INVESTIGATION OF A NEW APPROACH FOR REPRESENTING TRAFFIC VOLUMES IN HIGHWAY CRASH ANALYSIS AND FORECASTING

John N. Ivan and Chen Zhang

UNITED STATES DEPARTMENT OF TRANSPORTATION

REGION I UNIVERSITY TRANSPORTATION CENTER

PROJECT UCNR17-7

FINAL REPORT

July 8, 2008

Performed by
University of Connecticut
Connecticut Transportation Institute
Storrs, CT 06269

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle INVESTIGATION OF A NEW APPROACH FOR REPRESENTING TRAFFIC VOLUMES IN HIGHWAY CRASH ANALYSIS AND FORECASTING			**5. Report Date** July 8, 2008 **6. Performing Organization Code**
7. Author(s) John N. Ivan; Chen Zhang			**8. Performing Organization Report No.**
9. Performing Organization Name and Address Connecticut Transportation Institute University of Connecticut, Unit 2037 Storrs, CT 06260-2037			**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** DTRS99-G-0001
12. Sponsoring Agency Name and Address New England (Region One) UTC Massachusetts Institute of Technology 77 Massachusetts Avenue, Room E40-278 Cambridge, MA 02139			**13. Type of Report and Period Covered** Final Report for UTC Year 17, Project UCNR 17-7, Sep 1, 2004 – Aug. 31, 2006 **14. Sponsoring Agency Code**

15. Supplementary Notes
Supported by a grant from the US Department of Transportation, University Transportation Centers Program

16. Abstract

Accident prediction modeling studies are good for identifying correlations between crash risk and explanatory factors, but cannot give definite safety effects of countermeasures related to the significant covariates. Causality can only be proven using crash reconstruction methods or carefully constructed before-after studies, both of which require special data observation or analysis time, making them impractical to apply on large study data sets representing the variation of characteristics in the road network. This report proposes two advances in crash modeling: (1) a collision type categorization based on factors contributing to the occurrence of collisions to support estimation of prediction models that can better identify crash causality, and (2) definition of crash exposure to consider the traffic flow situation that is necessary for specific collision types to occur.

Generalized linear modeling assuming a negative binomial or scaled Poisson distribution is used to estimate prediction models. Evaluations of model goodness-of-fit and diagnostics results are discussed in comparing different model outcomes and assessing the effectiveness of adopting the newly defined collision type categories and crash exposure. Three studies are carried out to evaluate the new exposure definitions and the collision categorizing method based on contributing factors. In study 1, a new exposure, crash opportunities, defined as the number of times vehicles traveling in opposite directions meet, is proposed for opposite-direction collisions including head-on and sideswipe. However, models using crash opportunities do not perform better than those using traffic volumes. Study 2 uses K-means cluster analysis to categorize collisions into categories with similar patterns of contributing factors. The resulting categories are (1) rear-end collisions, (2) other same-direction collisions including sideswipe and turning, (3) intersecting-direction collisions including turning opposite-direction, turning same-direction, intersecting path and angle, and (4) segment collisions including head-on, opposite-direction sideswipe and single vehicle. Study 3 then proposes vehicle time spent following as a new exposure definition for same-direction collisions and evaluates this new exposure for the collision categories from study 2. The new exposure is found to have a linear relationship with the total same-direction crashes, especially rear-end crashes. This linearity shows that vehicle time spent following can well represent traffic flow intensity and state for rear-end crashes to occur. Also, this finding further indicates the value of categorizing collisions by their contributing factors instead of types of crashes.

17. Key Words	18. Distribution Statement
Highway Safety, traffic exposure, crash causality, contributing factors, rear-end crashes, head-on crashes	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages 56+ prefatory material	22. Price

Abstract

Through accident prediction models, researchers have identified correlations between crash risk at specific locations with many different explanatory factors such as traffic volume, roadway geometrics, temporal effects, driver characteristics, and land use. Many studies have been conducted to evaluate the correlation in different contexts, as some focus on segment accidents and others on intersection accidents; however, being correlative rather than causal studies, they cannot give definite safety effects of countermeasures related to the significant covariates. On the other hand, crash reconstruction methods can directly identify actual collision causes, but data demands and time consumption required for these methods cause difficulty for them to be widely applied on large study data sets representing the variation of characteristics in the road network. This eliminates the possibility of deriving clear predictions about the effect of roadway characteristics and other factors that may have influence on crash risk and can be transferred to other contexts, outside of narrowly focused before-after studies. To fill in the gap between the transferability of the results from crash prediction models and the strong causality connections of crash reconstruction analysis, this report proposes a methodology that accounts for crash causality by defining collision type categories, with the crashes in each sharing common contributing factors. A collision type categorization based on crash causality, or factors contributing to the occurrence of collisions, would support estimation of prediction models that offer a more accurate understanding of crash risk correlation for each collision type as well as suggesting appropriate countermeasures for reducing the predicted collisions.

Furthermore, after establishing causality-based collision type categories, this report intends to advance the state of crash prediction modeling by redefining crash exposure to consider the practical (or causal) relationship between traffic volume and crashes. It also considers the traffic flow situation that is necessary for specific collision types to occur; therefore, each crash category could conceivably use a different measure of exposure measurement for its distinct occurrence mechanism. For example, for head-on and sideswipe opposite-direction crashes to occur, two vehicles traveling in opposite directions must meet. Similarly, rear-end and other types of same-direction crashes can happen only when vehicles closely follow one another. Combining this with the idea of linking prediction models to crash causalities through the established collision type categories, the proposed exposure measurement can potentially explain with more accuracy the variation observed in crash risk due to the traffic flow state or the chances of being exposed to the particular type of collision.

Generalized linear modeling assuming a negative binomial or scaled Poisson distribution is used to estimate prediction models. Evaluations of model goodness-of-fit and diagnostics results are discussed in comparing different model outcomes and assessing the effectiveness of adopting the newly defined collision type categories and crash exposure. The collision type categorization is conducted using K-means cluster algorithm. Ten collision types are grouped into four different clusters each corresponding to a particular set of main contributing factors.

Three studies are carried out to evaluate the new exposure definitions and the collision categorizing method based on contributing factors. In study 1, a new exposure, crash opportunities, is proposed for opposite-direction collisions including head-on and sideswipe. Crash opportunities defined as the number of times vehicles traveling in opposite directions meet is compared to a commonly used exposure measurement. However, the results show that models using crash opportunities do not perform better and an additional variable representing traffic flow has to be added to the models to permit the model to predict adequately. Study 2 uses K-means cluster analysis to categorize collisions according to the contributing factors. Different types of collisions having similar pattern in their collision causes are grouped together. The results find the final categories are (1) rear-end collisions, (2) other same-direction collisions including sideswipe and turning, (3) intersecting-direction collisions including turning opposite-direction, turning same-direction, intersecting path and angle, and (4) segment collisions including head-on, opposite-direction sideswipe and single vehicle. Study 3 then proposes vehicle time spent following as a new exposure definition for same-direction collisions and evaluates this new exposure for the collision categories from study 2. The new exposure is found to have a linear relationship with the total same-direction crashes, especially rear-end crashes. This linearity shows that vehicle time spent following can well represent traffic flow intensity and state for rear-end crashes to occur. Also, this finding further indicates the need of categorizing collisions by their contributing factors instead of types of crashes.

Table of Contents

List of Tables

List of Figures

Glossary of Acronyms and Symbols

FHWA	Federal Highway Administration
FARS	Fatality Accident Report System
V/C	Volume over Capacity Ratio
LOS	Level of Service
VMT	Vehicle Miles Traveled
VKT	Vehicles Kilometers Traveled
VKTP	Vehicle Kilometers Traveled by Period
AADT	Annual Average Daily Traffic
ADT	Average Daily Traffic
ConnDOT	Connecticut Department of Transportation
TCLP	Traffic Count Locator Program
TAVS	Traffic Accident Viewing System
β	Expected Value of Exposure
B	Exposure
λ	Value of Accident Count
π	Accident Rate, λ/β
ρ	A Constant, Expected Value of π

1 Introduction

1.1 Background

In the United States, rural roads represent the largest individual class of highways in the national system, over 70 percent according to the 2003 summary of the Federal Highway Administration. In that same year, nearly 40 percent of the national total vehicle-miles were traveled on rural roads. Two-lane highways account for nearly 80 percent of rural roads and are critical for providing both mobility and accessibility to rural residents, agriculture and industry. The Fatality Accident Reporting System reported that in 2004, 22,000 fatal collisions occurred on rural roads, nearly 60 percent of the total. Similarly, about 60 percent of the fatalities along with 30 percent of injury crashes were reported to have occurred on two-lane highways in 2004.

These disturbing numbers have spurred transportation safety professionals to seek more efficient strategies to reduce the impact of collisions on the general public. As a result, many traffic safety researchers have turned to crash prediction models and opened discussions on roadway geometric design issues and other factors possibly correlated with crash risk. Statistical methodologies have been adopted for such model analysis, and many studies have identified significant correlation between some specific factors and the crash risk based on model results, leading to practical modifications that can be possibly applied to reduce the crash risk.

