transportation and the Highway Research Center of Federal Highway Administration, and researchers in all of the Transportation Institutes across the U.S. or other research institutes relating to transportation research are considered to be experts in this area. A total of 28 experts were chosen to provide estimates of the potential for a possible traffic event.

Based on the model structure depicted in figures 5.1 and 5.2, a questionnaire (shown in appendix A) was designed to collect expert knowledge. The questionnaire contains propositions for possible events in the traffic event system. A scale from 0 to 10 corresponding to the probability scale 0 to 1 was set as a degree of confirmation (or disconfirmation) for each body of evidence with (or without) a potential contributing effect on the proposition. The scale is one of increasing effect; that is, a score of 10 indicates the strongest effect on the proposition. The factors in the fourth level of the model, including SN, RUT, IPM, AGE, HC, VA, DD, ADT, and TW, were categorized into different levels to reduce measurement variations.

In general, the process of collecting expert knowledge is the most difficult stage in constructing a knowledge-based model. It is time-consuming and has a low ratio of response. For this collecting process, a total of 16 copies of the questionnaire were returned over a 5-month period. Based on the collected questionnaires, a body of expert knowledge was assembled. The procedure used for combining the expert knowledge is given in the next section.

5.3. Combination of Expert Knowledge

Three distinct schemes were considered for combining the collected expert knowledge. The first one is the min-max principle used in the fuzzy set theory (Zadeh 1965; Dubois and Prade 1980; Fung and Fu 1975); the second is the subjective Bayesian method for the rule-based system (Duda et al. 1976); the last is Dempster's rule of combination in the belief function theory (Dempster 1967; Shafer 1976).

The min-max principle is based on the concept of union and intersection operators in aggregating two sets. As an example, if A and B are two sets over a universe V, and $a \in A$ and $b \in B$, then the min-max principle states that:

$$a \cap b \triangleq a \wedge b = \min(a, b) \tag{5.1}$$

$$a \cup b \triangleq a \vee b = \max(a, b) \tag{5.2}$$

The difficulty in using this principle to combine the expert knowledge is that the determination of when one should use the maximum principle or the minimum principle for combination is very subjective. Also, the combination results using the min-max principle, in general, are highly approximate and imprecise.

The subjective Bayesian method assumes that the collected bodies of evidence for a proposition (hypothesis) are conditionally independent. If E stands for evidence and H stands for hypothesis, then a likelihood ratio is defined as

$$R = \frac{P(E|H)}{P(E|\bar{H})} \tag{5.3}$$

The likelihood ratio represents prior knowledge of a hypothesis. Experts are supposed to assign a value for this ratio to each hypothesis. In general, it is a difficult task for human experts to assign two values of *conditioned* probability at same time. Furthermore, a modification is necessary when bodies of evidence to a hypothesis are in conflict.

Dempster's rule of combination (1967) calculates the orthogonal sum of the bodies of evidence to a proposition. It considers both the heterogenous and the conflicting conditions of the bodies of evidence. Therefore, the rule was selected to be used in this subsection.

During the process of combining the expert knowledge, it is assumed that there is no difference between experts' perception in assigning a scale of confirmation or disconfirmation to each body of evidence. Also, the assigned scales by the chosen experts are assumed to be normally distributed for each specific proposition. Based on these two assumptions, the combination process proceeded as follows:

Step 1. Calculate the degree of support for each body of evidence. This was carried out by calculating the trim mean of the collected experts' assigned scales on each body of evidence. The trimming procedure excluded the most extreme 5% of the assigned scales of confirmation or disconfirmation. It provides a more representative scale for each body of evidence. Based on the collected questionnaires, the trim mean for each body of evidence was calculated and proportionally converted into a probability scale value between 0 and 1. Tables 5.1 through 5.4 present the calculated values.

Step 2. Identify frames of discernment for each level. For example, in the second level, the Θ = {Bad Roadway Section, Bad Driver/Vehicle}; the Θ = {Bad Pavement Condition (PC), Bad Geometric Condition (GC), Bad Traffic Condition (TC)} for the proposition of bad roadway section in the third level; the Θ = {Low SN, High RUT, High IPM, High AGE} for the proposition of bad pavement condition in the fourth level. By the same rule, the other frames of discernment can be identified. After

Table 5.1. Calculated degrees of support and basic probability numbers for the third level propositions.

			Confirm.	Disconfir.	m(A)
Bad Roadway Section	Bad PC		0.429	0.007	0.427
	Bad GC		0.657	0.014	0.654
	Bad TC		0.657	0.007	0.655
Bad Driver/ Vehicle Condition	Driver's Experience	Good	0.114	0.171	0.096
		Fair	0.314	0.057	0.301
		Little	0.564	0	0.564
	Driver's Personality	Normal	0.221	0.114	0.201
		Nervous	0.393	0.057	0.379
		Aggressive	0.643	0	0.643
	Driver's Physical Status	Tired	0.686	0	0.686
		Drug/Alcohol Influenced	0.879	0	0.879
		Alert	0.107	0.164	0.091
	Vehicle Condition	good	0.129	0.121	0.115
		fair	0.257	0.043	0.249
		poor	0.443	0.036	0.434

Notes: PC = pavement condition.
GC = geometric condition.
TC = traffic condition.
Confirm. = confirmation.
Disconfir. = disconfirmation.
m(A) = basic probability number--a measure of belief that committed exactly to each body of evidence.

Table 5.2. Calculated degrees of support and basic probability numbers for the factors in the proposition of bad pavement condition.

			Confirm.	Disconfir.	m(A)
Bad Pavement Condition	Low SN		0.728	0.007	0.727
	High RUT		0.579	0.007	0.577
	High IPM		0.5	0.021	0.495
	High AGE		0.314	0.036	0.306
	SN	<20	0.979	0	0.979
		20-25	0.871	0	0.871
		25-30	0.678	0	0.678
		30-35	0.407	0.007	0.405
		35-40	0.236	0.07	0.223
		>40	0.093	0.007	0.092
	RUT	>1.0"	0.721	0	0.721
		0.5-1.0"	0.379	0.007	0.377
		<0.5"	0.143	0.057	0.136
	IPM	>300	0.657	0	0.657
		250-300	0.579	0	0.579
		200-250	0.443	0	0.443
		150-200	0.336	0	0.336
		100-150	0.221	0	0.221
		<100	0.093	0	0.093
	AGE	>15	0.629	0	0.629
		10-15	0.393	0	0.393
		5-10	0.214	0.014	0.212
		2-5	0.093	0.114	0.083
		<2	0.043	0.207	0.034

Notes: Confirm. =	confirmation.
Disconfir. =	disconfirmation.

Table 5.3. Calculated degrees of support and basic probability numbers for the factors in the proposition of bad geometric condition.

			Confirm.	Disconfir.	m(A)
Bad Geometric Condition	Bad HC		0.333	0	0.333
	Bad VA		0.333	0	0.333
	Bad DD		0.333	0	0.333
	HC (Horizontal Curvature)	Slight	0.136	0.136	0.12
		Moderate	0.379	0.043	0.369
		Severe	0.693	0	0.693
	VA (Vertical Alignment)	Slight	0.1	0.129	0.088
		Moderate	0.3	0.021	0.296
		Severe	0.6	0	0.600
	DD (Driving Difficulty)	Slight	0.15	0.093	0.138
		Moderate	0.386	0	0.386
		Severe	0.714	0	0.714

Notes: Confirm. = confirmation.
Disconfir. = disconfirmation.

Table 5.4. Calculated degrees of support and basic probability numbers for the factors in the proposition of bad traffic condition.

			Confirm.	Disconfir.	m(A)
Bad Traffic Condition	Low ADT		0.571	0.029	0.564
	High TW		0.657	0.014	0.654
	ADT	>15,000	0.564	0	0.564
		10,000-15,000	0.407	0	0.407
		6,000-10,000	0.279	0	0.279
		3,000-6,000	0.171	0.05	0.164
		1,000-3,000	0.064	0.129	0.056
		<1,000	0.021	0.221	0.016
	TW	>20%	0.634	0	0.634
		15%-20%	0.5	0	0.5
		10%-15%	0.336	0	0.336
		5%-10%	0.221	0.043	0.214
		<5%	0.093	0.093	0.085

Notes: Confirm. = confirmation.
Disconfir. = disconfirmation.
ADT = average daily traffic (vehicles).
TW = percentage of time when road surface is wet(%).

identifying the frames of discernment, calculate the basic probability number m(A) for each body of evidence using the Dempster's rule of combination to combine the values of confirmation and disconfirmation obtained from step 1. The calculated basic probability numbers are shown in the last columns of tables 5.1 through 5.4. It should be noted that the calculated basic probability numbers of the factors in the lowest level represent the degree of support of these factors. They will be used directly in finalizing the model. The basic probability numbers for those factors in the third and fourth levels were then used to calculate their belief intervals.

Step 3. Calculate the belief intervals for the propositions in the third level, which include bad pavement condition, bad traffic condition, bad geometric condition and bad driver/vehicle condition. Tables 5.5 and 5.6 show the belief intervals for the propositions.

Step 4. Calculate the belief function numbers for the fourth level factors-- low SN, high RUT, high IPM, high AGE, HC, VA, DD, high ADT, and high TW. Results are shown in tables 5.7 through 5.9.

Step 5. Propagate the belief function numbers and the degrees of support from the lowest level to the top level to finalize the model.

Table 5.5. The belief intervals of the factors in the proposition of bad roadway section.

Bad Roadway Section	m(A)	Bel-1	Bel-2
Bad PC	0.427	0.135	0.314
Bad GC	0.654	0.341	0.519
Bad TC	0.655	0.344	0.523

Note: Bel-1 = lower probability. Bel-2 = upper probability.

Table 5.6. The belief intervals for the factors in the proposition of bad driver/vehicle condition.

Bad Driver/Vehicle Condition		m(A)	Bel-1	Bel-2
Driver's Experience	Good	0.096	0.007	0.058
	Fair	0.301	0.027	0.084
	Little	0.564	0.079	0.141
Driver's Personality	Normal	0.201	0.015	0.070
	Nervous	0.379	0.038	0.095
	Aggressive	0.643	0.111	0.172
Driver's Physical Status	Tired	0.686	0.134	0.196
	Drug/Alcohol influenced	0.879	0.446	0.508
	Alert	0.091	0.006	0.058
Vehicle Condition	Good	0.115	0.008	0.062
	Fair	0.249	0.020	0.079
	Poor	0.434	0.047	0.106

Table 5.7. Calculated belief function numbers for the factors in the proposition of bad pavement condition.

Bad Pavement Condition	m(A)	Bel(A)
Low SN	0.727	0.412
Hig RUT	0.577	0.212
High IPM	0.495	0.152
High AGE	0.306	0.068

Table 5.8. Calculated belief function numbers for the factors in the proposition of bad geometric condition.

Bad Geometric Condition	m(A)	Bel(A)
HC	0.333	0.198
VA	0.333	0.203
DD	0.333	0.199

Table 5.9. Calculated belief function numbers for the factors in the proposition of bad traffic condition.

Bad Traffic Condition	m(A)	Bel(A)
High ADT	0.564	0.312
High TW	0.654	0.459

5.4. An Interpretation of the Model

Essentially, the model hypothesizes that the occurrence of a traffic accident is due to a bad roadway section and/or bad driver/vehicle conditions. This is intuitively correct, since the traffic event system is commonly viewed as a human-vehicle-roadway system. The results based on the constructed expert knowledge base reveal the relative weight (proportion) of each proposition to the top event, a traffic accident. For the roadway section part, the belief intervals for the three propositions are:

- Bad pavement condition (PC): [0.135, 0.314]
- Bad geometric condition (GC): [0.341, 0.519]
- Bad traffic condition (TC): [0.344, 0.523]

As an illustration, the belief interval for the bad pavement condition is interpreted as the total belief, based on the human experts' judgement, that a bad pavement condition will contribute to a bad roadway section to a degree of 0.135 to 0.314 (on a scale of 0 to 1). The upper probability 0.314 represents the total belief of 1 -Bel({Bad GC, Bad TC}). It should be noted that there is a large variation between the lower and upper probability values. This indicates that there exists a large difference of recognition among human experts in considering this proposition. It is also noted that the sum of the three lower degrees of belief is less than 1. This is a feature of Shafer's theory. It reveals that for each frame of discernment there always exists an unassigned degree of belief.

The belief intervals for the factors in the proposition of bad driver/vehicle condition indicate that there is a high degree of belief interval, [0.446, 0.508], for a drug/alcohol-influenced driver. Likewise, a tired or aggressive driver is dangerous, too. An inexperienced driver is another possible cause of a traffic accident. In essence, these results coincide with this study's engineering judgements.

When considering suspected causal factors for the proposition of bad pavement condition, the low SN possesses a high degree of belief with a value of 0.412 and thus is considered as most significant factor. For the factors in the proposition of bad geometric condition, there is not much difference in the degree of belief between one factor and the next factor in the proposition. The factor of high TW has a higher degree of belief than the factor of high ADT in the proposition of bad traffic condition.

In order to give a clearer picture of this interpretation, several figures were plotted and are shown in appendix B. Essentially, they represent the results that displayed in tables 5.1 through 5.9.

5.5. Summary

In this chapter, a knowledge-based model for the accident identification and risk assessment was developed. The formulation of this model is based on a collected expert knowledge base and the belief function theory (Dempster 1967; Shafer 1976) introduced in the chapter 2. A questionnaire as shown in appendix

A was designed to collect the expert knowledge. The application of the belief function theory to the problem is considered to be a pioneering step.

The developed model is characterized by several features which can be summarized as follows:

- The model is flexible. It can be used under the conditions of data measurements being available or not available. It can provide a degree of belief for a specific factor or a combined belief for the top event--a traffic accident.
- The model can be used to identify significant causal factors for each proposition.
- The model can be updated when additional expert knowledge is available. The model is more mathematically rigorous than heuristic because Dempster's rule of combination and belief function theory provide a mathematical foundation for the combination and representation of expert knowledge.
- The model reveals unacceptable values of suspected causal factors such as SN, RUT, ADT, and so on, using calculated degree of belief.
- An important drawback of this model is that the construction of the model is based on human judgement and therefore is subjected to individual uncertainties.

Chapter 6

A CASE STUDY - DEVELOPMENT OF A WET PAVEMENT INDEX

6.1. Introduction

A wet pavement accident is a type of traffic accident that occurs on a wet pavement surface. The high accident risk of wet pavement accidents has been confirmed by Campbell (1971) and Brodsky and Hakkert (1988). Essentially, the reduced roadway surface traction and the restricted visibility due to rainy or snowy weather are the two main causes of increased risk of traffic accidents on wet roads. A study by Kulakowski and Harwood (1990) showed that roadway skid resistance can be reduced by 20 to 30% when the water film on the road surface is 0.05 mm. The objective of this case study is to develop a wet pavement index (WPI) using traffic and roadway characteristics to identify those segments of highway having a high potential for wet pavement accidents.

In order to estimate the accident risk of wet roads, the vehicle exposure M_i defined in chapter 2 was modified to account for the factor of percentage of time when a road surface is wet. WM_i is the new notation representing the wet vehicle exposure and is calculated as:

$$WM_i = SL_i * ADT_i * DAYS_i * TW_i \tag{6.1}$$

where:

SL = section length (miles)
ADT = average daily traffic (number of vehicles/day)
$DAYS$ = time duration (days)
TW = percentage of time when road surface is wet

It may be noted that when considering the wet pavement accident problem for the State of Pennsylvania, the value of WM_i is often in the range of 0.02 to 2.5.

To develop the wet pavement index, several methodologies, including classical regression methods, a direct Bayesian regression method (proposed in the author's thesis proposal), a hierarchical accident index method (Kulakowski et al. 1990), the improved empirical Bayes procedure developed in the previous chapters, and the knowledge-based model presented in chapter 5, were applied to estimate the risk of wet pavement accidents. An evaluation of these methodologies was performed using actual accident and roadway characteristics data collected from 308 road sections in Pennsylvania.

6.2. Preliminary Data Analysis

The Pennsylvania Department of Transportation has provided a data base for this study that consists of the accident records and other necessary information for 308 highway sections in Pennsylvania. These records were for the years 1983-88 and are given in a report by Kulakowski et al. (1990b). These sites had no substantial improvements during the 1983-88 period. It was decided that the data for 1983-85 would be used in evaluating the risk for wet pavement accidents and the results would then be used to predict the number of wet pavement accidents for the years 1986-88.

A preliminary data analysis of the real data set was carried out before evaluating those previously described methodologies. Figures 6.1 and 6.2 show the histograms of the wet pavement accidents for the periods 1983-1985 and 1986-1988, respectively. They essentially verify that the assumption of a Poisson random variable for the occurrence of traffic accidents is suitable when figures 6.1 and 6.2 are compared with figure 6.3, generated Poisson distributions. The relationship between wet pavement accidents and skid resistance is plotted in figures 6.4 and 6.5. It is obvious from figures 6.4 and 6.5 that no simple relationship can be assumed and that using the skid resistance alone as an index of accident potential for a roadway section is inadequate.

6.3. An Evaluation of Methodologies

In this section, a number of methodologies including the classical regression methods, the direct Bayesian regression method, the hierarchical accident index method, the modified empirical Bayes procedures, and the knowledge-based model were evaluated using the available real data set provided by PennDOT. A comparison of these methods is presented in the next section.

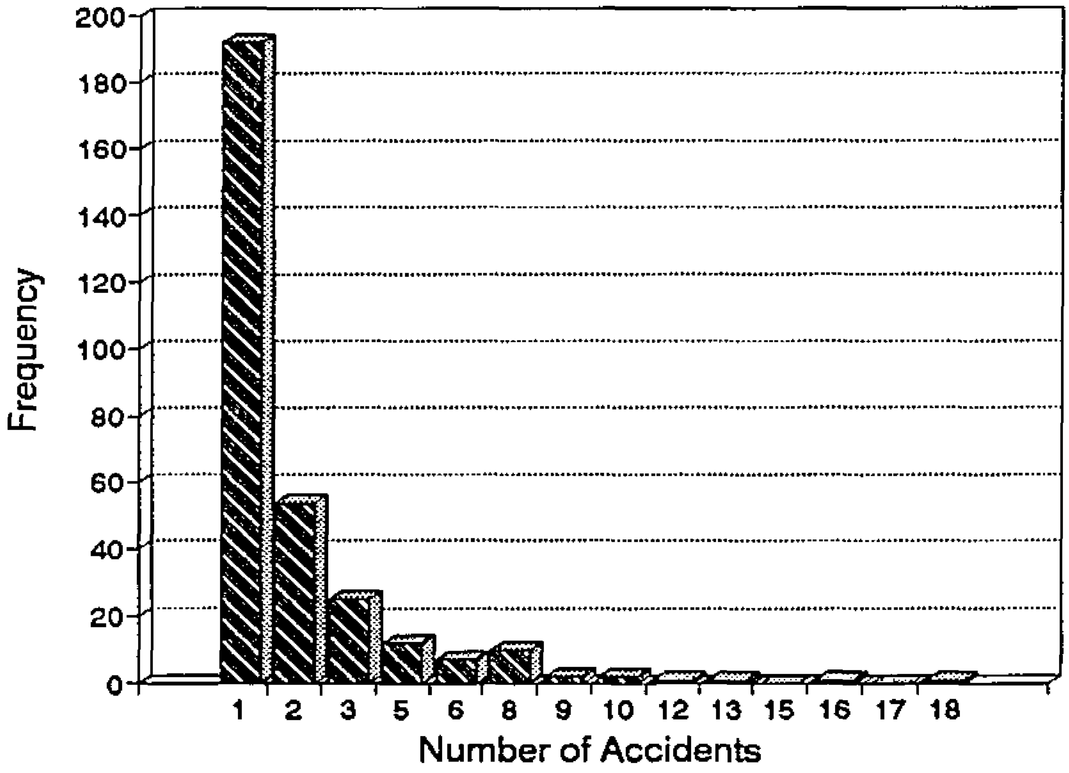


Figure 6.1. Histogram of wet pavement accidents for the state of Pennsylvania in 1983-1985.

