SINGLE AND MULTI-VEHICLE CRASH PREDICTION MODELS FOR TWO-LANE ROADWAYS

Raghubhushan K. Pasupathy
John N. Ivan
Paul J. Ossenbruggen

UNITED STATES DEPARTMENT OF TRANSPORTATION

REGION I UNIVERSITY TRANSPORTATION CENTER

PROJECT UCNR9-8

FINAL REPORT

February 24, 2000

Performed by

University of Connecticut
Connecticut Transportation Institute
Storrs, CT 06269-2037

and

University of New Hampshire
Department of Civil Engineering
Durham, NH 03824

Exhibit C

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
SINGLE AND MULTI-VEHICLE CRASH PREDICTION MODELS FOR TWO-LANE ROADWAYS	February, 2000
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Raghubhushan K. Pasupathy John N. Ivan Paul J. Ossenbruggen	NEUTC UCNR9-8

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
University of Connecticut Connecticut Transportation Institute Civil & Environmental Engineering, U-37 Storrs, CT 06269-2037	11. Contract or Grant No. DTRS95-G-0001

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
New England (Region One) UTC Massachusetts Institute of Technology 77 Massachusetts Avenue, Room 1-235 Cambridge, MA 02139	Final 1996/09/01-1998/05/31
	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by a grant from the US Department of Transportation, Univers ity Transportation Centers Program

16. Abstract

A scientifically based risk management method is needed for transportation decision-making processes. For instance, highway designers and planners use levels of service (LOS) and costs (construction and maintenance) as measures of effectiveness when evaluating competing alternatives. Since highway safety considerations are not empirically based, no predictive models exist. Therefore it is impossible to objectively evaluate or rank alternatives for safety in the design or planning processes. With today's methods, highway safety, or more appropriately, highway risk evaluation, typically becomes a concern only after construction when the consequences of unsafe design become vividly realized.

This work is aimed at fundamentally improving the way roadway risk evaluation is practiced. The roadway risk model described here is formulated using the same principles that have evolved over the past two decades for evaluating public health and environmental risks. The proposed model and the ones developed by EPA, OSHA and FDA use the same Measure of Effectiveness (MOE), the probability that an individual is exposed to some hazard and experiences some undesirable consequence from the exposure. This framework seems to be particularly apt under recent calls in the public arena for codifying cost-benefit-risk analysis into federal law and for fundamental changes in risk regulation.

A Poisson Regression (PR) model is proposed that incorporates the principles of traffic flow theory and the Highway Capacity Manual (HCM) level of service (LOS) ratings. This research could have a major impact in the manner in which risk analyses are performed and, in turn, in the manner in which highway safety is perceived and policy decisions are made.

17. Key Words	18. Distribution Statement
highway safety, risk, Poisson regression, public health, highway capacity, level of service, highway design, traffic exposure	No restrictions, This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified		

DISCLAIMER

This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers or University Research Institutes Program, in the interest of information exchange. The U. S. Government assumes no liability for the contents or use thereof.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Problem Statement

Motor vehicle crashes are one of the 10 leading causes of death in the United States ranking only after cancer, heart disease, stroke, and respiratory infections (The World Almanac and Book of Facts: 1996). About 450,000 people have died on our highways over the last decade (Table 1). Another 30 million were left injured in the 300 million crashes that occurred (Bureau of Transportation Statistics 1995). Americans spend over $500 billion per year on automotive transportation. They travel 2.1 trillion vehicle-miles and incur over $72 billion in crash costs every year (The World Almanac and Book of Facts: 1996). This $72 billion is almost equal to the total costs incurred by all levels of government for all highway construction, maintenance, and operations in one of the past years. The social and humanitarian costs are obviously also high - 60 percent of those dying and 70 percent of those injured are in the highly productive 15 to 45 year old age group (The World Almanac and Book of Facts: 1996). These figures portray the extent of the problem and the immediate need for corrective action.

Having recognized the need for corrective action, the first step towards an effective solution is to identify the primary causes for the current situation. The "nut behind the wheel" is usually perceived as the cause of most traffic crashes and it is not hard to find statistics to support this claim. This perception often leads to the assumption that crash countermeasures must concentrate on changes in driver attitudes and behavior. For instance, Henderson (1971) compared contemporary ways of thinking about traffic deaths and injuries to the way people used to think about cholera. Since cholera struck

Table 1

Fatality Rates on American Highways

Year	Fatalities	Fatality Rate per 100, 000 population	Vehicle Miles Traveled (Billions)	Fatality Rate per 100 Million VMT
1966	50,894	26.0	926	5.5
1967	50,724	25.7	964	5.3
1968	52,725	26.4	1016	5.2
1969	53,543	26.6	1062	5.0
1970	52,627	25.8	1110	4.7
1971	52,542	25.4	1179	4.5
1972	54,589	26.1	1260	4.3
1973	54,052	25.6	1313	4.1
1974	45,196	21.2	1281	3.5
1975	44,525	20.7	1328	3.4
1976	45,523	20.9	1402	3.2
1977	47,878	21.8	1467	3.3
1978	50,331	22.7	1545	3.3
1979	51,093	22.8	1529	3.3
1980	51,091	22.5	1527	3.3
1981	49,301	21.5	1553	3.2
1982	43,945	18.9	1595	2.8
1983	42,589	18.2	1653	2.6
1984	44,257	18.7	1720	2.6
1985	43,825	18.4	1774	2.5
1986	46,087	19.1	1835	2.5
1987	46,390	19.1	1921	2.4
1988	47,087	19.2	2026	2.3
1989	45,582	18.4	2096	2.2
1990	44,599	17.9	2144	2.1
1991	41,508	16.5	2172	1.9
1992	39,250	15.4	2240	1.8
1993	40,150	15.6	2297	1.7
1994	40,676	15.6	2347	1.7

Source: Bureau of Transportation Statistics (1995).

mostly poor people, the disease was attributed to their undesirable ways of life. People believed that Cholera would be obliterated only if the poor would change their ways. Henderson points out that control of the environment (i.e., purified water and construction of sewage systems), not changes in behavior, brought the disease under control. He concludes that focusing too much on the driver as the cause, and therefore the solution to crashes, often masks our ability to see other changes that could reduce injuries and deaths.

It is important to realize that crashes are generally the result of bad decisions by the driver made in an environment created by the engineer. Thus, the engineer has a good deal of influence on the likelihood of a driver making a bad decision. This prompted Anderson (1976) to say that engineers could attack the lion's share of the safety problem if they got beyond the driver error myth. Hence, the role of transportation engineers for setting a stage for the safe and efficient movement of goods and people cannot be overstated.

The Need for Highway Safety Research

Transportation Engineering can be defined as the study of the complex interactions among drivers, vehicles and roadways. Crashes result when there is a conflict between at least two of these elements. Hence, if one is to create an environment to reduce the likelihood of conflicts among these elements, one must have a thorough knowledge of the complex relationships.

Highway safety research aims at understanding these relationships better so that the basic objective of safety can be satisfied while still serving traffic demand. This is

achieved by trying to identify the elements responsible for the occurrence of crashes and by learning how to predict future crash occurrences. At the same time, the last decade has seen a decreasing trend in resources for highway improvements, whereas traffic demand continues to increase. Hence, there is a strong need for the judicious use of existing resources. Safety research helps to do this by identifying those features which, when improved, give the maximum safety benefit for the resources invested. These reasons make safety research indispensable under the present circumstances.

Since travel demand has risen consistently through the years (Table 1), one might expect highways in America to be much more dangerous now than in the past. Yet, statistics show that this is not true and that America's highways are much safer today than ever. Table 1 shows the trend in total number of fatalities and their corresponding rates on a yearly basis. Though the exposure is increasing, it can clearly be seen that rates are decreasing consistently over the years and so are the number of fatalities. For example, about 46,000 people died in crashes on America's highways in the year 1990. If however, fatal crashes occurred in 1990 at the rate they occurred in 1967, there would have been almost 120,000 deaths (Hensing 1991). This increased safety can be attributed to better highway planning and design, improved traveler information systems and innovative road safety devices - all the results of highway safety research in combination with advances in medical care and vehicle safety features.

Past research has given us information in a number of areas that has proved vital for the construction of safe roads in an economical way. For example, research has revealed that crash rates decrease with an increase in lane width up to about 12 feet (Zegeer 1981), after which the safety gains are either negative or very marginal. This

finding has been adopted as standard practice in the design of most modern roads. Apart from identifying the elements that are directly related to crash occurrence, research has also given us an idea of the typical conditions under which crashes occur. For instance, research shows that 70 percent of fatal crashes take place at night, of which 60 percent involve only one vehicle (Polanis 1995). This can be useful information when deciding what counter measures need to be taken.

Thus, the past two decades have seen important progress in providing a safer roadway environment for American road users. However, as statistics suggest, there is still room for improvement. We do not yet have a complete understanding of the nature of crashes and what causes them. Complete understanding of the nature of crashes could prove extremely useful in further reducing the crash occurrence on our roads.

Objectives and Scope

Highway safety research, as the above discussion suggests, could mean a very broad range of topics, of which crash prediction methods are a small but important part. It essentially deals with quantifying the relationship between the crashes observed at a site and the existing traffic and geometric conditions. These prediction models can give us an idea of which are the important variables and how much each of them contribute to causing the observed crashes at a site.

Traditionally, crash prediction models have tended to be macroscopic in nature. For instance, researchers have tended to use summary statistics on traffic such as annual average daily traffic (AADT), rather than microscopic measures such as hourly volume counts. Another example of a macroscopic model is one that considers all crashes

together rather than splitting crashes by type. Microscopic data such as hourly volume counts have generally been used only to quantify relationships between crashes and traffic flow at a single site. In the few cases in which researchers have used hourly volumes or have split crashes by type, importance has not been given to the geometric effects and hence, these studies cannot be applied to all sites. A useful study would look at how both the hourly volume counts and the geometric effects jointly contribute to the observed crashes at various sites.

Such is the focus of this thesis which incorporates traffic condition variables (level of service) along with geometric characteristics to predict crash frequencies at eight different two-lane state highway locations across the state of Connecticut. In order to do this, hourly volume counts on every day for the period between October 1990 to September 1996 were obtained for each of these eight sites. Poisson Regression was used to build separate models to predict single and multi-vehicle crash frequencies. It was found that traffic and geometric characteristics affect these two crash types in very different ways and hence a split such as this is very much warranted. For instance, it was found that while single vehicle crashes tend to decrease with an increase in shoulder width, the trend is reversed with multi-vehicle crashes. Similarly, level of service (LOS) seems to be a much more important predictor variable for single vehicle crashes than for multi-vehicle crashes.

Chapter 2 of this thesis is a literature review that traces a number of important studies that have contributed to knowledge in crash prediction and methods. The third chapter elaborates on the database and some of its limitations. Poisson regression is discussed in Chapter 4 with a brief account of the nature of the distribution and its

assumptions. Chapter 5 describes the model form used, its advantages and the preliminary models that were estimated. The final models are established and discussed in detail in Chapter 6. Chapter 7 concludes the thesis by giving the important conclusions from this study and suggesting future areas of research. It also includes a section that talks about the database and how it could be enhanced.

Chapter 2: Literature Review

On Relating Crashes to Traffic Flow

In performing an experiment, the number of successes achieved largely depends on the number of trials performed. Scientists call this "exposure," or the number of opportunities for success or failure. This can be extended to predicting crashes on a highway. The number of crashes observed on a highway naturally depends on the amount of traffic flowing on it. However, this relationship seems to be complex and is not understood thoroughly. Hence, we cannot assume a priori that the number of crashes observed is linearly related to the number of vehicles counted (like with a roll of dice).

Increasing traffic on a roadway can increase the chance of a crash disproportionately. First, the fact that there are more vehicles on the road means that we are increasing the number of trials performed, and hence there is a greater likelihood of a crash occurring. Additionally, since there are more vehicles on the road, there are more interactions between vehicles. This further increases the likelihood of a crash occurring. A number of researchers have investigated this complex interaction in the past.

One of the first such studies was by Gwynn (1967) who analyzed crashes and traffic flow on U.S Route 22 through the city of Newark, New Jersey. Hourly volumes on every day between the years 1959 and 1963 were classified into 100 volume ranges by magnitude. Crash rates were computed and plotted against volume class. He found a distinct "U" relationship, with more crashes observed at the higher and lower traffic volumes. It is important to note that absolute volumes were used in this study rather than a measure of congestion such as volume/capacity ratio.

Zhou and Sisiopiku (1997) performed a similar study on Interstate 94 in Michigan. This study was slightly different from the previous one in that it included volume/capacity (v/c) ratio instead of the absolute traffic volume. The results were very similar to that of Gwynn, as they found a distinct "U" relationship between v/c ratio and crash rates. In this case, since only one roadway segment was considered, v/c ratio is probably just the same as considering the absolute traffic volume where the capacity is not variable.

An interesting extension of these studies would be to consider multiple roadway segments in the study. In such a case, the absolute traffic volume might not be a good measure in predicting crashes since different segments can have different capacities. Instead, as Frantzeskakis (1983) suggests, congestion measures such as the v/c ratio or the level of service (LOS) may be better predictor variables.

Model Form

From the studies summarized above, it appears that the crash rate does not remain constant with changes in traffic flow. This suggests that a linear model might not be suitable for predicting crash frequencies. Apart from this, a number of other arguments have been made against linear regression models. Jovanis and Chang (1986) question the application of a continuous distribution such as the normal distribution in modeling the occurrence of crashes, a discrete process. Crash data are typically heteroscedastic (i.e., variance of the residual is not constant) and hence violate the equality of variance assumption of a linear model. Even variance stabilizing transformations (modeling a rate instead of the number of crashes) are shown to violate these assumptions regarding

variance and give inaccurate estimates. Finally, there is a chance that a linear model will predict negative crash rates, which are meaningless.

As an alternative to linear regression, Jovanis and Chang suggest Poisson regression for predicting crashes. The Poisson distribution is a right skewed distribution that replicates the occurrence of rare events such as crashes better than a normal distribution. Joshua and Garber (1990) used linear and Poisson regression models in predicting truck crashes in the state of Virginia. The study included a critical comparison of the two models. It was found through this study that Poisson models describe the relationship between truck crashes and associated traffic and geometric variables better than the linear models. A number of other studies (e.g., Miaou et al. 1992) concluded likewise and it is now widely accepted that Poisson regression is the most appropriate tool for crash modeling.

Extensions of the Poisson Model

A central assumption of the Poisson regression model is that the mean and variance of the distribution representing the error are equal. However, crash data frequently exhibit variances greater than the mean; this condition is called "over-dispersion." Miaou and Lum (1993) attributed this condition to possible inaccuracies in exposure data and to omission of some important variables in the model. As a result of over-dispersion, tests of the significance of variables in the models are rendered inaccurate, though the coefficient estimates remain reliable (Agresti 1990).

A number of solutions have been proposed for the over-dispersion problem but the one that is becoming increasingly popular is the use of a Negative Binomial model.

The Negative Binomial model is a simple extension of the Poisson model in that it relaxes the assumption regarding the equality of variance (Hadi et al. 1995). Hadi et al. used the Negative Binomial and Poisson models in estimating the effects of cross section design elements on crash rates. Negative Binomial models were shown to describe crash rates more effectively than the corresponding Poisson models. Over-dispersion can also be corrected using the "over-dispersion parameter" (Agresti, 1990). This parameter is defined as the ratio of the variance to the mean. To correct over-dispersion, Agresti suggests dividing the t-statistics by the square root of the over-dispersion parameter.

Geometric Variables

Among the most useful features of a good crash prediction model is its ability to tell us how much each of the variables present in the model contributes to the observed crashes. In most models, exposure is the most important variable in explaining the variation in crashes. However, a part of the variation is also explained by the geometric characteristics of the site. Apart from systemic influences, if one is interested in reducing crashes at a site, one can do so only by controlling the geometric factors. Hence, a thorough understanding of the contribution of geometric elements to crashes is necessary.

One of the most important early studies was that of Dart and Mann (1970). This study was aimed at identifying causes for high crash rates in Louisiana. Though this study applied linear regression models in trying to explain crashes, some important findings surfaced. For instance, it was found for the first time that cross slope was important, and crashes are associated with poor drainage. This is a useful finding for places like Louisiana, where the rainfall is heavy.

Zeeger et al. (1981) studied the effects of lane and shoulder width on crashes. This study was limited to two-lane roadways. It was found that lane width had a marked effect on crashes. Crash rates decreased with an increase in lane width until a width of about 12-ft, after which the rates started increasing again. The same trend was also observed with shoulder width though to a much lesser extent. A number of other studies have confirmed this trend (Milton and Mannering 1997).

Ivan and O'Mara (1997) studied two-lane roadways in the state of Connecticut. Just like a number of earlier studies, they found the frequency of intersections on the segment to be one of the most important predictor variables. This is probably because as the frequency of the intersections increases, so does the conflict opportunity. However, it is important to realize that crashes occurring at mid-block may be completely different from those at an intersection. The two locations can be operating at entirely different v/c ratios and exposure levels, and crashes can be caused by very different geometric characteristics (Frantzeskakis 1983). Bared and Vogt (1997) studied mid-block crashes apart from intersection crashes using different exposure variables for both. The measure used for the intersection exposure was the product of the major and minor AADTs. This study reported satisfactory results for both models.

Due to the research done over the last decade, we now have a fair idea of how the geometric elements on a highway influence the crashes occurring on it. For instance, the effects of shoulder width, lane width and access points are quantifiable to a fair degree of accuracy. What is needed now is modeling at a greater detail. Fortunately, there has been a trend toward microscopic modeling. In other words, researchers are increasingly modeling two-lane roads apart from multi-lane roads and separating single vehicle

crashes from multi-vehicle crashes. This is because these crashes seem to have very different causes (Persaud and Mucsi 1995). Persaud and Mucsi used derived hourly volumes to test the effect of light conditions on the two types of crashes. The study found that light conditions affect each of these types very differently. For instance, multi-vehicle crashes were found to be occurring during the daytime when the light conditions were good whereas the single-vehicle crashes were more likely to occur after sunset. Also, it was found that while single vehicle crashes were associated with narrow lanes and shoulders, multi-vehicle crashes were associated with wider lanes and shoulders. Hence, Persaud and Mucsi suggest that each of these types be modeled separately.