1.2 Causality-Based Collision Type Categorization

Traditional motor vehicle crash prediction modeling focuses on investigating the correlation between crashes and related explanatory variables, such as lane width, shoulder width, horizontal or vertical curvature, median type, traffic control type, and land use. Statistical methodologies have been widely applied in such analysis to identify important road features that are apparently correlated with crash risk. Many studies also have found that dividing crash prediction models into sub-models by collision type categories not only yields better prediction but also reveals significantly different correlations between crash risk and traffic flow and many other explanatory factors in different sub-models (Shankar *et al.* 1995, Ivan *et al.* 2000, Qin *et al.* 2004, Lord *et al.* 2005). Some studies have firmly demonstrated the importance of estimating crash prediction models by collision type (Shankar *et al.* 1995, Kim *et al.* 2006, 2007). However, because there is not yet agreement on a universal collision type taxonomy, it would be helpful to develop such a taxonomy that is both logical and offers clear advantages over others to support consistency among crash prediction models estimated by different researchers.

Single-vehicle viz. multi-vehicle involvement is the most commonly used criterion for distinguishing among collision types, as it has been consistently found that the likelihood of being involved in a single vehicle crash decreases with traffic volume, while the opposite holds for a multi-vehicle crash (Qin *et al.* 2004, Lord *et al.* 2005). Moreover, distinctive correlation between crash risk and many traffic flow condition-related variables, such as traffic density, volume to capacity (V/C) ratio, and level of service (LOS), have also been discovered, revealing the difference between single and multi-vehicle crashes as well as many other explanatory variables, such as roadway design features, environment condition, and driver characteristics (Zhou and Sisiopiku 1997, Ivan *et al.* 1999, Ivan *et al.* 2000, Lord *et al.* 2005). However, knowing such correlations alone is not sufficient to reduce the occurrence of crashes, which requires first understanding the collision mechanism. Since traditional statistical crash prediction models cannot relate crash occurrences to the actual causes, relying only on the model results one cannot learn about what actually causes the crashes.

Instead, crash causality is better determined through accident reconstruction analysis, which focuses on identifying events apparently causing an accident to occur (Davis 2006). For this research, the crash contributing factors assigned by reporting police officers are considered to be a surrogate assessment of crash causality. Then by defining a set of collision types, each associated with the same contributing factors, relating the dependent variables (collisions by type) to road, traffic and other relevant characteristics directly identifies how these characteristics relate not only to crash incidence but, more usefully, to the actual factors that contribute to the collisions occurring. This has the potential to offer clearer information about how to reduce the incidence of crashes.

Consequently, an empirical study, as described in the second study described in this report, is carried out by observing the contributing factors for a sample of motor vehicle crashes in Connecticut and categorizing their collision types using K-means clustering algorithm into groups with similar contributing factors. Because crash prediction by collision type is often used as a surrogate or a step towards prediction of crash severity, this report also proposes, as explained in the second paper, to examine whether or not the severity distributions of the collision types in each group are similar to one another, and more different from those in other groups. The outcome of these analyses is reported along with interpretations and suggestions as to how crash prediction modeling might be improved as a result for use in identifying crash causal factors.

1.3 Crash Opportunities as Exposure

1.3.1 Exposure and Crash Prediction Models

In road crash prediction models, crash exposure plays a critical role by normalizing the variation in expected crash risk according to the usage level of the roads under investigation. For more than 20 years, researchers have recognized that traffic flow and the extent of road usage are inextricably related to accident risk (Bruhning and Volker 1982), and this motivated the inclusion of accident exposure when estimating the accident risk of a roadway or network. Until recently, traffic safety researchers considered a linear relationship between crashes and exposure. According to Bruhning and Volker, if one assumes β is the expected value of exposure B and λ is the value of the accident count, the equation $E(\lambda|\beta) = \rho\beta$ accounts for a linear relationship between the number of accidents and exposure given a properly chosen value for β. By moving β to the other side of the equation, we then have $E(\lambda/\beta|\beta)=\rho$, where λ/β is the traffic crash rate π, a random variable, and ρ is its expected value. If one assumes that the mean-segment based crash rate, π, is independent of the crash exposure, we then have $E(\lambda/\beta)=E(\pi)=\rho$. This definition of exposure technically requires that the relationship between crashes and exposure should be linear, as the ratio of these two quantities is defined to be the crash rate per unit of exposure. As crash risk is usually correlated with many other covariates such as roadway geometrics, ρ can also be estimated by properly formulating the relevant covariates. If an appropriate measurement is chosen to be the exposure and an appropriate format is chosen for estimating ρ with the parameterized covariates, the prediction models could then reveal the covariates that are correlated with higher crash risk and give better crash count prediction that is based on all the estimated parameters and the exposure measurement.

However, this issue of linearity in crash rate λ/β gets complicated, as Bruhning and Volker showed that exposure can be defined in a great number of ways, such as vehicle population, length of road network, specific intersection or pedestrian crossings (when pedestrian crashes are considered in the study), and the number of vehicle miles or route length. Most of these exposure measurements are actually linearly related to traffic volume, which is not necessarily linearly related to crashes, as will be demonstrated presently. Recent studies (Persaud and Dzbik 1993, Hauer 1995, Maher and Summersgill 1996, Mountain *et al.* 1998, Ivan *et al.* 2000, Lord *et al.* 2005) have revealed a non-linear relationship between crashes and the traditionally defined exposure (traffic volume). Traditionally, exposure is defined to be linearly dependent on some representation of traffic volume, such as vehicle miles traveled (VMT) for a roadway segment or total entering traffic for an intersection. However, crashes usually do not simply increase (or decrease) linearly with the change in traffic volume. Hauer (1995) argued that the linearity assumption must pass two fundamental tests: First, the linearity has to be plausible; and second, it has to have empirical support. For many candidate measurements of exposure, such as vehicle population and length of roadway, it is plausible that there is a monotonous relationship between these measurements and crash counts at particular locations; however, whether this monotonous relationship is linear is a question left to be discovered through empirical results.

Again, Hauer (1995) pointed out that the chance of having an accident for a vehicle traveling a roadway segment is affected by the presence and density of other vehicles. Therefore, crash count does not necessarily increase (or decrease) linearly with changes in traffic flow, or exposure, when exposure is a linear function of traffic flow. As research studies in predicting crash models often do, the exposure measurement is modeled as a variable with a coefficient (or an exponent) to be estimated along with all other parameters in the model. Different research studies have used various definitions of exposure in crash prediction models, reporting different conclusions. Fridstrom *et al.* (1995) measured the contribution of randomness, exposure, weather, and daylight to road accident count variation. Using gasoline sales as

proxy for exposure or traffic volume, they found that for injury accidents the coefficient for gasoline sales came out almost equal to 1.0, which implied a linear relationship between the accident count and exposure. They also found a significantly smaller than 1.0 coefficient in fatal accident models; they suggested that as society becomes increasingly more used to motorized transport, knowledge on how to reduce the accident severity accumulates among individuals and institutions.

1.3.2 Crash Opportunities

In addition to using traffic volume in explaining variation in crash risk, other variables that reflect traffic flow characteristics can also be used as independent variables in crash prediction models, *i.e.*, AADT, V/C ratio, and traffic density. Studies have found that when considering these variables that represent the traffic flow state, crash models can offer richer explanations of crash risk variation (Zhou and Sisiopiku, 1997, Lord *et al.* 2005). Some of the variables need more specific data than just traffic volume, such as hourly traffic flow, directional proportions in traffic flow, and even vehicle travel speed. However, these traffic-related data are more difficult to obtain than the total volume and it is difficult to estimate them accurately enough. This is especially true for V/C ratio, since capacity is strongly impacted by adverse weather (Agarwal *et al.* 2005, Zhang *et al.* 2005), so that when V/C ratio is computed using normal traffic volume data and capacity estimates, it would not describe the actual traffic situation in each specific time period. In order to also take into consideration the actual traffic state while defining crash exposure, this report proposes crash opportunities for estimating prediction models to be specified differently for alternative collision scenarios. This new concept of exposure, or opportunities, accounts for traffic volume variation and is a function of traffic volume. The relationship between crashes and opportunities is examined based on empirical results described in the first and third studies included in this report.

As the previous discussion shows, many possible factors contribute to road crashes. As a result, it is difficult to have a model thoroughly explain the effect of each explanatory factor due to possible confounding among the factors. However, if one can define an appropriate exposure measurement that has a form plausibly and empirically linear to crash count and includes factors describing the traffic context, one may manage to reduce the cross-correlation between these factors and other factors that are also important to be considered in the model in explaining the variation in crash risk. For example, in vehicular stream models (Papacostas and Prevedouros 1993) speed is a strong indicator for the traffic flow condition; however, it is also found to be highly correlated with many other roadway characteristic variables (Jonsson 2005). Therefore, if speed can be added to the computation of the opportunity measure that is linear to crash count and replaces traditional exposure, the confounding between speed and other variables, such as roadway width, curvature and land use, can possibly be avoided while still considering the contribution of all of these factors along with speed in the model.

It is also hoped that dividing crash count by crash opportunities will produce a crash rate that does not vary with volume. In defining crash exposure, Hauer (1982) argued that a unit of exposure corresponds to a trial and the result of such a trial is the occurrence or non-occurrence of a crash. He also defined system risk to be the probability of crash occurrence in a trial. Based on Hauer's argument, trials should be defined differently for each distinct crash type according to the traffic conditions under which each occurs. For example, opposite direction crashes could not occur on a two-lane highway segment at a time when there are only vehicles traveling in one direction. Similarly, same direction crashes could not occur unless vehicles follow each other with unsafe headways. Following this reasoning, for opposite direction crashes, the number of trials should be a function involving traffic volumes separated by direction that estimates the number of possibilities for such a collision to occur. Similarly, for same direction crashes, the number of trials should consider not only the directional traffic volume in both directions individually, but also the interaction between the directional traffic with the need for vehicles to pass one another.