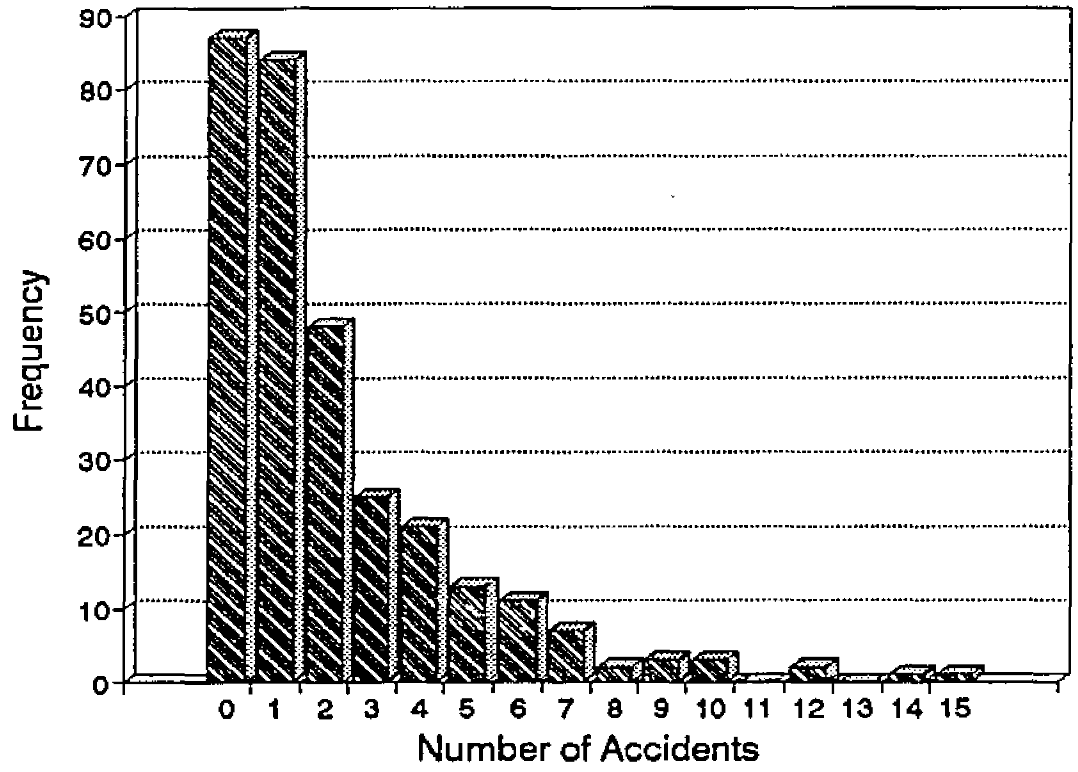


Figure 6.2. Histogram of wet pavement accidents for the state of Pennsylvania in 1986-1988.

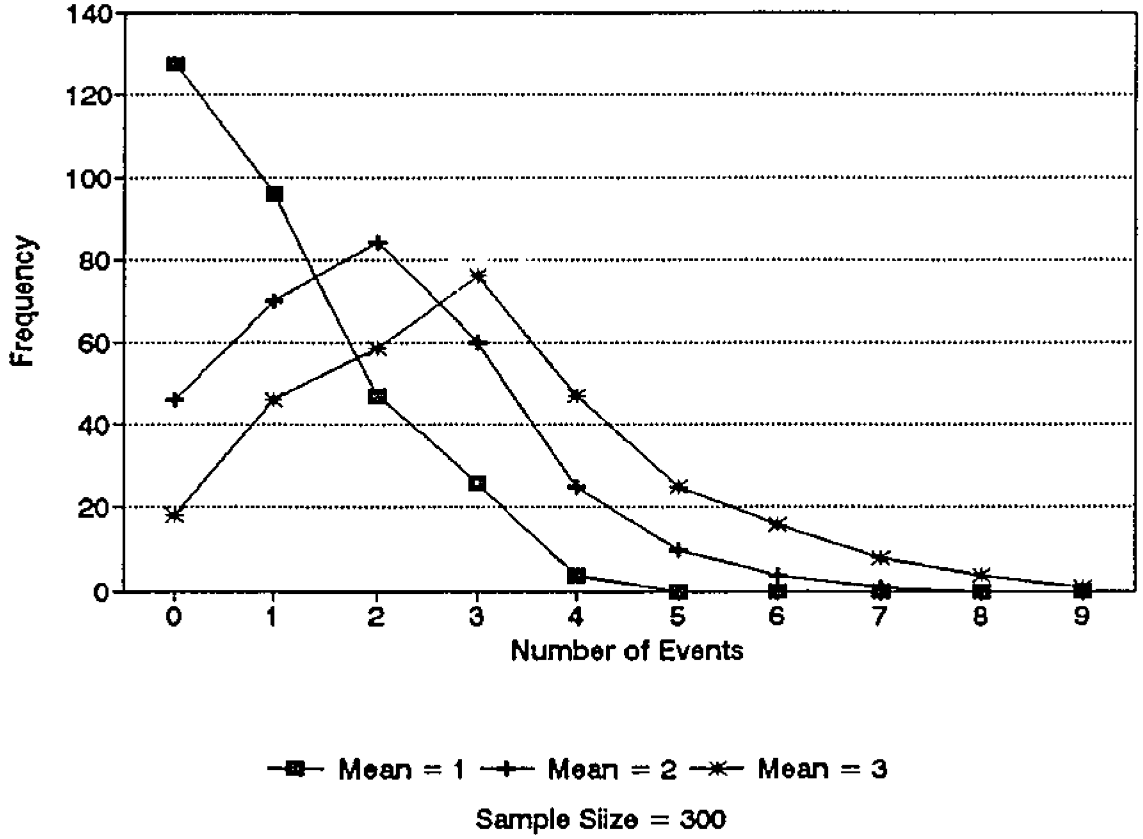


Figure 6.3. Generated Poisson distributions.

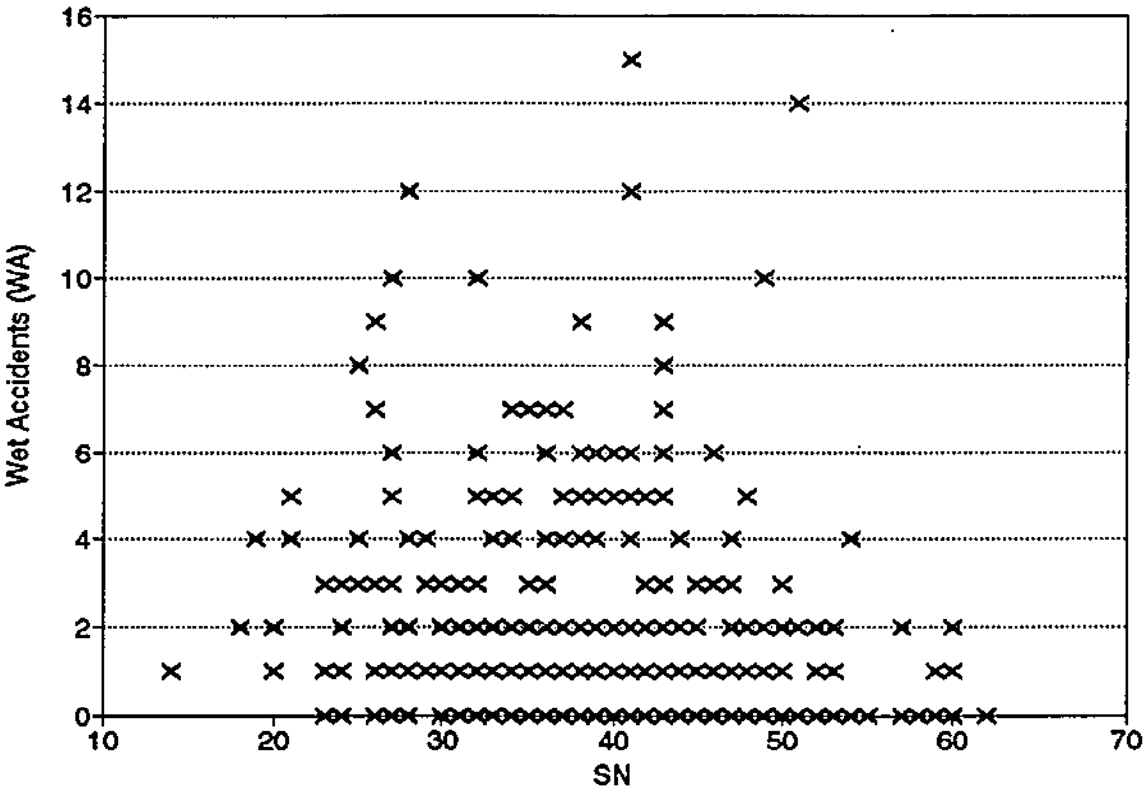


Figure 6.4. Plot of wet pavement accidents versus skid resistance in Pennsylvania in 1983-1985.

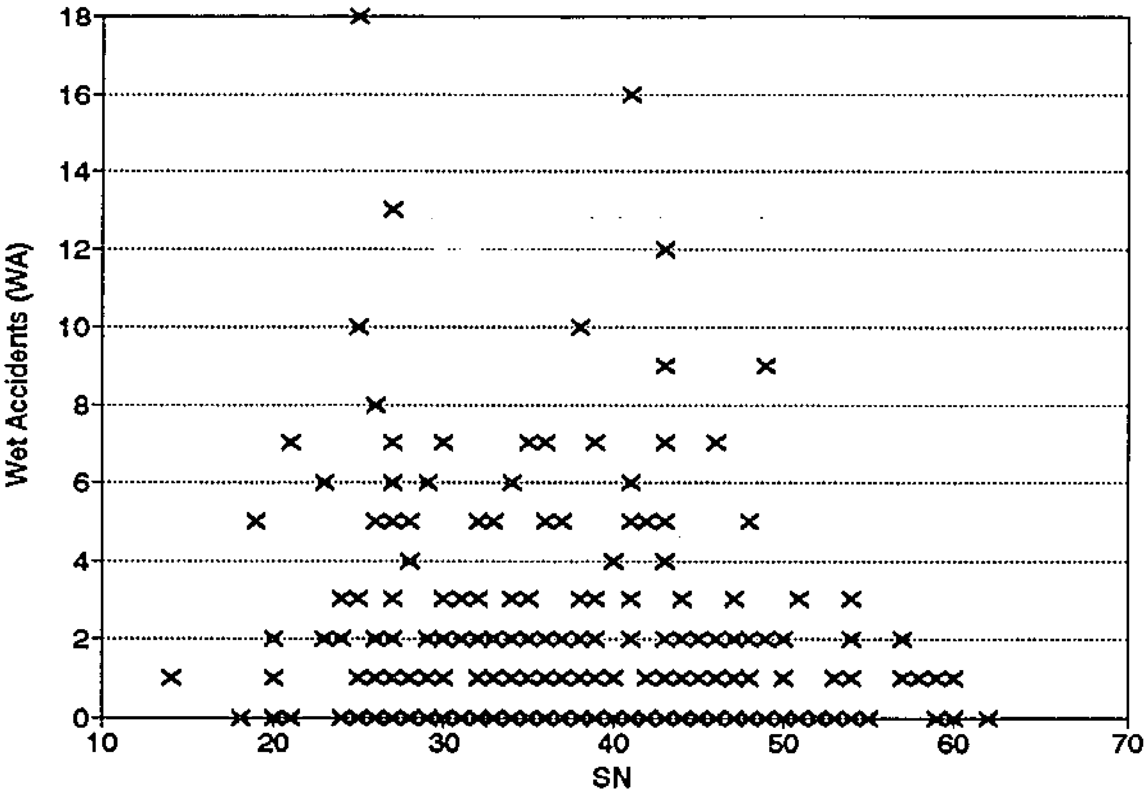


Figure 6.5. Plot of wet pavement accidents versus skid resistance in Pennsylvania in 1986-1988.

6.3.1. Classical Regression Methods

Regression analysis is a frequently used approach to construct a cause-and-effect relationship between attributing factors and traffic accidents. A generic form of the regression model can be represented by:

$$y_i = f(\underline{x}, \underline{\theta}) + \epsilon_i \qquad i=1,2,3,\cdots,N \tag{6.2}$$

where

y_i = observations (i.e., traffic accidents)

$\underline{x}$ = vector of attributing factors

$\underline{\theta}$ = vector of parameters

ϵ_i = random disturbances (errors)

A general assumption behind this model is that the random disturbances ϵ_i are uncorrelated with each other and are normally distributed with zero mean and constant variance. Also, the measurements of the attributing factors are assumed to be free from errors. Based on these assumptions, different types of the model such as linear additive type, nonlinear additive type, nonlinear multiplicative type, and so on, can be assumed.

The linear regression method using the least square technique is the simplest method for performing parameter estimation. It gives a general picture of the input-output relationship of a system. However, in most situations, the result may be misleading due to the existence of nonlinearity in the system. To alleviate this problem, nonlinear regression methods are applied.

Three of the most commonly-used types, including linear additive type, nonlinear additive type, and nonlinear multiplicative type, of regression model were selected and applied to the data set provided by PennDOT (Kulakowski et al. 1990b) to observe the effectiveness of regression methods in parameter estimation. The attributing factors were filtered first through the step-wise regression technique to choose the factors that were statistically and practically significant. Regression analyses were then performed on the three models.

The first model is a linear additive model and is expressed as:

$$E(Y) = b_0 + b_1(WM) + b_2(DD) + b_3(PS) + b_4(SN) + b_5(TP) \tag{6.3}$$

The second model is a nonlinear additive model and is represented by:

$$E(Y) = b_0 + b_1 WM + b_2 PS + b_3 SN + b_4 TP + b_5(PS)(SN) + b_6(SN)(DD) + b_7(SN)(TP) \tag{6.4}$$

The third model is a nonlinear multiplicative model:

$$E(Y) = b_0 \frac{(WM)^{b_1}(DD)^{b_2}(PS)^{b_3}}{(SN)^{b_4}} \tag{6.5}$$

where b_i are parameters to be estimated and WM is the value of wet vehicle exposure expressed in terms of millions of vehicle miles.

The parameters of the first and second models were estimated by using MINITAB (Ryan et al. 1989) software, whereas the SAS NLIN procedure based on the Gauss-Newton and Marquardt methods was applied to estimate the parameters of the third model. The regression analyses of these three models are shown in appendix B. Figures 6.6 and 6.7 show the prediction results of future (1986-88) wet pavement accidents on the same road sections for the linear and the nonlinear additive models. The coefficients of multiple correlation of these two models are around 0.51 (R^2=26%). A major problem for these two models is that they produce undesired *negative estimates* of future wet accidents. The prediction results of the third regression model (equation 6.5) are shown in figure 6.8. The coefficient of multiple correlation is 0.50 (R^2 = 25.2%), which may not be better than the other two regression models; however, the problem of producing negative estimates in the other two models is eliminated. A common problem in the use of regression methods is sample size. A small data sample may not produce good parameter estimates.

6.3.2. A Direct Bayesian Regression Method

The direct Bayesian regression method proposed in the thesis proposal assumes that for a location, there exists a cause-and-effect relationship between the expected (average) number of accidents and the attributing factors and the occurrence of the accidents is a Poisson random process. Since these attributing factors are likely to be interactive, a multiplicative model is assumed. The proposed approach uses a different technique from the classical regression

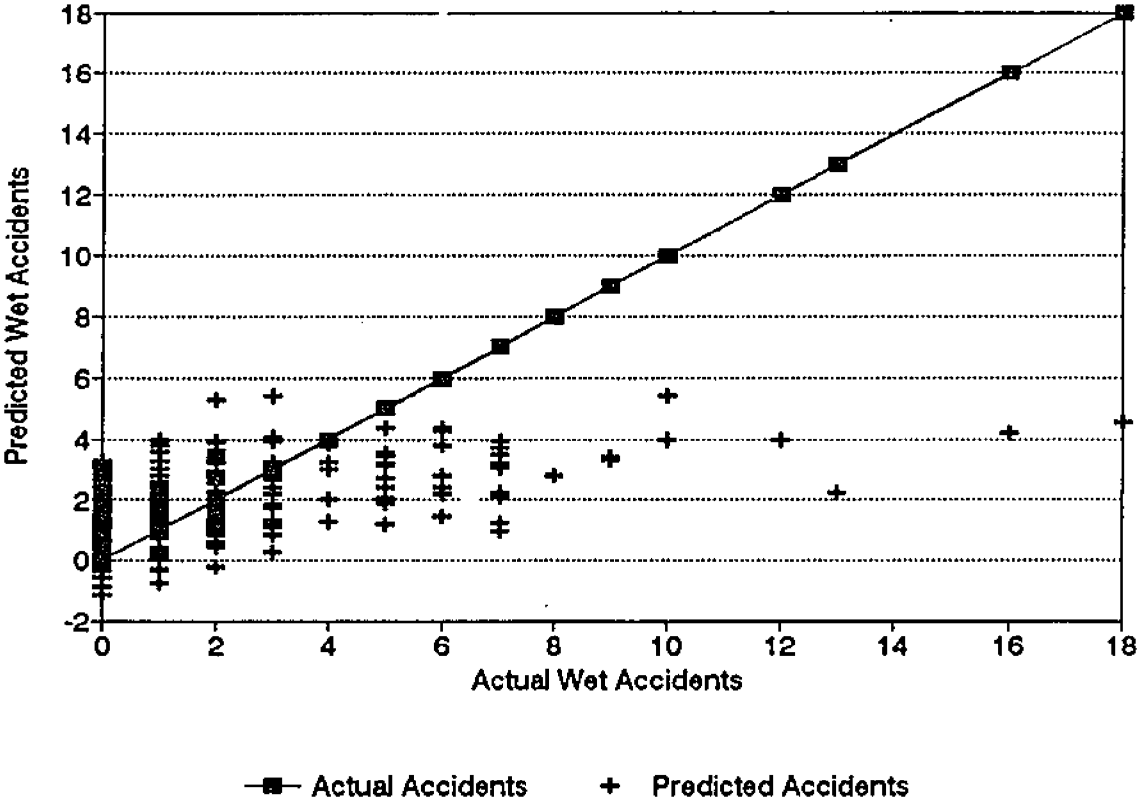


Figure 6.6. Results of the linear regression model (1986-1988).