An important drawback of most of these models is the lack of information about the traffic conditions under which the crashes occurred. This information can be extremely vital in explaining crashes and forms the first step if one is looking to model in greater detail. For instance, by just considering AADT, two roadways with the same AADT and geometric conditions but very different peaking characteristics cannot be differentiated. In such a case, a model considering just the AADT will predict the same crash rates at both these sites. However, the peaking characteristics can be captured easily if one has accurate hourly volume information and hence the exact LOS at which the trips were made. Secondly, incorporating this information about the traffic conditions along with the other geometric variables can bring out the true effect of the geometric variables. Since crashes are usually caused by the combined effect of traffic and geometric characteristics, absence of vital traffic conditions can confound the geometric effects.

Proposed Extensions

From a number of recent studies (Bared and Vogt, 1997) it appears that modeling in greater detail is much more powerful in bringing out the effects of specific variables. Another drawback of traditional models is that the analysis is localized to only the times and conditions when a crash occurred. For instance, in trying to study the effect of surface conditions on crash occurrence, traditional models have information on the surface condition only at the time when the crash occurred. In such a case, one does not have an idea of the number of successful trips made under the same surface conditions. However, it is important to have both the number of successful trips and the number of failed trips under specified conditions, for a better model.

The present study tries to incorporate a number of the above mentioned elements. Unlike a lot of other studies it was not assumed a priori that there exists a linear relationship between traffic volume and crash frequency. Instead, the exponent on the exposure term is estimated along with the other parameters in the model. In an attempt at microscopic modeling, multi-vehicle and single-vehicle crashes are investigated separately and accurate information about traffic conditions is incorporated in the form of LOS variables. Another important feature of this study is that all trips at a site are given equal importance. In other words, the analysis is not limited to those trips that ended in a crash. Finally, for the first time, LOS variables are used in combination with geometric variables in modeling crash occurrence. Since it was thought that crash occurrence on two-lane roads can be completely different from that on multi-lane roads, these types had to be separated while modeling crashes. This study considers only two-lane roads.

Chapter 3: The Database

The database used in this study consists of five individual smaller data sets obtained from a number of different sources. Following is a list of these data sets, which are subsequently discussed in detail.

1. ATR Database
2. HPMS Database
3. Crash Database
4. Light Database
5. Precipitation Database

ATR Database:

The Connecticut Department of Transportation (ConnDOT) maintains thirty-seven continuous count stations called automatic traffic recorders (ATRs) to record traffic data through the entire year. The ATR locations are essentially spot locations on the roadway, each of which is given a station number. Figure 1 shows the location of these sites. Since hourly volume counts are otherwise difficult to obtain, the ATR data predominantly dictated the sites to be chosen for the study. All ATR data were available in the form of ASCII text files. A page from a sample record of the ATR data is shown in Figure 2 and some important variables are explained below:

1. Station Identification: As mentioned above, each of the stations is given a station number by which it can be identified.

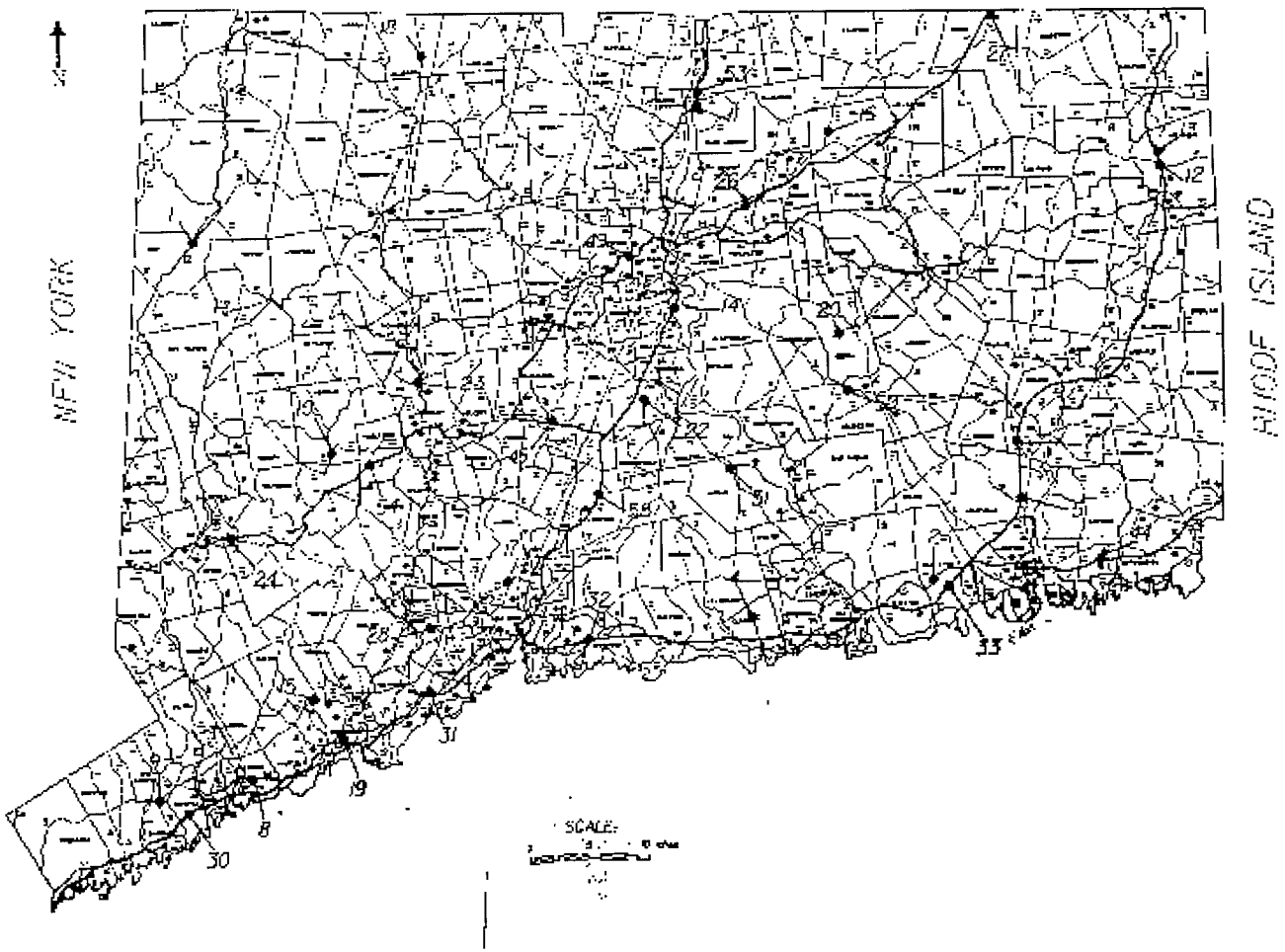

Figure1
Location of the Continuous Stations
Source: HPMS Database

Column	Field Length	Alpha/ Numeric	Description	Page
1	1	N	Record Type	6-3-1
2-3	2	N	FIPS State Code	6-2-1
4-5	2	N	Functional Classification	6-2-4
6-11	6	A	Station Identification	6-2-3
12	1	N	Direction of Travel	6-2-3
13	1	N	Lane of Travel	6-2-3
14-15	2	N	Year of Data	6-2-3
16-17	2	N	Month of Data	6-3-1
18-19	2	N	Day of Data	6-3-2
20	1	N	Day of Week	6-3-2
21-25	5	N	Traffic Volume Counted, 00:01 - 01:00	6-3-2
26-30	5	N	Traffic Volume Counted, 01:01 - 02:00	6-3-2
31-35	5	N	Traffic Volume Counted, 02:01 - 03:00	6-3-2
36-40	5	N	Traffic Volume Counted, 03:01 - 04:00	6-3-2
41-45	5	N	Traffic Volume Counted, 04:01 - 05:00	6-3-2
46-50	5	N	Traffic Volume Counted, 05:01 - 06:00	6-3-2
51-55	5	N	Traffic Volume Counted, 06:01 - 07:00	6-3-2
56-60	5	N	Traffic Volume Counted, 07:01 - 08:00	6-3-2
61-65	5	N	Traffic Volume Counted, 08:01 - 09:00	6-3-2
66-70	5	N	Traffic Volume Counted, 09:01 - 10:00	6-3-2
71-75	5	N	Traffic Volume Counted, 10:01 - 11:00	6-3-2
76-80	5	N	Traffic Volume Counted, 11:01 - 12:00	6-3-2
81-85	5	N	Traffic Volume Counted, 12:01 - 13:00	6-3-2
86-90	5	N	Traffic Volume Counted, 13:01 - 14:00	6-3-2
91-95	5	N	Traffic Volume Counted, 14:01 - 15:00	6-3-2
96-100	5	N	Traffic Volume Counted, 15:01 - 16:00	6-3-2
101-105	5	N	Traffic Volume Counted, 16:01 - 17:00	6-3-2
106-110	5	N	Traffic Volume Counted, 17:01 - 18:00	6-3-2
111-115	5	N	Traffic Volume Counted, 18:01 - 19:00	6-3-2
116-120	5	N	Traffic Volume Counted, 19:01 - 20:00	6-3-2
121-125	5	N	Traffic Volume Counted, 20:01 - 21:00	6-3-2
126-130	5	N	Traffic Volume Counted, 21:01 - 22:00	6-3-2
131-135	5	N	Traffic Volume Counted, 22:01 - 23:00	6-3-2
136-140	5	N	Traffic Volume Counted, 23:01 - 24:00	6-3-2
141	1	N	Footnotes	6-3-2

Figure 2

Sample Record of ATR Data

2. Direction of Travel: The volume is recorded for each direction of travel on the roadway. Directions are given in codes as defined by ConnDOT.
3. Year, Month, Date and Day: This is the exact date and day (day of the week) on which the count was taken is contained in this field.
4. Traffic Volume: These variables contain the actual volume counted for each of the 24 hours on the specified date.

ATR data were obtained for each of the eight two lane ATR locations for the period from October 1990 to September 1996. This database was very important since it provided the exact exposure for each hour during this period on each of the roadways considered. Also, by matching the crash database (to be described later) with the ATR database by site, date and time, we could get an idea of the traffic conditions under which each of the crashes occurred.

HPMS Database:

The geometric characteristics of the sites under study are contained in this database. ConnDOT maintains Highway Performance Monitoring Sites (HPMS) all over the state of Connecticut. These are randomly selected segments of roadway that are continuously monitored for existing conditions of the roadway, environment and traffic factors. The HPMS sites have uniform geometric characteristics throughout their length.

Unfortunately, there were a number of ATR locations that did not coincide exactly with any of the HPMS sites. Of the 37 ATR locations, we could find 28 HPMS segments that were close to an ATR location, eight of which were two lane roadways that

Table 2

Study Site Locations

HPMS Id	Town Name	**Location**
A001093490	Waterford	From Oil Mill Road to Niantic River Road
A005047810	East Windsor	From 0.06 mi North of Route 140 to East Windsor – Enfield Town Line
A007054780	Kent	From North Kent Road #2 to Kent Cornwall Town Line
A008063640	Colebrook	From Sandy Brook Road to 0.31 mi North of Sandy Brook Road
A012037490	Killingly	From North Street to 0.08 mi North of Peckham Lane
A066025050	Hebron	From 0.05 mi West of Country Lane to South Bound Route 85
A081002310	Clinton	From Egypt Lane to Opposite Indian River
A124002030	Darien	From Pembrook Road to Darien New Canaan Town Line

came under this study. The exact locations of each these eight sites is shown in Table 2. Care was taken to see that there were no major intersections between the ATR location and the corresponding HPMS site. This was done to ensure that the approximation about traffic volumes at the site was valid. A sample sheet from the HPMS database is shown in Figure 3. The database contains about 70 variables; those relevant to this study are explained below:

DATE: 970115 STATE: CONNECTICUT 1994 HPMS FHWA: SUBMITTAL SOFTWARE 94 V:3

LISTING OF CODED VALUES FOR SECTION
RURAL/URBAN

COUNTY(ITEM 6): 011 SECTION ID(ITEM 7A): A001093490 LRS IDENT(ID)(ITEM 7B): 000000001000 LRS MPT/KMPT(ITEM 8): 00934900094700
STATE CONTROL FIELD(ITEM 1): 013152WATERFORD FROM: OIL MILL RD TO: NIANTIC RIVER RD W

ITEM NO		CODED VALUE	ITEM NO		CODED VALUE	ITEM NO		CODED VALUE
9	RURAL/URBAN DESIGNATION	3	16	OFFICIAL INTERSTATE ROUTE NUM	00000	27	STANDARD AADT VOLUME GROUP	03
2	REPORTING UNITS	C	17	ROUTE SIGNING	2	28	AADT	009000
3	YEAR	94	18	ROUTE QUALIFIER	0	29	AADT DERIVATION	2
4	STATE CODE	09	19	SIGNED ROUTE	00000001	30	NUMBER OF THROUGH LANES	02
5	TYPE OF SECTION	1	20	GOVERNMENTAL OWNERSHIP	01	31	URBAN LOCATION	5
10	URBANIZED AREA SAMPLING/CODE	0259	21	SPECIAL SYSTEMS	00	32	ACCESS CONTROL	3
11	NONATTAINMENT AREA CODE	047	22	TYPE OF FACILITY	2	33	MEDIAN TYPE	4
12	FUNCTIONAL SYSTEM	14	23	DESIGNATED TRUCK RT/PARKWAY	4	34	MEDIAN WIDTH	000
13	G/RATE FUNCTIONAL SYSTEM CODE	3	24	TOLL	1	35	ROUGHNESS (IRI)	085
14	NATIONAL HIGHWAY SYSTEM	0	25	SECTION LENGTH(XXX.XXX)	001210	36	PAVEMENT CONDITION(PSR X.X)	46
15	UNBUILT FACILITY	1	26	DONUT AREA AADT VOLUME GROUP	C	38	RECORD TYPE	1000

SAMPLE NUMBER(ITEM 39): 152H124 SAMPLE SUBDIVISION(ITEM 40): 1

41	DONUT EXP FACT(XXX.XXX)	000000	55	RIGHT-OF-WAY WIDTH	075	69	VOLUME/SERVICE FLOW RATIO (X.XX)	066
42	STD EXPANSION FACT(XXX.XXX)	002455	56	WIDENING FEASIBILITY	4	70-A	TURNING LANES: LEFT	4
43	SURFACE/PAVEMENT TYPE	62	57	HORIZONTAL ALIGNMENT ADEQUACY	0	70-B	RIGHT	4
44	PAVEMENT SECTION	1	59	TYPE OF TERRAIN	0	71	PREVAILING SIGNALIZATION	2
45	SN OR D	053	60	VERTICAL ALIGNMENT ADEQUACY	0	72	PERCENT GREEN TIME (%)	41
46	TYPE OF BASE	9	62	% SIGHT DISTANCE (500 FEET	000	73	FUTURE AADT	012800
47	TYPE OF SUBGRADE	1	63	SPEED LIMIT	040	74	FUTURE AADT YEAR	15
48	OVERLAY/PAVE THICKNESS(XX.X)	020	64	WEIGHTED DESIGN SPEED	070	75	CLIMATE ZONE	02
49	YEAR OF SURFACE IMPROVEMENT	1991	65-A1	SINGLE UNIT TRUCKS: % PEAK	02	76	DRAINAGE ADEQUACY	1
50	TYPE OF IMPROVEMENT	78	65-A2	% AVERAGE DAILY	03	77	TYPE OF DEVELOPMENT	0
51	LANE WIDTH	12	65-B1	COMBINATION TRUCKS: % PEAK	01	78	NUM GRADE-SEPARATED INTERCHANGES	00
52	SHOULDER TYPE	1	65-B2	% AVERAGE DAILY	01	79-A	NUM AT-GRADE INTERSECT: SIGNALS	01
53-A	SHOULDER WIDTH: RIGHT	00	66	K-FACTOR (%)	11	79-B	STOP SIGNS	00
53-B	LEFT	00	67	DIRECTIONAL FACTOR (%)	050	79-C	OTHER/NONE	05
54	PEAK PARKING	2	68	PEAK CAPACITY	00750	80	NUMBER AT-GRADE RR CROSSINGS	00

CLASSES:	A	B	C	D	E	F	G	H	I	J	K	L	M
58-CURVES:	0600748	0200115	0100041	0400097	0200209	0000000	0000000	0000000	0000000	0000000	0000000	0000000	0000000
61-GRADES:	1200323	1600440	0600260	0300137	0100050	0000000							

Figure 3

Sample Record of HPMS Data

- HPMS ID: This refers to the 10 character code given to each of the HPMS sites (e.g. A001093490). Positions 2-4 represent the route number in which the segment is located (in this case Route 1). The last five digits represent the starting milepost of the segment (in thousandths of a mile).
- Functional System: This refers to functional classification given to each of the roadways. This functional class can be one of 12 categories listed below:

1 - Rural Principal Arterial -- Interstate
2 - Rural Principal Arterial -- Other
6 - Rural Minor Arterial
7 - Rural Major Collector
8 - Rural Minor Collector
9 - Rural Local
11 - Urban Principal Arterial -- Interstate
12 - Urban Principal Arterial -- Other Freeways and Expressways
14 - Urban Principal Arterial Other
16 - Urban Minor Arterial
17 - Urban Collector
19 - Urban Local

Of these categories, the eight segments in this study fall in 2, 6, 7 and 14.

- Section Length: The length of the segment is expressed in miles with three decimal places (XXX.XXX). Each of the eight segments measures approximately 0.5 to 1.5 miles in length.
- Lane Width: The prevailing lane width to the nearest foot is reported in this data item. Each of the sites in this study has lane widths of either 11 feet or 12 feet
- Right Shoulder: Since the present study considers only two lane roadways, only the right shoulder width is of relevance. The shoulder width is reported to the nearest foot. The shoulder widths range from 0 to 8 feet amongst the sites under consideration.

- Sight Distance: This represents the percent of length of the roadway segment with sight distance of at least 1500 ft. This is the measure of the passing sight distance as prescribed by the Highway Capacity Manual (Transportation Research Board 1994).
- Speed Limit: This represents the posted speed limit on the segment.
- Peak Capacity: This represents the hourly capacity expressed as the total of both directions for two lane roadways. This data item in combination with the hourly volume was used to compute the v/c ratio and hence the LOS during each hour at each of the sites.
- Drainage Adequacy: This variable is a qualitative measure of the drainage conditions at the site. Each of the sites is given one of the three drainage ratings shown below:

  1 - Good Drainage implying no flooding, erosion or other damage.
  2 - Fair Drainage implying that a little maintenance effort is required.
  3 - Poor Drainage implying that there is severe flooding and other drainage problems.