This argument raised by Hauer is similar to the concept of opportunities proposed in this report. Here the definition of exposure is extended beyond the notion of a "trial" to also consider the opportunity for crashes to occur as well as the effect of traffic flow conditions on the risk of crashes occurring. By accounting for the mechanistic interactions of different traffic flow streams relevant to each specific type of collision, the concept of crash opportunities is a self-explaining functional form and has an explicit, understandable meaning. The goal of this report is to define exposure functions based on opportunities for several crash types, investigate the validity of a linear relationship between crashes and these definitions of crash opportunities and evaluate possible improvement in crash prediction that is possible by using these definitions to replace traditional representations of traffic volume in crash prediction models.

1.3.3 Variable Selection

In addition to the improvement that an appropriate exposure measurement can bring to the crash prediction models, proper selection of covariates also helps the prediction models to reveal the situations which correlate to higher accident risk level. It is probable that different exposure measurements might result in different sets of parameter estimates for the same covariates. Therefore, to interpret the meaning of these parameter estimates, with respect to how these covariates relate to driver behavior, also supports more accurate evaluation of the performance of exposure measurements and therefore reveal the actual more risky situations. Suppose one is able to control for a wide and properly selected range of the crash-related variables regarding traffic flow state, roadway, weather, and driver behavior, then it is only the random effects that cause the variation in crash risk and can be well explained by a Poisson or negative binomial distribution. Such variables include the commonly available roadway geometric variables, lane width, shoulder width, and speed-related design values, and other less commonly available details such as the roadway curvature, land use pattern and observed driver behavior.

The time period in which the collisions occurred can be included in the models as a potential indicator accounting for driver behaviors. Driver behavior is expected to vary by many different factors including age distribution of the driver population, road type, land use type, time of day, etc. In addition to the collision contributing factor-based crash types, in this report, time period indicator will be used to substitute part of the effects that driver behavior has on crash risk. Also, land use intensity has been found to be associated with the type, the purpose and the amount of the transportation-related activities conducted (Ossenbruggen *et al.* 2000). Consequently, including some measure of land development intensity may help to explain the variation in crash risk while included in the prediction models. This report assesses this quantity by observing the number of access points along each segment, *i.e.* the number of driveways and minor roads intersecting it. The number of access points provides a measure of the traffic volume that turns on and out of the selected segments, therefore explaining part of crash risk for especially same-direction collisions when involving turning, merging and decelerating vehicles.

Paved roadway width is another variable that is linked to highway safety. Studies have found that narrow paved roadway width is correlated with a higher likelihood of accident involvement (Abdel-Aty 2000). However, this does not necessarily suggest that widening roadways is an efficient way at reducing accident risk, as accident impacts due to changing a road characteristic can only be assessed using before-after studies. This report includes pavement width as a variable in sub-models divided by collision type categories and further discusses the width correlation with crash risk. Speed limit is also included frequently in crash prediction models (Finch 1994, Elvik 2004, Jonsson 2005). Although extra attention needs to be paid to account for potential confounding between the effect of speed limit and many other design variables, its correlation with roadway safety level is carefully evaluated and interpreted through the model results.

1.4 Scope and Organization of Report

The process of conducting modeling analysis in achieving the goals of this report is shown in Figure 1.1. This procedure focuses on discussion about two major topics, crash exposure and collision type categories, four crash prediction contexts, and three modeling comparison steps.

The two subjects are collision type categorization and exposure definition. Shown as coordinates in Figure 1.1, two different collision taxonomies are discussed in this report: (1) the traditionally defined or commonly employed crash categories, and (2) the new categories proposed in this report; and two types of exposure measurements, respectively, (1) the traditionally used vehicle miles or kilometers traveled (VMT or VKT) and (2) crash opportunities as defined in this report. Partitioning the two-dimensional space created by these alternative exposure definitions and collision type categorizations, four different combinations of exposure and collision type categorization emerge, representing four contexts in which to estimate crash prediction models. Context A, corresponding to VMT or VKT as exposure and the traditional collision type categories, is the one most frequently used in crash prediction models. This report will show the changes and the potential improvements that adopting crash opportunities and the new category definitions can make to crash model prediction results by moving to context B, which corresponds to using the newly defined crash exposure and collision type categories in the third paper.

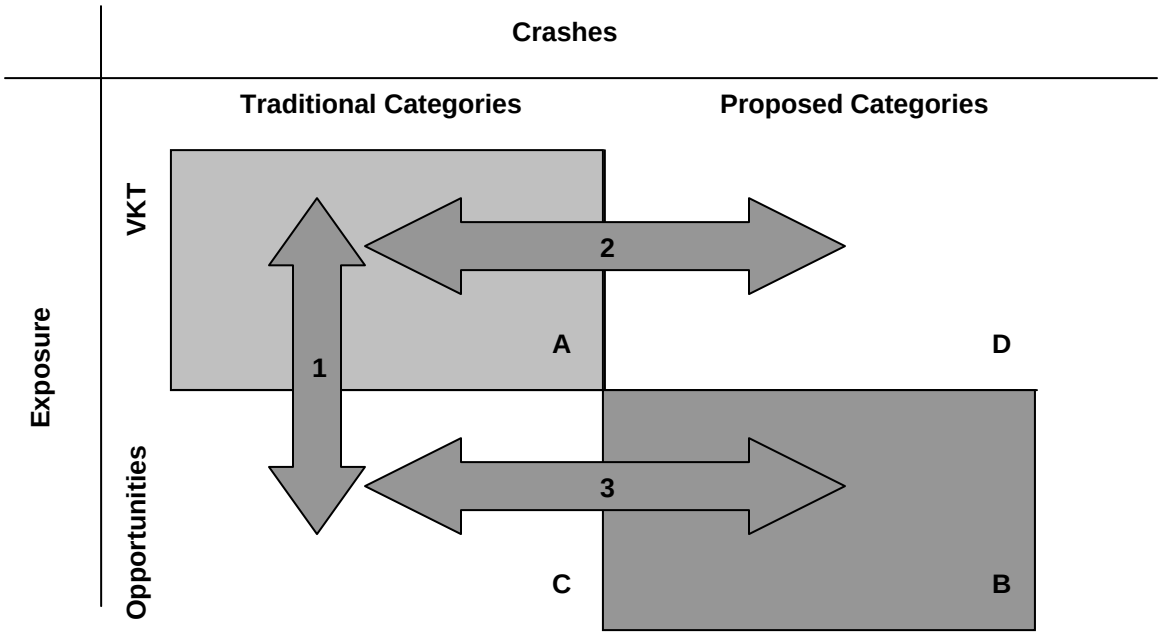


Figure 1.1 Modeling Framework and Procedure

To better assess the potential benefits of using different exposure measurements and collision type categorization on model performance, the first paper and third studies compare models that use VKT or crash opportunities with those using crash categories defined by vehicle direction of travel or categories based on contributing factors, respectively. Among the three steps marked in Figure 1.1, step 1 is carried out to show the improvement in crash prediction model results achieved by replacing VKT with crash opportunities as exposure, that is, moving from context A to C. Step 2 analysis is used to demonstrate, first, how the contributing factors-based crash categorization can help connect the interpretation of model results with collision causality and potential remedies using traditional exposure measures; and second, whether the models do perform better while predicting crashes which were caused by similar contributing factors. In step 3, this report aims to show whether the step 2 analysis would yield enhanced results when at the same time using opportunities as exposure.

The above procedure is described in the three papers included in this report. The first paper focuses on the step 1 comparison particularly for opposite-direction crashes. The second paper discusses in detail the causalities-based collision categorization. Then in the third paper, the step 2 and step 3 analysis is carried out for predicting same-direction crashes.

2 Opposite-direction Crash Prediction Models Using Opportunities as Exposure

2.1 Introduction

There has been considerable research conducted in recent years into including traffic volumes in crash estimation models for assessing risk on roadway segments. In crash estimation models, traffic volume serves two major roles: first, to represent exposure to crashes; second, to help explain variation in expected crash risk. Vehicle-miles-traveled (VMT), or vehicle-kilometers-traveled (VKT), is the most commonly used measure of exposure in crash prediction models. However, as demonstrated by Ivan (2004), to simply divide the number of crashes observed during a time period by the VMT observed during that same time period cannot capture all of the variation in the occurrence of roadway crashes, because the variation in crash risk is related to the type of crash and the time of day as well as the traffic volume. Therefore, to more accurately estimate crash exposure, different crash types should be treated separately, and the effect of exposure should be considered by time of day.

In defining crash exposure, Hauer (1982) argued that a unit of exposure corresponds to a trial and the result of such a trial is the occurrence or non-occurrence of a crash. He also defined system risk to be the probability of crash occurrence in a trial. Based on Hauer's argument, trials should be defined in a different format for each separate crash type according to the traffic conditions under which crashes occur. For example, opposite direction crashes could not occur on a two-lane highway segment at a time when there are only vehicles traveling in one direction. Following this reasoning, for opposite direction crashes, the number of trials should be a function involving traffic volumes in both directions.

Consequently, in this study, we evaluate a new approach for representing the effect of traffic volume in crash prediction. This approach extends the definition of exposure beyond the notion of a "trial" to also consider the opportunity for crashes to occur (Ivan 2004, Hauer 1982) as well as the effect of traffic conditions on the possibility of crashes occurring. We define this new measurement as crash "opportunities" and define it differently for each type of crash. In this preliminary study, we specify the crash opportunity functions for opposite-direction sideswipe, head-on and single vehicle crashes on two-lane highways, computed by time of day.