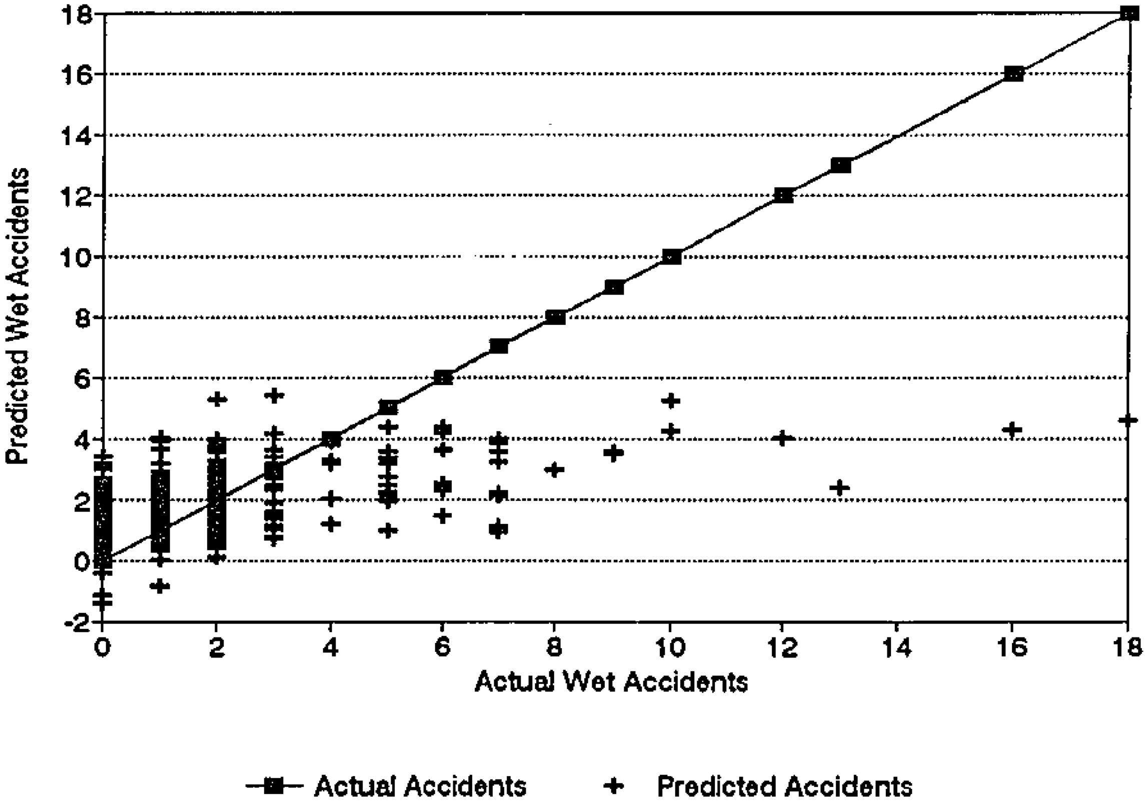


Figure 6.7. Results of the nonlinear additive regression model (1986-1988).

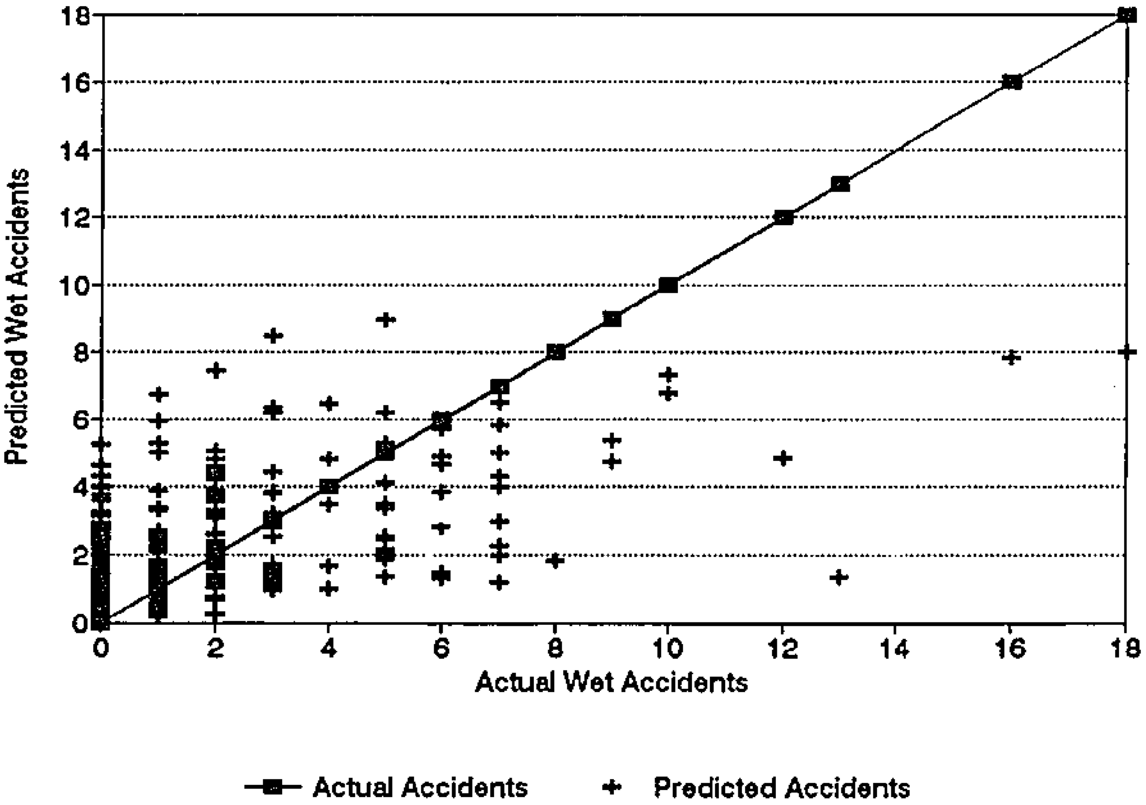


Figure 6.8. Results of the nonlinear multiplicative model using SAS NLIN procedure (1986-1988).

approach in performing the parameter estimation task. The procedure of the approach is as follows:

1. Use the step-wise regression (or the best subset regression) technique to identify the most significant attributing factors from the accident history data.
2. Estimate the expected number of accidents λ_i (the Poisson mean) of the Poisson model by using the multiplicative model (loglinear model) as:

$$E\,(\Lambda) = \prod_{j=1}^{N} C_0 Z_j^{C_k} \qquad for\ k{=}1{,}2{,}3{,}... \tag{6.6}$$

   Here Λ represents the random variable of expected number of accidents λ_i; Z_j are attributing factors such as skid number, traffic volume, driving difficulty, and so on; and C_0, C_k are coefficients to be estimated.
3. Define a range of initial guess value for each parameter based on engineering judgement to formulate a nested parameter space.
4. Use the Bayes theorem to update the probability of each possible value of the parameters.
5. Obtain the expected values of the parameters using the ideal Bayes estimator.

6. Check the probability distribution of each combination of the parameters. Shift the range of the parameter if it is necessary, then repeat steps 4 through 6.

In essence, the procedure is different from conventional nonlinear regression procedures such as Gauss-Newton, Marquardt, and Gradient methods because the coefficients of the model are assumed to be random variables rather than fixed parameters. To make a comparison, the procedure was evaluated on the real data set to estimate the number of wet accidents using the model assumed in equation 6.5. The results are plotted in figure 6.9. From this figure, it may be observed that the method presents a close result to that obtained from the SAS NLIN procedure. However, it is noted that the method depends on the sample size of the data and the initial guess values and levels of the parameters. Modifications are necessary to improve the estimation efficiency.

6.3.3. A Hierarchical Accident Index Method

The hierarchical accident index method was proposed in my thesis proposal and applied to the wet pavement index project (Kulakowski et al. 1990b). It is a combination of subjective fuzzy reasoning and a probabilistic-type approach. The model is constructed as a hierarchy with accident risk index (ARI) at the top level and three indices--accident experience index (AEI), generalized skid resistance index (SNI), and driving difficulty index (DDI)--at the second level. The lowest

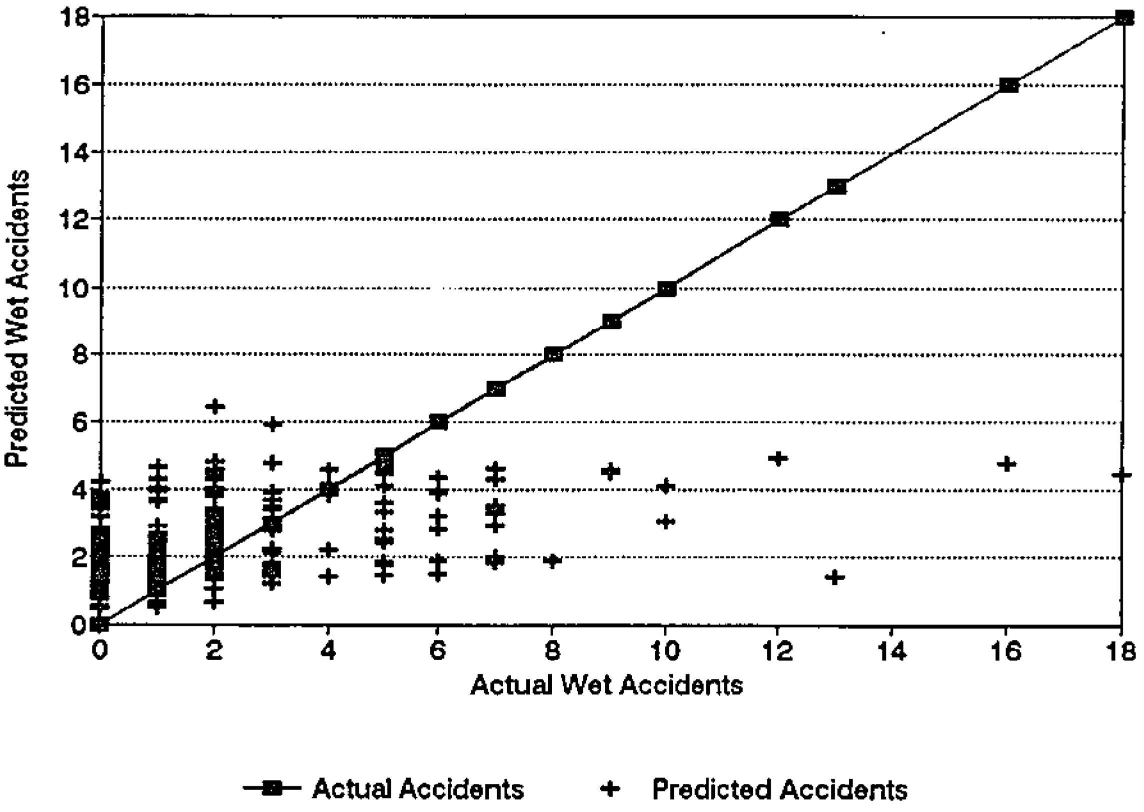


Figure 6.9. The results of the direct Bayesian regression method (1986-1988).

level is the attributing factors. Figure 6.10 depicts the concept of the hierarchical structure. The combination of the three indices used the fuzzy eigenweight method (Saaty 1977) to assign specific weight for each index. The definitions and formulations of the three indices are shown as follows:

- Accident Experience Index

The accident experience index (AEI) should give, using the previous accident statistics, a measure of relative hazardousness of the roadway condition for each location of interest. Basically, the determination of the AEI is based on the Rate Quality Control Method (Norden et al. 1956). A *modification* was made to incorporate the information of accident severity and wet weather exposure.

Following the assumption of a Poisson distribution for the accident frequency, critical accident rates for different highway groups can be calculated by approximating the upper control limit of the number of accidents from "Poisson's Experimental Binomial Limit" table (Molina 1942) under a desired coefficient of confidence. Statistically, if the coefficient of confidence is 0.995, the probability of the observed number of accidents being greater than or equal to the upper limit is 0.005. This coefficient is subjectively chosen to set a control interval. The approximate formula for the upper control limit H_u of the accident rate ARX is represented by:

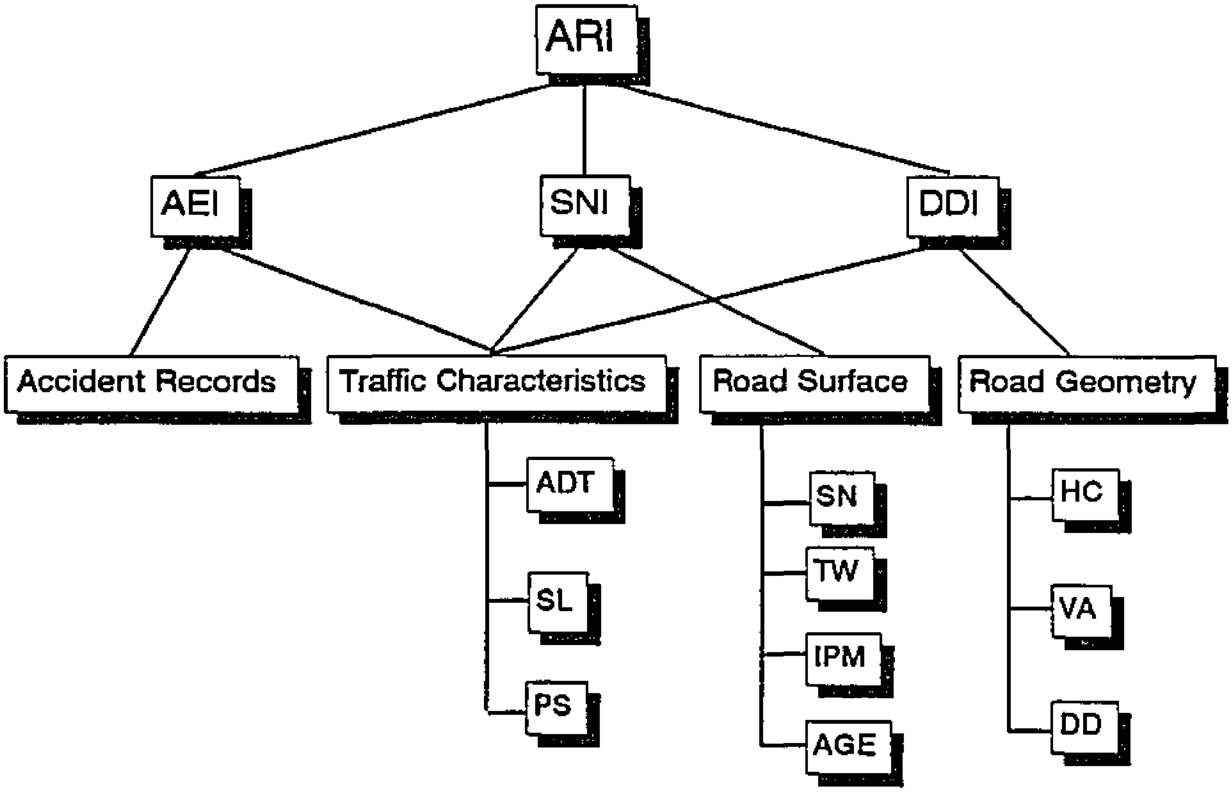


Figure 6.10. Schematic diagram of hierarchical index method.

$$H_u = ARX_{avg} + c\sqrt{\frac{ARX_{avg}}{m_i} + \frac{1.0}{2m_i}} \tag{6.6}$$

The first two terms are obtained from approximating the Poisson distribution by a normal distribution; the last term is due to the fact that only an integer number of accidents can be observed. The constant c is selected for different confidence intervals; for instance, the c value is 2.576 for a 99.5% confidence interval. The procedure for determining the AEI is described as follows:

1. Identify the highway types: 1--intersections; 2--sections.
2. Calculate the average accident rate (RAavg) and average severity rate (RSavg) for each collection of highway sections (a group):

   2.1. Calculate the accident rate (ARX_i) for each site, using the maximum likelihood estimate for h_i ($= y_i / m_i$).

   2.2. Calculate the severity rate ($SEVX_i$) for each site. The severity rate is defined as the total number of injuries and fatalities divided by vehicle exposure.

   2.3. Determine the average accident rate, RAavg, and average severity rate, RSavg, by using:

$$RA_{avg} = \frac{1}{N}\sum_{i=1}^{N} ARX_i \tag{6.7}$$

$$RS_{avg} = \frac{1}{N}\sum_{i=1}^{N} SEVX_i \tag{6.8}$$

3. Calculate the accident rate, ARX_i, and severity rate, $SEVX_i$, for each location of interest. It should be noted that the data used here may differ from the data set used in step 2 for calculating the *RAavg* and *RSavg*.

4. Calculate the critical accident rate (RCAT) and critical severity rate (RCSEV):

$$RCAT_i = RA_{avg} + c\sqrt{\frac{RA_{avg}}{m_i} + \frac{1.0}{2m_i}} \tag{6.9}$$

$$RCSEV_i = RS_{avg} + c\sqrt{\frac{RS_{avg}}{m_i} + \frac{1.0}{2m_i}} \tag{6.10}$$

Here, c=2.576 for 99.5% confidence interval.

5. Calculate the accident experience index, AEI, for each site:

   5.1. Normalize the accident rate and severity rate as:

$$ARnorm_i = \frac{ARX_i}{RCAT_i};$$

$$SEVnorm_i = \frac{SEVX_i}{RCSEV_i} \tag{6.11}$$

5.2.1. If the total number of accidents and number of accident severity are to be considered, then

$$AEI_i = \max(ARnorm_i \, , SEVnorm_i \,) \tag{6.12}$$

5.2.2. If the number of wet accidents and number of wet severity are to be considered, then:

$$WAnorm_i = \frac{(ARnorm_i \,)(wet\ accident\ ratio)_i}{TW_i \, / \, 100.} \ ;$$

$$WSEVnorm_i = \frac{(SEVnorm_i \,)(wet\ accident\ ratio)_i}{TW_i \, / \, 100.} \tag{6.13}$$

and the AEI is calculated by

$$AEI_i = \max(WAnorm_i \, , WSEVnorm_i \,) \tag{6.14}$$

- Generalized Skid Resistance Index

The generalized skid resistance index (SNI) should take care of seasonal and short-term variations of skid resistance due to environmental conditions. This can be done by a normalization procedure (Wambold et al. 1988). The procedure is designed to use a nonlinear regression model to normalize the skid resistance measurement at the site of interest with respect to standard test conditions. It requires weather information, dates of measurements, and other environmental conditions, including the air temperature for the site. Another normalization procedure was then taken by first identifying the maximum value (SNMAX) of adjusted skid numbers for each highway group and normalization was performed by dividing the SNMAX by the skid number of each site of interest. The ratio is then called SNI. It represents a relative measure of the roadway surface traction. A site with a high SNI is supposed to represent a condition of low surface traction.

- Driving Difficulty Index

The last index, driving difficulty index (DDI), is composed of three variables, the rating of horizontal curvature (HC), the rating of vertical alignment (VA), and the rating of driving difficulty. Detailed definitions of these three variables are shown in the report by Kulakowski et al. (1990b). The driving difficulty index is then formulated by a unweighted sum of these variables as shown below:

$$DDI = \frac{1}{3}((HC\ rating) + (VA\ rating) + (DD\ rating)) \tag{6.15}$$

- Accident Risk Index

Frequently, situations are encountered in which no precise measurements or information on objects are available and comparisons among the objects must be made. The subjective assignment of weights to the objects is a natural way to solve the problem. However, consistent assignments of the weights may not always be possible. Saaty's (1977) eigenweight method provides a solution to this kind of problem. The application of this method to obtain an accident risk index was initiated by formulating a pair-wise comparison matrix of the three indices--AEI, SNI, and DDI. A scale from 1 to 9, representing the intensity of importance, was chosen for each pair of indices. It is defined (Saaty 1977) that a rating of 1 on the scale stands for equal importance of the two indices; 3 represents weak importance--one index is slightly favored over another one; 5 is called strong importance--one index is strongly favored over the other index; 7 is called demonstrated importance--one index is strongly favored and its dominance is demonstrated in practice; and 9 represents absolute importance--one index is absolute in its importance over the other. The positions 2, 4, 6, and 8 are intermediate ratings.

Appropriate weights can be obtained from a set of normalized eigenvector corresponding to the maximum real eigenvalue of the pair-wise comparison matrix. The main requirement to assure consistent assignments of the weights is that the maximum real eigenvalue of the pair-wise comparison matrix must be equal to or very close to the dimension of the comparison matrix. For the problem of estimating the risk of wet pavement accidents, the scales for the three indices were subjectively selected based on engineering judgement. The rating selected for SNI was between the weak importance and the strong importance over the AEI and was selected as possessing demonstrated importance over the DDI; AEI was selected a rating between the equal importance and the weak importance over the DDI. The pair-wise comparison matrix was then formulated (see table 6.1).

Table 6.1. Pair-wise comparison matrix.

Index	AEI	SNI	DDI
AEI	1.0	0.25	2.0
SNI	4.0	1.0	7.0
DDI	0.5	0.143	1.0

The maximum eigenvalue of the comparison matrix was 3.002, which is very close to the dimension of the matrix, 3. The corresponding normalized eigenvector was then determined:

AEI: $w_1 = 0.187$

SNI: $w_2 = 0.715$

DDI: $w_3 = 0.098$

Then the accident risk index was obtained by:

$$ARI_i = w_1 AEI + w_2 SNI + w_3 DDI \tag{6.16}$$

The hierarchical accident index method was then evaluated on the real data set to obtain ARI, which is called WPI for this wet pavement accidents problem, for each site. A sorting procedure was carried out on the obtained WPI's to prioritize the accident potential with respect to the number of wet accidents in 1986-1988. Results are shown in figure 6.11. Care should be taken when comparing the predicted WPI's with the actual number of wet accidents since they are not in the same scale. Thus, figure 6.11 is shown to provide an implication of this method.