- Signals: This represents the number of intersections with a signal controlling the route being inventoried.
- Stop Signs: This represents the number of intersections with a stop sign controlling the route being inventoried.
- Other or no Controls: This represents the number of intersections where the route being inventoried is not controlled by either a signal or a stop sign -- or is controlled by other type of signing or has no controls.

Crash Database:

This database contains detailed information about all crashes that occurred between October 1990 and September 1995 in each of the eight selected HPMS sites. Crash experience on Connecticut's roadways is maintained in a computer database and is stored as ASCII text files. The crash information about the eight sites was extracted from this database and formatted. A sample output from this database is shown in Figure 4. Some of the important information obtained through this database is discussed below:

- Date and Time: The date and the time at which the crash took place.
- Location: The exact mileage on the segment where the crash took place.
- Light Condition: The light conditions when the crash took place.
- Surface Condition: The condition of the surface when the crash took place (ice, wet, dry etc.)
- Crash Type: The type of the crash. This should be one of the fourteen crash types as defined by ConnDOT.
- Cause of the Crash: The cause of the crash as perceived by the officer reporting the crash.
- Vehicles Involved: The type of vehicles involved in the crash and the direction in which each of them was proceeding.

```
ACCIDENT EXPERIENCE
TOWN OF WATERFORD          ROUTE NUMBER     1                    LOCATION
093.49 094.70
PREPARED 09 30 96     PERIOD FROM 01 01 89 TO 09 30 95          MON DA

LIGHT  SURF       COLLISION  INJURIES RAMP TOT
MILEAGE ALPHA DESCRIPTION OF ACC. LOCATION   RDWY. FACT.  CASE # DAY  TH TE YR
HOUR   COND  COND WEATH   TYPE      K A B C TYPE INJ
*************************************************************************
**************************************************
093.49 INT OIL MILL RD                    TW TN RD UNDIVD  002918 SAT JAN 27 90
1400 DAYLT   DRY  CLEAR TURN-INTS
 DRIVER VIOLATED TRAFFIC CONTROL
 WB VAN                                    VEHICLE GOING STRAIGHT
 SB AUTO PAS                               VEH TURNING LEFT FROM PROPER LANE
 - - - 1      1
093.49 INT OIL MILL RD                    MW TN RD UNDIVD  008473 THU MAR 23 89
1305 DAYLT   DRY  CLEAR HD-ON TRN
 DRIVER FAILED TO GRANT RIGHT OF WAY
 WB AUTO PAS                               VEH TURNING LEFT FROM PROPER LANE
 EB AUTO-SW                                VEHICLE GOING STRAIGHT
 NB JEEP-TYP                       STOPPED FOR TRAFFIC SIGNALS
093.49 INT OF OIL MILL RD                 PO TN RD UNDIVD  071611 SAT SEP 01 90
1651 DAYLT   DRY  CLEAR REAR END
 DRIVER FOLLOWING TOO CLOSE
 EB AUTO PAS                               VEHICLE GOING STRAIGHT
 EB TRUCK ST                       STOPPED FOR TRAFFIC SIGNALS
093.49 INT OLD MILL RD                    KS TN RD UNDIVD  103787 SUN JAN 27 91
1055 DAYLT   DRY  CLEAR FIXED OBJ
 DRIVER INCAPACITATED
 WB TRUCK ST                               VEHICLE GOING STRAIGHT
   STRUCK GUIDE RAIL                OFF RD LEFT
093.49 INT OIL MILL RD                    DF TN RD UNDIVD  104203 MON JAN 25 93
1519 DAYLT   DRY  CLEAR REAR END
 DRIVER FOLLOWING TOO CLOSE
 WB TRUCK ST                               VEHICLE GOING STRAIGHT
 WB AUTO PAS                       STOPPED FOR TRAFFIC SIGNALS
093.49 INT OLD MILL RD                    PA TN RD UNDIVD  114295 SAT MAR 28 92
1027 DAYLT   DRY  CLEAR REAR END
 DRIVER FOLLOWING TOO CLOSE
 EB AUTO PAS                               VEHICLE GOING STRAIGHT
 EB AUTO PAS                       STOPPED FOR TRAFFIC SIGNALS
 - - - 1      1
093.49 ON OIL MILL RD                     RJ TN RD UNDIVD  130843 SUN JUN 13 93
1735 DAYLT   DRY  CLEAR BACKING
 DRIVER INATTENTIVE
 SB AUTO PAS                               VEH BACKING ALONG ROADWAY
 SB AUTO PAS                       STOPPED FOR TRAFFIC SIGNALS
                          CONNECTICUT DEPARTMENT OF TRANSPORTATION

ACCIDENT EXPERIENCE
TOWN OF WATERFORD          ROUTE NUMBER     1                    LOCATION
093.49 094.70
PREPARED 09 30 96     PERIOD FROM 01 01 89 TO 09 30 95          MON DA

LIGHT  SURF       COLLISION  INJURIES RAMP TOT
```

Figure 4

Sample Record of Crash Data

Light Data:

This data set contains the sunrise and sunset times for each site on every day of the period in consideration. Information about sunrise and sunset times was obtained in order to get an idea of the light conditions under which the trips were made. Data were obtained from the Applied Environmetrics Meteorological Table developed by the National Bushfire Research Unit (1995). This software calculates the sunrise and sunset times at a site whose geographic locations are known (latitude and longitude). Since the sunrise/sunset times did not change drastically from day to day on the same site, the timings were calculated only every 10 days. The exact locations of each of the sites were deciphered using a detailed map of Connecticut. The three categories for the light condition variable are as defined below:

Light - The time of the day between one hour after sunrise and one hour before sunset.

Dark - The time of the day after sunset and before sunrise.

Dusk - The hour after sunrise and the hour before sunset.

Precipitation Database:

This database obtained from the National Meteorological Center contains hourly precipitation levels at various stations all over the state. Precipitation data were obtained in an attempt to get an idea of the surface conditions when the trips were made. For each of the HPMS sites, a corresponding precipitation station was selected (based on distance) and the surface condition during each hour was classified as wet or dry based on the precipitation information. However, this information was not incorporated in the final

database because these data were available only on selected days of the year, resulting in an extremely small sample of cases that could include these data.

Database Merging:

Each of these individual databases explained above (except the precipitation database) were then merged together by site, date, time and direction of travel to constitute the final comprehensive data set. Thus, each individual case in this combined data set represented one hour of data from a selected site, and had information from each of the four data sets discussed above.

The selection of sites for study as noted above was based largely on the availability of data. Data that were obtained were checked for adequate representation in important variables. For instance, they were checked to see if there were a substantial number of sites from both rural and urban areas. Similarly, they were checked to see if there was reasonable representation as far as the functional system class was concerned. The current database seemed overall to satisfy these conditions.

Chapter 4: Methodology

The normal distribution is a convenient approximation to the binomial distribution when the number of trials in an experiment is sufficiently large. In particular, for this approximation to be valid, the sample size must conform to the following restriction:

$$n >= \frac{5}{\min(\pi, 1-\pi)}$$

Where n is the required number of observations and π is the probability of observing a success in a trial. However, in many instances, the value of π can be so small that n (number of observations) needs to be extremely large for a normal approximation to be valid. For instance, in crash modeling, π can be as small as 10^{-6} and thus can require extremely large sample sizes for normal approximations. Thus, using the normal approximation can often be erroneous in modeling real world data such as crashes.

The Poisson Regression

The Poisson distribution is a discrete probability distribution that provides a good approximation to the binomial when π is small and n is large but $n\pi$ is less than 5. According to this distribution, the probability of observing y successes in n trials is given by the formula:

$$P(y) = \lambda^y e^{-\lambda} / y!$$

where λ is the average rate of success (here $n\pi = \lambda$)

In particular, the Poisson distribution has been useful in finding the probability of y occurrences of an event that occurs randomly over an interval of time provided certain assumptions are met:

1. Events occur one at a time; therefore, two or more events cannot occur at the same time.
2. The occurrence of an event in a given period does not change the probability of an event occurring in some later period.

Traditionally, crashes have been modeled with a normal distribution (linear regression) but linear regression has a number of disadvantages as far as modeling crashes is concerned. First, the use of a continuous distribution for modeling a discrete process such as crash occurrence is not very suitable. Second, the normal distribution is a symmetric distribution and does not replicate crash occurrence accurately. A right skewed distribution such as the Poisson distribution has been shown to be better. A number of researchers have elaborated on this topic and have suggested the Poisson distribution as a superior alternative in crash modeling as discussed in Chapter 2. Based on these arguments, a Poisson regression analysis was chosen to model crashes in the present study.

The most critical step in a Poisson regression analysis is obtaining an estimate of λ. The Poisson distribution assumes the presence of an average crash rate (λ) that remains constant over time. This average crash rate is estimated by a regression analysis with the geometric and traffic conditions as candidate variables. Thus, we estimate an average rate for every unique combination of these independent variables. We then go on

to use this estimated rate in calculating the probability of observing a fixed number of crashes, over the conditions for which the average rate applies.

In order to estimate the rate for the given conditions, we need to assume a model form that relates this average rate to the explanatory variables. A number of model forms can be found in the literature. For instance, some of the model forms suggested are

1. $\lambda = \beta_0(X_1)^{\beta_1}(X_2)^{\beta_2}(X_3)^{\beta_3}(X_4)^{\beta_4}...\varepsilon$

2. $\lambda = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + ... + \varepsilon$

3. $\lambda = \beta_0(\beta_1)^{X_1}(\beta_2)^{X_2}(\beta_3)^{X_3}(\beta_4)^{X_4}...\varepsilon$

4. $\lambda = e^{\beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + ... + \varepsilon}$

where

β_0 = intercept,

$\beta_1, \beta_2, \beta_3, \beta_4$ = regression coefficients to be estimated

X_1, X_2, X_3, X_4 = geometric variables

λ = the crash rate

ε = error term

The choice of a model form is largely arbitrary (Valavanis 1959) and the present study uses the model form 4. Model form is discussed in greater detail in the next chapter.

Over-dispersion

A common criticism of the Poisson distribution is its assumption that the variance of the data be equal to the mean, since in modeling real world data, we frequently encounter situations where the observed variance is different from the mean. Usually, the

observed variance is greater than the mean (over-dispersion) though some studies have also reported the reverse condition (under-dispersion). In Poisson regression, the violation of the variance assumption does not change parameter estimates (Agresti 1990). However, all t-statistics are rendered inaccurate and thus need to be corrected. Agresti suggested dividing the t-statistics by the square root of the over-dispersion parameter ($\tau^{0.5}$) in order to obtain the correct significance levels of the regression coefficients. Here, the over-dispersion parameter is given by:

$$\tau = \frac{\chi^2}{N - p}$$

where: χ^2 = computed chi – squared value of the model.

N = number of observations

p = number of coefficients considered in the model.

In order to address the over-dispersion problem, researchers are also resorting to the use of a negative binomial model instead of a Poisson model. A negative binomial model is similar to the Poisson model in every respect other than the equality of the mean and variance assumption; In other words, the only difference is that this stipulation is relaxed. A number of recent studies have used negative binomial regression with promising results.

Another solution to the over-dispersion problem is the use of quasi-likelihood estimation techniques to model estimation. Unlike maximum likelihood estimation, quasi-likelihood estimation does not need to make any distribution assumption. The maximum-likelihood and quasi-likelihood methods use the same transformation functions while modeling data that follow a Poisson distribution. The advantage of the quasi-

likelihood method is that it allows for separate mean and variance structures. Quasi-likelihood estimation computes the dispersion parameter and assumes that the variance is equal to the product of the mean and the dispersion parameter. Thus, the over and under-dispersion problems are automatically taken care of. The present study uses the quasi-likelihood estimation techniques in model estimation. The software package used for this purpose is S-PLUS (1995).

Model Form

As described in Chapter 2, the exposure term explains most of the variation in crash prediction models. This is because this term gives vital information about the number of trips made in the site. This being the case, an important step in the model building process is the decision regarding the place where the exposure term can be incorporated in the model. A number of different forms have been used in the past. For instance, we could have the exposure term on either the right or the left side of the equation as illustrated below:

$$Crashes = Ve^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots}$$

$$Rate = Crashes/V = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots}$$

where V = exposure and other symbols are as defined earlier.

These two model forms do not convey the same meaning. For instance, the first equation uses the exposure as information for predicting the number of crashes. The second model, by trying to model the rate, is indirectly modeling exposure also. Thus, it

does not use exposure as information, unlike the first model. Also, since it seems more reasonable to assume that the number of crashes (being an integer) rather than the crash rate is Poisson distributed, the first model form is used here.

The next important decision that was made with respect to model form was about the exponent on the exposure variable. The above two models assume that the crash frequency is linearly related to the exposure. However, a number of studies have proved that this assumption is false. In fact, the crash frequency seems to have a non-linear "U" shaped relationship with exposure. Keeping these points in mind, the model form was changed to incorporate the possibility of a non-linear relationship between crashes and exposure. Thus, along with the coefficients of the geometric and traffic elements, the exponent on the exposure was also estimated. The revised model form is shown below:

$$Crashes = V^{a} e^{\beta_0+\beta_1 X_1+\beta_2 X_2+....}$$

Where = exponent on exposure to be estimated

In an attempt to model at greater detail, the possibility of modeling single and multi-vehicle crashes separately was investigated. It was hypothesized that single and multi-vehicle crashes occur under very different traffic and site conditions. For instance, Figure 5 shows the dependence of single and multi-vehicle crash rates on the shoulder width.

From the graph, it appears that single vehicle crashes are associated with narrow shoulders, while the relationship appears much more complex in the case of multi-vehicle crashes. It must be noted however, that this graph might not reveal the true relationships, since these are plots of total average crash rates versus shoulder width. In order to capture the true relationship, all elements that affect the crash rates need to be controlled, to

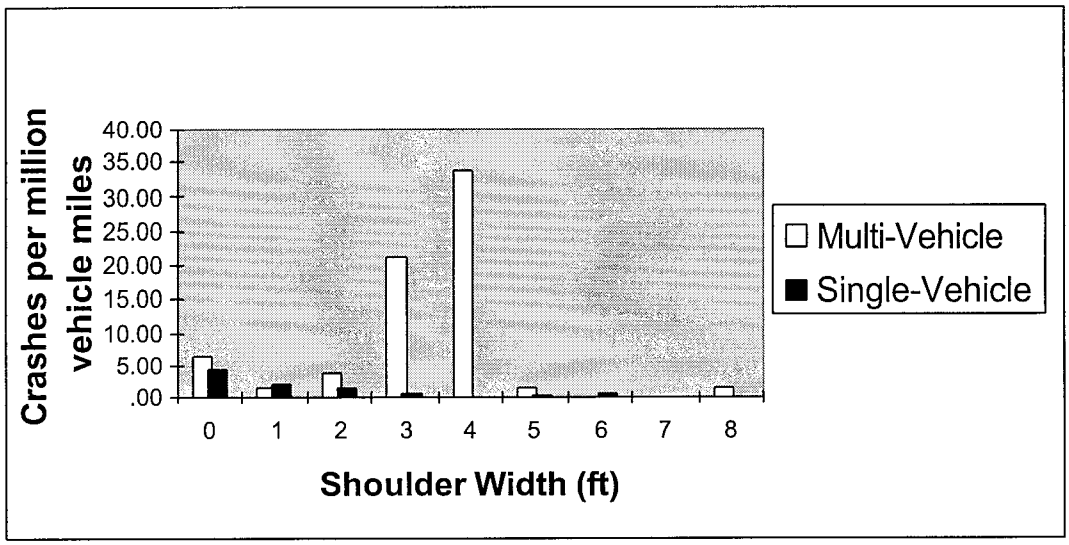

Figure 5. Crash Rate versus Shoulder Width

isolate the effect of shoulder width. In spite of this, there seems to be sufficient evidence to suspect that single and multi-vehicle crashes have different causal factors.

In order to shed more light on the above issue, preliminary model estimation was performed. This estimation is discussed in more detail in the next chapter. The preliminary analysis was essentially aimed at grouping sites that were similar with respect to the occurrence of single and multi-vehicle crashes. Since the traffic and light conditions were controlled for, the similarity in sites was purely based on geometry and land use.

Model and Variable Selection

Model selection is perhaps the most important step in the modeling process. When modeling phenomena such as crashes, where there are a high percentage of zeros, traditional goodness of fit tests such as Pearson's chi-square or the likelihood ratio

statistic are no longer valid. This is because these statistics are poorly approximated by the chi-squared distribution (Agresti 1990). A useful alternative is the Akaike's Information Criterion (AIC). AIC recognizes that the objective of a statistical forecasting model is to convey as much information as possible with a minimum number of variables. Invariably, the information conveyed by a model increases as you add more variables to the model. Hence, there is a clear conflict between trying to incorporate more information into a model while at the same time reducing its complexity. AIC helps to trade off between these two competing objectives in evaluating competing models, and is given by:

$$\text{AIC Value} = -2(\text{maximum likelihood}) + 2p$$

where p is the number of parameters in the model. Models with a lower AIC are preferred.

As mentioned previously, the statistical software package S-PLUS was used for the Poisson regression analysis. S-PLUS uses a procedure that is similar to the forward selection procedure in linear regression. As a first step, the user specifies a base model, which should include all those variables that the user wishes to force into the model. Typically, a null model (containing no variables) is specified as the base model. Next, the user also specifies the pool of variables to choose from for the final model. Thereafter, S-PLUS adds or deletes variables one at a time until the model cannot be improved any further. The AIC is used as the criterion for evaluating the model at every step. Apart from the AIC, two other important factors were used in checking the correctness of a model:

Engineering judgment: This is just to make sure that the variables selected into the model have coefficients and signs that make intuitive sense and that their contribution to the model can be explained logically. As mentioned previously, it is common that a regression analysis sometimes chooses variables that merely explain the current crash dataset, rather than selecting those variables that physically cause the crash. The only solution to this problem is to specify a meaningful set of variables to choose from, and to check if the variables selected for a model make engineering sense. In this study, all variables in the candidate models were checked for their signs and magnitude to make sure that they are meaningful.