2.2 Background

Hakkert and Mahalel (1982) are among the earliest researchers to base prediction models on the assumption of a Poisson distribution. In their study estimating intersection crashes as a function of the traffic flows on the approaches, they defined the vehicle exposure as the number of occasions for crashes, which was computed by summing the products of traffic flow at the 24 points of an intersection where vehicle paths cross or merge. Based on two-year data from 250 urban and interurban intersections in Israel, they found that the expected number of crashes had a good correlation of 0.8 with vehicle exposure. However, they didn't separate crashes by collision type, so accordingly they didn't have a specific vehicle exposure for different collision types, which might have limited the ability of their models to explain the crash risk variation even better than that.

More recently, Lord *et al.* (2005) reported that for both rural and urban freeway segments, traffic flow itself might not adequately characterize the crash process on freeway segments, and other variables such as volume to capacity ratio might offer a better fit. Furthermore, they concluded that separate models for single and multi-vehicle crashes should be estimated instead of a single model for all crashes. Though Lord *et al.* grouped their data by hour of the day, they did not report how the hour of day affects the crash risk variation.

Fridstrom *et al.* (1995) also reported findings about the relationship between crashes and traffic volumes in the four Nordic countries. They used either traffic volume estimated from traffic counts or gasoline sales (which they considered to be a proxy for traffic volume – the dataset also included monthly crashes from counties in each individual country) as the crash exposure in their models. After including the other variables regarding weather, daylight, legislation and speed limit, they found an almost proportional relationship between injury crashes and traffic volume, but a less proportional relationship in the case of fatal crashes and number of fatalities.

2.3 Crash Opportunities

In the above-mentioned studies, none of the researchers focused on relating a specific crash type to both the traffic volume-based exposure and traffic condition when crashes occur. In this study, we combine the two by introducing the concept of crash opportunities.

Since head-on and opposite-direction sideswipe crashes both involve vehicles approaching one another from opposite directions, the opportunity function was defined as the number of meetings for vehicles approaching one another from opposite directions on a selected segment during a particular time of day, over a defined period of time (several years). For example, suppose we know hourly traffic volumes in both directions on a two-lane highway segment; then, the crash opportunities for time of day t during a time period p would be

$$O_t = \frac{2L}{u}\sum_{h\in t} V_{h1}V_{h2} \tag{2.1}$$

where O_t is the number of crash opportunities (number of vehicle meetings) during time of day t, L is the segment length, u is the space mean speed of traffic in both directions, and V_{h1}, V_{h2} are hourly traffic flows at each hour h during time of day t for direction 1 and direction 2, respectively, during the time period p. Equation (2.1) is derived assuming a constant traffic volume over each 1-hour period, and that all vehicles travel at the space mean speed.

Equation (2.1) is derived in Figure 2.1. During hour h, one vehicle in direction 1 will on average meet $(2L/u)V_{h2}$ vehicles, because V_{h2} is the rate of arrivals and $2L/u$ is the time required for all vehicles met by a vehicle traveling in direction 1 to clear the section. As a result, in this hour, the number of opportunities for two vehicles to possibly collide or sideswipe each other on the same segment O_h is

$$O_h = \frac{2L}{u}V_{h1}V_{h2}, \tag{2.2}$$

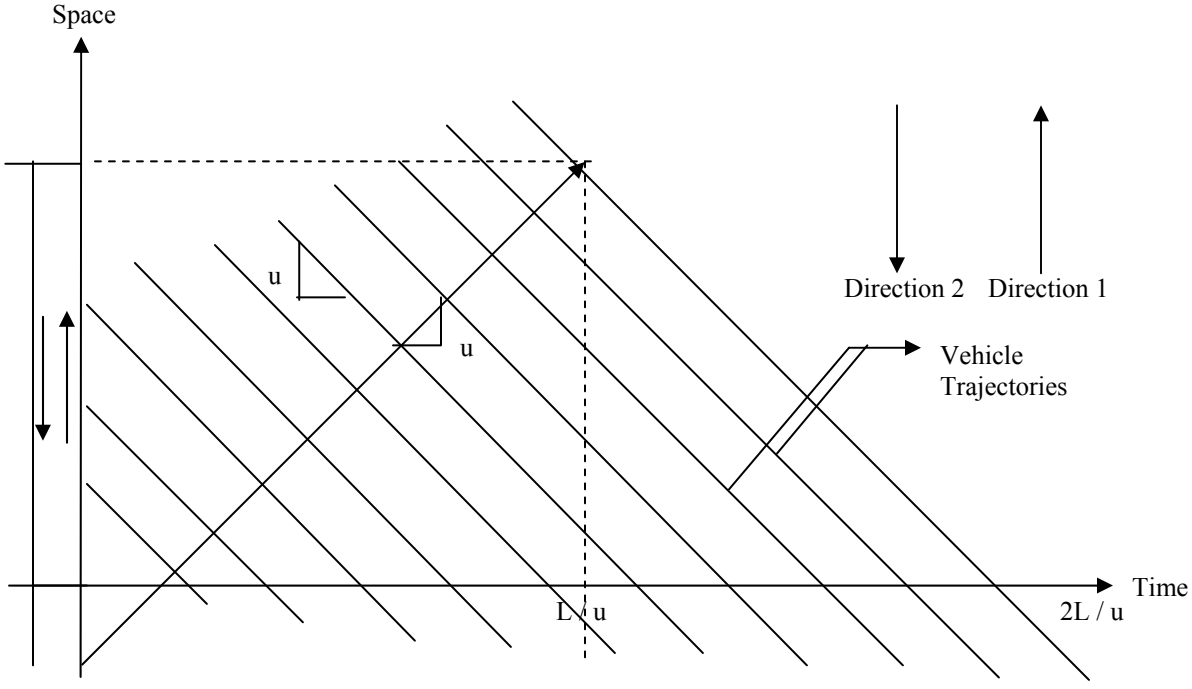


Figure 2.1 Space-Time Diagram

which multiplies the number of meetings per vehicle by the number of vehicles traveling in direction 1 in one hour. If we sum this opportunity by hour for a particular time of day t during the time period p, we obtain equation (2.1). Ideally, if we know continuous hourly traffic counts for a location, we then will be able to compute the head-on and opposite-direction sideswipe crash opportunities for that location during any time period, since these two types of opposite direction crashes can only occur when there are vehicles passing by each other at the road segment.

Defining the opportunity function for single vehicle crashes is more complex. At the most basic level, any time a vehicle traverses a road segment there is an opportunity for a single vehicle crash. However, if a vehicle loses control on a two-lane road at the exact moment when a vehicle is approaching from the opposite direction, there is a greater possibility of a head-on rather than a single vehicle crash occurring. Consequently, we considered both the classical vehicle kilometers traveled and the opportunity function defined above for single vehicle crashes, too.

2.4 Study Design

This study requires a unique data set. The most restrictive requirement was for directional traffic volumes by hour of day, which were necessary for computing the opportunity function. Other data required included characteristics of the roadway and crash records by collision type and time of day over a selected study period. Based on these requirements, we selected 95 segments randomly from 24 two-lane rural highways in Connecticut. For each of these sites, a 24-hour traffic count with directional split on a weekday during the study period was collected using the "Traffic Count Locator" (TMSADT) program provided by Connecticut Department of Transportation (ConnDOT). Crash data for the selected roadway segments were then obtained for the entire period through Traffic Accidents Viewing System (TAVS) program, also from ConnDOT and contained information about crash type and the time of day.

We selected segments with a uniform length of 1 kilometer and similar traffic control conditions, *i.e.*, no signalized or stop sign controlled (on the major road approaches) intersections and no thickly settled areas. Directional traffic counts also had to be available at each location. The study period covered 7 years from 1995 to 2001, inclusively. The data were then grouped two ways: by the daily time period (defined as "early morning" from 2 am to 6 am; "morning peak" from 6 am to 10am; "midday" from 10 am to 2 pm; "afternoon peak" from 2 pm to 6 pm; "evening" from 6 pm to 10 pm; and "overnight" from 10 pm to 2 am), and for the whole day. In the end, under the two different groupings, respectively, there were 570 (95 segments times 6 times of day) and 95 observations.

Because the observed counts were observed during a single 24-hour period, we then needed to compute a crash opportunity count that would be representative of the entire study period, not just for the day the count was taken for each segment. ConnDOT provided factors for expanding 24-hour counts (ADT data) to Annual Average Daily Traffic (AADT) volumes; these factors made it possible to find a representative opportunity based on just 24-hour traffic counts by correcting the monthly effects and day-of-week effects for the day the count was observed. After adjusting the monthly effects and day-of-week effects on the 24-hour directional traffic counts by applying the adjustment factors from ConnDOT, we obtained the hourly traffic volumes on a yearly average basis. Since no speed data were available for any of the segments in our collection, we used the posted speed limit to approximate the average speed for each segment. Then, by using equation (2.1), the crash opportunities during the 7-year study period were computed for the above mentioned two groupings.

Exploratory analyses were conducted on the data for different crash types. Table 2.1 presents the distribution of crashes by collision type and by time and hour of day. Figures 2.2 and 2.3 illustrate the data in Table 2.1. The trends of crashes versus opportunities vary greatly by crash type and time of day. Figure 2.4 and 2.5 show the trends for opposite direction sideswipe, head-on and single vehicle crashes by crash opportunities at morning peak time period and overnight time period. For instance, it can be seen that during the morning peak time period, as opportunities increase, the number of single vehicle crashes decreases while that of sideswipe and head-on crashes increases. However, during the overnight time period, there is a much flatter trend observed for both opposite-direction sideswipe and head-on crashes. Single vehicle crashes seem to have an upside down "U" shaped relationship with opportunities, but the curve shape might be biased by having too many cases with low opportunities. It deserves attention that during the two time periods, the ranges of opportunities are quite different, and the maximum value of opportunities during the overnight period is only one tenth of that during morning peak period. However,

since in the models, time of day is used as a dummy (binary) variable, each time period is treated individually to possibly separate the effects of traffic volume by times of day.