6.3.4. Modified Empirical Bayes Approach

In general, because of significant differences in traffic conditions that exist on different highway sections, it is desirable that the various highway sections be grouped into classes that *should* have similar accident rates and then be subjected to an empirical Bayes procedure as separate classes. Effective use of this procedure will require that each class have a large number of sections. These are, of course, conflicting recommendations, and some compromise must be made. The results of computer simulation in chapters 3 and 4 showed that the class size should be at least 60 (and preferably more than 100). When these procedures are applied to the highway sections of an entire state, this requirement can be easily attained. For the Pennsylvania data set the entire group of 308 highway sections is treated as one class because of the relatively small size of the data set. The results of applying the modified AA procedure include the maximum likelihood,

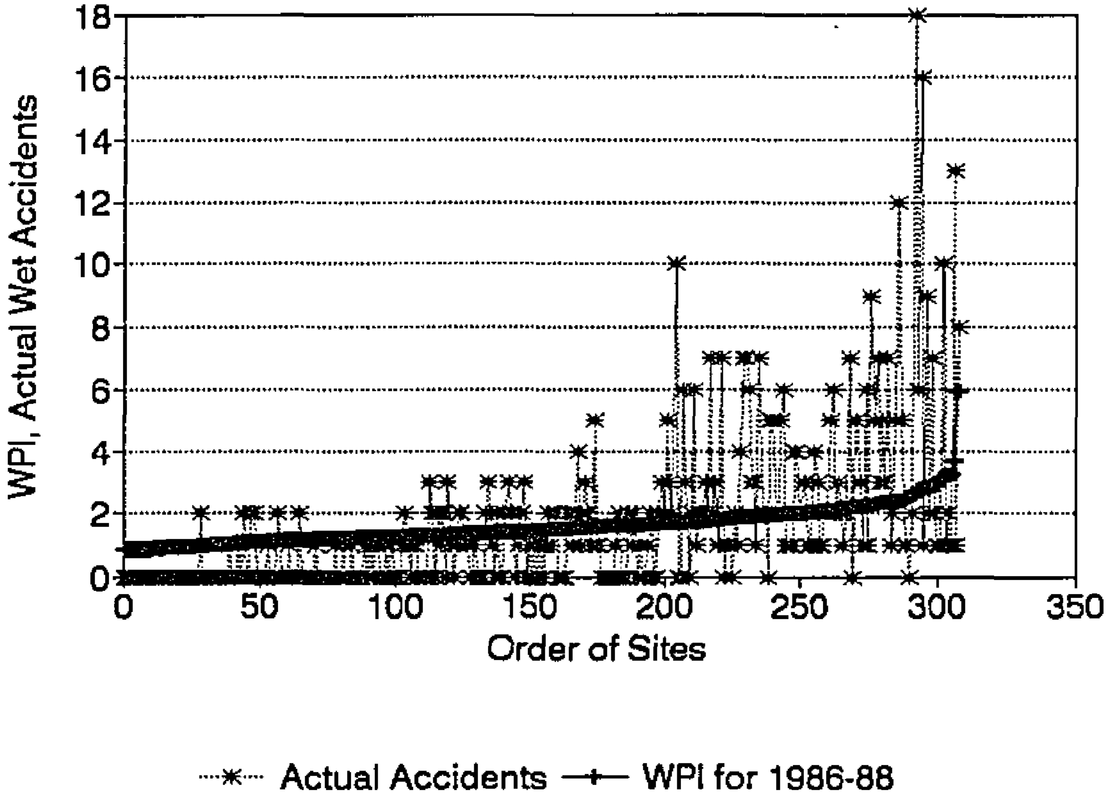


Figure 6.11. Results of the hierarchical accident index method (1986-1988).

and the two new estimators are presented in table 6.2. Table 6.2 indicates that the new estimators have provided some improvement over the modified AA procedure; additionally, all of the empirical Bayes procedures performed considerably better than the maximum likelihood. It should be noted that the true accident rate for each site of interest, based on the assumption of a Poisson-gamma model for the occurrence of the accidents, can never be known precisely. Thus, for this evaluation, the errors in predicting the numbers of wet pavement accidents were used in the sum of absolute errors loss function. It should be also noted that the AA procedure without the modification would have resulted in $\hat{\alpha} =$ -38.74, and accordingly an $\hat{\alpha}$ of 1.5 was used in this evaluation. An exploratory analysis of the real data set using the rule of fixed value of α is shown in table 6.3. This analysis provides evidential support for the selected modified rule in chapters 3 and 4.

A sorting procedure was then carried out on the estimated results from the modified AA procedure and two new estimators to prioritize the predicted number of wet pavement accidents in ascending order. The results are shown in figures 6.12, 6,13, and 6.14. As these figures illustrate, the predicted number of wet accidents smoothly follows the increasing trend of actual wet accidents. This reveals that the modified empirical Bayes procedures are effective approaches to the estimation of accident potential for road sites.

Table 6.2. Results of evaluation on the real data set.

Estimator	Sum of Absolute Errors
MLE: Maximum Likelihood	535.12
AA: Modified Arnold Antle	453.98
L1: New Estimator $\hat{h} = (\hat{\alpha}_p - 0.21) \hat{\beta}_p$	438.56
L2: New Estimator $\hat{h} = 0.92(\hat{\alpha}_p \hat{\beta}_p)$	431.82

Table 6.3. Results of AA procedure using different values of α.

α Value	1.0	1.5	2.0	3.0	5.0
Sum of Absolute Errors	454.85	453.98	457.68	467.85	486.84

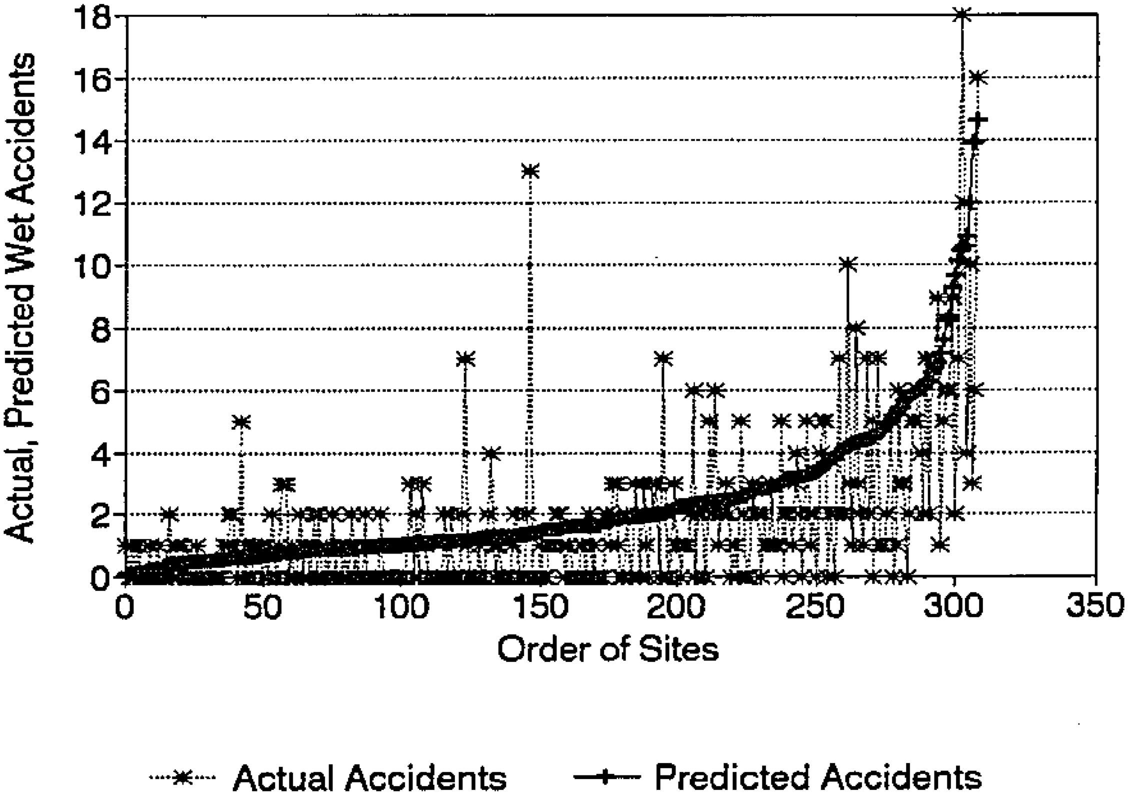


Figure 6.12. The prioritized number of wet accidents using the modified AA procedure (1986-1988).

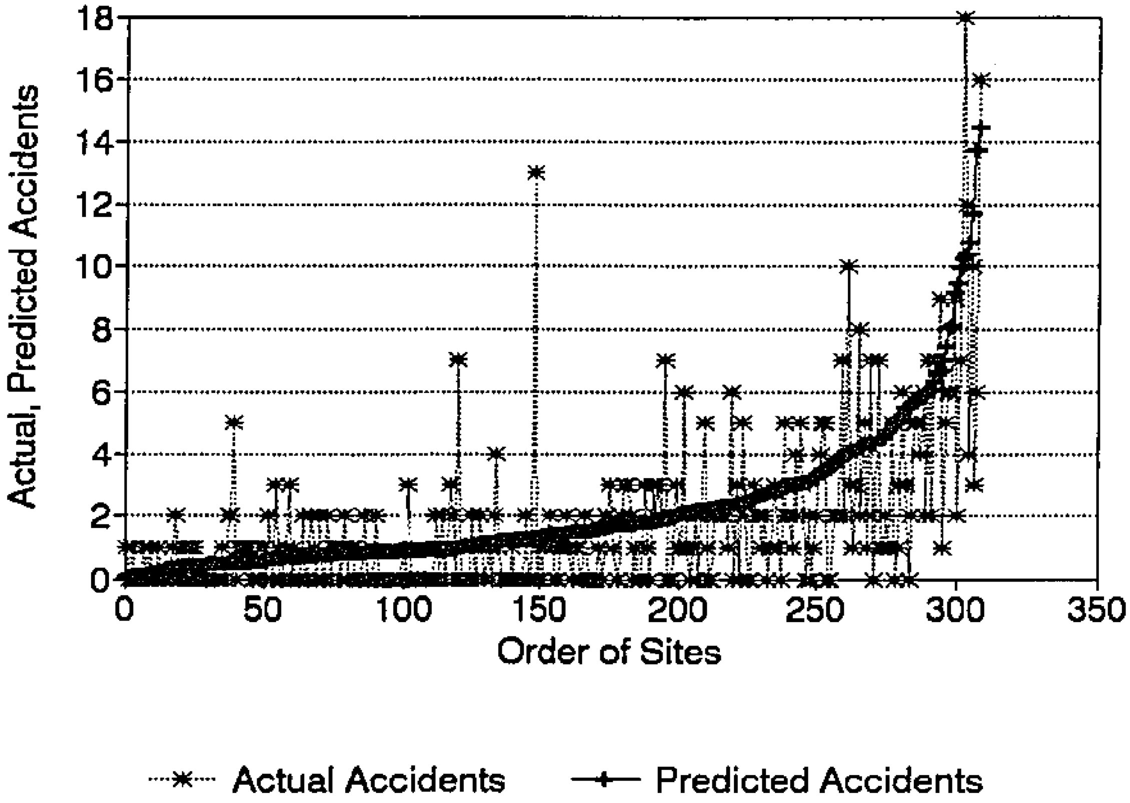


Figure 6.13. The prioritized number of wet accidents using the L1 estimator (1986-1988).

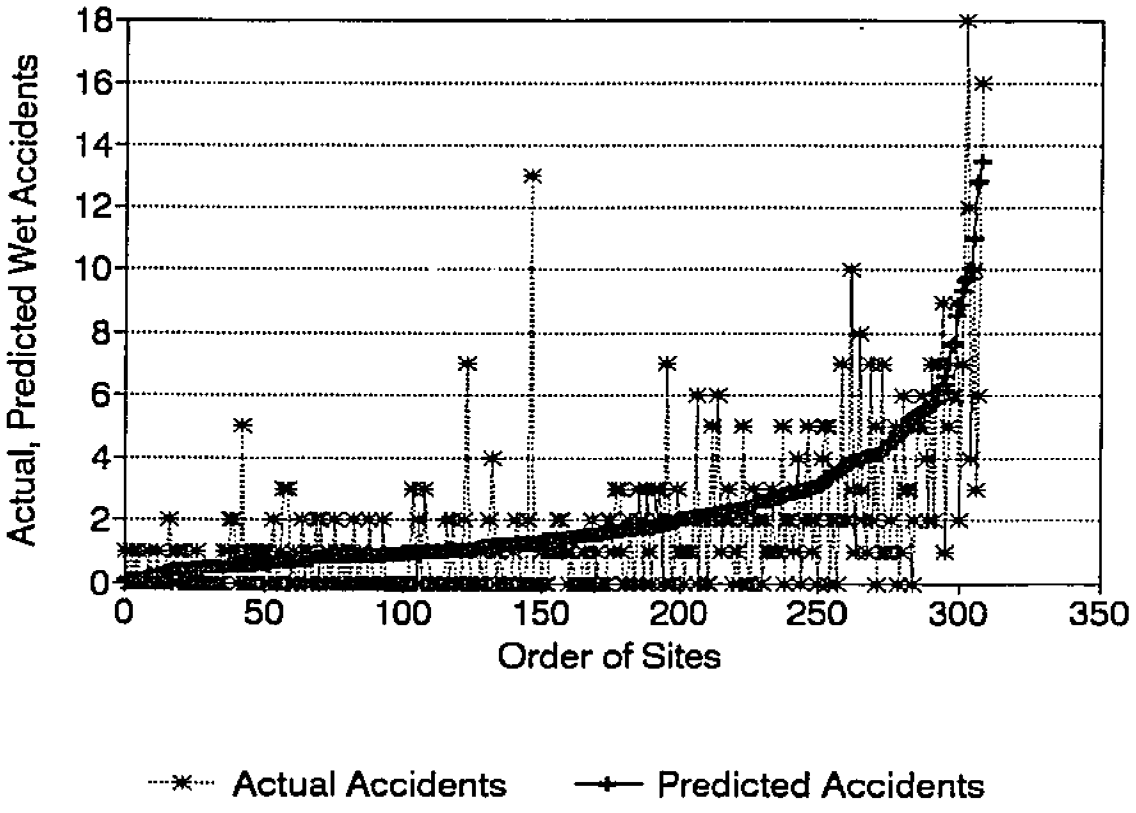


Figure 6.14. The prioritized number of wet accidents using the L2 estimator (1986-1988).

6.3.5. Knowledge-Based Model Approach

Two important features of the knowledge-based model developed in chapter 5 need to be pointed out before proceeding with the evaluation. First, the knowledge-based model can be used to predict the accident potential, a probability interval in which a traffic accident is likely to take place, for a site of interest if the measurements of the attributing factors become available. In reality, difficulties exist in defining and collecting the driver/vehicle information. Second, the calculated belief interval representing the accident potential for the road site of interest does not specifically refer to any one type of accident.

For the real data set provided by PennDOT, data measurements are available for the roadway and traffic characteristics only. Consequently, the evaluation of the knowledge-based model will concentrate on predicting the accident potential for roadway sections, that is, calculating the belief interval for the proposition of the bad roadway section. It should be also noted that due to the absence of driver/vehicle information, the calculated belief interval for the proposition of bad roadway section represents *solely* the portion of total belief committed to the roadway section. It may not be used for predicting the number of accidents occurring in the future. Hence, a site identified as a bad roadway section with a high degree of belief does not necessarily have a large number of accidents. In order to evaluate the model on the real data set, it was decided that the wet pavement index should be defined as:

$$[WPI\text{-}1,\ WPI\text{-}2] = 10*[Bel\text{-}1,\ Bel\text{-}2] \tag{6.17}$$

where WPI-1 and WPI-2 are the lower and upper bounds of WPI, respectively; while Bel-1 and Bel-2 are the *lower probability* and the *upper probability* of the belief interval, respectively. The converting factor 10 was used to comply with the confirmation scale set in the questionnaire shown in appendix A.

It is noted that the lower probability of the calculated belief interval for the real data set ranges from 0.09 to 0.28 while the upper probability of the belief interval ranges from 0.15 to 0.45. Histograms of these two probabilities are shown in figures 6.15 and 6.16.

Since the proposition of bad roadway section is combined from three bodies of evidence, the bad PC, the bad GC, and the bad TC, a high degree of belief of a bad roadway section implies that at least one of its bodies of evidence is in bad condition. This provides a simple and effective way to identify significant casual factors.

For the purpose of identifying accident-prone road sites, the belief intervals for the road sites, which represent the *latent* accident potential of a traffic accident, were used to prioritize the road sites. However, a critical value of the accident potential must be determined. According to the accident records reported by the Accident Record System of PennDOT, shown in figure 6.17, the roadway- and environment-related factors in the period of 1985-1989 account for only 13.2% of all accidents. Definitions of these factors are shown in Mason et al.

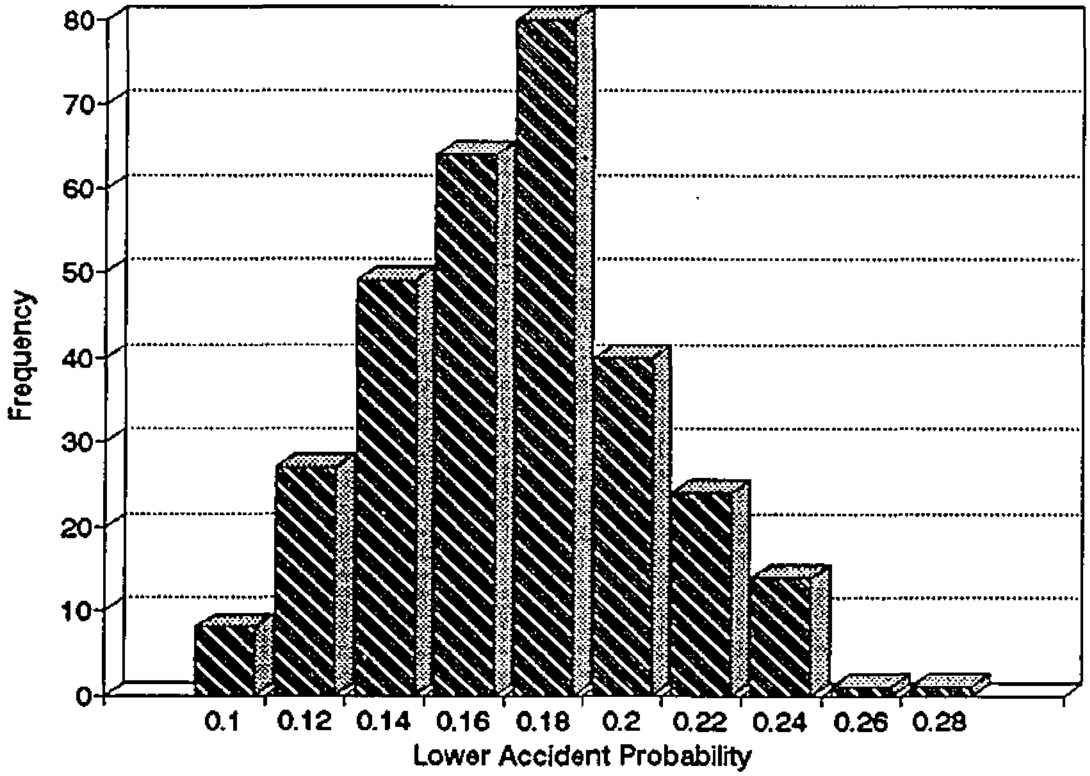


Figure 6.15. Histogram of the lower accident probabilities for roadway sections in Pennsylvania in 1983-1985.