Significance: Another important check is to ensure that the model includes only variables that are statistically significant. In other words, it must be known with reasonable confidence (95%) that the coefficient of every variable is not actually zero.

These checks were performed on every model that was considered in the preliminary and the final analysis.

Chapter 5: Preliminary Models

In crash studies, the candidate variables for a regression analysis are usually highly correlated. This is probably because we can rarely find roads that are designed with highly differing design standards with respect to different geometric features. For example, we seldom find a site that has high design standards for shoulder width but very poor design standards for lane width. This being the case, the researcher has the important role of helping the regression analysis select the variables that actually cause the crash, as opposed to allowing it to merely pick those variables that seem to explain crashes statistically. The researcher can do this by carefully picking out the pool of candidate variables considered for regression. In order to do this, it would be helpful to have an idea of which geometric variables are likely to cause crashes.

Prior to building the final crash models, a preliminary model estimation was undertaken. This step had two main objectives:

1. To group similar sites (with respect to crash rates) and thus get an idea of the important geometric variables affecting crashes.
2. To check if the grouping of sites is similar with respect to single and multi-vehicle crash rates. A difference in grouping for each crash type would validate a split modeling in crashes (since it would mean a difference in the causal factors).

In this preliminary analysis, the actual geometric variables were not included, but each site was given a dummy id that represented its unique combination of geometric characteristics. A new qualitative variable called "site_id" was introduced, which had the eight dummy ids as its classes. In order to isolate the geometric variables causing the two

types of crashes, the traffic and the light conditions had to be controlled. The three variables used in this analysis are explained below in more detail.

Site id (SITE_ID): As mentioned earlier, "site_id" is a categorical variable that represents the unique combination of geometric characteristics at each site, and takes the following values:

1. Rte 1
2. Rte 5
3. Rte 7
4. Rte 8
5. Rte 12
6. Rte 66
7. Rte 81
8. Rte 124

Level of Service (LOS): This variable is based on the Highway Capacity Manual (HCM) 1994 methodology for computing LOS on two-lane highways. The HCM prescribes the computation of LOS based on the available sight distance, type of terrain and the prevailing traffic condition (v/c ratio). Here, the capacity depends on the existing geometric conditions at the site and is also computed according to the methods prescribed by the HCM. Following are the values it takes:

1. LOS A
2. LOS B
3. LOS C
4. LOS D
5. LOS E

Light Conditions (L_Conds): "L_Conds" was introduced to capture the prevailing light conditions when each trip was made. The computation of this variable including a

detailed description of each of the classes was given in Chapter 3; they are assigned values as follows:

1. Light
2. Dark
3. Dusk

To aid understanding of the preliminary analysis, an example of the results from an estimated model are displayed and discussed in detail in the following section.

Shown in Figure 6 are results from a model estimated with the independent variables site id, LOS and light conditions. The dependent variable in this case is the total crash frequency. All three independent variables are categorical, so the regression coefficients in the model were estimated as treatment contrasts. When considering treatment contrasts, S-PLUS considers every variable level (except the base) as a dummy. The base levels for the model in Figure 6 are Rte. 1, LOS A and light condition "light." Thus, the intercept represents the crash rate for the base level, i.e. for Rte. 1 at LOS A and during daylight conditions. The coefficient shown on Rte. 7 is with respect to Rte. 1 and the coefficient on LOS B is with respect to LOS A. This means that the crash rate on Rte 7 is $(e^{1.32} - e^{1.00})$ units less than that on Rte 1, all other conditions being the same.

The t-statistics shown are for the null hypothesis that the difference in means between the particular class and its base case is zero ($\mu_{Rte7,LOSA,Dark} - \mu_{Rte1,LOSA,Light} = 0$). Thus, since the t-value is –4.51, at 95% confidence we can reject the null hypothesis that Rte. 1 is not different from Rte. 7, all other conditions being the same. In this manner, each of the sites can be statistically compared to the base case (Rte. 1). However, only comparisons with the base case can be performed with these results. In order to make all pair-wise comparisons, paired t-tests need to be performed between all pairs of sites. This

$$Accidents = V^a e^{\beta_0 + \sum_{i=1}^{8} \beta_{1i} site\_id_i + \sum_{j=1}^{5} \beta_{2j} LOS_j + \sum_{k=1}^{3} \beta_{3k} L\_Conds_k}$$

Coefficients:

	Value	Std. Error	t value
(Intercept)	-10.22	1.16	-8.79
site_id 2 (Rte 5)	-1.96	0.51	-3.83
site_id 3 (Rte 7)	-1.32	0.29	-4.51
site_id 4 (Rte 8)	-1.30	0.49	-2.66
site_id 5 (Rte 12)	0.71	0.28	2.57
site_id 6 (Rte 66)	-1.71	0.31	-5.58
site_id 7 (Rte 81)	-1.63	0.49	-3.33
site_id 8 (Rte124)	-0.92	0.39	-2.35
LOS 2 (B)	-0.48	0.29	-1.64
LOS 3 (C)	-0.64	0.36	-1.77
LOS 4 (D)	-0.57	0.46	-1.24
LOS 5 (E)	-0.46	0.53	-0.86
L_Conds 2 (Dark)	0.14	0.25	0.57
L_Conds 3 (Dusk)	0.22	0.29	0.76
(Exp on Exposure) a	0.94	0.09	10.12

(Dispersion Parameter for Quasi-likelihood family taken to be 1.253302)

Figure 6

Sample Results from S-PLUS

is essentially equivalent to changing the base level each time. As mentioned before, these comparisons are an important part of the analysis in this chapter.

Single Vehicle Crash Models

Four models were estimated for the single-vehicle crashes; their results are displayed in Table 3. Important results and interpretations are discussed below.

- Each of the first two models (Models 1 and 2) has just one explanatory variable. Neither of these models performs as well as Models 3 or 4 (Models 1 and 2 have high AIC and deviances). This suggests that both LOS and the site characteristics (represented by site id) are very important together in explaining single vehicle crashes.
- Based on the AIC criterion, Model 3 is the best model. In fact, adding the light variable to Model 3 (i.e Model 4) seems to have no effect. This is because LOS and light conditions may be highly correlated, because most of the sites experience similar traffic conditions at similar times of the day, particularly on weekdays.
- Having chosen Model 3 as the best model, all pairwise comparisons were performed to test differences between sites (Table 4). As explained earlier, a t-value of less than 1.96 means that there is no evidence to suggest dissimilarity between sites. Simply using this criteria alone leads to no clear groupings; the following grouping is based on the smallest t-statistics.

Group 1:	Rtes 5, 7, 12, 66
Group 2:	Rtes 1, 81
Group 3:	Rtes 8, 124

Table 3
Preliminary Models for Single Vehicle Crashes

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	-11.00[1] (-10.03)[2]	-13.57 (-10.88)	-13.09 (-10.75)	-12.78 (-7.74)
Rte 1	Base		Base	Base
Rte 5	-2.50 (-2.71)		-1.12 (-1.35)	-1.19 (-1.42)
Rte 7	-0.38 (-1.11)		-0.66 (-2.14)	-0.65 (-2.10)
Rte 8	-0.97 (-1.41)		-1.66 (-2.74)	-1.64 (-2.62)
Rte 12	-7.49 (-0.70)		-6.91 (-0.78)	-6.96 (-0.79)
Rte 66	-1.42 (-3.70)		-0.68 (-1.89)	-0.70 (-1.94)
Rte 81	-0.76 (-1.51)		-0.01 (-0.03)	-0.06 (0.12)
Rte 124	-0.72 (-2.29)		0.87 (1.91)	0.79 (1.69)
LOS A		Base	Base	Base
LOS B		-1.32 (-3.75)	-1.67 (-5.18)	-1.60 (-4.52)
LOS C		-2.54 (-5.17)	-2.83 (-6.08)	-2.74 (-5.42)
LOS D		-2.99 (-4.14)	-3.80 (-5.34)	-3.67 (-4.93)
LOS E		-1.91 (-3.85)	-3.49 (-5.30)	-3.31 (-4.55)
LIGHT				Base
DARK				0.02 (-0.08)
DUSK				-0.23 (-0.47)
Exponent on Exposure	0.86 (9.52)	1.14 (9.82)	1.16 (10.51)	1.13 (8.21)
Dispersion	0.80	0.90	0.60	0.60
Null Deviance	363.50	363.50	363.50	363.50
Residual Dev.	218.72	204.73	185.98	185.80
AIC Value	236.72	216.73	211.98	215.80

1 -> Variable coefficient 2 -> t-statistic corrected for over-dispersion

Table 4
Comparing Sites with Respect to Single Vehicle Crashes
(paired t-statistics)

	Rte 1	Rte 5	Rte 7	Rte 8	Rte 12	Rte 66	Rte 81	Rte 124
Rte 1	-	1.04	1.65	**2.11**	0.61	1.46	0.02	-1.48
Rte 5	-1.04	-	-0.43	0.41	0.51	-0.42	-0.99	-1.83
Rte 7	-1.65	0.43	-	1.27	0.55	0.04	-1.02	**-2.47**
Rte 8	**-2.11**	-0.41	-1.27	-	0.46	-1.17	-1.80	**-2.83**
Rte 12	-0.61	-0.51	-0.55	-0.46	-	-0.55	-0.61	-0.68
Rte 66	-1.46	0.42	-0.04	1.17	0.55	-	-1.50	**-2.61**
Rte 81	-0.02	0.99	1.02	1.80	0.61	1.052	-	-1.30
Rte 124	1.48	1.83	**2.47**	**2.83**	0.68	**2.61**	1.30	-

Bold face *indicates failed t-tests*

- A close examination of similar sites revealed that the shoulder width and truck percentage could be important variables. However, a clear picture about the individual geometric features causing crashes did not emerge. This was probably because the relative amounts by which each characteristic affects crashes could not be judged merely by observation.

The level of service (LOS) is very important in predicting single vehicle crashes, as are the geometric characteristics. It is suspected that the LOS itself is highly correlated with the geometric characteristics, because it is the geometry that determines the capacity of a roadway, and LOS is computed from the capacity. Also, it should be noted that the single vehicle crash rate decreases monotonically as the LOS becomes poorer. This is probably because drivers are much more careful when there is lot of traffic on the road.

Another plausible reason is that the better levels of service are more likely to be observed during very early and very late hours of the day, when the drivers can be very tired or sleepy.

Multi-Vehicle Crash Models

Results from the regression for the multi-vehicle crashes are displayed in Table 5. Important results and interpretations are discussed below.

The best model is Model 1 since it has the lowest AIC value (391.28). It should be noted that this model includes only "site id" amongst the three candidate variables. The fact that LOS was not included in the best model was very surprising. However, it must be remembered that the site LOS has been computed as the LOS for the section, ignoring any intersections on the segment. Thus, the actual LOS for the section may be vastly different from the LOS based on the computed capacity. Since most multi-vehicle crashes are likely to take place at intersections, the LOS variable does not give any vital information in explaining these crashes.

The table of t-statistics for the pair-wise comparison of sites is shown in Table 6.

- Based on these t-statistics for Model 1, we arrived at the following grouping of sites:

Group 1:	Rte. 1
Group 2:	Rtes. 5, 7, 66, 81
Group 3:	Rte. 12
Group 4:	Rtes. 8,124

Table 5
Preliminary Models for Multi-Vehicle Crashes

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	-10.08[1] (-11.53)[2]	-9.88 (-12.04)	-9.98 (-10.72)	-11.31 (-7.33)
Rte 1	Base		Base	Base
Rte 5	-2.03 (-3.72)		-2.28 (-3.86)	-2.30 (-3.78)
Rte 7	-1.86 (-4.19)		-1.84 (-4.12)	-1.81 (-3.99)
Rte 8	-1.17 (-2.10)		-1.11 (-1.90)	-0.94 (-1.50)
Rte 12	0.88 (3.13)		0.86 (3.00)	1.01 (3.18)
Rte 66	-1.95 (-5.70)		-2.09 (-5.45)	-2.21 (-5.35)
Rte 81	-2.59 (-3.15)		-2.67 (-3.22)	-2.63 (-3.05)
Rte 124	-0.99 (-4.00)		-1.54 (-3.18)	-1.66 (-3.13)
LOS A		Base	Base	Base
LOS B		0.64 (1.67)	0.15 (0.38)	0.14 (0.32)
LOS C		-0.14 (-0.32)	0.21 (0.49)	0.33 (0.67)
LOS D		-0.18 (-0.39)	0.63 (1.16)	0.76 (1.24)
LOS E		0.46 (1.02)	0.83 (1.28)	0.86 (1.21)
LIGHT				Base
DARK				0.26 (0.79)
DUSK				0.45 (1.27)
Exponent on Exposure	0.88 (12.20)	0.76 (10.37)	0.86 (10.73)	0.96 (7.96)
Dispersion Null Deviance Residual Dev. AIC Value	1.37 788.97 373.28 391.28	1.44 788.97 504.44 516.44	1.38 788.97 370.26 396.26	1.40 788.97 367.97 397.97

1 -> Variable coefficient
2 -> t-statistic corrected for dispersion

Table 6

Comparing Sites with respect to Multi-Vehicle Crashes

(paired t-tests)

	Rte 1	Rte 5	Rte 7	Rte 8	Rte 12	Rte 66	Rte 81	Rte 124
Rte 1	-	**4.35**	**4.90**	**2.45**	**-3.66**	**6.67**	**3.69**	**4.69**
Rte 5	**-4.35**	-	-0.31	-1.36	**-6.09**	-0.16	0.67	**-2.17**
Rte 7	**-4.90**	0.31	-	-1.18	**-6.79**	0.20	0.94	**-2.21**
Rte 8	**-2.45**	1.36	1.18	-	**-4.29**	1.46	1.72	-0.36
Rte 12	**3.66**	**6.09**	**6.79**	**4.29**	-	**8.66**	**4.91**	**7.03**
Rte 66	**-6.67**	0.16	-0.20	-1.46	**-8.66**	-	0.87	**-3.13**
Rte 81	**-3.69**	-0.67	-0.94	-1.72	**-4.91**	-0.87	-	**-2.24**
Rte 124	**-4.69**	**2.17**	**2.21**	0.36	**-7.03**	**3.13**	**2.24**	-

Boldface *indicates a failed t-test*

- From a close examination of similar sites, shoulder width and truck percentage seemed to be important again. However, as we had noted in the case of single vehicle crashes, a clear picture of the elements of roadway geometry affecting crashes could not be obtained.
- From the four models shown, it can be observed that the site id is much more important than the level of service in explaining multi-vehicle crashes. This is much different from findings with single-vehicle crashes, where the level of service was very important. This will be discussed in more detail with the final models.

Thus, we have seen that the present exercise did not reveal a clear picture of the important geometric features causing crashes. However, it revealed important differences between single and multi-vehicle crashes that have validated the categorization of the crashes. For instance, the grouping of sites for multi-vehicle crashes is completely different from that for single vehicle crashes. This suggests that the geometric elements affecting single-vehicle crashes are completely different from those causing multi-vehicle crashes. Secondly, it can be observed from multi-vehicle Model 3 (though not the best model) that the multi-vehicle crash rate increases as the level of service becomes poorer. This is completely different from the relationship we saw in the case of single vehicle crashes.

Chapter 6: Final Models

After the preliminary analysis for the single and multi-vehicle crashes, the next step was to incorporate the actual site variables in place of the site ids. Variables describing LOS and the light conditions (L.COND) were retained in their original form for the final analysis. The list of geometric variables and their abbreviations as used in the models are shown in Table 7. All of these variables and their classes were explained in detail in Chapter 3. The models for the single and multi-vehicle crashes included the same geometric variables as candidates for the final model.

As mentioned previously, the statistical software package S-PLUS was used for the Poisson regression analysis. The null model was formed using the original base values for categorical variables and the pool of candidate variables contained all the geometric, traffic and light variables, and all possible two way interactions of these variables.

Single Vehicle Crash Models

Table 8 shows the results for the single-vehicle crash models; the following paragraphs include a detailed explanation of the steps involved in selecting the best models. The variables included in the model at each step and the pool of candidate variables are outlined for clarity.

Table 7
Site Variables Used

Variable Name	Explanation	Type of Variable	Range or Classes *(found in data)*
CLIM_ZON	Climate Zone	Categorical	1 = Freeze Cycle 2 = Freeze-Thaw Cycle
FUNC_SYS	Functional System of the Roadway	Categorical	2 = Prin. Arterial 6 = Minor Arterial 7 = Major Collector
LAN_WID	Lane Width	Continuous	11 and 12 ft
P_PAR	Peak Parking	Categorical	1 = Present 2 = Absent
RUR_URB	Location Type	Categorical	1 = Rural Area 3 = Urbanized (50 – 199)* 4 = Urbanized (> 200)
SHO_RIGH	Shoulder Width	Continuous	0 – 8 ft
SIGHT_DI	Percent of Roadway with Sight Distance >1500 ft	Continuous	0, 10 and 80 %
SIGNALS	No of Signalized Intersections	Continuous	0 – 1
SIN_TR_D	Percentage of Trucks in Daily Traffic	Continuous	0 – 4 %
SIN_TR_P	Percentage of Trucks in the Peak Hour	Continuous	0 – 2 %
SPEED_LI	Speed Limit	Continuous	35, 40, 45 mph
UNSIG	No of Unsignalized Intersections	Continuous	0 – 6

* *Population in 1000s*

Table 8
Final Models: Single Vehicle Crashes

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-16.26 (-11.63)	-13.97 (-12.91)	-15.60 (-12.11)	-14.65 (-11.27)	-14.78 (-11.74)	-12.80 (-11.34)
LOS A	Base	Base	Base	Base	Base	Base
LOS B	-1.66 (-5.20)	-1.57 (-4.96)	-1.69 (-5.16)	-1.64 (-5.08)	-1.65 (-5.22)	-1.60 (-4.78)
LOS C	-2.87 (-6.35)	-2.44 (-5.84)	-2.91 (-6.33)	-2.73 (-5.95)	-2.81 (-6.27)	-2.51 (-5.39)
LOS D	-3.76 (-5.51)	-3.15 (-5.01)	-3.99 (-5.73)	-3.71 (-5.28)	-3.90 (-5.72)	-3.10 (-4.59)
LOS E	-3.34 (-5.65)	-2.57 (-5.73)	-3.81 (-6.33)	-3.43 (-5.53)	-3.70 (-6.31)	-2.35 (-5.01)
P_PAR	2.38 (5.56)					
SIN_TR_P	1.21 (3.47)		1.30 (3.17)	1.61 (3.69)	1.47 (3.68)	
SHO_RIGH	1.41 (4.03)		-0.30 (-4.83)	-0.33 (-4.56)	-0.27 (-4.23)	-0.11 (-2.14)
CLIM_ZON			0.96 (3.20)			
SIGHT_DI				-0.01 (-1.64)	-0.02 (-2.30)	-0.02 (-2.45)
LAN_WID						
UNSIG		0.34 (4.21)				
SIGNALS				0.54 (1.92)		
Exponent on Exposure	1.16 (-11.63)	1.09 (10.92)	1.21 (11.64)	1.14 (10.80)	1.18 (11.71)	1.13 (10.83)
Dispersion	0.67	0.67	0.65	0.66	0.64	0.73
Null Deviance	363.50	363.50	363.50	363.50	363.50	363.50
Residual Dev.	182.25	192.28	183.81	183.54	185.95	193.96
AIC Value	200.25	206.28	201.81	203.54	203.95	209.96

Values in parentheses are t-statistics after correction for dispersion

Step 1: *All Variables Included*
Model 1: *LOS, P_PAR, SIN_TR_P, SHO_RIGH*

Initially, when the entire list of candidate variables was available for regression, Model 1 was selected as the best model through the step-wise search procedure. This model included LOS, peak parking, percentage single unit trucks in the peak hour, and shoulder width. Though the AIC value was very good for this model, a number of important flaws were observed. For instance, the inclusion of peak parking as an important variable in predicting single vehicle crashes did not make any engineering sense. As mentioned in Chapter 3, this variable has two classes ("parking allowed" and "not allowed"), but seven of the eight sites came under the second category ("not allowed") for this variable. Hence, it is possible that this variable was just acting as a site specific dummy (for the only site that allowed parking – Rte. 1). Thus, this variable did not have sufficient representation of its classes. Another important drawback of the model is that shoulder width has a positive sign, suggesting that single vehicle crashes are associated with wider shoulders. This result is unexpected and therefore rejected.