Table 2.1 Number of Crashes by Time, Hour of Day and Collision Type

Time of day	Sideswipe Opposite Direction	Head-on	Single Vehicle	Beginning hour	Sideswipe Opposite Direction	Head-on	Single Vehicle
Early Morning	6	2	91	2AM	3	1	32
				3AM	0	0	21
				4AM	2	0	18
				5AM	1	1	20
Morning Peak	22	5	132	6AM	7	0	28
				7AM	4	2	45
				8AM	7	2	29
				9AM	4	1	30
Midday	22	9	115	10AM	4	0	23
				11AM	6	0	33
				12PM	8	3	33
				1PM	4	6	26
Afternoon Peak	32	13	131	2PM	4	3	40
				3PM	14	6	38
				4PM	9	3	30
				5PM	5	1	23
Evening	26	8	123	6PM	10	3	37
				7PM	6	2	25
				8PM	5	2	23
				9PM	5	1	38
Overnight	15	6	134	10PM	1	0	28
				11PM	7	1	38
				12AM	7	4	34
				1AM	0	1	34
All Day	123	43	726				

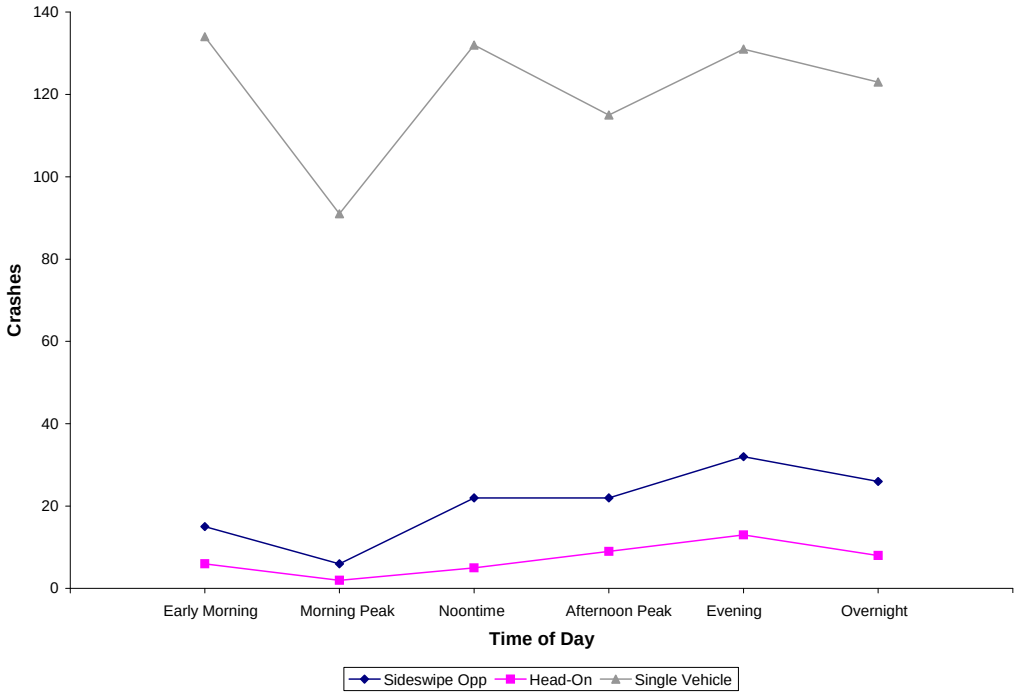


Figure 2.2 Crashes by Time of Day

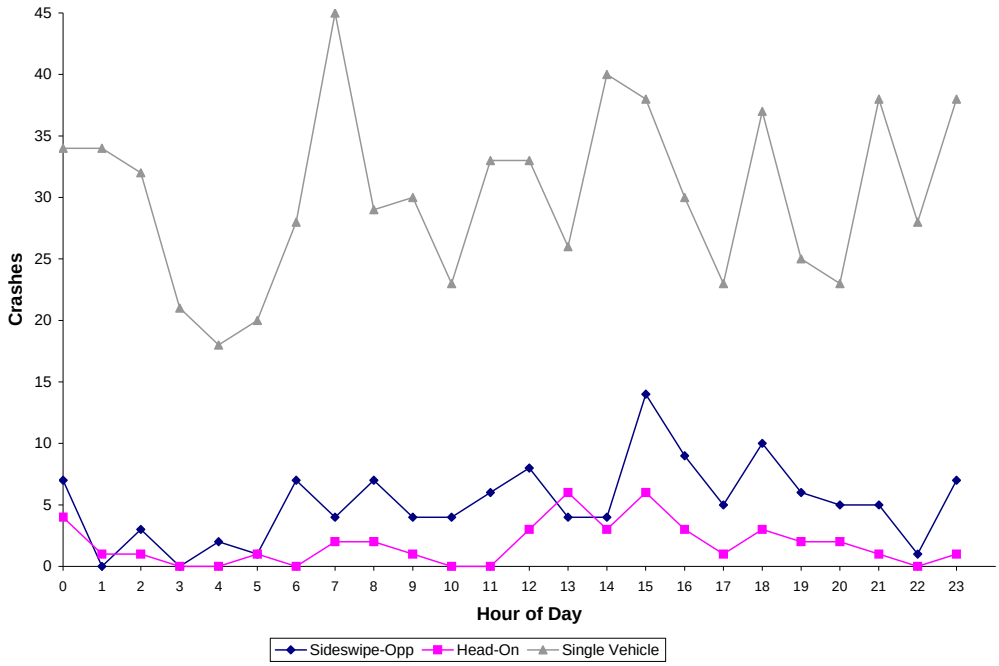


Figure 2.3 Crashes by Hour of Day

2.5 Statistical Modeling

Two sets of predictive models were developed for the 95 two-lane highway segments. One set of models used vehicle kilometers traveled (VKT) as crash exposure, and the other set used the previously defined opportunities as crash exposure. The probabilistic structure used for developing the models was the following: the number of crashes Y_{it} during the 7 year period at the ith segment, tth time of day, follows a Poisson distribution and is independent over all segments and all times of day when conditioned on its mean μ_{it}, *i.e.*,

$$Y_{it} \mid \mu_{it} \approx Poi(\mu_{it}),\ i = 1, 2, \ldots, 95;\ t = 1, 2, 3, 4, 5 \text{ and } 6. \tag{2.3}$$

Though one might question the assumption of constant crash probability through the study period at each segment, the invariance-under-summation property of the Poisson distribution provides a solid basis for the validity of the assumption of Poisson distributed number of crashes in our modeling. According to Fridstrom et al (1995), for a very short time interval we can assume that the crash probability is constant and that events occurring during the disjoint time intervals are probabilistically independent. Therefore, the number of crashes occurring at the same location during a given time period will be Poisson distributed with some unknown parameter.

In equation (2.3), the mean μ_{it} is specified as

$$\mu_{it} = f(V)e^{X'\beta} \tag{2.4}$$

where f is a function of traffic flow V, X' stands for a vector of roadway characteristic variables, such as lane width, shoulder width, etc. including intercept, and β represents the corresponding parameter vector. Following the accepted practice, $e^{X'\beta}$ follows a Gamma distribution with shape parameter $1/\phi$, so the number of crashes at each segment and at different times of day Y_{it} follows a negative binomial distribution with mean μ_{it} and variance $\mu_{it}(1+\mu_{it}/\phi)$, where ϕ is the overdispersion parameter of the negative binomial distribution. Similar to a previously study done by the first two authors (Zhang and Ivan 2005), When $1/\phi$ is not significantly different from 0, the NB distribution is approximately equivalent to a Poisson distribution.

Although it is not obvious, the formula for crash opportunities is actually a function of AADT. Supposing $\hat{O}_t$ is the opportunities estimated for time of day t. Then defining h_{1t} and h_{2t} respectively to be the hourly proportions of AADT in either direction during the hours at that time of day, we then derive the following equation:

$$\hat{O}_t = \frac{2L}{u}(AADT^2)\sum_t h_{1t}h_{2t}\ . \tag{2.5}$$

Equation (2.5) shows that the formula for computing crash opportunities at any time of day is a function of AADT squared and the hourly traffic proportions by direction. Because the directional distribution varies greatly throughout a whole day at a single location, and this variation is not identical at all locations, to properly represent the crash opportunities, it is necessary to model crash rate by time period. Therefore, using equation (2.5) as crash exposure in the prediction models, we can see how crash rate is actually related with traffic volumes. Furthermore, by estimating the models both with and without AADT as an additional explanatory variable, we can investigate whether or not it is necessary to include an extra traffic volume term in the model to represent the effect of the traffic flow state beyond the contribution of opportunities.

In order to focus the study on investigating the importance of the role that crash opportunity plays in predicting crash rate, in this preliminary study, we only included roadway width as an explanatory variable in the models (in addition to the volume related variables already discussed). We also accounted for a potential time of day effect (e.g., late night driver inattentiveness) by including 5 dummy variables: period 1 through period 5 for early morning period to evening period respectively, in the models. The overnight period was set to have a coefficient of 0, and the effect of any other different time of day on crash rate can be obtained by comparing its coefficient with that of the overnight period. So if the coefficient of any dummy variable is less than 0, the predicted number of crashes is less during that time period of day than during the overnight period, and vice versa.