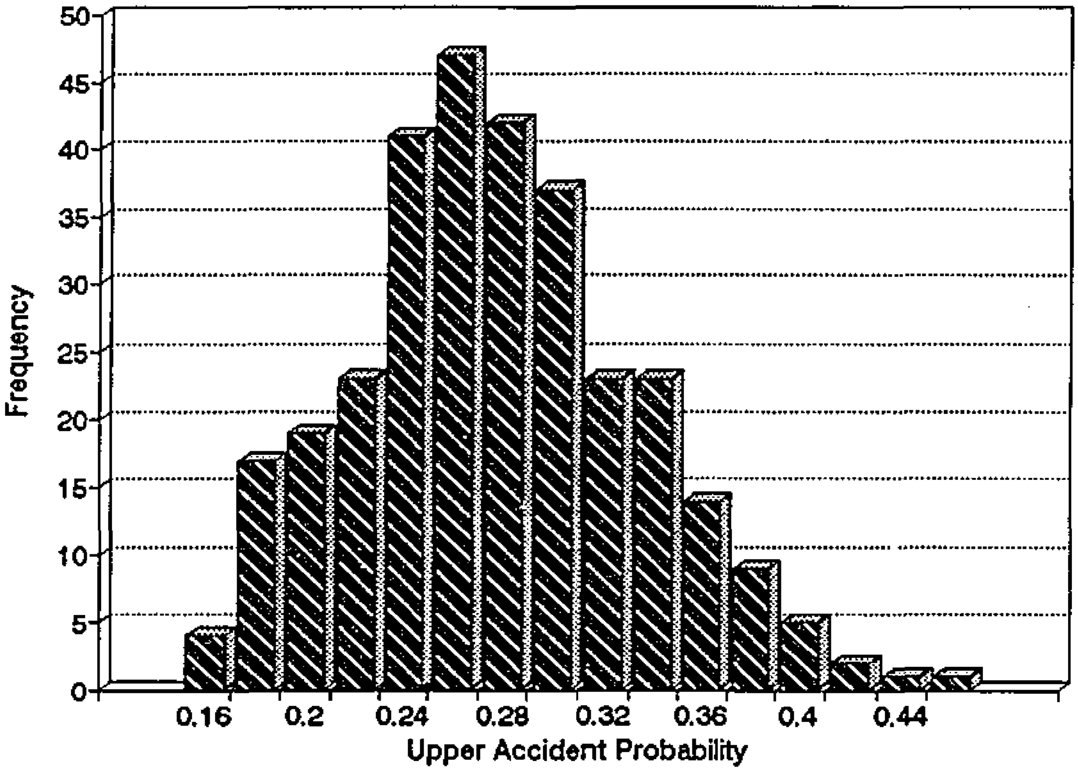
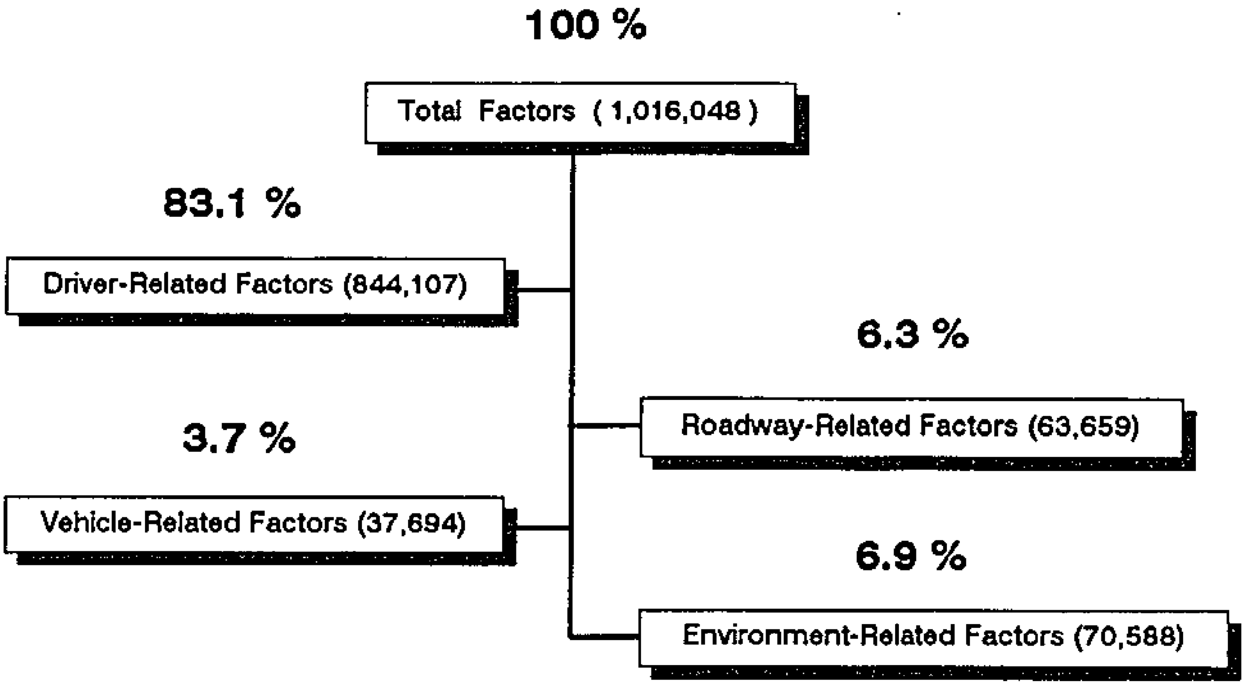


Figure 6.16. Histogram of the upper accident probabilities for roadway sections in Pennsylvania in 1986-1988.


100 %
Total Factors (1,016,048)
83.1 %
Driver-Related Factors (844,107)
6.3 %
Roadway-Related Factors (63,659)
3.7 %
Vehicle-Related Factors (37,694)
6.9 %
Environment-Related Factors (70,588)

Source: Accident Record System, Pennsylvania Department of Transportation.

Figure 6.17. Schematic diagram of the factors in all accidents in Pennsylvania (1985-1989).

(1991). This finding was therefore used as a basis to determine the critical value of the accident potential of road sites. The belief intervals of the driver, vehicle, and roadway section factors were then calculated. The results are displayed in table 6.4. It may be noted that the belief interval for the proposition of bad roadway section is [0.025, 0.189]. Based on this belief interval, a road site is identified to be hazardous when the lower probability of its belief interval is greater than 0.189. Frankly speaking, this critical value is not absolute. It can be changed when new evidence becomes available. A sorting procedure is then carried out on the estimated WPI's to prioritize the road sites. Results are shown in figure 6.18.

Table 6.4. Calculated belief intervals for the factors in a traffic accident.

Traffic Accident	m(A)	Bel-1	Bel-2
Bad Roadway Section	0.132	0.025	0.189
Poor Driver	0.831	0.805	0.969
Poor Vehicle	0.037	0.006	0.170

Source: Accident Record System, Pennsylvania Department of Transportation.

6.4. A Comparison of the Methodologies

After evaluating those developed methodologies on the real data set, a comparison was made to determine the best method for the wet accident problem.

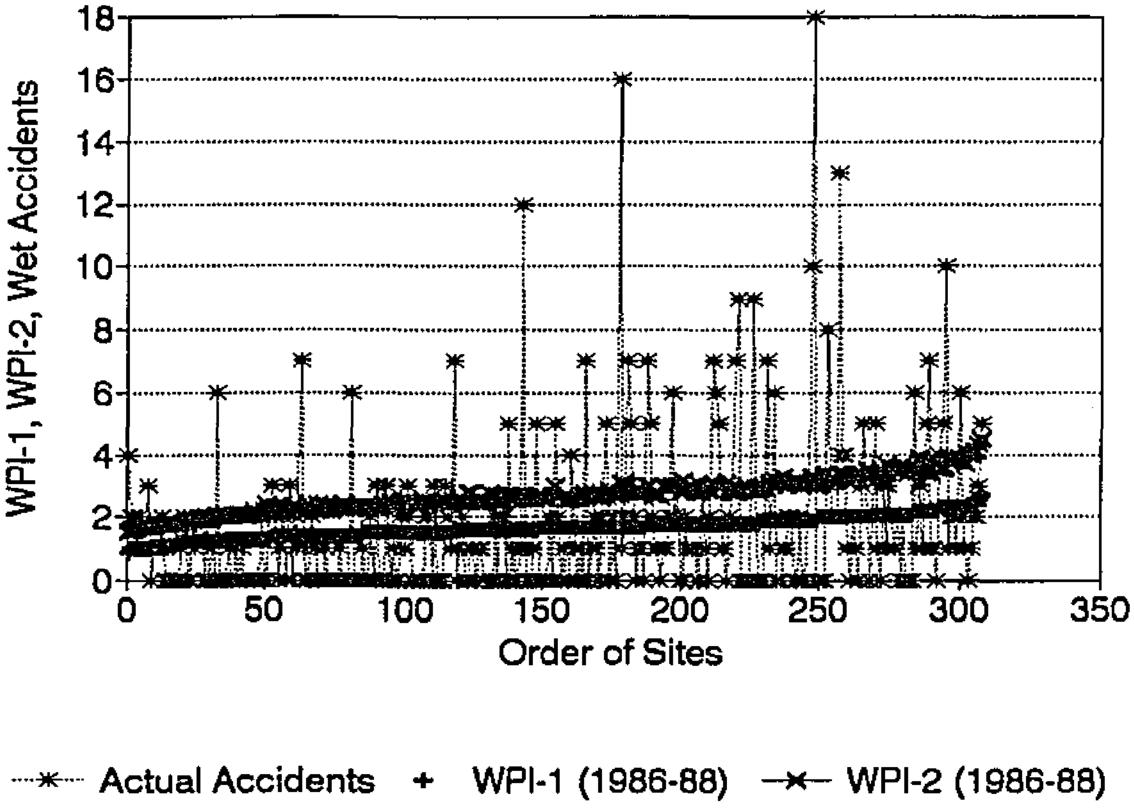


Figure 6.18. The results of the knowledge-based model approach.

Basically, the criteria to justify these methodologies are based on their estimation efficiency and/or based on subjective judgement. The classical regression methods, the direct Bayesian regression method, and the modified empirical Bayes procedures can be justified from their estimation efficiency. On the other hand, the hierarchical accident index method and the knowledge-based model approach would be justified subjectively.

The classical linear and nonlinear additive regression methods, as discussed earlier, produce negative estimates for the number of accidents. Hence, these two methods are not considered. The nonlinear multiplicaptive model using the SAS NLIN regression procedure or the direct Bayesian regression procedure, however, provides a feasible approach to the wet pavement accident problem. Table 6.5 presents a comparison of all of the developed methods.

It is noted that the sum of absolute error between the actual and the predicted number of wet accidents when using the SAS NLIN procedure is 512.72, whereas the sum of absolute error for the direct Bayesian regression procedure is 539.56. A slightly better performance of the SAS NLIN procedure is observed.

When the SAS NLIN procedure is compared with the modified empirical Bayes procedures, it is observed from the table 6.5 that the modified Bayes procedures, especially the two new median estimator L1 and L2, are far better than the SAS NLIN procedure.

The hierarchical accident index method is a combination of fuzzy reasoning technique and probabilistic type of approach. It calculates the WPI's based on the

Table 6.5. Summarized results of the evaluation on the real data set.

Type of Approach	Methods	Sum of Absolute Errors	Output	Required Data
Objective	Direct Bayesian Regression	539.56	Predicted Accidents	Accident Records and Factor Measurements
	Maximum Likelihood	535.12	Predicted Accidents	Accident Records and Factor Measurements
	Nonlinear Procedure (SAS NLIN)	512.72	Predicted Accidents	Accident Records and Factor Measurements
	Modified AA Procedure	453.98	Predicted Accidents (Accident Rates)	Accident Records and Factor Measurements
	New Median Estimator L1	438.56	Predicted Accidents (Accident Rates)	Accident Records and Factor Measurements
	New Median Estimator L2	431.82	Predicted Accidents (Accident Rates)	Accident Records and Factor Measurements
Combined	Hierarchical Accident Index	N/A*	Accident Risk	Accident Records and Factor Measurements
Subjective	Knowledge-based Model	N/A*	Accident Risk (Significant Factors)	Expert Knowledge and Factor Measurements

*: Not applicable.

accident records and the measurements of roadway and traffic characteristics. On the other hand, the knowledge-based model approach calculates the total belief that is committed to the proposition of a bad roadway section for each location of interest. A comparison of these two methods with the Bayesian methods is difficult since the WPI's obtained by these two methods are the accident potential for the road sites but not the predicted number of wet pavement accidents. It is observed that, from figure 6.11, the hierarchical accident index method performed almost as well as the modified empirical Bayes procedures shown in the figures 6.12, 6,13, and 6.14.

The main disadvantage of the empirical Bayes procedures, however, is that the procedures rely on accident records and data measurements of the roadway sections to estimate the parameters in the prior distribution. If those data are not available, the empirical Bayes methods cannot be applied. The same drawback exists in the hierarchical accident index method. This disadvantage, however, does not exist in the knowledge-based model approach, since an expert knowledge base was constructed. The model provides a simple and easy way to identify bad roadway sections and significant causal factors through the knowledge base. Its performance is not fully justified on this real data set. Table 6.6 presents a comparison of identifying significant causal factors for the modified empirical Bayes procedures, the hierarchical accident index method, and the knowledge-based model approach. Since all of the factor effects are represented by only one parameter, the Poisson mean λ, the wet vehicle exposure (defined in equation

Table 6.6. Results of identifying significant causal factors on the real data set.

Methods	Significant Factors
Modified AA Procedure	Wet Vehicle Exposure (WM)
New Median Estimator L1	Wet Vehicle Exposure (WM)
New Median Estimator L2	Wet Vehicle Exposure (WM)
Hierarchical Accident Index	SN, DD
Knowledge-based Model	Driver, SN, TW, DD

6.1) is considered to be significant in estimating the accident risk when the modified empirical Bayes procedures were used. The SN and DD are considered to be significant when using the hierarchical accident index method. Based on calculated belief intervals and belief function numbers presented in chapter 5, the driver, SN, TW, and DD are identified as significant factors for a traffic accident.

6.5. Summary

In this chapter, a case study of developing a wet pavement index to evaluate the accident risk of wet pavement accidents is presented. Several methodologies including the classical regression methods, the direct Bayesian method, the hierarchical accident index method, the modified empirical Bayes procedures, and the knowledge-based model were developed and evaluated using the real data set provided by PennDOT. Essentially, three types of approach are considered. One is the objective (direct) type regression method and Bayesian methods; the second is the combined (indirect) type hierarchical accident index method, and the third is the subjective (indirect) type knowledge-based model. Based on an absolute error loss function, the modified empirical Bayes procedures are considered to be superior to the other approaches for the risk assessment problem. However, the knowledge-based model approach should be considered if, in addition to predicting accident risk for roadway sections, an identification of significant causal factors for an accident is desired. A further justification of the capability of the knowledge-based model can be carried out if the driver/vehicle information is available.

Chapter 7

CONCLUSIONS AND RECOMMENDATIONS

In this research, three types of approach to the problem of identification and risk assessment for a traffic accident were explored--one is called an objective type of approach, which includes classical regression techniques, a direct Bayesian regression method, and modified empirical Bayes procedures; the second is a subjective type of approach using a developed knowledge-based model; the third is a combined (hybrid) approach using a hierarchical accident index method that was proposed in the thesis proposal. A summary of the development of new methods and conclusions based on research findings are given in the following subsections.

7.1. Summary

In this thesis, four new methods--the modified AA procedure, the two new median estimators, the knowledge-based model, and the hierarchical accident index method--were developed to assess the accident risk and identify significant causal factors for a traffic accident.

The development work begins with an introduction of the Bayesian methods and the belief function theory (Shafer 1976) in chapter 2. The problem of nonpositive and large values of parameter α encountered in the AA procedure was solved using a modified rule presented in chapter 3. The modified rule, α = 1.5 for $\alpha \leq 0$ and $\alpha = 10$ for $\alpha > 10$, proved to be very effective and efficient in estimating the accident rate under simulated conditions. The random effect of vehicle exposure represented by a Weibull random variable in the computer simulation was investigated. A sample size of 100 was recommended for parameter estimation.

In chapter 4, two new median estimators for a gamma distribution were developed for the traffic accident problem in which an absolute error loss function is considered. Based on computer simulation, the values of the two constants k_c and k_b used to determine the two new median estimators, L1 and L2, were 0.21 and 0.92. An evaluation of the L1 and L2 on a simulated data set reported by Morris et al. (1991) indicated that the L1 and L2 are very efficient. If the median estimators and the modified rule presented in chapter 3 are combined, a new rule for the AA procedure can be represented by setting $\hat{\alpha} = 1.5$ if $\hat{\alpha} < 0.3$ and estimating β by the equation $\hat{\beta} = SY / (1.5*SM)$; if $\hat{\alpha}$ is greater than 10, then $\hat{\alpha}$ would be set equal to 10 and, accordingly, β would be estimated by $\hat{\beta} = SY / (10*SM)$.

A knowledge-based model was developed in chapter 5 to eliminate the disadvantage of relying on accident records and measurement data to assess the accident risk. The model is based on a collected expert knowledge base and the belief function theory. A questionnaire shown in appendix A was designed to collect the expert knowledge. The model can provide a degree of belief for a single factor in a proposition or a combined degree of belief for a traffic accident. It can be updated when additional expert knowledge is available. The model can be used to identify significant causal factors by using a calculated belief function number for each factor. The uncertainties introduced by human experts in constructing the model, however, are an important drawback to the model.

Validation of these newly developed methods in addition to classical regression methods was performed in chapter 6 on a real data set provided by Pennsylvania Department of Transportation. The data set contains 308 highway sections in Pennsylvania for the period of 1983-1988. Accident records and measurements of site-specific characteristics such as SN, RUT, IPM, and so on, were available for the 6-year period. Results of the validation showed that the modified empirical Bayes procedures are superior to the other approaches based on an absolute error loss function. The hierarchical accident index method performed almost as well as the modified empirical Bayes procedures in estimating the accident risk of wet pavement accidents. The knowledge-based model can predict accident risk and identify significant causal factors for wet accidents simultaneously. Its performance can be fully justified if driver/vehicle information is available.

7.2. Research Findings and Conclusions

A number of research findings presented in this thesis can be summarized as follows:

1. Classical linear regression methods are not suitable for risk assessment in traffic event systems due to the fact that a given system is nonlinear and these methods may produce negative estimates.
2. Modifications to the direct Bayesian method are necessary in order to improve its computational efficiency and accuracy.
3. The hierarchical accident index method provides an alternative means for determining the accident potential for road sites by using both accident records and roadway and traffic characteristics. It performs almost as well as the modified empirical Bayes procedures on the real data set.
4. The fundamental assumptions made in developing the modified empirical Bayes procedures are appropriate for the traffic event system.
5. Based on an absolute error loss function, the modified empirical Bayes procedures perform almost as well as the ideal Bayes procedure as long as the sample size of the simulated data is equal to or greater than 100. It should be noted that the ideal Bayes procedure is an optimal estimator when the prior distribution is precisely known.

6. When considering the random effect of vehicle exposure, the modified empirical Bayes procedures proved to be effective for different levels of vehicle exposure during computer simulation.
7. The modified empirical Bayes procedures result in a smaller sum of absolute errors between the actual and the predicted number of wet accidents than the other methods on the real data set.
8. The SN of each location in the real data set remained unchanged during the 1983-1988 period. This potentially increased the difficulty of identifying significant causal factors for the wet pavement accident problem.
9. The knowledge-based model provides detailed information on the occurrence of traffic accidents. It predicts the accident risk for traffic accidents. It is suitable for daily or routine surveys of highway systems to identify significant causal factors or hazardous road sites.
10. The knowledge-based model can be updated and expanded when new bodies of evidence are available.
11. Based on the evaluation of the real data set, wet vehicle exposure is considered to be significant for the modified empirical Bayes procedures. The SN and DD were identified as significant factors in estimating the accident risk when the hierarchical accident index method was applied. The significant causal factors are driver, SN, TW, and DD when using the knowledge-based model approach.