As explained in the previous chapters, these problems probably arose because the candidate variables are highly correlated, and in such a case, the regression procedure may not pick those variables that physically cause the crash. Hence, since it is known that peak parking is not a desirable variable, it was removed from the pool of candidate variables. The step-wise procedure was performed again without this variable.

Step 2: *P_PAR excluded*
Model 2: *LOS, UNSIG*

In Model 2, the trends in level of service are similar to those found in Model 1. UNSIG has a positive coefficient suggesting that single-vehicle crashes increase with an increase in the number of unsignalized intersections. This relationship between the number of unsignalized intersections and single-vehicle crashes cannot be explained intuitively, so UNSIG was removed so that the model could include more meaningful variables.

Step 3: *P_PAR and UNSIG excluded*
Model 3: *LOS, SIN_TR_P, SHO_RIGH, CLIM_ZON*

Model 3 includes CLIM_ZON in addition to the variables in Model 1. The only two classes of CLIM_ZON that were present in the database were 1 and 2. Class 1 represents sites that come under a "freeze" cycle while class 2 represents those that have a "freeze-thaw" cycle. Since only two classes of climate zone are present, there is not enough variation in the dataset for this variable. Also, the difference in the single vehicle crash rates between these two classes could not be sufficiently explained. Hence, this variable was the next to be removed from the pool of candidate variables.

It can be seen that Model 3 is very similar to Model 1 in the variables that it includes. However, SHO_RIGH has a more meaningful negative sign in Model 3 unlike Model 1. LOS and SIN_TR_P have very similar coefficients in both Models 1 and 3. Thus, the step-wise regression procedure was continued with CLIM_ZON removed.

Step 4: *P_PAR, UNSIG and CLIM_ZON excluded*
Model 4: *LOS, SIN_TR_P, SHO_RIGH, SIGHT_DI, SIGNALS*

It should be observed that this model is similar to the Models 1 and 3 in the variables included except that Model 4 includes SIGHT_DI and SIGNALS rather than P_PAR or CLIM_ZON. Most of the variables included in this model take predictable signs. For instance, sight distance was included for the first time, and has a negative coefficient suggesting that single vehicle crashes are associated with sites where the sight distance is poor. The main shortcoming of this model was the inclusion of the number of signals. Similar to UNSIG, the relationship between the single-vehicle crashes and the SIGNALS could not be explained and hence, this variable was removed in the next step. SIN_TR_P was again included in this model with a positive sign and though not as serious as the inclusion of UNSIG, this is another minor drawback of the model. Apart from this, LOS and SHO_RIGH take expected trends.

Step 5: *P_PAR, UNSIG, CLIM_ZON and SIGNALS excluded*
Model 5: *LOS, SIN_TR_P, SHO_RIGH, SIGHT_DI*

Model 5 appears to be the best model, with all variables taking expected signs and all their coefficients being significantly different from zero. Though Model 5 appears the best, it still includes SIN_TR_P whose presence is troublesome. The preliminary estimation had revealed that single-vehicle crashes were associated with better levels of service and hence, the inclusion of a peak hour variable is questionable. For instance, the model would have made much more sense if SIN_TR_D was included rather than SIN_TR_P. Hence, a number of other steps were performed before the final model was selected. The first step was to exclude the peak hour truck variable.

Step 6: *SIN_TR_P excluded from models*
Model 6: *LOS, SHO_RIGH, SIGHT_DI*

Model 6 shows the results after removing the peak hour truck variable. The model performs very well as far as the signs of the variables and their significance are concerned and appears to be the best model from an engineering point of view. The model, however, has a much higher AIC value than Model 5. Thus, on the basis of goodness of fit and engineering judgement criteria, the best single-vehicle crash models were identified to be Models 5 & 6. However, both these models had some minor drawbacks as mentioned above and hence were scrutinized further.

In all the models considered to this point, the level of service was included as an important variable. From the coefficients of each of the classes of LOS, it can be observed that there is a decreasing trend. In other words, the single vehicle crash rate seems to decrease as the level of service becomes poorer, up to LOS D, after which there is a slight increase. As observed with the preliminary estimations, this is probably because single-vehicle crashes generally occur during less congested conditions.

However, this trend in LOS is not conclusive since the t-values tell us only the significant difference of a class as compared to its base class (LOS A). For instance, we do not know if there is a significant difference between the levels of service D and E though the coefficients suggest an increasing trend. Thus, in order to evaluate the differences between the other levels of service (apart from LOS A), paired t-tests need to be performed. Having chosen models 5 and 6 for further investigation, paired t-tests for LOS were performed for each of these models.

Table 9
Comparing the Levels of Service (Single Vehicle Models, 5 and 6)
(paired t-tests)

	LOS A	LOS B	LOS C	LOS D	LOS E
LOS A	-	5.22[1] 4.78[2]	6.26 5.39	5.71 4.59	6.31 5.01
LOS B	5.22 4.78	-	2.99 2.20	3.56 2.34	4.29 2.18
LOS C	6.26 5.39	2.99 2.20	-	1.64 **0.85**	1.71 **-0.32**
LOS D	5.71 4.59	3.56 2.34	1.64 **0.85**	-	**0.32** **1.08**
LOS E	6.31 5.01	4.29 2.18	1.71 **0.32**	**0.32** **1.08**	-

Bold face *indicates failure to reject Null-Hypothesis of no difference between coefficient values (at 90% confidence)*

1 Model 5
2 Model 6

Table 9 shows the results from the paired t-tests performed for Models 5 and 6. It can be seen for both models that the higher levels of service (A, B) are significantly different from each of C, D and E. For Model 5, LOS C is significantly different from D and E at 90% confidence. However, for the same model, there is no evidence to suggest any difference between LOS D and E. In Model 6, there is no evidence to suggest any difference between levels of service C, D, E. Hence, in order to get significantly different LOS, the following grouping was performed:

Model 5: LOS A LOS B LOS C LOS (D & E)

Model 6: LOS A LOS B LOS (C, D & E)

With this grouping of the levels of service, three models were tested. Table 10 summarizes the results of each of these models with the merged levels of service. Models 7 and 9 correspond to Models 5 and 6 with the levels of service merged. As mentioned before, it was thought that a variable such as SIN_TR_D would make Model 5 more meaningful than SIN_TR_P. Model 8 shows the results after replacing SIN_TR_P with SIN_TR_D in Model 7. As can be observed, SIN_TR_D is not significantly different from zero and the model performs very poorly with a high AIC value. Model 10 is a further extension of Model 9 with SPEED_LI included. As can be observed from Table 10, this model has the lowest AIC value. Thus, Models 7, 9 and 10 remain as the final competing models, so further analysis was limited to these models.

After merging LOS, paired t-tests were performed for Models 7 and 9 again, to check if all classes were significant. The results are summarized in Tables 11 and 12. We can observe that with such a grouping of the LOS, we obtain classes that are statistically different from each other.

Looking at the coefficients on LOS in Models 7 and 9, it can be seen that the single vehicle crash rate decreases monotonically with a decrease in the level of service. Since the levels of service were all shown to be significantly different, we can be confident that we have captured the true relationship between the single vehicle crash rate and the level of service. As the last step, the variable SIN_TR_P was removed from the candidate variables and the search for the best model was continued. The next best model (Model 10) was very similar to models 7 and 9 in that it included level of service, sight distance and shoulder width as important variables. Apart from these, speed limit was also included in the final model. The signs of the coefficients are very similar to Models

Table 10

Final Models: Single Vehicle Crashes

Variable	Model 7	Model 8	Model 9	Model 10
Intercept	-14.88 (-12.22)	-13.06 (-10.48)	-13.06 (-11.84)	-8.91 (-4.13)
LOS A	Base	Base	Base	Base
LOS B	-1.66 (-5.30)	-1.66 (-4.89)	-1.65 (-4.91)	-1.59 (-4.90)
LOS C	-2.83 (-6.31)	-2.53 (-5.36)		
LOS (D, E)	-3.77 (-7.01)	-2.59 (-5.90)		
LOS (C, D, E)			-2.56 (-6.57)	-2.76 (-6.86)
SIGHT_DI	-0.02 (-2.30)	-0.02 (-2.31)	-0.02 (-2.44)	-0.02 (-2.49)
SHO_RIGH	-0.27 (-4.36)	-0.12 (-2.32)	-0.12 (-2.33)	-0.15 (-2.84)
SIN_TR_P	1.49 (3.80)			
SIN_TR_D		-0.01 (-0.06)		
SPEED_LI				-0.09 (-2.10)
Exponent on Exposure	1.19 (12.13)	1.16 (11.26)	1.16 (11.50)	1.13 (11.53)
Dispersion Null Deviance Resid. Deviance AIC Value	0.64 363.50 186.01 202.01	0.75 363.50 194.94 210.94	0.75 363.50 194.96 206.96	0.71 363.50 191.44 205.44

Values in parentheses are t-statistics after correction for dispersion

Table 11
Comparing Levels of Service for the Final Model (Single Vehicle Model 7)
(paired t-tests)

	LOS A	LOS B	LOS C	LOS (D, E)
LOS A	-	5.30	6.32	7.01
LOS B	5.30	-	2.99	4.83
LOS C	6.32	2.99	-	1.96
LOS (D, E)	7.01	4.83	1.96	-

All classes are significantly different from each other at 95% confidence

Table 12
Comparing Levels of Service for the Final Model (Single Vehicle Models, 9 and 10)
(paired t-tests)

	LOS A	LOS B	LOS C, D, E
LOS A	-	4.91[1] 4.90[2]	6.57 6.86
LOS B	4.91 4.90	-	3.26 3.85
LOS C, D, E	6.57 6.86	3.26 3.85	-

All classes are significantly different from each other at 95% confidence

1 -> Model 9
2 -> Model 10

7 and 9. Speed limit is observed to have a negative coefficient, indicating that roads with a higher speed limit have a lower single-vehicle crash rate. This does not necessarily mean that increasing the speed limit on a roadway will decrease the crash rate. Instead such a trend was observed probably because roads with higher speed limits are generally designed better. It should also be noted that other researchers (Bared and Vogt 1997 and Ivan and O'Mara 1997) have observed similar trends.

At this stage, the final models chosen are 7, 9 and 10. Since the relationship of peak truck percentage to single-vehicle crashes is not understood clearly, the validity of Model 7 can be questioned. In comparing Models 9 and 10, important questions about speed limit arise because speed limit is not a basic variable, but is a function of the existing conditions on the roadway and hence does not directly affect the crash rate. Hence, the presence of such a variable is not very desirable.

The above discussions suggest Model 9 to be the best single vehicle crash model. The variables included are LOS, shoulder width and sight distance. All variables are statistically significant and take signs that make intuitive sense. The AIC value, though not as impressive as the other models considered, still suggests a good fit with a minimum number of variables.

Some Important Observations

In modeling single-vehicle crashes, it can be observed that the data were underdispersed. This condition, though not very common, is not very surprising. This condition may be associated with the large number of cases observed to have no single-vehicle crashes. Also, the cases with non-zero observations had a small number of single

vehicle crashes and thus produced very little variation. In essence, a larger dataset might have mitigated this problem but underdispersion could be an innate characteristic of single vehicle crashes.

The exponent on exposure has been uniformly observed to be slightly greater than 1.00. This further confirms the non-linear relationship between crashes and traffic flow. Light conditions did not emerge as an important variable in predicting single-vehicle crashes probably because of its correlation with the level of service

Multi-Vehicle Crash Models

In modeling multi-vehicle crashes, a procedure similar to that adopted for single vehicle crashes was followed. The same candidate variables were used in modeling multi-vehicle crashes and the base model was specified to be a null model in step-wise regression. Table 13 summarizes the results from the multi-vehicle crash modeling.

Table 13

Final Models: Multi Vehicle Crashes

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-1.25 (-0.57)	-17.44 (-9.43)	-2.14 (-1.18)	-23.24 (-14.01)	-21.21 (-6.75)	-22.23 (-14.78)
RUR_URB (3)	1.12 (4.43)					
RUR_URB (4)	-0.24 (-0.36)					
SPEED_LI	-0.13 (-2.36)		-0.12 (-2.59)			
SHO_RIGH	0.36 (4.14)	0.86 (5.93)	0.39 (4.60)	0.59 (5.03)	0.67 (3.81)	0.47 (5.15)
ROW_WID	-0.04 (-3.55)	-0.04 (-3.76)	-0.04 (-4.32)		-0.01 (-0.73)	
FUNC_SYS (6)		6.71 (8.22)		6.50 (8.96)	6.51 (8.01)	5.98 (9.12)
FUNC_SYS (7)		5.83 (7.26)		5.69 (7.64)	5.67 (7.08)	5.35 (7.52)
FUNC_SYS (14)		4.55 (6.85)		6.65 (9.03)	5.92 (5.11)	6.11 (9.51)
SIGNALS		1.41 (3.79)		2.18 (7.13)	1.96 (3.88)	2.36 (7.73)
UNSIG		0.38 (2.23)	0.35 (4.51)		0.12 (0.53)	
P.PAR		1.60 (2.82)		0.80 (1.57)	1.03 (1.61)	
SIN_TR_P	-2.60 (-8.68)		-2.61 (-9.27)			
SIN_TR_D				1.39 (6.36)	0.95 (1.57)	1.39 (5.94)
Exponent on Exposure	0.88 (11.96)	0.88 (12.69)	0.88 (12.19)	0.89 (12.99)	0.88 (12.71)	0.87 (13.06)
Dispersion Null Deviance Residual Dev. AIC Value	1.47 788.97 379.20 395.20	1.29 788.97 367.78 387.78	1.41 788.97 379.39 393.39	1.31 788.97 365.38 383.38	1.30 788.97 364.57 386.57	1.31 788.97 368.44 384.44

Values in parentheses are t-statistics after correction for dispersion

Step 1: *All Variables Included*
Model 1: *RUR_URB, SPEED_LI, SHO_RIGH, ROW_WID, SIN_TR_P*

When all the variables were included in the pool of candidate variables, Model 1 was chosen by the software to be the best model. The model has a number of drawbacks. First, the inclusion of a variable such as speed limit is not desirable for reasons discussed in the previous section. Also, from the sign of SIN_TR_P, the model suggests that the multi-vehicle crash rate decreases with an increase in the percentage trucks in the peak hour. This trend, though possible, is very unlikely. Another shortcoming is the inclusion of the variable ROW_WID. Variables such as lane width may be preferred to right of way width, since the effect of lane width on crashes is more direct. Lastly, the model has a high AIC value and a high residual deviance. As the next step in the analysis, either ROW_WID or SPEED_LI had to be removed from the pool of candidate variables to permit more meaningful geometric variables to be included. Since the inclusion of SPEED_LI was considered to be a bigger problem, this variable was removed from the list of candidate variables

Step 2: *SPEED_LI excluded*
Model 2: *RUR_URB, SHO_RIGH, ROW_WID, FUNC_SYS, SIGNALS, UNSIG, P_PAR*

After speed limit is excluded from the analysis, Model 2 becomes the best model and includes the variables shown. P_PAR is one of the variables that are included in this model. It should be noted that the presence of peak hour variables is acceptable when modeling multi-vehicle crashes unlike single vehicle crashes. This is because, intuitively, multi-vehicle crashes should be associated with congested traffic conditions unlike what was found with single vehicle crashes. However, the inclusion of P_PAR has other

problems. Peak parking as described previously has "presence" or "absence" as its levels. Amongst the eight sites that were considered in the present study, only one allowed parking. Hence, there is definitely not sufficient representation from all classes of this variable. It appears that this variable is acting as a dummy id for the Rte 1 site. Hence, P_PAR was not considered for further analysis. It should be noted that this model includes ROW_WID just as in Model 1 and this remains another shortcoming. Apart from these arguments, the model performs well. The variables take expected signs, are all significantly different from zero and the model has a low AIC value. An important observation from Models 1 and 2 is that the coefficient on shoulder width is positive. This suggests that multi-vehicle crashes are associated with wider shoulders. This point will be discussed in greater detail later on in this chapter.