One-side roadway width effect was modeled by aggregating the data into 3 levels, to avoid constraining the model to return a linear relationship with crash risk. According to the observed distribution of the values for this variable, we set level 1 to correspond to a one-side roadway width of 3 m (9.8 ft), 3.5

m (11.5 ft), or 4 m (13.1 ft); level 2 consists of a roadway width of 4.5 m (14.8 ft); and level 3 consists of roadway widths from 5 m (16.4 ft) up to 6 m (19.7 ft). Using the dummy variables in the same way as for time of day effect, in our models, roadway width category 3 is set to have a coefficient of 0. Therefore, the effect of either width category 1 or 2 on crash rate can be obtained by comparing their parameter coefficients with that of the roadway width category 3. If the coefficient on the dummy variable is less than 0, the predicted number of crashes is less given that width category than width category 3, and vice versa. The cross-tab results of each roadway width category with the number of segments and the number of crashes by type are shown in Table 2.2.

Table 2.2 Number of Segments and Crashes by Collision Type and One-side Roadway Width

One-side road width m (ft)	Group	Number of segments	Sideswipe	Head-on	Multi-vehicle	Single Vehicle
3 (9.8)	1	1	0	0	0	1
3.5 (11.5)	1	9	3	1	4	55
4 (13.1)	1	16	15	5	20	115
4.5 (14.8)	2	41	70	25	95	323
5 (16.4)	3	21	27	8	35	166
5.5 (18.0)	3	5	5	0	5	35
6 (19.7)	3	2	3	4	7	31

Since AADT is often correlated with roadway geometric variables (because heavily traveled roads are generally designed to a higher standard), when including AADT in the models, we also add terms to consider interactions between AADT and roadway width to reduce the confounding effect between them. As a result, in the models with AADT included, AADT 1, 2 and 3 respectively show the effect of AADT on crash rate by the three roadway width categories. An additional model is also obtained for each crash type not including this interaction between AADT and roadway width category. Statistical software SAS (1999) was used in this study for estimating crash models. Opportunities and VKT were then used as exposure in the two sets of crash prediction models; the goodness-of-fit results for the models are summarized in Tables 2.3 and 2.4, while the resulting parameter values and statistics are summarized in Tables 2.5 and 2.6. Tables 2.3 and 2.4 provide the resulting log-likelihood for each model along with the log-likelihood for the corresponding intercept-only model and the Pearson X^2, defined according to McCullagh and Nelder (1989), where

$$r_P = \frac{y - \mu}{\sqrt{V(\mu)}} \tag{2.6}$$

And the Pearson X^2 statistic = $\sum r_P^2$.

Though Pearson X^2 does not tell much in terms of model goodness-of-fit, it can be used as a measure of residual variation, and further it is useful for outlier detection and assessing the influence of single observations on the fitted model (SAS 2004). In SAS, the deviance, defined to be twice the difference between the maximum attainable log likelihood and the log likelihood of the model under consideration, is computed by $D^*(y,\mu) = 2(l(y,y) - l(y,\mu))$, where $l(y,y)$ is the maximum attainable log likelihood of the model that has a parameter for every observation, and $l(y,\mu)$ is the log likelihood of the estimated model (2004). Deviance difference is then defined to be the difference of the deviance values of two nested models. By comparing the deviance difference with a chi-square statistic with degree of freedom p_1-p_2, which is the number of additional variable(s) in the model including more factors, the significance level of the added variable(s) can then be determined based on the test result: if the deviance difference is significantly larger than the chi-square statistic (this research uses a 95 percent confidence level), *i.e.* the p-value of the test is less than 0.05, the added variable(s) then have significant effects on crashes; and vice versa. Respectively for models using opportunities and models using VKT, the three types of models, model without AADT, model with AADT, and model with AADT as well as the interaction of AADT and roadway width factors, are considered as nested models, and therefore the

significance level of the effects of AADT and the interaction terms of AADT and roadway width factors can be assessed through the deviance difference chi-square tests. The test results are discussed in the following section.

Table 2.3 Goodness-of-Fit Results for Models with Opportunities as Crash Exposure

Crash Type	Model Type	Log-likelihood	Pearson X^2	Log-likelihood of Intercept Model
Opposite-direction Sideswipe	Width	-307.67	2415.40	-347.55
	Width & AADT (interaction)	-283.44	923.97	
	Width & AADT (no interaction)	-282.71	852.02	
Head-on	Width	-144.71	713.71	-160.71
	Width & AADT (interaction)	-136.52	453.62	
	Width & AADT (no interaction)	-136.20	469.54	
Multi-vehicle	Width	-357.59	1856.84	-408.38
	Width & AADT (interaction)	-328.40	795.45	
	Width & AADT (no interaction)	-327.95	750.36	
Single-vehicle	Width	-690.09	1405.61	-846.36
	Width & AADT (interaction)	-522.23	676.87	
	Width & AADT (no interaction)	-516.05	658.83	

Table 2.4 Goodness-of-Fit Results for Models with VKT as Crash Exposure

Crash Type	Model Type	Log-likelihood	Pearson X^2	Log-likelihood of Intercept Model
Opposite-direction Sideswipe	Width	-281.39	810.44	-304.22
	Width & AADT (interaction)	-279.08	675.62	
	Width & AADT (no interaction)	-278.90	663.14	
Head-on	Width	-138.57	466.66	-144.47
	Width & AADT (interaction)	-138.32	446.23	
	Width & AADT (no interaction)	-137.54	460.08	
Multi-vehicle	Width	-327.72	687.80	-354.35
	Width & AADT (interaction)	-325.55	605.13	
	Width & AADT (no interaction)	-325.26	600.98	
Single-vehicle	Width	-559.22	819.14	-656.05
	Width & AADT (interaction)	-493.57	621.57	
	Width & AADT (no interaction)	-489.34	598.11	

2.6 Discussion

When looking at log-likelihood and Pearson X^2 statistics from Tables 2.3 and 2.4, we can see that the models using VKT as crash exposure perform better than those using opportunities. Also, adding AADT as a covariate improves the log-likelihood of the models using opportunities as exposure much more than for the models using VKT as exposure. There is no significant improvement for the models stratifying the AADT exponent by roadway width. Then by comparing the Pearson X^2 values from different models, we find that including AADT in models results in much lower Pearson X^2 values.

Tables 2.5 and 2.6 show the estimated parameters that define the relationship between crashes and traffic volumes considering the time of day effect, one-side roadway width effect, and the interaction effect of roadway width and traffic volume for each model. When number of opportunities is used as crash exposure, roadway width category 2 has no significantly different effect on predicted crashes from category 3 for any crash type (except for single vehicle crashes under some conditions), which means that there is no significant difference regarding safety between the selected segments belonging to roadway width category 2 and 3. Furthermore, for head-on crashes, even the coefficients for width category 1 and category 3 are not significantly different for either exposure measure. This finding is consistent with conclusions drawn in a previous study by the first two authors (Zhang and Ivan 2005), where roadway width variables were insignificant for predicting head-on crash incidence.

In the crash prediction models for opposite-direction sideswipe crashes, a negative coefficient is obtained for each time of day dummy variable. Furthermore, morning peak and afternoon peak time periods seem to have the lowest parameter coefficients, which implies that during these two time periods the opposite-direction sideswipe crash rate is the lowest compared with other times of day. The effect of the early morning period is not significantly different from that of the overnight period, and these two have the highest coefficients of the dummy variables. Figures 2.4 and 2.5 illustrate the trend of different types of crashes against opportunities during morning peak time period and overnight time period.

In the predicted head-on crash models, similar negative parameter coefficients were observed for the time period dummy variables. As for opposite-direction sideswipe crashes, the early morning time period doesn't have a significantly different effect on predicted crash rate from the overnight time period, and during the evening time period the predicted head-on crash rate again is higher than that during the morning peak, midday, and afternoon peak time periods. During the morning peak period, the head-on crash rate is higher than during the afternoon peak period, unlike the opposite-direction sideswipe crash rate for which the two time periods have nearly identical effects.

For predicting single vehicle crashes, again the early morning period is not significantly different from the overnight period, and during the evening period the predicted crash rate is higher than during the overnight and early morning time periods. The lowest predicted single vehicle crashes occur during the afternoon peak period, and the morning peak period has the second lowest predicted crash rate. Similar to opposite-direction sideswipe and head-on crashes, the midday period has less single vehicle crashes than overnight and early morning time periods but higher than the morning peak and afternoon peak time periods.

Of greater importance to this report are the findings relative to the traffic volume variables. As noted above, in all cases using VKT for exposure results in a better fit than using opportunities. Furthermore, adding AADT as a covariate (with an exponent) always increases the log-likelihood, though more substantially for the models with number of opportunities as exposure. What is most interesting is the resulting exponents estimated for AADT in each model, which are all significantly different from 0. It is first important to remember that opportunities is really a function of the AADT squared (as demonstrated earlier), and VKT is just the AADT multiplied by the number of days in the study period and the segment length. Consequently, the negative exponents estimated for AADT are really offsetting an exponent of 2 for the opportunities models and 1 for the VKT models. This results in effective exponents of about 0.77 and 0.66 for the opposite direction sideswipe models with opportunities and VKT, respectively, 0.9 and 0.87 for the head-on models, and 0.12 and 0.72 for the single vehicle models. These effective exponents are consistent with previous research by the second author (Qin *et al.* 2005); a similar modeling effort using data from four different states permitting the exponent on AADT to be estimated freely found exponents close to 1.0 for head-on crashes (including both head-ons and sideswipes), and much less than 1.0 for single vehicle crashes.