Based on the findings stated above, it is concluded that the modified empirical Bayes procedures and the knowledge-based model are the methods that should be considered for the problem of identification and assessment of risk in traffic event systems. The modified empirical Bayes procedures should be used when objective information--accident records and measurements of roadway and traffic characteristics--is available. However, the knowledge-based model approach should be applied if the objective information is not available.

7.3. Recommendations for Future Research

From the results of the simulations and the applications of the different methods for estimating the risk of a wet pavement accident using the Pennsylvania data, the following recommendations can be formulated:

1. The highway sections should be grouped into classes that number at least 60 each, where the highway sections within classes are as similar as possible in the properties that affect the risk of a wet pavement accident. The quality of the measurements such as SN, TW, ADT, RUT, AGE, and so on, should be improved.
2. The new estimators presented in this thesis, L1 and L2, performed better than other known estimators on sets of computer-simulated and actual accident data. More research on robustness should be conducted to fully evaluate these new estimators.

3. A study of the robustness of the selected model (gamma-Poisson) should be performed. Perhaps some procedures for model development and evaluation should be developed. This may be especially true in regard to the form for the prior distribution.

4. More research should be conducted on the knowledge-based model to explore and evaluate its capabilities and limitations. The feasibility of expanding the knowledge-based model to become an expert system should be investigated.

BIBLIOGRAPHY

Abbess, C., Jarrett, D., and Wright, C. C., "Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the 'Regression-To-Mean' Effect," *Traffic Engineering and Control*, Oct. 1981.

Arnold, S. F. and Antle, C. E., "An Empirical Bayes Solution For the Problem Considered by Williford and Murdock," *Accident Analysis and Prevention*, Vol. 10, 1978.

Baker, J. S., "An Evaluation of the Traffic Conflicts Technique," *Highway Research Record*, No. 384, pp. 1-8, 1972.

Barnett, J. A., "Computational Methods for a Mathematical Theory of Evidence," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, BC, Vol. 2, pp. 868-875, August 1981.

Bayes, T., "An Essay Towards Solving a Problem in the Doctrine of Chance," *Philosophical Transaction of the Royal Society*, 53, pp. 370-418, 1763.

Boswell, M. T., TULSIM version 3.1, The Statistics Department, University Park PA: The Pennsylvania State University, Sept. 1987.

Brodsky, H. and Hakkert, A. S., "Risk of a Road Accident in Rainy Weather," *Accident Analysis and Prevention*, Vol. 20, No. 3, pp. 151-176, 1988.

Brüde, U. and Larsson, J., "The Use of Prediction Models for Eliminating Effects Due To Regress-To-The-Mean in Road Accident Data," *Accident Analysis and Prevention*, Vol. 20, No. 4, pp. 299-310, 1988.

Campbell, M. E., "The Wet-Pavement Accident Problem: Breaking Through," *Traffic Quarterly*, 15, pp. 209-214, 1971.

Caples, G. B. and Vanstrum, R. C., The Price of Not Walking, Minnesota Mining and Manufacturing Company, Sept. 1969.

Dempster, A. P., "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals Mathematical Statistics*, 38, pp. 325-339, 1967.

Dempster, A. P., "A Generalization of Bayesian Inference," *Journal of Royal Statistical Society*, Ser. B 30, pp. 205-247, 1968.

Dubois, D. and Prade, H., Fuzzy Sets and Systems: Theory and Applications, New York, Academic Press, 1980.

Duda, R. O., Hart, P. E., and Nilsson, N. J., "Subjective Bayesian Methods for Rule-Based Inference Systems," American Federation of Information Proceeding Societies (AFIPS), New York, NY: *National Computer Conference*, Vol. 45, pp. 1075-1082, 1976.

Dunlap, J. W., Orlansky, J., and Jacobs, H. H., Manual for the Application of Statistical Techniques for use in Accident Control, U.S. Department of Commerce, Office of Technical Service, Washington, DC, June 1958.

Fine, W. T., "Mathematical Evaluations for Controlling Hazards," *Journal of Safety Research*, Vol. 3, No. 4, pp. 157-166, 1971.

Fung, L. W. and Fu, K. S., "An Axiomatic Approach to Rational Decision Making in a Fuzzy Environment," *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*, New York, Academic Press, 1975.

Garrett, J. W. and Tharp, K. J., "Development of Improved Methods for Reduction of Traffic Accidents," *National Cooperative Highway Research Program*, Report No. 79, 1969.

Garvey, T. D., Lowrance, J. D., and Fischler, M. A., "An Inference Technique for Integrating Knowledge from Disparate Sources," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, BC, pp. 319-325, Aug. 1984

Gaver, D. P. and O'Muircheartaigh, I. G., "Robust Empirical Bayes Analyses of Event Rates," *Technometrics*, Vol. 29, Feb. 1987.

Giles, C. G., and Sabey, B. E., "A note on the Problem of Seasonal Variation in Skidding Resistance," *Proceedings, First International Skid Prevention Conference*, Part I and Part II, pp. 563-568, Charlottesville, VA, 1959.

Gunaratne, M., Altschaeffl, A. G., and Chameau, J. L., The Use of Fuzzy Sets Mathematics in Pavement Evaluation and Management, Interim Report, FHWA/IN/JHRP-84/18, Indiana Department of Highways, Indianapolis, IN 46204.

Harwood, D. W., Blackburn, R. R., Kulakowski, B. T., and Kibler, D. F., Wet Weather Exposure Measures, Final Report No. FHWA/RD-87, Washington, DC, Federal Highway Administration, 1987.

Hauer, E. and Persaud, B. N., "Problem of Identifying Hazardous Locations Using Accident Data," *Transportation Research Record*, No. 975, 1984.

Heimbach, C. L., Allred, P. W. , Anderson, R. L., Atkins, J. R., and Hooper, J. W., Developing Traffic Flow Indices for the Detection of High Accident Potential Highways in North Carolina, Raleigh, NC: North Carolina State University, Oct. 1968.

Henry, J. J., Saito, K., and Blackburn, R., Predictor Model for Seasonal Variations in Skid Resistance: Comprehensive Final Report, Report No. FHWA/RD-83/005, Washington, DC, Federal Highway Administration, 1983.

Higle, J. L. and Witkowski, J. M., "Bayesian Identification of Hazardous Locations," *Transportation Research Record*, No. 1185, pp. 24-36, 1988.

Ivey, D. L. and Griffin III, L. I., Development of Wet Weather Safety Index, Texas Transportation Institute, Report No. FHWA TX77-2211F, Washington, DC, Federal Highway Administration, 1977.

Kulakowski, B. T. and Digiovanni, M. A., "Measurement and Modelling of the Water Films of Road Surfaces," *Instrument Society of America*, Vol. 27, No. 1, pp. 9-14, 1988.

Kulakowski, B. T. and Harwood, D. W., "Effect of Water-Film Thickness on the Tire-Pavement Friction," *ASTM Special Technical Publication*, No. 1031, 1990a.

Kulakowski, B. T., Wambold, J. C., Antle, A. E., Lin, C., and Mason, J. M., Jr., Development of a Methodology to Identify and Correct Slippery Pavements, Report No. FHWA-PA90-002+88-06, Washington DC: Federal Highway Administration, Nov. 1990b.

Laughland, J. C., Haefner, L. E., Hall, J. W., and Clough, D. R., "Methods for Evaluating Highway Safety Improvements," *National Cooperative Highway Research Program*, Report No. 162, Transportation Research Board, 1975.

Lemmon, G. H. and Krutchkoff, R. G., "An Empirical Bayes Smoothing Technique," *Biometrika*, 56, No. 2, pp. 361-365, 1969.

Lin, C., Antle, C. E., and Kulakowski, B. T., "An Application and Evaluation of Empirical Bayes Methods for the Problem of Evaluating the Risk of Wet Pavement Accidents," submitted to *Journal of Accident Analysis and Prevention*, Aug. 1991.

Maher, M. J. and Mountain, L. J., "The Identification of Accident Blackspots: A Comparison of Current Methods," *Accident Analysis and Prevention*, Vol. 20, No. 2, pp. 143-151, 1988.

Maritz, J. S., "Smooth Empirical Bayes Estimation for One-Parameter Discrete Distribution," *Biometrika*, 53, pp. 417-429, 1966.

Maritz, J. S., "Empirical Bayes Estimation for the Poisson Distribution," *Biometrika*, 56, 2, pp. 349-359, 1969.

Maritz, J. S., Empirical Bayes Method, London: Methuen and Company, Ltd., 1970.

Mason, J. M., Jr., Patten, M. L., Plummer, C. W. Jr., and Micsky, R. J., Safer Truck Travel, Pennsylvania Transportation Institute, Draft Final Report, Report No. 9219, University Park, PA: The Pennsylvania State University, Nov. 1991.

Molina, E. C., Possion's Exponential Binomial Limit, New York, D. Van Nostrand C. Inc, 1942.

Morin, D. A., "Application of Statistical Concepts to Accident Data," *Highway Research Record Bulletin*, No. 188, 1967.

Morris, C. N., Christiansen, C. L., and Pendleton, O. J., Application of New Accident Analysis Methodologies: Vol III--Theoretical Development of New Accident Analysis Methodology, Report No. FHWA-RD-91-015, College Station TX: Texas Transportation Institute, The Texas A&M University, May 1991.

National Safety Council, Manual of Classification of Motor Vehicle Traffic Accidents, 3rd ed., ANSI D16.1, 1976.

Norden, M. L., Orlansky, J., and Jacobs, H. H., "Application of Statistical Quality-Control Techniques to Analysis of Highway-Accident Data," *Highway Research Record Bulletin*, No. 117, 1956.

Perkins, S. R. and Harris, J. I., "Traffic Conflict Characteristics--Accident Potential at Intersections," *Highway Research Record*, No. 225, pp. 35-43, 1968.

Persaud, B. N. and Hauer, E., "Comparison of Two Methods for Debiasing Before-and-After Accident Studies," *Transportation Research Record*, No. 975, 1984.

Rice, J. M., "Seasonal Variations in Pavement Skid Resistance," *Public Roads*, Vol. 40, No. 4, pp. 160-166, Mar. 1977.

Robbins, H. E., "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp. 157-164, 1955.

Rockwell, T. H. and Treiterer, J., "Sensing and Communication Between Vehicles," *National Cooperative Highway Research Program*, Report No. 51, 1968.

Rutherford, J. R. and Krutchkoff, R. G., "Some Empirical Bayes Techniques in Point Estimation," *Biometrika*, 56, No. 1, pp. 133-137, 1969.

Ryan, B. F., Joiner, B. L., and Ryan, T. A. Jr., Minitab Reference Manual: Release 7, State College, PA: Minitab Inc., Apr. 1989.

Saaty, T. L., "A Scaling Method for Priorities in Hierarchical Structures," *Journal of Mathematical Psychology*, No. 15, 1977.

SAS Institute, SAS User's Guide: Statistics, 5th ed., NC, 1985.

Shafer, G., A Mathematical Theory of Evidence, Princeton, NJ: Princeton University Press, 1976.

Walpole, R. E. and Myers, R. H., Probability and Statistics for Engineers and Scientists, 1st ed., New York, NY: Macmillan Publishers, Co., 1972.

Wambold, J. C., Henry, J. J., Antle, C. E., Kulakowski, B. T., Meyer, W. E., Stocker, A. J., Button, J. W., and Anderson D. A., Pavement Friction Measurements Normalized for Operational, Seasonal, and Weather Effects, Pennsylvania Transportation Institute, Report No. 8711, University Park, PA: The Pennsylvania State University, 1988.

Winkler, R. L., An Introduction to Bayesian Inference and Decision, 1st ed., New York, NY: Holt, Rinehart and Winston, Inc., 1972.

Zadeh, L. A., "Fuzzy Sets," *Information and Control*, No. 8, 338, 1965.

Zadeh, L. A., "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3, 28, 1973.

Appendix A

A QUESTIONNAIRE FOR IDENTIFICATION AND RISK ASSESSMENT IN TRAFFIC EVENT SYSTEM

by

Chunming Lin

June 1991

The Pennsylvania Transportation Institute
The Pennsylvania State University
University Park, PA 16802

PROPOSITION: ACCIDENT POTENTIAL FOR THE HIGHWAY SECTION IS HIGH

Body of Evidence	Scale of Confirmation (0-10)	Scale of Disconfirmation(0-10)
Pavement condition (PC) is **bad.**		
Geometric condition (GC) is **bad.**		
Traffic condition (TC) is **bad.**		

Note:

1. The scale of confirmation (or disconfirmation) represents an assigned degree of support to each body of evidence that has a (or has no) contributing effect on the proposition. It is a scale of increasing effect, that is, a rating of 10 represents the strongest effect on the proposition.

2. There is no compelling reason to assign a scale of disconfirmation (or confirmation) to each body of evidence, although it would be helpful. If you have no idea what rating to assign on this scale, please assign a rating of zero.

PROPOSITION: ACCIDENT POTENTIAL FOR DRIVER /VEHICLE IS HIGH

Body of Evidence	Scale of Confirmation (0-10)	Scale of Disconfirmation (0-10)
Driver's experience is good.		
Driver's experience is fair.		
Driver's experience is little.		
Driver's personality is normal.		
Driver's personality is nervous.		
Driver's personality is aggressive.		
Driver is tired or sleepy.		
Driver is drug or alcohol influenced.		
Driver is alert.		
Vehicle condition is good.		
Vehicle condition is fair.		
Vehicle condition is bad.		

Note:

1. The scale of confirmation (or disconfirmation) represents an assigned degree of support to each body of evidence that has a (or has no) contributing effect on the proposition. It is a scale of increasing effect, that is, a rating of 10 represents the strongest effect on the proposition.

2. There is no compelling reason to assign a scale of disconfirmation (or confirmation) to each body of evidence, although it would be helpful. If you have no idea what rating to assign on this scale, please assign a rating of zero.

PROPOSITION: PAVEMENT CONDITION (PC) IS BAD

Body of Evidence	Scale of Confirmation (0-10)	Scale of Disconfirmation (0-10)	The Unacceptable Value (or Range)
Skid resistance is **low**.			
Rutting is **high**.			
Roughness is **high**.			
Pavement age is **high**.			

Note:

1. The scale of confirmation (or disconfirmation) represents an assigned degree of support to each body of evidence that has a (or has no) contributing effect on the proposition. It is a scale of increasing effect, that is, a rating of 10 represents the strongest effect on the proposition.

2. There is no compelling reason to assign a scale of disconfirmation (or confirmation) to each body of evidence, although it would be helpful. If you have no idea what rating to assign on this scale, please assign a rating of zero.

3. The unacceptable value (or range) for each body of evidence is designed to estimate its critical value (or range).

PROPOSITION: SKID RESISTANCE IS LOW

Ranges of Skid Resistance	Degree of Confirmation (0-10)	Degree of Disconfirmation (0-10)
< 20		
20 - 25		
25 - 30		
30 - 35		
35 - 40		
> 40		

PROPOSITION: RUTTING IS HIGH

Ranges of Rutting (in)	Degree of Confirmation (0-10)	Degree of Disconfirmation (0-10)
> 1.0		
0.5 - 1.0		
< 0.5		

PROPOSITION: ROUGHNESS IS HIGH

Ranges of Roughness (in/mi, IPM)	Degree of Confirmation (0-10)	Degree of Disconfirmation (0-10)
> 300		
250 - 300		
200 - 250		
150 - 200		
100 - 150		
< 100		

PROPOSITION: PAVEMENT AGE IS HIGH

Ranges of Pavement Age (years)	Degree of Confirmation (0-10)	Degree of Disconfirmation (0-10)
> 15		
10 - 15		
5 - 10		
2 - 5		
< 2		

PROPOSITION: TRAFFIC CONDITION (TC) IS BAD

Body of Evidence	Scale of Confirmation (0-10)	Scale of Disconfirmation (0-10)	The Unacceptable Value (or Range)
Average daily traffic is **high.**			
Weather condition is **bad.**			

Note:

1. The scale of confirmation (or disconfirmation) represents an assigned degree of support to each body of evidence that has a (or has no) contributing effect on the proposition. It is a scale of increasing effect, that is, a rating of 10 represents the strongest effect on the proposition.

2. There is no compelling reason to assign a scale of disconfirmation (or confirmation) to each body of evidence, although it would be helpful. If you have no idea what rating to assign on this scale, please assign a rating of zero.

3. The unacceptable value (or range) for each body of evidence is designed to estimate its critical value (or range).

PROPOSITION: AVERAGE DAILY TRAFFIC IS HIGH

Range of Average Daily Traffic	Degree of Confirmation (0-10)	Degree of Disconfirmation (0-10)
> 15,000		
10,000 - 15,000		
6,000 - 10,000		
3,000 - 6,000		
1,000 - 3,000		
< 1,000		

PROPOSITION: WEATHER CONDITION IS BAD

Range of Wet Time (TW)	Degree of Confirmation (0-10)	Degree of Disconfirmation (0-10)
> 20 %		
15% - 20%		
10% - 15%		
5% - 10%		
< 5%		

Note: TW represents the percentage of time when the road surface is wet. It includes rainy, foggy, and snowy days.

PROPOSITION: GEOMETRIC CONDITION (GC) IS BAD

Body of Evidence	Scale of Confirmation (0-10)	Scale of Disconfirmation (0-10)
HC is slight.		
HC is moderate.		
HC is severe.		
VA is slight.		
VA is moderate.		
VA is severe.		
DD is slight.		
DD is moderate.		
DD is severe.		

Notes:

1. HC = horizontal curvature; VA = vertical alignment; and DD = driving difficulty. Detailed definitions of these three quantities are shown in tables A.1 through A.3 (Kulakowski et al. 1990b).

2. The scale of confirmation (or disconfirmation) represents an assigned degree of support to each body of evidence that has a (or has no) contributing effect on the proposition. It is a scale of increasing effect, that is, a rating of 10 represents the strongest effect on the proposition.

3. There is no compelling reason to assign a scale of disconfirmation (or confirmation) to each body of evidence, although it would be helpful. If you have no idea what rating to assign on this scale, please assign a rating of zero.

Table A.1. Horizontal curvature rating.

Criterion	Rating		
	Slight	Moderate	Severe
Warning Signs	No curve signs present	Curve signs with advisory speed plates	Presence of the following curve warning signs:
			W1-1R or 1L = Turn sign where recommended speed is 30 mi/h or less.
			W1-3R or 3L = Reverse turn sign used to mark two turns in opposite directions that are separated by a tangent of less than 600 ft.
			W1-5R or 5L = Winding road sign used where there are three or more curves separated by a tangent of 600 ft.
			W1-6 = Large arrow sign used to give notice of a sharp change of alignment in the direction of travel.
			W1-8 = Chevron alignment sign used to give notice of a sharp change of alignment with the direction of travel.
			W1-20R or 20L = Horseshoe curve sign used to mark a curve that produces a central angle of 135° or more. (Pennsylvania Title 67, Pub. 68 -- official traffic control devices)
Degree of Curvature	< 3°	4° - 8°	> 8°
Other	No evidence of braking or slowing down upon entering curve		Evidence of hard braking, tire markings, on pavement or shoulder while rounding the curve; or an unexpected, moderate curve by 1/2 mile or more of tangent/flat curves.

Table A.2. Vertical alignment rating.