Step 3: *P_PAR excluded*
Model 3: *SPEED_LI, SHO_RIGH, ROW_WID, UNSIG, SIN_TR_P*

Model 3 drops out functional system but includes speed limit, shoulder width, right of way width, number of unsignalized intersections and the peak percentage of single unit trucks. This model is similar to Model 1 in the variables included and their signs (except location type), and suffers from the same drawbacks. The right of way width and speed limit take expected signs (negative). It is observed again that the shoulder width has a positive coefficient. It appears from these three models that the functional system is an important variable for the performance of the multi-vehicle crash model. Hence, as the next step, FUNC_SYS was retained but SPEED_LI and ROW_WID were excluded from the analysis in order to improve the model.

Step 4: *FUNC_SYS included, SPEED_LI and ROW_WID excluded*
Model 4: *SHO_RIGH, FUNC_SYS, SIGNALS, P_PAR, SIN_TR_D*

Model 4 is among the best models for multi-vehicle crashes. The model has a very low AIC value and the variables take predictable signs. For instance, the "percent daily trucks" and the "number of signalized intersections" variables have positive coefficients. The only drawback of this model is the inclusion of the peak parking variable but it turns out that this variable is not significantly different from zero, and hence was removed from further analysis. Before this step was performed, it was checked if UNSIG or SPEED_LI could make Model 4 better.

Step 5: *ROW_WID, UNSIG included*
Model 5: *SHO_RIGH, ROW_WID, FUNC_SYS, SIGNALS, UNSIG, P_PAR, SIN_TR_D*

Model 5 considers right of way width and the number of unsignalized intersections in addition to the variables considered in Model 4. This model has neither of the two variables right of way width and unsignalized intersections, significantly different from zero. Hence, it was established that these variables do not add any more information to Model 4, thereby making Model 4 better. As mentioned earlier, the only drawback of model 4 is the presence of the parking variable. This variable was thus removed in the next step.

Step 6: *P_PAR, ROW_WID and UNSIG excluded*
Model 6: *SHO_RIGH, FUNC_SYS, SIGNALS, SIN_TR_D*

Model 6 appears to be the best model in predicting multi-vehicle crashes on two-lane roads. It includes shoulder width, functional system, number of signalized

intersections, and percent daily trucks as its selected variables. As observed with the previous models, shoulder width has a positive coefficient. The signalized intersection variable and the percent daily truck variable have positive coefficients conveying an increasing relationship with multi-vehicle crash rates. The model has all continuous variables statistically significant and a low AIC value. On these bases, this model was chosen for further investigation.

Functional system happens to be the only non-continuous variable included in the final multi-vehicle crash model. The next step in reforming the model was to check if this categorical variable had all its classes significantly different from each other. Thus, paired t-tests were performed between the classes of the functional system variable for Model 6. The results are displayed in Table 14.

Table 14

Comparing Difference Between Levels of Functional System Class
(Multi Vehicle Accident Models)

	Rural Principal Arterial (2)	**Rural Minor Arterial (6)**	**Rural Major Collector (7)**	**Urban Principal Arterial (14)**
Rural Principal Arterial (2)	-	8.96	7.64	9.02
Rural Minor Arterial (6)	8.96	-	**1.51**	**0.48**
Rural Major Collector (7)	7.64	**1.51**	-	**1.57**
Urban Principal Arterial (14)	9.02	**0.48**	**1.57**	-

Bold face indicates no significant difference between indicated classes at 95% confidence.

The functional system variable for these data has four classes coded as 2, 6, 7 and 14. Each of these is explained in detail in Chapter 3. From the table of t-statistics, it appears that class 2 is significantly different from the other classes, but there is no evidence to suggest any differences among the rest of the classes. Thus the variable classes were merged into two groups (class 2 and all others) and Model 6 was re-evaluated. Results from this estimation are displayed in Table 15 as Model 7; the t-statistic for level 2 of the functional system suggests that it is significantly different from the class representing the rest of the levels for this variable. As the results indicate, the other variables do change in sign, and remain statistically significant even after the merging of the levels of functional system.

Model 7 for multi-vehicle crashes includes shoulder width as one of its variables. Shoulder width has been included in this model as a continuous variable of the first degree (exponent = 1.00). A problem we often face with continuous variables of the first degree is that non-linear relationships cannot be captured. For instance, by including shoulder width in the present form, we are forcing the regression to fit a straight line (by estimating a single coefficient) to the observed data. However, as Chapter 2 elaborates, previous studies have indicated that the crash rates might not be linearly related to shoulder width. In order to incorporate the possibility of a non-linear relationship with crash rate, models were estimated with shoulder width included as a categorical variable along with the same variables as Model 7. Based on Figure 6, which plots crash rate versus shoulder width, a number of different categorizations were tried and the results from this analysis are displayed in Table 15 as Models 8, 9 and 10.

Table 15

Final Models: Multi Vehicle Crashes

Variable	Model 7	Model 8	Model 9	Model 10
Intercept	-15.87 (-14.69)	-14.72 (-14.07)	-15.32 (-11.13)	-18.57 (-12.89)
FUNC_SYS (6, 7, 14)	Base	Base	Base	Base
FUNC_SYS (2)	-5.79 (-9.83)	-3.34 (-8.79)	-4.05 (-7.55)	-7.98 (-7.09)
SIGNALS	2.31 (7.78)	1.50 (4.40)	2.53 (6.39)	3.55 (8.53)
SIN_TR_D	1.26 (6.84)	1.01 (5.39)	1.40 (4.92)	2.24 (8.17)
SHO_RIGH	0.44 (5.15)			
SHO_RIGH (<= 2)		Base		
SHO_RIGH (> 2)		1.16 (5.07)		
SHO_RIGH (<= 3)			Base	
SHO_RIGH (> 3)			0.71 (1.84)	
SHO_RIGH (<= 4)				Base
SHO_RIGH (> 4)				3.79 (4.48)
Exponent on Exposure	0.88 (13.83)	0.90 (13.57)	0.87 (11.67)	0.91 (13.22)
Overdispersion Null Deviance Resid. Deviance AIC Value	1.30 788.97 370.88 382.88	1.32 788.97 374.07 386.07	1.69 788.97 400.26 412.26	1.33 788.97 384.67 396.67

Values in parentheses are t-statistics after correction for overdispersion

Model 8 includes shoulder width in the form of two classes (<=2 ft and > 2ft). This model, though it has significantly different shoulder classes, performs poorly. The deviance and hence the AIC value is very large for this model. Model 9 is an attempt at another classification of shoulder widths based on the graph in Figure 6. From this figure, it appears that the multi-vehicle crash rate increases with shoulder width up to a width of 3 ft after which it starts decreasing again. This model, though expected to perform well, also has high deviance and AIC values and performs much worse than any of the models. This is probably because the trends within the classes are not constantly increasing or decreasing. Model 10 tries another classification of shoulder widths (<=4ft and > 4ft) but also performs poorly.

Surprisingly, none of these models perform as well as the original model, where shoulder width was a continuous variable. It was thought that the models with shoulder classes would perform at least as well as the model with shoulder width as a continuous variable. However, it appears that none of the classifications that were attempted seemed to improve the existing model. This is probably because, though there might be internal fluctuations in the trends, there seems to be an overall increase in crash rates with an increase in shoulder widths.

Some Important Observations

It can be observed that in modeling multi-vehicle crashes, the data were now overdispersed. This condition is contrary to what was observed with single vehicle crashes. Two reasons could be cited for this condition. First, multi-vehicle crashes by nature exhibit large variances in their occurrence. In other words, their occurrence is

more unpredictable than single vehicle crashes that seem to behave in a much more orderly fashion. Another reason might be there were not enough data on single vehicle crashes. Overdispersion in crashes is more in agreement with previous experiences in crash modeling. Also, the exponent on the exposure term was uniformly less than 1.00. This again confirms the non-linear relationship between crashes and traffic flow.

It is a little surprising that the LOS did not turn out to be an important variable for multi-vehicle crashes. This may be because the poor LOS was not sufficiently represented in the dataset. This might also have to do with the fact that a large portion of the multi-vehicle crashes occur at intersections and the LOS variable was computed for the section between the intersections, not at the intersections themselves.

It was noted that shoulder width had a positive coefficient indicating that multi-vehicle crashes were associated with wider shoulders. This may be because the presence of wider shoulders encourages drivers to attempt to use shared lanes as protected left and through lanes (Persaud and Mucsi 1995) causing crashes even when it is not safe to do so. To be sure, the data suggest that crash rate decreases with an increase in shoulder width up to about 3 feet, then increases again. Bared and Vogt (1997) also reported this finding.

Functional system was an important variable in explaining multi-vehicle crashes. Though this is not a fundamental variable like shoulder width or lane width, it is important because it can be thought of as representing driver expectation. For instance, based on the design and the location of a particular road segment, the driver automatically expects to drive at a particular speed and this may be completely different from the speed limit or safe speed at the site. This perception by the driver and the

roadside environment can be thought of as being represented by the FUNC_SYS variable.

Chapter 7: Conclusions and Recommendations

The current study involved the development of prediction models for single and multi-vehicle crashes on two lane roadways. Data on geometric characteristics, hourly traffic volumes, number and type of crashes etc. were obtained from the Connecticut Department of Transportation. In addition to this, light data was also incorporated, by computing the sunrise and sunset times at each of the eight sites included in this study.

A Poisson regression analysis was performed to obtain separate models for single and multi-vehicle crashes. In both of these models, traffic volume turned out to be the most important predictor. Specifically, traffic volume was found to be related to both single and multi-vehicle crashes in a non-linear fashion. Single and multi-vehicle crashes seemed to occur under quite different conditions and were caused by different factors. For instance, it was found that single vehicle crashes mostly occurred during less congested times of the day (high levels of service) and decreased as congestion increased (low levels of service). On the other hand, multi-vehicle crashes did not show any such clear correlation with the congestion levels.

As far as geometric characteristics go, single vehicle crashes increased with decreasing shoulder width while multi-vehicle crashes showed the exact opposite trend. The other important causal geometric factor for single vehicle crashes was sight distance with the crashes more likely at lower sight distances. For multi-vehicle crashes, in addition to shoulder width, an increase in the number of signals and the percentage of single unit trucks seemed to increase the number of multi-vehicle crashes observed.

Model form was found to be a critical factor that determines how well a prediction model performs. This has been proved time and again in the past and this current research further confirms this. For instance, we might know that a particular explanatory variable is very powerful in explaining crashes at a site but what is also important is the way in which this variable is introduced inside the model. This needs knowledge of the exact relationship between the dependent variable (crashes) and the predictor (say traffic flow). Unfortunately, in most cases we are not aware of the exact nature of this relationship.

We faced this exact problem with respect to traffic flow. It was seen that crashes were very highly correlated with respect to traffic flow. By modeling crash rate instead of crashes or by assuming that the exponent of exposure in the model is 1.00, we are implicitly assuming a linear relationship between crashes and exposure. Since this was not a proven fact, an assumption such as this might not be valid. In order to test this, and to give the model more flexibility, exponent on exposure was computed through the modeling process. It was found that this exponent was statistically different from 1.00 thereby rendering the linear relationship hypothesis invalid.

The volume of traffic as was mentioned earlier, seems to be the single most important predictor of crashes. This was because in both single and multi-vehicle crash models, traffic flow (exposure variable) explained most of the variance in observed crashes. This being the case, a good model requires accurate data about the traffic conditions in the site, preferably even hourly volume counts. Use of Annual Average Daily Traffic (AADT) to approximate the vehicle miles traveled at a site might reduce the natural variance that exists in exposure data and this might result in heavy

underdispersion. Apart from retaining the natural variation in traffic data, the hourly volumes also give us important clues about the congestion levels at the site in the form of the level of service variable (LOS). This proved to be an important variable in predicting single-vehicle crashes.

Surprisingly, multi-vehicle crashes did not seem to depend on LOS. This is probably because, most of the multi-vehicle crashes occurred at intersections and LOS at a controlled intersection can be completely different from that computed at mid-block. However, the current study did not have enough data to compute the LOS at intersections separately. This being the case, the LOS computed at mid-block was substituted for the intersection LOS. This was probably why the analysis did not pick up a strong relationship between the multi-vehicle crashes and LOS.

It has been suggested (Frantzeskakis 1983) that the exposure at intersections should be expressed as the product of the volumes on the major and the minor approaches. Again, due to lack of data on the minor street volumes and information about the exact location of accidents, intersection accidents could not modeled separately. This might have been a large source of variation in the accidents observed.

The final models for single and multi-vehicle accidents emerged after a series of steps that involved the removal of a number of variables. These variables were included in course of the initial variable selection done by the statistical package. However, these variables had to be examined carefully by the researcher to make sure that they made engineering sense. For instance, the percentage single unit trucks in the peak hour (SIN_TR_P), was one of the most important variables in predicting single vehicle accidents. This made no engineering sense because "SIN_TR_P" happens to be a peak

hour variable and single vehicle crashes are unlikely to occur during peak hour conditions. Similarly, the number of signals (SIGNALS), was another variable that emerged as one of the important variables in predicting single vehicle accidents. This again made no engineering senese because single vehicle crashes are much more likely to occur at mid-block. Apart from correlation between candidate variables, no other reason could be cited for the above phenomenon. Similar trends were also observed in modeling multi-vehicle accidents.

A major pitfall of a regression analysis with categorical variables is the lack of adequate representation of the different classes of a variable. For instance, peak parking was an important variable in predicting multi-vehicle accidents. It had two different classes, "Parking Allowed" and "Parking Not Allowed". The variable had to be removed because the lack of variation precluded its true effect to be computed. Instead, it was likely acting as a site specific dummy variable for Rte. 1, which happened to be the only site that allowed parking. Similarly, there were some variables such as climate zone and lane width which were also not considered because there was not enough variation in the variables themselves.

Among the site variables, shoulder width was an important variable for both single and multi-vehicle accident prediction. Interestingly, single vehicle crashes were found to be associated with narrow shoulders while multi-vehicle crashes were found to be associated with wider shoulders. These finding might have some important implications. Since single vehicle crashes are mostly "run-off road", this might mean that a wider shoulder is actually reducing the risk of such a crash, probably because drivers have more time to react. Also, these wider shoulders are causing more multi-vehicle

crashes probably because drivers are tempted to go around other stopped vehicles. This is probably leading to rear-end collisions.

At every step, it was found that single-vehicle crashes had completely different characteristics than multi-vehicle crashes. This validated the split modeling that was hypothesized. Strangely, light conditions did not seem to affect any of these crashes possibly due to a correlation with LOS or other variables.

Suggestions for Future Research

An important area in which the dataset was lacking was the number of sites considered and the number of years of crash data. Because of these limitations, crashes could be split only into single and multi-vehicle categories and not any further. With additional data, accidents could further be split into their types such as rear–ends, angle collisions, head-ons, side swipes and run-off road. However, if such a modeling is undertaken, it should be ensured that there is sufficient data for each type of accident. At the same time, data should not be manually picked to ensure this, since it might cause a bias. It was found that the two types considered (single and multi-vehicle) were completely different; a further split might produce even better results.

As stated previously, this study did not model intersection accidents separately. If information on the exact location of the crashes on the site and the location of intersections is obtained, this modeling can be done more effectively. Further, it would be more useful if one could obtain the volume counts on the minor roadway also. Apart from

this, it might help to get data on the accurate percentage of trucks and other types of vehicles also.

Finally, an empirical Bayesian approach that takes into account the history of the roadway and the crashes observed on it could produce much better models. As we know, the Poisson process is a "memoryless" process. In other words, just the fact that a site has experienced a high crash rate is not enough reason for the process to assume that the site will observe a high crash rate in the future. However, this historical information about the site is made use of in a Bayesian approach. Probably, a combination of Poisson regression and an empirical Bayesian method may be considered. This way, the extraneous influences on the roadway that cannot be represented by any variable can be accounted for.

References

Agresti, A. (1984) Analysis of Ordinal Categorical Data. New York, NY: John Wiley.

Agresti, A. (1990) Categorical Data Analysis. New York, NY: John Wiley.

American Association of State Highways and Transportation Officials (1994). A Policy on Geometric Design of Highways and Streets: 1994. Washington, D.C.

Anderson, Howard L. (1976) "Let's Try to Dispel Some Highway Safety Myths," Traffic Engineering, Vol. 46, No. 12, July, pp. 47-53.

Barbaresso, James and Brent Bair (1983) "Accident Implications of Shoulder Width on Two-Lane Roadways." Transportation Research Record 923, pp.90-97.

Bared, Joe G. and Vogt, A. (1997) "Highway Safety Evaluation System for Planning and Preliminary Design of Two Lane Rural Highways"

Bozdogan, H. (1987) "Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions." Psychometrika, Vol. 52, No. 3, September, pp. 345-370.

Bureau of Transportation Statistics (1995) Transportation Data Sample CD-ROM disk., U.S. Department of Transportation.

Cameron, A. and Trivedi, P. (1990) "Regression Based Tests for Overdispersion in the Poisson Model." Journal of Econometrics, Vol. 46, July, pp. 347-364.

Dewar, Robert E. (1991) "The Driver: Improving Performance to Improve Safety." ITE Journal, Vol. 61, No. 7, July, pp. 33-37.

Dobson, Annette J. (1996) An Introduction to Generalized Linear Models. New York, NY: Chapman and Hall.

Eliason, Scott R. (1993) Maximum Likelihood Estimation: Logic and Practice. Newbury Park, CA: Sage Publications Inc.

Foley, James L. Jr. (1991) "Traffic Safety Retrospective." ITE Journal, Vol 61, No.7 , July, pp. 61-64.

Frantzeskakis, John M. (1983) "Accident Analysis on Two Non-Controlled Access National Highways in Greece." ITE Journal, Vol. 65, No. 10, October, pp. 32-34.

Greene, William H. (1992) LIMDEP Version 6.0, Bellport, NY: Econometric Software Inc.

Gwynn, D.W. (1967) "Relationship of Accident Rates and Accident Involvements with Hourly Volumes." Traffic Quarterly, Vol 23, July, pp. 407-418.

Hadi, Mohammed A. et al. (1995) "Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression." Transportation Research Record, No.1500, pp. 169-177.

Henderson, Michael (1971) "Human Factors In Traffic Safety: A Reappraisal," Traffic Accident Research Unit, Department of Motor Transport, New South Wales, Australia.

Hensing, David J. (1991) "The Roadway Environment: Progress in Making It Safer." ITE Journal, Vol 61, No.7 pp. 26-31, July.