2.7 Summary and Conclusions

These findings suggest that the number of opportunities is not as effective for explaining the variation in crash experience as was hoped. This is demonstrated most strongly in that the models using the new opportunity measure as exposure do not perform as well as models with the same covariates using VKT as exposure. Also, the log-likelihood improves much more when AADT is added to the models with the new opportunity measure than the models with VKT as exposure, indicating that including AADT is more critical for the opportunity models for explaining crash occurrence.

These results do not necessarily preclude the applicability of the opportunity function for replacing VKT as the exposure measure. First of all, we acknowledged at the beginning that while we expected the proposed new function to be more logical for accounting for the number of opportunities for a crash to occur, the actual risk of a crash occurring is likely to depend upon the traffic volume as well. These poor results suggest that it may be more efficacious to enter the traffic volume in the crash risk function in a different form, perhaps as a covariate in the exponential risk function rather than raised to an exponent and multiplied by the risk function.

Second, it is important to remember that we assumed the hourly volume distribution and directional splits for the entire study period to be represented by traffic counts taken on a single day for each study location. Clearly, these distributions of traffic volumes are not constant through the entire year; unfortunately, there has been insufficient research to date describing exactly how much variation there is. Perhaps the most important question is whether this variation is greater within or between highway locations. If the latter, then using one day counts may be sufficient; however, if the former, then it would not be.

This research has apparently raised more questions than it answered. One question, whether or not alternative representations of traffic volume in the crash risk function may help models with the opportunity function perform better than traditional models using VKT as exposure, is the subject of ongoing research by the authors. A second question, related to the variation of the hourly and directional traffic distribution at and among highway locations, is another subject of interest to the authors, and deserves concentrated research. This issue is of importance not only to motor vehicle crash forecasting, but also to transportation planning and air quality forecasting, where traffic volumes by time of day are important due to differences in the impacts of vehicle emissions through the day at a single location, as well as spatially in the road network.

Table 2.5 Model Results with Opportunities as Crash Exposure

Crash Type	Covariate (AADT)	Width 1	Width 2	Intercept	Early morning	Morning Peak	Midday	Afternoon Peak	Evening
Opposite-direction sideswipe	N/A	1.43 (<0.0001)	0.17 (0.5859)	-15.80 (<0.0001)	-0.84 (0.1535)	-2.98 (<0.0001)	-2.66 (<0.0001)	-2.94 (<0.0001)	-2.04 (<0.0001)
	-1.23 (<0.0001)	1.00 (0.0002)	0.17 (0.5302)	-4.65 (0.002)	-0.45 (0.3811)	-2.83 (<0.0001)	-2.48 (<0.0001)	-2.92 (<0.0001)	-1.88 (<0.0001)
	AADT 1 -1.00 (0.0003)	-3.90 (0.3477)	-2.33 (0.5531)	-1.76 (0.6004)	-0.48 (0.3469)	-2.86 (<0.0001)	-2.51 (<0.0001)	-2.94 (<0.0001)	-1.90 (<0.0001)
	AADT 2 -1.27 (<0.0001)								
	AADT 3 -1.55 (<0.0001)								
Head-on	N/A	0.90 (0.0535)	0.44 (0.3290)	-17.01 (<0.0001)	-0.67 (0.4563)	-3.36 (<0.0001)	-2.25 (0.0002)	-2.66 (<0.0001)	-1.92 (0.0018)
	-1.10 (<0.0001)	0.43 (0.3222)	0.37 (0.3395)	-6.73 (0.0039)	-0.61 (0.4584)	-3.28 (<0.0001)	-2.33 (<0.0001)	-2.79 (<0.0001)	-2.04 (0.0002)
	AADT 1 -1.43 (0.0051)	2.98 (0.6596)	-1.33 (0.81)	-6.36 (0.2148)	-0.60 (0.4667)	-3.26 (<0.0001)	-2.32 (<0.0001)	-2.78 (<0.0001)	-2.03 (0.0002)
	AADT 2 -0.96 (0.0032)								
	AADT 3 -1.14 (0.0378)								
Multi-vehicle	N/A	1.30 (<0.0001)	0.27 (0.3231)	-15.54 (<0.0001)	-0.78 (0.1220)	-3.06 (<0.0001)	-2.50 (<0.0001)	-2.82 (<0.0001)	-1.96 (<0.0001)
	-1.18 (<0.0001)	0.88 (0.0001)	0.24 (0.2982)	-4.83 (0.0002)	-0.45 (0.3036)	-2.90 (<0.0001)	-2.39 (<0.0001)	-2.85 (<0.0001)	-1.89 (<0.0001)
	AADT 1 -1.05 (<0.0001)	-2.46 (0.4894)	-2.20 (0.5060)	-2.55 (0.3671)	-0.48 (0.2761)	-2.93 (<0.0001)	-2.42 (<0.0001)	-2.87 (<0.0001)	-1.90 (<0.0001)
	AADT 2 -1.16 (<0.0001)								
	AADT 3 -1.42 (<0.0001)								
Single-vehicle	N/A	0.76 (0.0002)	0.43 (0.0359)	-12.74 (<0.0001)	0.32 (0.2857)	-3.34 (<0.0001)	-3.00 (<0.0001)	-3.83 (<0.0001)	-2.40 (<0.0001)
	-1.88 (<0.0001)	0.44 (0.0012)	0.07 (0.5987)	3.76 (<0.0001)	0.16 (0.3948)	-3.47 (<0.0001)	-3.24 (<0.0001)	-3.96 (<0.0001)	-2.69 (<0.0001)
	AADT 1 -1.55 (<0.0001)	-7.18 (0.0009)	-4.81 (0.0135)	8.56 (<0.0001)	0.15 (0.4345)	-3.48 (<0.0001)	-3.23 (<0.0001)	-3.96 (<0.0001)	-2.69 (<0.0001)
	AADT 2 -1.86 (<0.0001)								
	AADT 3 -2.42 (<0.0001)								

Number in parentheses is the significance level for the parameter

Table 2.6 Model Results with VKT as Crash Exposure

Crash Type	Covariate	Width 1	Width 2	Intercept	Early morning	Morning Peak	Midday	Afternoon Peak	Evening
Opposite-direction sideswipe	N/A	1.25 (<0.0001)	0.24 (0.3779)	-16.00 (<0.0001)	-0.62 (0.2149)	-1.38 (0.0001)	-1.19 (0.001)	-1.20 (0.0004)	-0.76 (0.03)
	-0.34 (0.029)	1.14 (<0.0001)	0.23 (0.3961)	-12.90 (<0.0001)	-0.59 (0.2386)	-1.38 (<0.0001)	-1.20 (0.0008)	-1.22 (0.0003)	-0.76 (0.0268)
	AADT 1 -0.30 (0.2560)	-1.06 (0.7957)	-2.07 (0.5976)	-11.04 (0.001)	-0.60 (0.2334)	-1.40 (<0.0001)	-1.20 (0.0007)	-1.22 (0.0002)	-0.76 (0.0264)
	AADT 2 -0.29 (0.1989)								
	AADT 3 -0.54 (0.1403)								
Head-on	N/A	0.60 (0.1554)	0.42 (0.2834)	-16.79 (<0.0001)	-0.75 (0.3582)	-1.90 (0.0017)	-1.10 (0.0364)	-1.16 (0.0189)	-0.97 (0.0721)
	-0.17 (0.4794)	0.55 (0.1991)	0.41 (0.2846)	-15.21 (<0.0001)	-0.75 (0.3606)	-1.91 (0.0016)	-1.12 (0.0343)	-1.18 (0.0174)	-0.99 (0.0681)
	AADT 1 -0.71 (0.1425)	5.92 (0.3734)	-0.69 (0.9069)	-15.87 (0.0022)	-0.73 (0.3715)	-1.89 (0.0019)	-1.10 (0.0373)	-1.16 (0.0196)	-0.97 (0.0724)
	AADT 2 0.02 (0.9513)								
	AADT 3 -0.10 (0.8566)								
Multi-vehicle	N/A	1.11 (<0.0001)	0.30 (0.1896)	-15.64 (<0.0001)	-0.66 (0.1301)	-1.50 (<0.0001)	-1.15 (0.0002)	-1.17 (<0.0001)	-0.80 (0.0075)
	-0.94 (<0.0001)	0.62 (<0.0001)	0.10 (0.41)	-5.12 (<0.0001)	0.00 (0.9839)	-1.84 (<0.0001)	-1.76 (<0.0001)	-2.06 (<0.0001)	-1.40 (<0.0001)
	AADT 1 -0.36 (0.1273)	0.46 (0.8967)	-1.85 (0.5764)	-11.82 (<0.0001)	-0.62 (0.1460)	-1.50 (<0.0001)	-1.15 (0.0001)	-1.19 (<0.0001)	-0.80 (0.0063)
	AADT 2 -0.18 (0.3313)								
	AADT 3 -0.42 (0.1771)								
Single-vehicle	N/A	0.80 (<0.0001)	0.20 (0.1628)	-13.43 (<0.0001)	0.03 (0.8987)	-1.78 (<0.0001)	-1.65 (<0.0001)	-2.00 (<0.0001)	-1.30 (<0.0001)
	-0.28 (0.0338)	1.02 (<0.0001)	0.29 (0.1997)	-13.06 (<0.0001)	-0.62 (0.1463)	-1.50 (<0.0001)	-1.15 (0.0001)	-1.19 (<0.0001)	-0.80 (0.0062)
	AADT 1 -0.76 (<0.0001)	-4.83 (0.0141)	-4.39 (0.0134)	-1.15 (0.4588)	-0.01 (0.9745)	-1.84 (<0.0001)	-1.75 (<0.0001)	-2.06 (<0.0001)	-1.40 (<0.0001)
	AADT 2 -0.88 (<0.0001)								
	AADT 3 -1.38 (<0.0001)								