CRITERIA	RATING		
	SLIGHT	MODERATE	SEVERE
Percent Gradient	Gently rolling, flat grades (< 2%)	Moderate grades (2% - 5%)	Steep grades (> 5%)
Available Sight Distance	Unlimited sight distance (> 1000 ft)	Somewhat restrictive sight distance (400 - 800 ft)	Very restrictive sight distances (< 400 ft)
Length of Grade	Length of grade has little effect on truck speeds (< 5 mi/h speed differential)	Length of grade has some effect on truck speeds 5 to 15 mi/h speed differential)	Length of grade has a *major* effect on truck speeds (> 15 mi/h speed differential)

Table A.3. Driving difficulty rating.

	Rating		
Criterion	Slight	Moderate	Severe
Access Control in Study Segment	Less than 10 access points per segment	Between 10 and 30 access points per segment	More than 30 access points per segment
Turn Lane Presence	Separate turn lanes or turns not permitted	Center lane left turn	Turns made from thru lanes
Surrounding Land Use	Primarily residential/farming land use	Residential/commercial land use	Commercial characteristics of strip shopping development in urbanized areas. In rough topography, characterized by farming activity along the roadside environment
Signalization	Uncontrolled intersections	Less than three signalized intersections within segment	Major intersections controlled by traffic signals at > 3 locations within segment

Appendix B

A GRAPHICAL INTERPRETATION OF THE KNOWLEDGE-BASED MODEL

In this appendix, figures B.1 through B.5 represent calculated belief intervals and belief function numbers for the propositions in the second and the third level of the knowledge-based model, respectively. Figure B.6 depicts the degree of belief of a significant causal factor, low skid resistance (SN), in the proposition of bad pavement condition (PC). For the proposition of bad geometric condition (GC), there is not much difference in the degree of belief between one factor and the next factor. Therefore, only figure B.7 that shows the degree of belief for driving difficulty (DD) is provided. For the factors in the proposition of bad traffic condition (TC), the factor of high percentage of wet time (TW) possesses a higher degree of belief than the factor of high average daily traffic (ADT). Figure B.8 shows the degree of belief for high TW.

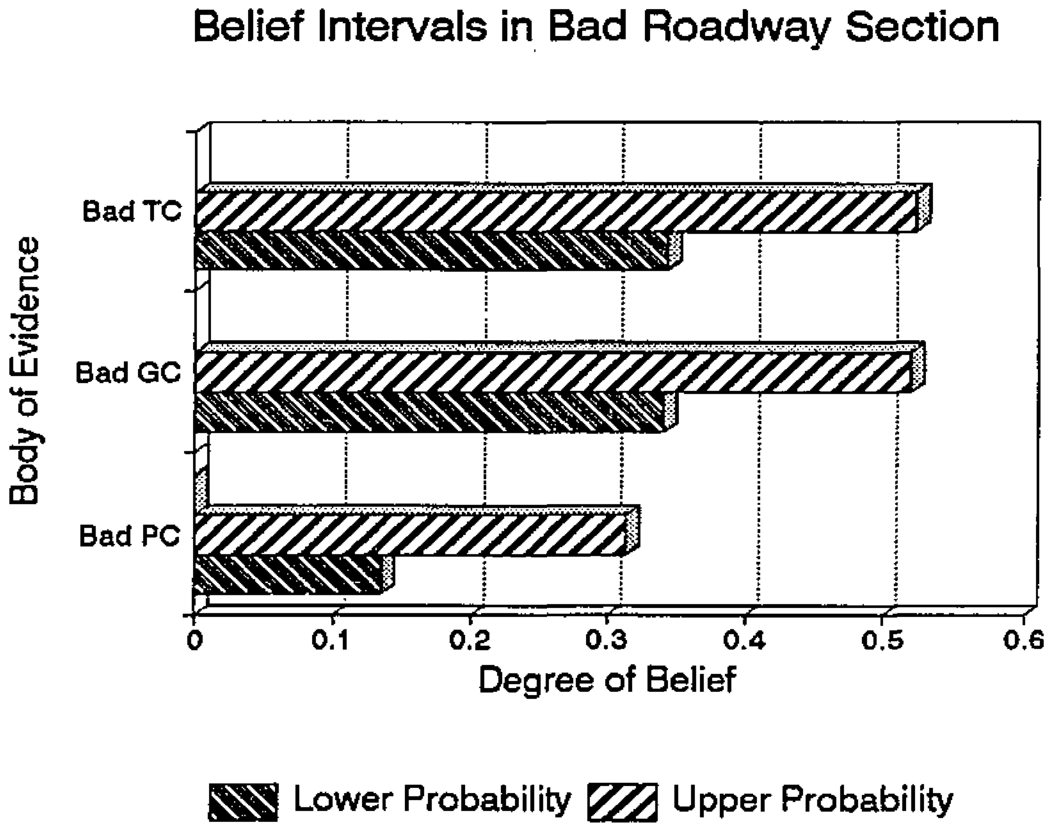


Figure B.1. Calculated belief-intervals for the factors in the proposition of bad roadway section.

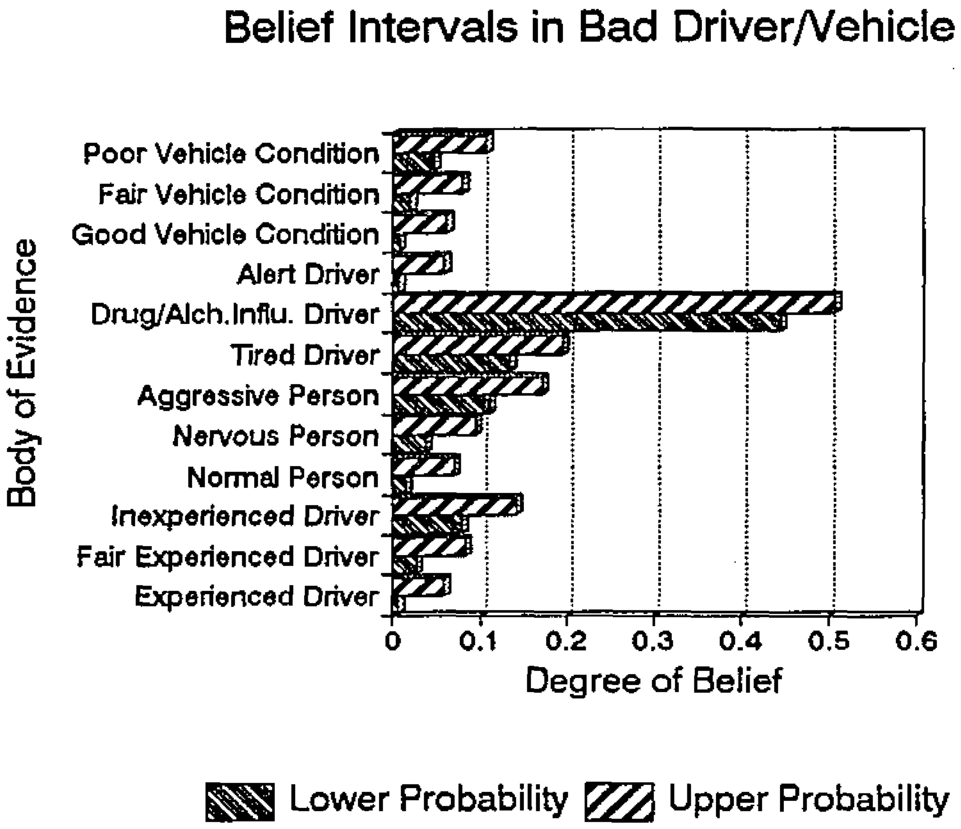


Figure B.2. Calculated belief intervals for the factors in the Proposition of bad driver/vehicle condition.

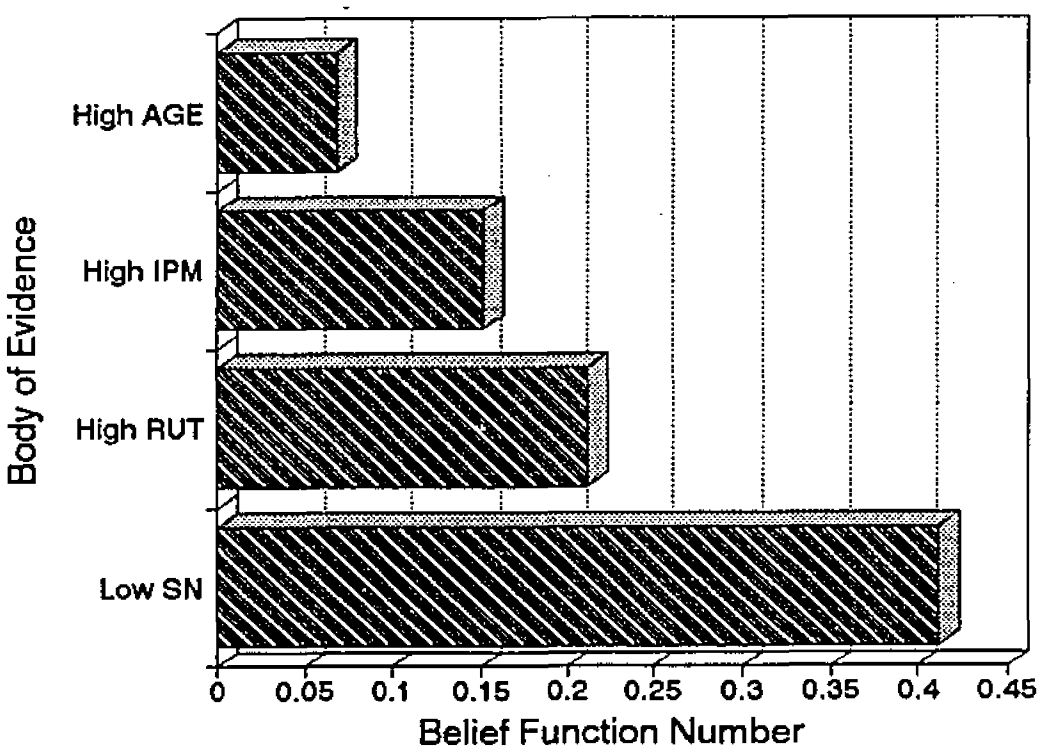


Figure B.3. Belief function numbers for the factors in the proposition of bad pavement condition.

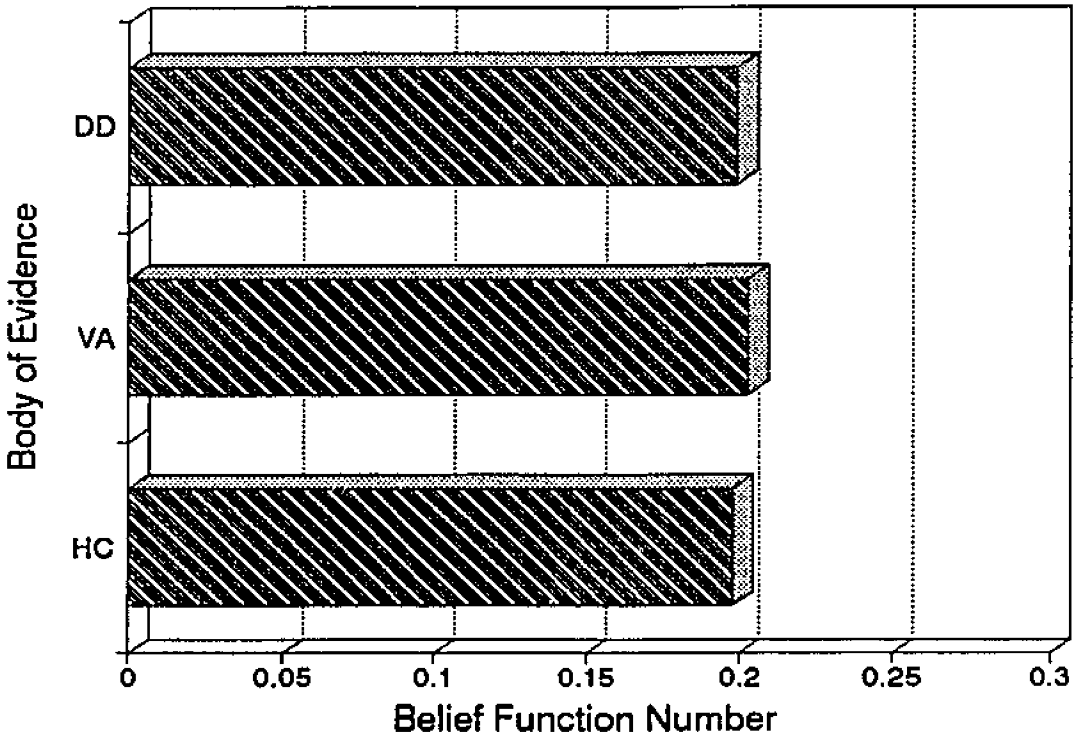


Figure B.4. Belief function numbers for the factors in the proposition of bad geometric condition.

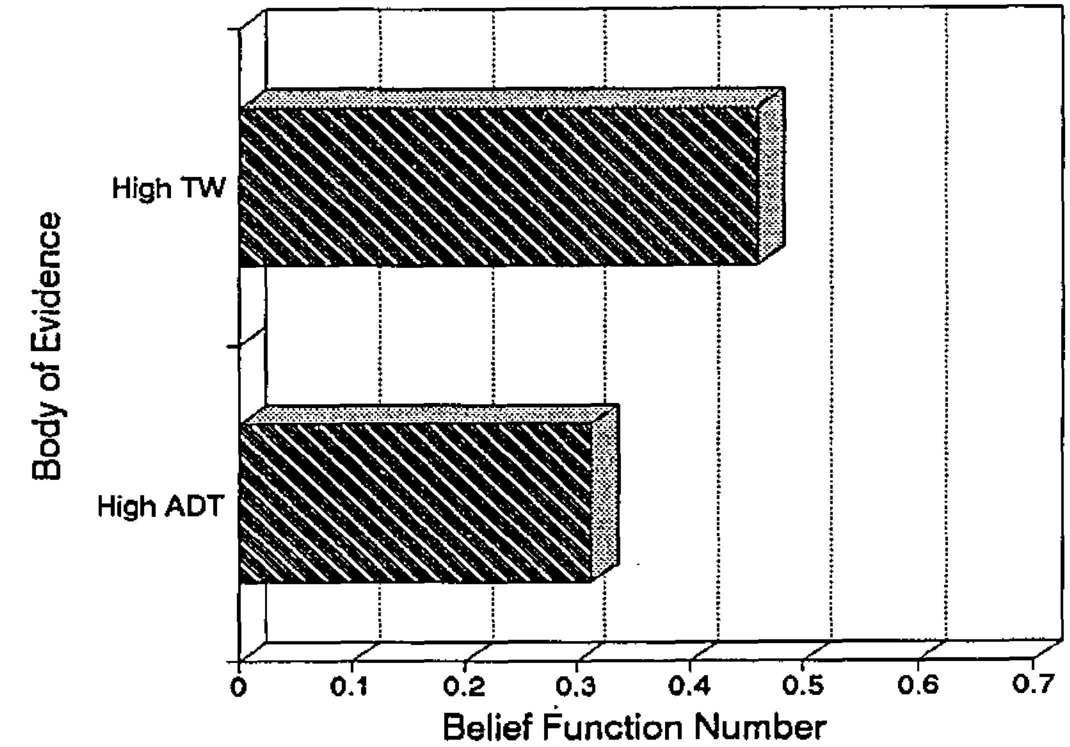


Figure B.5. Belief function numbers for the factors in the proposition of bad traffic condition.

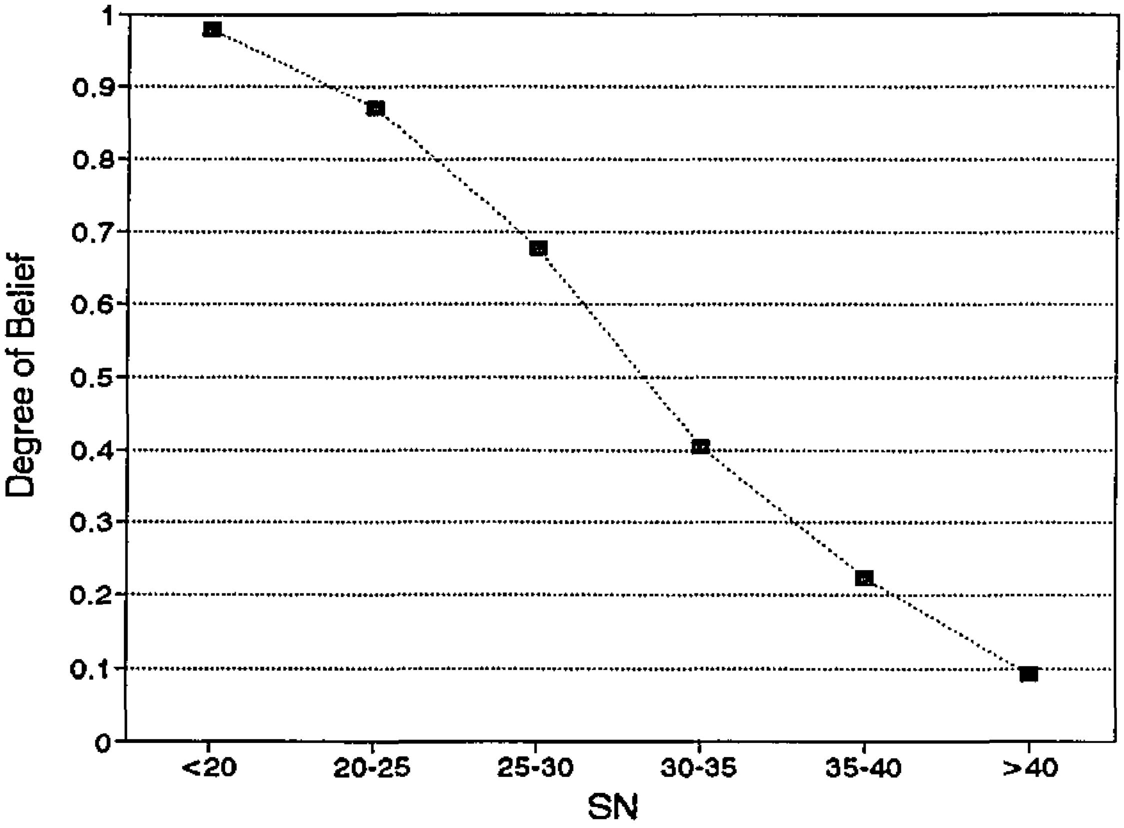


Figure B.6. Degree of belief for low skid resistance (SN).

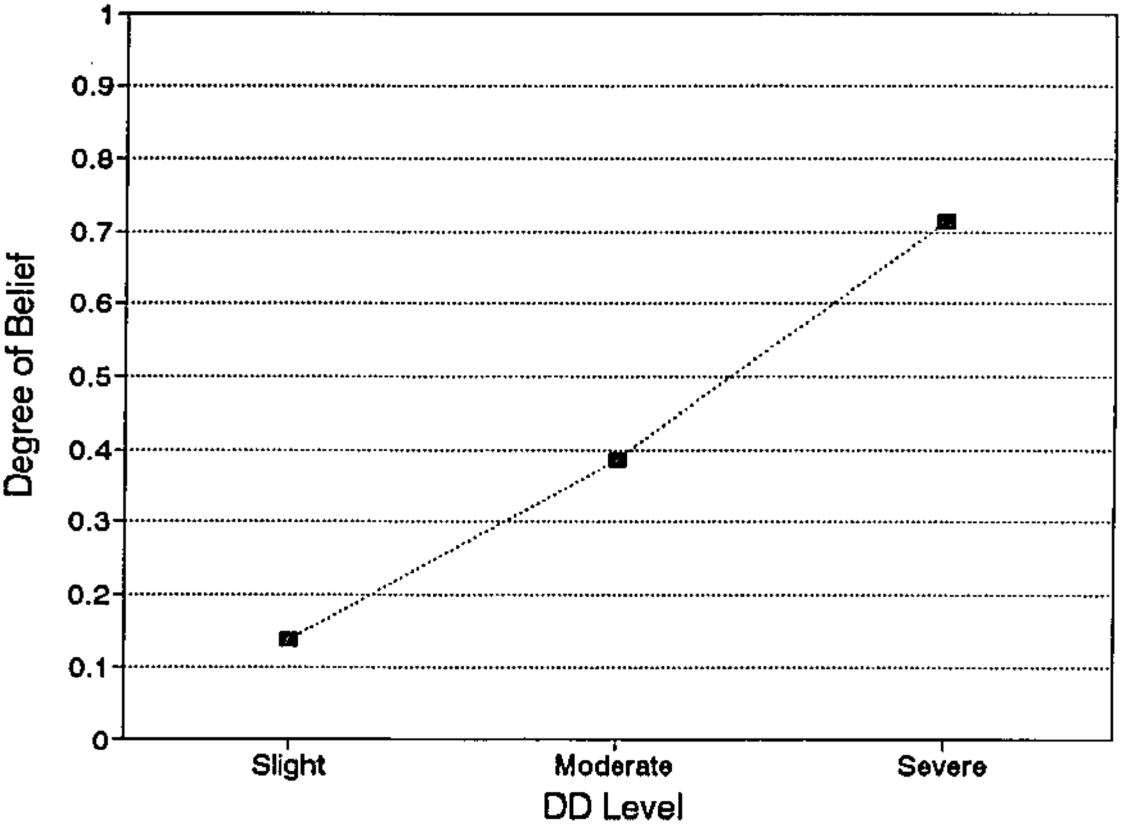


Figure B.7. Degree of belief for driving difficulty (DD).