Ivan, John N. and O'Mara, Patrick J. (1997) "Prediction of Traffic Accident Rates Using Poisson Regression." Presented at Transportation Research Board Annual Meeting, Washington, D.C., Paper No. 970861.

Joshua, Sarath, and Nicholas Garber (1990) "Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models." Transportation Planning and Technology Vol. 15, pp. 41-58.

McCullagh, P. and J.A. Nelder (1989) Generalized Linear Models. New York, NY: Chapman and Hall.

McShane, William and Roger Roess (1990) Traffic Engineering. New Jersey, NJ: Prentice Hall.

Miaou, Shaw-Pin, et al. (1992) "Relationship Between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach." Transportation Research Record 1376, pp. 10-18.

Mohamedshah, Yusuf M. et al. (1993) " Truck Accident Models for Interstates and Two-Lane Roads." Transportation Research Record 1407, pp. 35-41.

Ott, Lymann R. (1993) An Introduction to Statistical Methods and Data Analysis. Belmont, CA: Duxbury Press.

Polanis, Stanley F. (1995) "Some Thoughts About Traffic Accidents, Traffic Safety and the Safety Management System." ITE Journal, Vol. 65, No. 10, pp. 32-34, October.

Robinson, Carlton C. (1991) "Safety – An Important Responsibility." ITE Journal, Vol. 61, No. 7, pp. 21-24, July.

S-PLUS User's Manual, Version 3.3 for Windows (1995). Seattle, WA: Statistical Sciences Division, Mathsoft, Inc.

SPSS for Windows, Base System User's Guide, Release 6.0 (1993). Chicago, IL: SPSS Inc.

Valavanis, Stefan. (1959) Econometrics: An Introduction to Maximum Likelihood Methods. New York, NY: McGraw-Hill Book Company, Inc.

The World Almanac and Book of Facts: 1996. New Jersey: World Almanac Books.

Wright, Paul H. and Radnor J. Paquette (1987) Highway Engineering 5th ed. New York: Wiley and Sons.

Zeeger, C., Deen, R., and Meyes, J. (1981) "Effect of Lane and Shoulder Widths on Accident Reduction on Rural, Two-Lane Roads." Transportation Research Record 806, pp.33-43.

1994 Highway Capacity Manual (1994). TRB Special Report 209, National Research Council, Washington, D.C.

Appendix:

A Method of Identifying Hazardous Highway Locations

Using the Principle of Individual Lifetime Risk

A Method of Identifying Hazardous Highway Locations Using the Principle of Individual Lifetime Risk*

Paul J. Ossenbruggen**

Introduction

A scientific method for identifying hazardous highway locations is presented. The method employs the basic principles of probability and expected value theory, in which motor vehicle accidents are treated as random events. The risk R is defined as the expected loss or damage associated with the occurrence of a harmful event and is calculated as the product of $R = h\theta$ where h is the number of individuals exposed to a given harmful event, and θ is the probability of the event taking place. For the purpose of identifying a hazardous highway location, θ is the probability that an individual will be killed in a motor vehicle accident within a given year and h is the number of vehicle trips made at a given location. The highway risk R is therefore the expected number of fatal accidents per year for a given highway location.

A primary source of accident information is a report form that contains over thirty items.[1] These items collectively characterize a crash, its outcome, and the possible cause. State regulations vary, but typically a crash involving property damage over $1,000, injury or death must be reported. The form includes several items describing the motor vehicle(s), driver(s), occupants, and crash location, with space provided for a description of the accident and a collision diagram. Photographs of the event and surroundings are often attached. To illustrate the use of the hazardous highway location identification method, counts of fatality and injury producing collisions were

* The work was supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program.

** Dr. Ossenbruggen is Professor of Civil Engineering, University of New Hampshire. He holds B.S., M.S., and Ph.D. (Cvil Engineering) from Syracuse University. Email: *pjo@ciunix.unh.edu.*

1 Uniform Police Accident Reporting Form.

obtained from the Town of Durham, New Hampshire (NH), Police Accident records from 1990 to 1997. Values for the highway exposure h were obtained from traffic count records,[2] and individual demand for highway services was obtained from the National Personal Transportation Survey.[3]

Table 1 contains a list of factors describing motor vehicle accidents in Durham and throughout the nation.[4] Investigation of the accident records generally shows traffic accidents are rare. For example, an estimated 30 million trips were made in Durham each year, yet, during the same time period, only about 250 injury and property damage accidents and one fatal accident are reported. Expressed as a probability, the chance of an accident resulting in either injury or property damage is about 8 in 1,000,000. The probability of a fatal collision, estimated to be 3 in 100 million, is much smaller. For this reason, probability theory is used to derive a lifetime highway risk model and to develop a method for hazardous highway location identification.

The use of the term "hazardous highway location" might suggest that the purpose of developing the method is solely to identify poorly designed highways. Clearly, given the factors listed in Table 1, fatal accidents may occur on even the best designed highways. The objective of the model and identification method is to identify those locations that have an incidence of fatal crashes which is higher than what is considered acceptable. Once a hazardous highway is identified, factors including poor design, driver error, traffic congestion, poor weather conditions, and lax law enforcement, can be investigated to determine the cause or causes of the accident.

The Lifetime Highway Risk Model

A model for calculating θ, the probability that an individual will be killed in a fatal crash over his or her lifetime, is derived from geometric and Poisson probability distributions.

---

2 N.H. Dept. Transp., Bur. Transp. Planning, **Automatic Traffic Recorder Data for 1990-1994**.

3 **Stat. Abstract U.S.**, *National Personal Transportation Survey — Summary of Travel Trends 1969 to 1990,* 86 (115th ed. 1995).

4 Nat'l Highway Traffic Safety Admin. (NHTSA), **Traffic Safety Facts** 14, 43, 144–45 (1994).

The trip number in which an individual is killed is assumed to have a geometric distribution. The probability that an individual will be killed on trip number T, where t = 1,2, ... ∞, is expressed as

$$P(T = t) = \omega(1 - \omega)^{t-1} ,$$

where ω is the probability that an individual will be killed in a single motor vehicle trip. In other words, P(T = t) is the probability that an individual will make t - 1 trips without being killed and then will be killed on trip t.

Table 1
Motor Vehicle Crash Factors and Driver Characteristics

- *Crashes are Rare*. The Durham Police report about 250 motor vehicle crashes per year. It is estimated that over 30 million trips are made annually in Durham, NH.
- *Crashes Vary with Traffic Conditions.* Crashes are often assumed to be related to traffic congestion and episodic events. An episodic event, such as those triggered by special functions, e.g., sporting events and concerts, is suspected of causing traffic shock waves that frequently surprise drivers and cause a chain reaction of crashes.
- *Crashes Vary by Collision Type.* Accidents involve either single motor vehicles, two or more vehicles and pedestrians. NHTSA for 1994 reported that, nationwide, 20,505 fatality crashes involved a single vehicle; 15,718 fatality crashes involved multiple vehicles; and 5,472 fatality crashes involved pedestrians.
- *Crashes Vary Spatially.* In 1994, approximately 60% of crashes in Durham occurred on in-town, high-volume roads and parking lots. The remainder occurred on out-of-town, high-speed roads.
- *Crashes Vary Temporally.* Accidents are reported at different times of the day and in different seasons. In 1994, the number of weekday fatalities reached a nationwide peak of 6 fatalities per hour between 3:00 PM and 3:59 PM. The number of weekend fatalities reached a peak of 6.5 fatalities per hour between 1:00 AM and 2:59 AM.
- *Crashes Vary with Driving Conditions.* Accidents are reported for wet, dry and icy pavements.
- *Crashes Vary with Drivers' Physical Condition.* Drivers are involved in accidents when sober or under the influence of alcohol or drugs. NHTSA reports that 41% of all fatal crashes in 1994 involved alcohol. A driver's age can also affect his or her reaction time.
- *Crashes Vary with Driver Attitude.* NHTSA reports that young drivers tend to speed, and twice as many males as females are involved in accidents.
- *Crashes Vary with Driver Experience.* In 1994, NHTSA reported that 16–20 year-olds had the highest fatality rate (30.7 per 100,000) and 55– 64 year-olds the smallest fatality rate (10.7 per 100,000).

An individual is assumed to make a total of n trips in a lifetime. Mathematically, an individual is a survivor if the total number of trips T exceeds the total number of trips n an individual can make in a lifetime. The probability that an individual is a survivor is denoted by P(T > n)

and is determined by summing $P(T = t)$ over all trip numbers t greater than or equal to $n + 1$. After simplifying, the survival probability, which is expressed as a conditional probability since n is given, is

$$P(T > t|N = n) = \omega^n \quad .$$

The number of trips that an individual makes in a lifetime is assumed to be a random variable N with Poisson distribution,

$$P(N = n) = (e^{-\eta}\eta^n)/n! \quad ,$$

where η is the mean number of trips made by an individual in a lifetime. The probability that an individual is a survivor, expressed in terms of N and $P(T > N)$, is calculated by summing the product $P(T > t|N = n) \times P(N = n)$ for n equal to and greater than zero. After simplifying this expression, the product reduces to $P(T > N) = \exp(-\eta\,\omega)$. Since $P(T > N) + P(T \leq N) = 1$, the probability that an individual will be killed in a motor vehicle crash is given by the compound distribution[5]

$$\theta = P(T \leq N) = 1 - \exp(-\eta\,\omega) \quad . \qquad (1)$$

This lifetime highway risk model forms the basis of the hazardous highway location identification method.

A Safety Compliance Standard Using Individual Lifetime Risk

The lifetime highway risk model is a function of an individual's demand for highway services η, and the probability of a fatal crash in a single trip ω. To develop a method of hazardous highway location identification, a "statistical traveler" will be defined and the traveling behavior of the "statistical traveler" will be used to assign the model parameter η. Concepts of public health risk assessment of chronic low-level exposure to chemical contaminants and the public's perception of highway risk will be used to assign θ and, in turn, to determine ω.

The "Statistical Traveler." According to the National Personal Transportation Surveys,[6] the average number of daily trips per household for 1990 was reported to be 4.66. Given that there were 2.56 persons per household, the average person traveled about nine miles per day while making 1.82 trips. In 1990, the average person made about about 664.4 trips and traveled slightly less than 6,000 miles per year.

5 Marcel F. Neuts, **Probability** 224 (1973).

6 *See supra* note 3 at 636.

For the purposes of hazardous highway location classification, 1990 is assumed to be the base year and the "statistical traveler" makes η = 664.4 trips per year.

Public Health Considerations: Since a lifetime highway risk probability is the same measure of effectiveness as that used in the public health risk assessment of chronic low-level exposure to chemical contaminants,[7][8] public health and highway risks are therefore comparable. As a result, the assignment of an acceptable lifetime risk probability θ^* for a toxic chemical will be used as a guide for assigning an acceptable lifetime risk probability θ^* for highways.

A national public health goal is to minimize the probability that an individual will die prematurely from chronic low-level exposure to a toxic chemical. For the purposes of risk assessment, a premature death occurs when an individual dies from such low level exposure before reaching 70 years of age. The probability that an individual dies prematurely from chemical exposure is generally accepted to be on the order of θ^* = 1 in 1,000,000. The aim of a public health regulator is to determine an acceptable daily intake (ADI) for humans such that a premature death occurring has a probability of θ^*.

Animals are typically exposed to heavy dosages of chemicals relative to the animal's weight. The data are used to develop a dose response function, which is used to determine a virtual safe dose (VSD). Once known, an acceptable daily intake is determined from ADI = VSD/sf where sf is a safety factor dealing with uncertainties associated with the use of simple mathematical model structures; extrapolation of animal response data from high to low chemical doses; biological, intake and weight differences between animals and humans; and unknown chemical effects on humans. Depending on the level of uncertainty, safety factor assignments range in magnitude from 10 to 1,000.

The procedure adopted for highway risk will adopt the assumption that a premature death is one that occurs before 70 years of age; i.e., the "statistical traveler" is assumed to have the same life span of 70 years. Given a fixed lifespan and θ^*, the annual and single trip risk probabilities of θ and ω can be determined.

7 Leonardo Ortolano, **Environmental Regulation and Impact Assessment** 385-392 (1997).

8 Paul Ossenbruggen, **Fundamental Principles of Systems Analysis and Decision-Making** 193-200 (1994).

A Highway Safety Compliance Standard: The lifetime highway risk probability will not be assigned a value as small as given for chemical exposure: $\theta^* = 1$ in 1,000,000. Society will generally accept a higher level of highway risk than chemical risk. Society's perception and acceptance of these risks are summarized in Table 2. For these reasons, chemical risks, particularly those associated with carcinogenic chemicals, are considered dread risks. Despite public awareness of their dire consequences, motor vehicle accidents are considered less threatening than dread risk. Consequently, a highway safety compliance standard of $\theta^* = 1$ in 1,000 is considered to be a reasonable assignment of risk. Statistical evidence will illustrate that this assignment is sufficiently rigorous because, if met, there would be a six-fold decrease in the number of fatal collisions reported nationally.

Given $\theta^* = 1$ in 1,000 and η = (664.4 trips per person per year) x (70 years per lifetime), or 46,508 trips in a lifetime, a value of 2.2 in 100 million is obtained for ω using the lifetime highway risk model. Substituting $\eta = 664.4$ and $\omega = 2.2$ in 100 million into the lifetime highway risk model once again, an annual value of $\theta = 1.4$ in 100,000 is obtained for the highway compliance standard.

The same highway safety compliance standard of $\theta = 1.4$ in 100,000 is assumed to apply to all categories of highway systems. That is, freeways, two lane undivided highways, local roads, etc. are expected to provide the same level of safety. The assumption of a universal standard differs from current practice of hazardous highway identification, which categorizes highways by highway system type, location (urban or rural), and other features. This point will be explored in greater detail in the Discussion section.

The Method of Hazardous Highway Location Identification

Given the definition of risk $R = h\theta$ and the highway safety compliance standard θ, the numerical value of R can be calculated. The value of R is assumed to be an acceptable (or critical) number of fatal crashes per year for a given location. Similarly, a highway safety standard for injury accidents, R_I, will also be established. Given R and R_I and the fatality and injury counts, C and C_I, it is a simple matter to identify a highway location as being either safe or hazardous. In this

section, the focus is on developing a fundamental understanding of the definition of risk and how it applies to the method of hazardous highway location identification.

Fatal Accidents: Each vehicle that passes a specific spot on a highway is considered to be a candidate for a fatal motor vehicle accident. Consequently, the average daily traffic (ADT) level is considered the best and most practical measure of exposure; therefore, the exposure h is assumed to be equal to ADT.

The number of fatal accidents occurring at a given spot within a given year is represented by a random variable X. The probability of an individual being killed in a fatal collision is assigned to be θ. The probability of x events in h trials, $P(X = x)$, is typically assigned a binomial distribution. However, since is $h = ADT >> 100$, $\theta << 0.01$, and $h\theta \leq 20$, the distribution of X can be approximated by a Poisson distribution[9] with mean, $\lambda = h\theta = ADT\ \theta$. The acceptable number of fatal crashes at a given location for a given time span is estimated to be

$$R = ADT\ \theta = ADT\ [1 - \exp(-\eta\ \omega)] \qquad (2)$$

If the fatal crash count for a given location C exceeds the expected number of fatal crashes R, then the location is classified as hazardous; otherwise, the location is considered safe.

Injury Accidents: The hazardous highway location method can also be extended to injury crashes. If $C_I > R_I$, then the location is identified as hazardous. The principle of conditional probability and national highway injury and fatal crash counts are used to establish a safety compliance standard θ_I for injury crashes and, in turn, R_I.

According to NHTSA, motor vehicle accidents, 1988–94 ranged from 6 million to almost 7 million annually. During this period, the percentages of fatality and injury causing crashes remained almost constant at 0.6% and 32%, respectively. These data will be used to estimate the probability of a injury crash in a single trip ωI.

The conditional probability δ that, given an injury producing crash, it will be fatal is estimated to be the ratio of the number of fatal accidents to the number of injury producing accidents, or $\delta = 18/1{,}000$. The probability of a fatal crash is the product of its conditional probability given an injury-producing crash times the probability of an injury-producing crash or $\omega = \delta\omega_I$. Given $\omega = 2.2$/million and $\delta =$

9 Jay L. Devore, **Probability and Statistics** 114 (1987).

18/1,000, the value of ω_I is calculated to be ω_I = 1.2/million. Substituting h = 664.4 and ω_I into the lifetime highway risk model, $\theta_I = 1 - \exp(-\eta\ \omega_I)$, a value of the highway safety compliance standard for injury crashes is calculated to be θ_I = 7.8/10,000. An acceptable number of injury crashes at a given location is then:

$$R_I = ADT\ \theta_I = ADT\ [1 - \exp(-\eta\ \omega_I)] \quad . \qquad (3)$$

Classification: If either C > R or $C_I > R_I$, then the highway location is classified as hazardous; otherwise, it is classified as safe.

Table 2

Public Perception of Highway and Public Health Risks [10]

Category	*Highway*	*Chemical*	*Generalizations*
Degree of Fear	Little	Great	In the U.S., fear of lingering death from chemical exposure death is greater than the fear of sudden death from a vehicle crash.
Controllability	Great	Little	In comparison to a driver, individuals exposed to toxic chemical have little or no control.
Blame and Injustice	Individual	Someone else	A negligent driver can be blamed for damages to involuntary victims and himself. An injustice may have occurred when involuntary victims are involved and the driver is unharmed. When the negligent driver only harms himself, it can be argued that justice has been served. In both cases, it is reasonable to assume that no financial gain is received by the driver. A negligent chemical manufacturer can be blamed for exposing involuntary victims to toxic chemicals while receiving financial benefits from the sale of products. In comparison to a negligent driver, an injustice is perceived to have occurred in this incident.
Exposure Benefits	Great	Little	The personal automobile is considered essential to the economy of the U.S. In comparison, the benefits derived from a chemical tend to affect fewer individuals or com panies.

10 Adapted from Adam M. Finkel, *Comparing Risks Thoughtfully,* 7 **Risk** 325 (1996).

Case Study

Injury and fatal accident counts for Routes 4 and 108, both two-lane, undivided highways in Durham, are used to illustrate the identification method. Route 4 is a primary east-west corridor connecting the capital, Concord, in the middle of the state to Portsmouth on the Atlantic. Route 108 runs north-south.