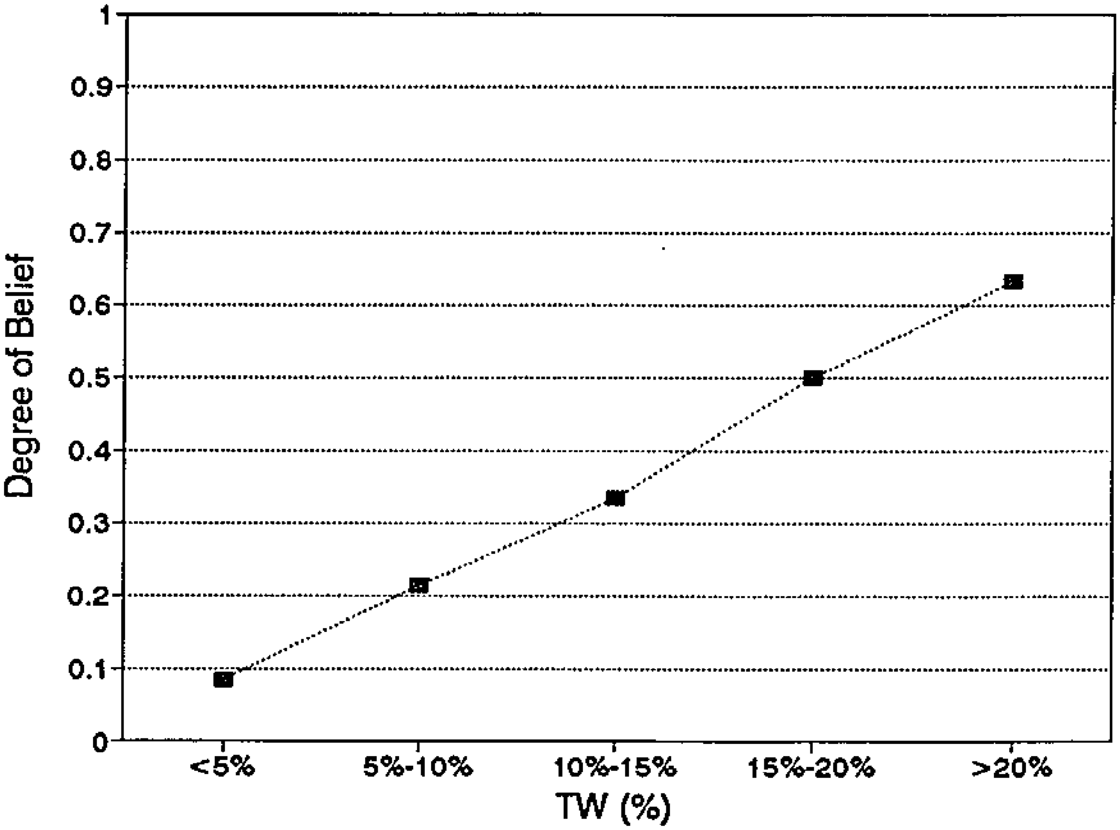


Figure B.8. Degree of belief for high percentage of wet time (TW).

Appendix C

A REGRESSION ANALYSIS OF THE REAL TRAFFIC ACCIDENT DATA SET

C.1. Linear Regression Results for 1983-1985 Data Set

- The regression equation for a total of 12 factors is

```
WA = 0.51 + 4.44 SL - 0.0538 SN + 0.317 RUT + 0.00157 IPM - 0.00234 AGE + 0.289 TW - 0.0532 PS
+0.000114 ADT - 0.0519 TP - 0.070 HC - 0.175 VA + 0.335 DD

Predictor         Coef         Stdev     t-ratio        p
Constant         0.508         1.959        0.26    0.796
SL               4.439         1.038        4.28    0.000
SN            -0.05380       0.01591       -3.38    0.001
RUT             0.3174        0.6801        0.47    0.641
IPM           0.001569      0.003962        0.40    0.692
AGE          -0.002343      0.008887       -0.26    0.792
TW              0.2893        0.1673       1.73    0.085
PS            -0.05322       0.01611       -3.30    0.001
ADT         0.00011360    0.00002432        4.67    0.000
TP            -0.05194       0.01837       -2.83    0.005
HC             -0.0701        0.1722       -0.41    0.684
VA             -0.1748        0.1963       -0.89    0.374
DD              0.3354        0.1835        1.83    0.069

s = 2.198        R-sq = 26.2%     R-sq(adj) = 23.2%
```

- Analysis of Variance

```
SOURCE        DF          SS          MS         F        p
Regression    12     505.555      42.130      8.72    0.000
Error        295    1424.715       4.830
Total        307    1930.270

SOURCE        DF      SEQ SS
SL             1      57.868
SN             1      39.941
RUT            1       1.289
IPM            1       0.460
AGE            1      17.902
TW             1      34.835
PS             1      90.421
ADT            1     190.411
TP             1      48.760
HC             1       3.593
VA             1       3.933
DD             1      16.141
```

- Lack of fit test

```
Possible interactions with variable ADT (P = 0.011)
Possible interactions with variable TP  (P = 0.063)
Possible lack of fit at outer X-values (P = 0.000)
Overall lack of fit test is significant at P = 0.000
```

- STEPWISE REGRESSION OF WA ON 12 PREDICTORS, WITH N = 308

```
    STEP          1         2         3         4         5         6         7
CONSTANT     1.1693    3.6024    1.7683    3.5882    3.6344    1.1236   -0.1319

ADT         0.00014   0.00014   0.00014   0.00014   0.00014   0.00014   0.00012
T-RATIO        6.17      6.46      6.72      6.76      6.93      6.71      5.72

PS                     -0.057    -0.061    -0.067    -0.058    -0.062    -0.054
T-RATIO                 -3.68     -4.00     -4.43     -3.86     -4.09     -3.42

SL                                  4.0       4.3       4.4       4.3       4.4
T-RATIO                            3.79      4.14      4.25      4.16      4.25

SN                                         -0.045    -0.047    -0.060    -0.056
T-RATIO                                     -3.20     -3.41     -3.96     -3.69

TP                                                   -0.057    -0.058    -0.057
T-RATIO                                               -3.17     -3.25     -3.21

TW                                                               0.31      0.33
T-RATIO                                                          2.01      2.18

DD                                                                         0.34
T-RATIO                                                                    1.86

S              2.37      2.32      2.27      2.24      2.21      2.19      2.19
R-SQ          11.08     14.85     18.69     21.35     23.88     24.90     25.75
```

- STEPWISE REGRESSION OF WA ON 10 PREDICTORS, WITH N = 308

```
    STEP          1         2         3         4
CONSTANT      1.112     3.702     5.661     5.732

WM             1.50      1.56      1.58      1.62
T-RATIO        7.47      7.92      8.15      8.49

PS                     -0.061    -0.067    -0.058
T-RATIO                 -4.04     -4.47     -3.86

SN                               -0.045    -0.048
T-RATIO                           -3.25     -3.50

TP                                         -0.062
T-RATIO                                     -3.48

S              2.31      2.25      2.22      2.18
R-SQ          15.42     19.71     22.41     25.39
```

Based on the above results, the statistically significant attributing factors are: WM, ADT, SN, PS, TP, DD.

- The regression equation for the selected factors is

```
WA = 5.05 + 1.54 WM - 0.0443 SN - 0.0520 PS - 0.0607 TP + 0.216 DD

Predictor         Coef        Stdev     t-ratio        p
Constant         5.048        1.040        4.85    0.000
WM              1.5440       0.1991        7.75    0.000
SN            -0.04434      0.01393       -3.18    0.002
PS            -0.05198      0.01558       -3.34    0.001
TP            -0.06073      0.01774       -3.42    0.001
DD              0.2158       0.1773        1.22    0.225
```

```
s = 2.178        R-sq = 25.7%     R-sq(adj) = 24.5%
```

- Analysis of Variance

```
SOURCE        DF          SS          MS         F        p
Regression     5     497.031      99.406     20.95    0.000
Error        302    1433.239       4.746
Total        307    1930.270

SOURCE        DF      SEQ SS
WM             1     297.556
SN             1      36.643
PS             1      98.369
TP             1      57.438
DD             1       7.025
```

- Lack of fit test

```
Possible interactions with variable WM (P = 0.018)
Possible interactions with variable PS (P = 0.000)
Possible interactions with variable TP (P = 0.000)
Possible interactions with variable DD (P = 0.073)
Possible lack of fit at outer X-values (P = 0.000)
Overall lack of fit test is significant at P = 0.000
```

C.2. Nonlinear Additive Regression Results for 1983-1985 Data Set

- The regression equation of the selected factors is

```
 WA = 6.72 - 0.0957 SN - 0.0715 PS - 0.127 TP + 0.00860 SN*DD + 0.00061 SN*PS+ 0.00173SN*TP + 1.49
WM

Predictor         Coef        Stdev    t-ratio        p
 Constant        6.720        2.733       2.46    0.014
 SN           -0.09568      0.07139      -1.34    0.181
 PS           -0.07150      0.06220      -1.15    0.251
 TP           -0.12667      0.06669      -1.90    0.058
 SN*DD        0.008601     0.004595       1.87    0.062
 SN*PS        0.000609     0.001634       0.37    0.710
 SN*TP        0.001727     0.001662       1.04    0.300
 WM             1.4903       0.2007       7.42    0.000

 s = 2.174        R-sq = 26.6%     R-sq(adj) = 24.8%
```

- Analysis of Variance

```
SOURCE        DF          SS          MS         F        p
Regression     7     512.646      73.235     15.50    0.000
Error        300    1417.623       4.725
Total        307    1930.269

SOURCE        DF      SEQ SS
SN             1      31.576
PS             1      73.544
TP             1      42.462
SN*DD          1      95.798
SN*PS          1       0.155
SN*TP          1       8.661
WM             1     260.450
```

- Lack of fit test

```
Possible interactions with variable PS (P = 0.000)
Possible interactions with variable TP (P = 0.000)
Possible interactions with variable SN*TP (P = 0.001)
Possible interactions with variable WH (P = 0.034)
Possible lack of fit at outer X-values (P = 0.000)
Overall lack of fit test is significant at P = 0.000
```

C.3. Linear Regression Results for 1986-1988 Data Set

- The regression equation of a total 12 factors is

```
WA = - 6.15 + 3.76 SL - 0.0609 SN + 0.486 RUT + 0.00380 IPM + 0.00674 AGE + 0.695 TW - 0.0536 PS
+0.000151 ADT - 0.0236 TP + 0.073 HC - 0.105 VA + 0.174 DD

Predictor        Coef        Stdev     t-ratio       p
Constant       -6.150        3.268       -1.88   0.061
SL              3.759        1.038        3.62   0.000
SN           -0.06093      0.01428       -4.27   0.000
RUT            0.4860       0.6781        0.72   0.474
IPM          0.003802     0.003851        0.99   0.324
AGE          0.006743     0.008901        0.76   0.449
TW             0.6955       0.2191        3.17   0.002
PS           -0.05359      0.01605       -3.34   0.001
ADT        0.00015108   0.00002408        6.27   0.000
TP           -0.02359      0.01890       -1.25   0.213
HC             0.0734       0.1719        0.43   0.670
VA            -0.1053       0.1973       -0.53   0.594
DD             0.1735       0.1832        0.95   0.344

s = 2.191        R-sq = 27.9%     R-sq(adj) = 25.0%
```

- Analysis of Variance

```
SOURCE        DF         SS          MS          F       p
Regression    12    549.133      45.761       9.54   0.000
Error        295   1415.786       4.799
Total        307   1964.919

SOURCE        DF     SEQ SS
SL             1     31.362
SN             1     73.896
RUT            1      3.696
IPM            1      0.054
AGE            1     24.861
TW             1     68.478
PS             1     64.834
ADT            1    267.291
TP             1      8.721
HC             1      0.194
VA             1      1.441
DD             1      4.305
```

- Lack of fit test

```
Possible curvature in variable VA (P = 0.019)
Possible lack of fit at outer X-values (P = 0.000)
Overall lack of fit test is significant at P = 0.000
```

- STEPWISE REGRESSION OF WA ON 12 PREDICTORS, WITH N = 308

```
    STEP         1        2        3        4        5
CONSTANT    0.6826   2.9337   5.3059  -3.5424  -5.8770

ADT        0.00015  0.00015  0.00015  0.00015  0.00015
T-RATIO       6.80     7.07     7.16     7.14     7.37

PS                   -0.053   -0.060   -0.057   -0.061
T-RATIO               -3.40    -3.91    -3.81    -4.15

SN                            -0.054   -0.058   -0.064
T-RATIO                        -3.82    -4.16    -4.61

TW                                       0.69     0.75
T-RATIO                                  3.53     3.91

SL                                                 3.8
T-RATIO                                           3.71

S             2.36     2.32     2.27     2.23     2.19
R-SQ         13.11    16.29    20.12    23.26    26.62
```

- STEPWISE REGRESSION OF WA ON 10 PREDICTORS, WITH N = 308

```
    STEP         1        2        3        4
CONSTANT    0.6472   3.0174   5.4449   5.4969

WM            1.32     1.36     1.36     1.37
T-RATIO       7.73     8.11     8.30     8.40

PS                   -0.056   -0.063   -0.056
T-RATIO               -3.67    -4.21    -3.72

SN                            -0.056   -0.057
T-RATIO                        -4.01    -4.18

TP                                     -0.044
T-RATIO                                 -2.43

S             2.32     2.27     2.22     2.20
R-SQ         16.32    19.86    23.89    25.34
```

- The regression equation for the selected factors is

```
WA = 5.35 + 1.35 WM - 0.0568 SN - 0.0548 PS - 0.0434 TP + 0.047 DD

Predictor       Coef       Stdev     t-ratio       p
Constant       5.351       1.049        5.10   0.000
WM            1.3547      0.1708        7.93   0.000
SN          -0.05679     0.01405       -4.04   0.000
PS          -0.05481     0.01578       -3.47   0.001
TP          -0.04338     0.01795       -2.42   0.016
DD            0.0467      0.1795        0.26   0.795

s = 2.204       R-sq = 25.4%     R-sq(adj) = 24.1%
```

- Analysis of Variance

```
SOURCE        DF          SS          MS         F         p
Regression     5     498.256      99.651     20.52     0.000
Error        302    1466.663       4.856
Total        307    1964.919

SOURCE        DF      SEQ SS
MVM            1     320.709
```

```
SN              1       61.518
PS              1       87.147
TP              1       28.554
DD              1        0.328
```

- Lack of fit test

```
Possible interactions with variable PS (P = 0.010)
Possible interactions with variable TP (P = 0.011)
Possible lack of fit at outer X-values (P = 0.000)
Overall lack of fit test is significant at P = 0.000
```

C.4. Nonlinear Additive Regression Results for 1986-1988 Data Set

- The regression equation for the selected factors is

```
WA = 11.2 - 0.216 SN - 0.179 PS - 0.0954 TP + 0.00217 SN*DD + 0.00340 SN*PS + 0.00137 SN*TP + 1.36
WM

Predictor        Coef        Stdev     t-ratio        p
Constant       11.157        2.755        4.05    0.000
SN           -0.21563      0.07197       -3.00    0.003
PS           -0.17933      0.06283       -2.85    0.005
TP           -0.09541      0.06745       -1.41    0.158
SN*DD        0.002172     0.004627        0.47    0.639
SN*PS        0.003404     0.001649        2.06    0.040
SN*TP        0.001368     0.001679        0.81    0.416
WM             1.3592       0.1714        7.93    0.000

s = 2.191         R-sq = 26.7%     R-sq(adj) = 25.0%
```

- Analysis of Variance

```
SOURCE       DF          SS          MS         F        p
Regression    7     524.191      74.884     15.59    0.000
Error       300    1440.728       4.802
Total       307    1964.919

SOURCE       DF      SEQ SS
SN            1      65.042
PS            1      65.783
TP            1      25.129
SN*DD         1      45.783
SN*PS         1      14.300
SN*TP         1       6.210
WM            1     301.943
```

- Lack of fit test

```
Possible interactions with variable PS (P = 0.032)
Possible interactions with variable TP (P = 0.003)
Possible interactions with variable SN*TP (P = 0.019)
Possible lack of fit at outer X-values (P = 0.000)
Overall lack of fit test is significant at P = 0.000
```

C.5. Nonlinear Multiplicative Regression Results

The following results were obtained from the SAS NLIN procedure (Gauss-Newton method)

- The regression equation for the selected factors is

```
WA = 3318.25(WM**0.661)(DD**0.11385)(PS**(-1.387))/(SN**0.4476)
```

- Analysis of Variance

```
SOURCE        DF         SS         MS
Regression     4   1838.125    459.531
Error        303   1442.880      4.760
Total        307   1930.269

 The R-sq = 1 - Error/Total = 0.252
```

VITA

I was born on Nov. 28, 1954, at a small town in Taiwan. In July 1973, I graduated from the Taiwan Provincial Taichung First High School and enrolled in National Taiwan College of Marine Science and Technology at Keelung. My major was Marine Engineering. After graduating from college, I entered the Navy's Radar Corp. to serve as a Reserved Officer. Two years later, I joined the Taiwan Machinery Manufacturing Corporation at Kaohsiung. During my stay in the company, I was a design engineer for engine system, piping system, and HVAC (heating, ventilating and air conditioning) systems. In 1986, I decided to pursue graduate study in order to advance my knowledge. I enrolled at The Pennsylvania State University as a Master's student in Mechanical Engineering. I earned my Master of Science degree in 1988 and continued my study as a Doctoral student in Mechanical Engineering at The Pennsylvania State University. I am a member of ASME (American Society of Mechanical Engineers) and SAE (Society of Automotive Engineers) and I worked for the Pennsylvania Transportation Institute as a graduate assistant. My advisor is Dr. B. T. Kulakowski and my publications include:

1. Kulakowski, B. T., Henry, J. J., and Lin, C., " Development of a Closed-Loop Calibration Procedure for a British Pendulum Tester," American Society of Testing and Materials (ASTM) STP 1031, 1990.
2. Lin, C., Antle, C. E., and Kulakowski, B.T., "An Application and Evaluation of Empirical Bayes Methods for the Problem of Evaluating the Risk of Wet Pavement Accidents," submitted to the Journal of Accident Analysis and Prevention in August 1991.
3. Kulakowski, B. T. and Lin, C., "Effect of Design Parameters on Performance of Road Profilographs," Transportation Research Board, National Research Council, Washington, D.C. 1990, in press.