The fatality and injury counts listed in Table 3 are divided into four groups. The highways are similar, yet each stretch possesses some distinctive characteristics. Route 4 West is a 2.25 mile stretch of roadway with limited access and freeway-type features, including two road-separated interchanges. A signalized intersection is located midway between interchanges. The intersection has a generous right-of-way, having paved breakdown lanes 9.5 feet in width and guardrails located 10 feet from the edge of the driving lane. In contrast, Route 4 East is a three mile section of highway with a narrow right-of-way. Its paved breakdown lanes range in width from 2–9.5 feet, with guardrails located in some places as close as two feet from the edge of the driving lane. Routes 108 North and South have highway characteristics most similar to those of Route 4 East. However, Route 108 does not have paved breakdown lanes.

Table 3

Motor Vehicle Fatal and Injury Accident Counts for Durham, NH [11]

	Route 4 E		*Route 4 W*		*Route 108 N*		*Route 108 S*	
Year	C	C_I	C	C_I	C	C_I	C	C_I
1990	1	8	0	3	0	7	0	3
1991	0	4	0	3	0	3	0	9
1992	2	12	1	2	0	4	0	3
1993	2	8	1	8	0	1	0	3
1994	0	3	0	3	0	3	0	5
1995	0	4	0	3	0	3	0	1
1996	1	6	0	2	0	3	0	6
Ave.	0.86	6.4	0.29	3.4	0	3.4	0	4.3

The average speeds on all these highways are estimated to be least 40 mph. The only exception is the one-half mile portion of Route 4 North, which is a business district with an average speed of about 35 mph. The ADT for Routes 4 East and West is 15,470 vehicles per day.

[11] Durham, NH Police Dept., *TIPS Accident Statistics Report* (computer output sheets, 1990–96).

The ADT values for Route 108 North and South are 10,000 and 9,250 vehicles per day, respectively.

The values of C and C_I for Routes 4 and 108 are shown in Table 3. The seven-year averages of C and C_I given at the bottom of the table are used for classifying a highway location as either safe or hazardous.

Table 4 contains the results of analyses obtained using the hazardous highway location identification method for stretches of highways of length L, expressed in miles. All classifications were made using the procedures described in the previous section.

After further evaluation of the spatial distribution of collisions, Route 4 West shown in Table 4 was reclassified. The method of hazardous highway location identification is derived for a spot location, but can also be applied to stretches of highway, as illustrated in Table 4. Classifying stretches of highways has important practical significance, but it should be realized that classifying crashes for long stretches can inflate the counts of C and C_I, thereby increasing the likelihood that a given stretch of highway will be classified as hazardous. For example, the average C and C_I values for Route 4, a 5.25 mile stretch of highway, are 1.14 and 9.8, respectively. In this case, the inequalities of $C > R$ and $C_I < R_I$ remain the same, but these inequalities may artificially give the impression that the 5.25 mile stretch of Route 4 is hazardous. Spatial distribution of crashes should be therefore considered in such cases.

Table 4
Hazardous Highway Location Classifications for Fatal Crashes in Durham

Location	*ADT*	*L*	*R*	*C*	R_I	C_I	Classified
4 E	15,470	2.25	0.22	0.86	12.1	6.4	Hazardous
4 W	15,470	3	0.22	0.29	12.1	3.4	Safe*
108 N	10,000	1	0.14	0.0	7.8	3.4	Safe
108 S	9,290	3	0.13	0.0	7.2	4.3	Safe

* Note that $C > R$; therefore, according to the hazardous highway location method, the stretch of Route 4 W is classified as hazardous. However, after considering spatial distribution of crashes, Route 4 W was reclassified. See text for explanation.

Investigation of the two fatal accident reports for Route 4 West shows that one crash occurred at a signalized intersection and the other at an interchange. Given this, the C averages for Route 4 West in Table 4 have been modified. The average values at the signalized intersection

and interchange are reduced to C = 0.15. No injury crashes were reported at the interchange. All injury crash counts (C_I = 3.4 per year including a total of eight injury crashes in 1993) are located at the signalized intersection. Since C < R and $C_I < R_I$, the two locations on Route 4 West satisfy the condition for a safe highway location, and the entire 2.25 mile stretch of Route 4 West is therefore classified as safe.

In comparison, all six fatal crashes on Route 4 East listed in Table 3 occurred on Route 4 at four different local street intersections. Two intersections on Route 4 were each the location of two fatal crashes. Given this information and the fact that all four intersections within the three-mile stretch of highway have similar design characteristics, the entire stretch of Route 4 East is classified as hazardous. The average number of injury crashes meets the highway safety standard, but the number of fatal accidents exceeds the safety compliance standard by a factor of four. While the average number of motor vehicle crashes on Route 4 East may be considered small, the crashes that have occurred on this stretch of highway have been extraordinarily violent.

Discussion

A Rigorous Safety Criterion: Since the same safety compliance standard of θ = 1.4 in 100,000 is assumed to be applicable to all highway classifications, the total number of fatal accidents satisfying the highway safety compliance standard can be estimated and compared to the reported number of fatal collisions that occurred nationwide. NHTSA reported for 1990 that there were 39,836 fatal crashes with 47,151 deaths, and 2,122,000 injury producing crashes.

Given 93 million households and 4.66 daily trips per household in 1990, the total number of trips per day is estimated to be TPD = (93 million) (4.66) = 433.4 million. The acceptable number of fatal crashes for θ = 1.4 in 100,000 is TPD θ = (433.4 million) (1.4/100,000) = 6,214. Likewise, the acceptable number of injury crashes for θ_I = 7.8 in 10,000 is TPD θ_I = (433.4 million) (7.8/10,000) = 340,355. The reported numbers of fatal and injury producing collisions exceed the number of fatal and injury producing accidents deemed acceptable by the safety compliance standards by factors of 6.4 and 6.2, respectively. These data give assurance that the highway compliance standards of θ = 1.4 in 100,000 and θ_I = 7.8 in 10,000 are rigorous and, at the same

time, suggest that more work is needed to reduce the number of motor vehicle crashes on the nation's highways.

Highway Safety Trends: In accordance with the lifetime highway risk model, the trip exposure η affects θ and R and, in turn, affects the safety classification of a highway location. The impact of exposure can be most vividly illustrated by example.

In 1969, the U.S. population was 226 million, compared to 249 million in 1990. During this 31 year period, however, the number of motor vehicle trips made per person increased by 50%. According the National Personal Transportation Survey, the average person in 1990 made daily 1.21 trips. The "statistical traveler" of 1969 made $\eta = 442$ trips per year, compared to the "statistical traveler" of 1990, who made $\eta = 664.4$ trips per year. The average trip length of about nine miles per trip has remained constant over this period.

Since an individual's trip exposure was less in 1969, the highway risk R is obviously less than the 1990 value. Given the same value of $\omega = 2.2$ in 100 million as in 1990 and $\eta = 442$ trips per year, the highway safety compliance standard for 1969 is calculated to be $\theta = 9.5$ in 1,000,000, a value less than $\theta = 1.4$ in 100,000 for 1990. Given 62.5 million households and 3.83 daily trips per household, TPD = (62.5 million) (3.83) = 239.4 million trips per day. The acceptable number of fatal crashes for 1969 is TPD θ = (239.4 million) (9.5/1,000,000) = 2,280. The reported number of fatalities for 1969 is 53,543 and the number of fatal accidents for 1969 was estimated to be about 50,000. Both greatly exceeded the acceptable number of 2,280 fatal crashes per year.

The decline in the reported number of fatal crashes from 50,000 in 1969 to 39,836 in 1990 is an indication that the steps taken to improve safety have been effective. Some of the most notable steps have been providing motor vehicles with standard safety equipment such as safety belts and collapsible steering wheel columns; making driving while intoxicated a criminal offense; passing mandatory seat belt laws; and educating the public to drive more responsibly.

Average Accident Rates: Various measures of the average accident rate are used to describe highway safety and identify hazardous locations.[12] The accident rate per 100 million vehicles miles traveled

(RMVM) is widely used in the analysis of accident data and for highway safety comparison. For fatal accidents, the RMVM used for stretches of highway is given by

$$RMVM = (100 \cdot C) / (365 \cdot L \cdot ADT)$$

and for injury crashes, $RMVM_I$ is given by

$$RMVM_I = (100 \cdot C_I) / (365 \cdot L \cdot ADT)$$

Comparing RMVM and $RMVM_I$ values for the Durham highways with various highway system categories for NH and the U.S. shows that the RMVM values for stretches of Route 4 East and West are larger than the values for all of the highway categories listed in Table 5. With the exception of Route 4 West, which was classified as safe after the spatial distribution of crashes was considered, the Durham highway classifications are consistent with the RMVM statistics for the system categories given in Table 5.

Table 5

RMVM Measures for Durham and U.S. Highways [13]

Location	*C*	*RMVM*	C_I	$RMVM_I$
4 W	0.29	2.25	3.43	27.0
4 E	0.86	5.06	6.43	38.0
108 N	0.0	0.0	3.43	93.93
108 S	0.0	0.0	4.29	42.31
	System Category: Urban Principal Arterial for 1992			
NH	16	1.59	1,195	118.91
U.S.	5,246	1.52	488,228	141.85
	System Category: Urban Total Systems for 1992			
NH	29	0.78	1,722	171.34
U.S.	15,202	1.12	781,631	227.09
	System Category: Total Systems for 1992			
NH	110	1.09	6,850	68.04
U.S.	34,928	1.56	2,216,245	98.95

The Critical Accident Rate Factor Method: This method is used to identify possible hazardous highway locations. If the critical accident rate at a location is significantly higher than the average for that highway system type, then the location is considered hazardous. To

[12] Nicholas J. Garber & Lester A. Hoel, **Traffic Highway Engineering** 138-42 (1997); Conn. Dept. Transp., *Accidents Records and Statistics Manual* (1993).

[13] Fed. Highway Admin., **Highway Safety Performance** — *1992* 5, 6, 16, 42 (1995).

illustrate, RMVM is used as a measure of effectiveness. The critical accident rate is calculated as an upper-level confidence level using statewide accident and traffic statistics, calculated from

$$RMVM_{cr} = \overline{RMVM} + Z \cdot S_{RMVM}$$

where $\overline{RMVM}$, S_{RMVM}, and Z are the average, standard deviation, and standard normal random variable for the sample, respectively. The values of Z, for example, are 1.645 for 95% and 2.576 for 99.5% upper confidence levels. A highway segment average is denoted as RMVM; therefore, if RMVM > $RMVM_{cr}$, then the segment is classified as hazardous; otherwise, it is classified as safe.

In practice, the critical accident rate factor method compares the accident history of a highway segment or intersection with the state accident history of the same type. The data are carefully sorted by highway system type, land use (rural and urban), geometric design, and traffic control characteristics. The goal of this method is to identify hazardous highway locations by category. In contrast, the goal of this paper is to identify hazardous highways independent of system type or any other type of categorization.

Sorting the data by highway category may lead to inconsistency and confusion in classification. For example, Routes 4 and 108 are designated to be urban principal arterial highways because the Durham population of over 10,000 people exceeds the required minimum population of 5,000. Given an urban designation, the RMVM values of Durham are compared to areas with much greater population densities. The NH statewide averages of $\overline{RMVM}$ and $\overline{RMVM}_I$ are 1.37 and 35.49 for rural principal arterial highways, and 1.59 and 118.91 for urban principal arterial highways, respectively.

For simplicity, Z = 0 and RMVMcr = $\overline{RMVM}$ and $RMVM_{Icr}$ = $\overline{RMVM}_I$. Given $RMVMI_{cr}$ = 35.49 and $RMVM_I$ = 38.0, Route 108 North is classified as hazardous when designated to be a rural principal arterial highway and, given $RMVM_{Icr}$ = 118.91, it is classified as safe when designated a urban principal arterial highway .

Consider another example dealing with sample variability. Because it reported a value for $\overline{RMVM}$ = 2.05, which exceeds the national average of $\overline{RMVM}$ = 1.56 for 1992, South Carolina may be considered one of the most dangerous states in the nation to drive. Ironically, for

all urban principal arterial highways in South Carolina, no fatal accidents were reported in 1992; therefore, $\overline{RMVM} = 0$, $Z = 0$, and $RMVM_{cr} = 0$. Clearly, this statistic has little or no practical value in hazardous highway identification.

In contrast, the hazardous highway identification method does not lead to these types of anomalies because the method uses the same fatality and injury compliance standards for all highway system types.

Abnormal Accident Rate Experience: The concepts of individual lifetime risk are adapted to identify hazardous locations with abnormal accident rate experience. In lieu of using statistical summaries employed by the critical accident rate factor method, the upper confidence level is calculated using the Poisson distribution. The adaptation makes use of the following steps for a given highway location: (1) estimating $\hat{\theta}$ using a risk definition of $\hat{\theta} = C/TPD$, where C is the number of reported fatal accidents, and (2) calculating the critical X_{cr} for a given confidence level and the Poisson probability distribution with mean $\lambda = ADT \cdot \hat{\theta}$. The individual lifetime risk method departs from the critical accident rate method of sorting accident and traffic data by highway system type, land use, geometric design, and traffic control characteristics. The same steps are used for injury producing accidents.

Table 6
Abnormal Accident Rates for the Highways in Durham, NH

	National Statistics			*Fatal*	*Injury*
Year	*C*	C_I	TPD	θ	θ_I
1990	39,836	2,122,000	433.4M	9.2/100,000	4.9/1,000

		X_{cr}*		X_{Icr}*	
Location	ADT	95%	99.5%	95%	99.5%
Route 4 W	15,470	4	5	90	99
Route 4 E	15,470	4	5	90	99
Route 108 N	10,000	3	4	61	68
Route 108 S	9,250	3	4	57	64

* At two confidence levels

Tables 6 and 7 contain the critical values of X_{cr} and X_{Icr}, for 95% and 99.5% confidence levels, obtained for Routes 4 and 108. Table 6

uses accident counts and TPD for the entire nation, whereas Table 7 uses accident counts and TPD only for NH. The Poisson distribution is a discrete probability distribution; therefore, X_{cr} and X_{Icr} are integers. Since $C < X_{cr}$ and $C_I < X_{Icr}$, Durham highways are not classified as locations with abnormal accident rates. Comparison of $\hat{\theta}$ and $\hat{\theta}_I$, as well as other statistics Tables 6 and 7, shows that, relative to national experience, NH is a safer place to drive.

Table 7
Abnormal Accident Rates for the Highways in Durham, NH

	New Hampshire Statistics			*Fatal*	*Injury*
Year	C	C_I	TPD	θ	θ_I
1992	110	1,978	417,000	5.7 /100,000	1/1,000

		X_{cr}*		X_{Icr}*	
Location	ADT	95%	99.5%	95%	99.5%
Route 4 W	15,470	3	4	23	27
Route 4 E	15,470	3	4	23	27
Route 108 N	10,000	2	3	16	19
Route 108 S	9,250	2	3	15	18

* At two confidence levels

Risk Communication: A most difficult task facing transportation professionals is presenting scientific and technical facts to the public, particularly when it is often hostile and suspicious because a proposal may affect the status quo of a particular community. The results of traffic safety analyses, such as statistical results of the critical accident rate methods, cost-benefit analysis, and other planning tools are usually not appreciated. Expressing highway safety in terms of the number of accidents per RMVM or the loss of a life in monetary terms are often neither understood nor easily accepted. Values of RMVM, for example, are considered by transportation professionals to be valuable for comparing and ranking the safety of different highway systems and studying safety trends. At the other extreme, presenting a proposal without reference to accident counts or other highway-related statistics trivialize the importance of safety.

A benefit of using lifetime risk is that it can be expressed as a probability θ or an expected value measure R. Most people have been exposed to the fundamental ideas of chance. Lotteries are

commonplace. A person can appreciate the notion that an outcome is a rare event if its chance of occurring over one's lifetime is expressed as 1 in 1,000 or, on an annual basis, 1.4 in 100,000. Using the definition of risk as an expected value, the highway safety compliance standard can be restated in terms that may be more easily understood by some lay people. For example, a compliance standard for Route 4 East was determined to have an expected value of R = 0.22 and was used to classify this highway stretch as hazardous. The same classification is obtained by using whole numbers for the expected value of R and rephrasing the definition of a safe highway. In other words, a highway location is defined to be safe if no more than one fatal accident occurs in a five year period. According to the data in Table 3, Route 4 East had two fatal crashes in two successive years and is therefore classified as a hazardous location because it does not meet the above definition.

Recently, the concern that some public health risks are trivial has led to a debate on regulatory risk reform.[14] Highway risk has mostly played a minor role in the debate. When it is discussed, the focus is generally directed at the highway safety cost-benefit analyses that tend to use figures which underestimate the value of life.[15] The concepts for describing highway safety using individual lifetime risk and a highway safety compliance standard of $\theta^* = 1/1{,}000$ brings a different perspective and hopefully better insight to the analysis and discussion of a common risk in life that affects virtually everyone daily. By introducing these highway risk concepts and a goal of achieving a highway safety compliance standard into the regulatory risk reform debate, some of the barriers preventing effective risk communication between highway safety experts and the public may be overcome.

---

14 Stephen Breyer, **Breaking the Vicious Cycle: Toward Effective Risk Regulation** 10-29 (1992). *See also,* John D. Graham, *Edging Toward Sanity on Regulatory Risk Reform* 11 **Issues Science & Tech.** 61–64 (1995).

15 **Risk, Costs and Lives Saved, Getting Better Results from Regulation** 137–49 (Robert W. Hahn, ed.1996).

Conclusions

A scientific framework for hazardous highway location identification is presented that considers both fatality and injury-producing accidents, the concept of individual lifetime risk and incorporates a safety compliance standard. A lifetime risk of 1 in 1,000 was chosen and defended by adopting principles from public health regulation and the public's perception of highway risk. The same safety standard is assumed to apply to all highway system categories. All highways are therefore expected to provide the same level of safety.

Using national accident counts, it was demonstrated that the selection of the value of lifetime risk is a rigorous standard. The application of the method was demonstrated by a case study of undivided two-lane highways in Durham, NH. It was shown that classifying highways with the hazardous highway location method is consistent with others used in practice. Since it employs the same measure of effectiveness used in public health, highway and public health risk can be compared and ranked. Further, since the method employs both probability and expected numbers of fatality and injury-producing crashes as measures of effectiveness, the results may be more easily understood by the lay public.