Number in parentheses is the significance level for the parameter

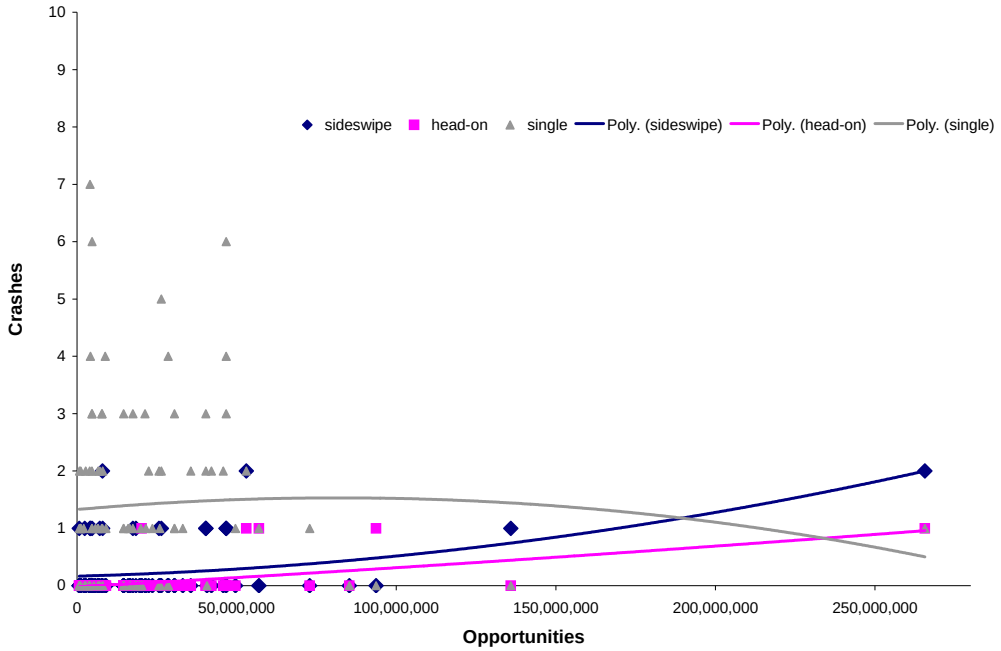


Figure 2.4 Crashes by Opportunities during Morning Peak Time period

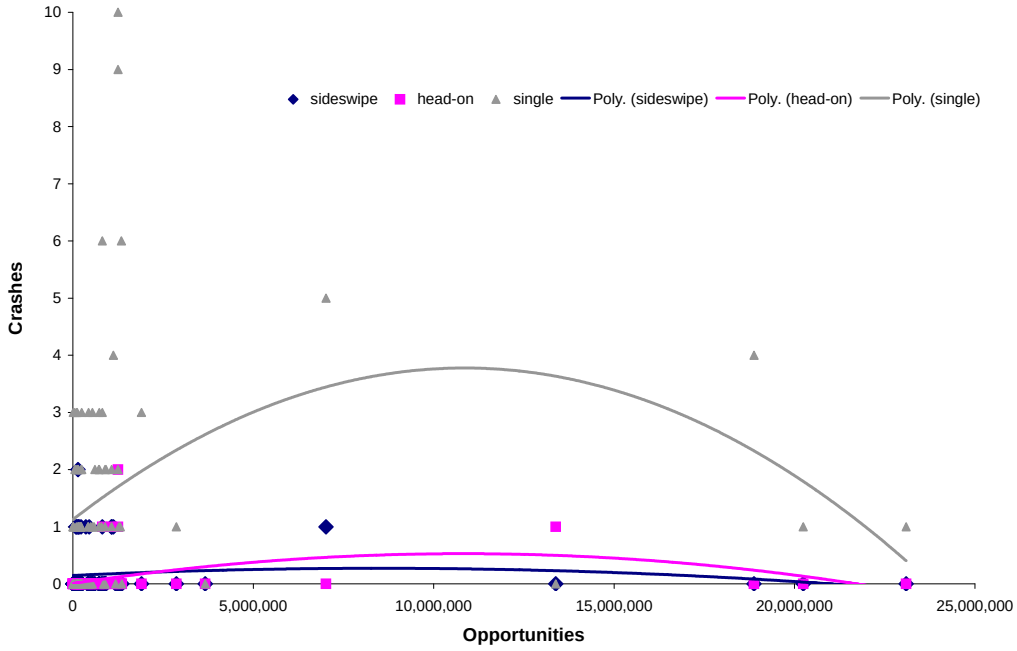


Figure 2.5 Crashes by Opportunities during Overnight Time Period

3 Collision Type Categorization Based on Crash Causality and Severity Analysis

3.1 Introduction

Traditional motor vehicle crash prediction modeling has mainly focused on investigating the correlation between crashes and related explanatory variables, such as lane width, shoulder width, horizontal/vertical curvature, median type, traffic control type, and land use. Statistical methodologies have been widely applied in such analysis, and studies have identified important road features which are apparently correlated with crash risk. However, knowing such correlations alone is not sufficient to reduce the occurrence of crashes, which requires first understanding the causes of the collisions. Since traditional statistical crash prediction models cannot relate crash occurrences to the actual causes, relying only on the model results one cannot learn about what actually causes the crashes.

Instead, crash causality is better determined through accident reconstruction analysis, which focuses on identifying events necessary for accident occurrence (2006). Here, the crash contributing factors assigned by reporting police officers are considered to be a surrogate account of crash causality. This approach is to relate road, traffic and other relevant characteristics to the observation of specific contributing factors rather than to the observation of crashes. Thus, instead of simply identifying correlation between road characteristics and crash incidence, we can identify correlation between road characteristics and crash contributing factors, which can hopefully give clearer information about how to reduce crashes.

Specifically, we observed the contributing factors for a sample of motor vehicle crashes in Connecticut and categorized their collision types using K-means clustering into groups with similar contributing factors. Because crash prediction by collision type is often used as a surrogate or a step towards prediction of crash severity, we also examined whether or not the severity distributions of the collision types in each group are similar to one another, and more different from those in other groups. This paper describes the outcome of these analyses and offers interpretations and suggestions as to how crash prediction modeling might be improved as a result for use in identifying crash causal factors.

3.2 Background

In motor vehicle crash analysis, predicting crash outcome or severity is often as important as predicting the crash count, particularly from a public health or benefit/cost analysis viewpoint. Many studies have thus investigated the correlation between crash severity and related factors, such as driver characteristics, environmental factors, roadway conditions, and vehicle-related variables. Khorashadi *et al.* (2005) studied the difference in driver severities between rural and urban accidents involving large trucks, and found road geometry, weather type, vehicle type, and driver action all to be influential. Delen *et al.* (2006) identified significant predictors for accident severities using neural networks and found that age is an important predictor for driver severities in severe injury crashes. Kim *et al.* (1995) studied personal and behavioral effects on crash and severity, and found that head-on and roll-over crashes strongly increase the odds of severe injuries. Vehicle travel speed is always recognized as the most important factor for determining crash severity. In Nilsson's study about the effects of speed on traffic safety (2004), he confirmed that the vehicles' collision speed gives rise to a kinetic energy which is absorbed by the vehicle constructions, passive safety measures and the involved car occupants, and the forces caused by the deceleration due to the sudden transformation of the kinetic energy leads to the occupants' injuries. The kinetic energy is proportional to the square of the speed; therefore, the higher the (relative) speed in a collision, the higher the probability that someone will get injured. Other than that, for both single- and multi-vehicle crashes, the direction in which a vehicle collides with a fixed object or another vehicle is also strongly correlated with the crash outcome or severity. The involved driving maneuvers vary by collision type, thus, different collision types potentially have different distributions of crash severity due to the differences in how the vehicles collide.

As Davis (2004) has argued, the circumstances under which road accidents occur are complex due to the random nature of accidents and the multiplicity and diversity of factors that combine to result in the accident. For example, many drivers negotiate their way through the same point on the road under the same physical and traffic conditions without experiencing a collision. A collision occurs when other factors come into play–factors that are difficult if not impossible to measure and predict, such as the physical, medical

and emotional state of the driver, the weather conditions and the driver's relative aggressiveness. All of these factors must be accounted for by the random element of crash prediction models. Transferring such models to contexts other than those in which they were estimated (*i.e.*, another geographic jurisdiction or even time period), must attempt to account for differences in these factors between the estimation and application contexts. Any correlation between road characteristics and crash occurrence simply indicates that the same combination of unobserved effects (driver, weather and other factors) is more likely to result in a collision under that combination of covariates. This does not necessarily imply causality; at best it suggests that the indicated combination of covariates exacerbates the driving task.

In other words, crash causality can only be confidently established through much more microscopic analysis of actual collisions as is done in accident reconstruction analysis. On the other hand, commonly available crash data can also provide hints at crash causality, namely the "contributing factor" usually determined by the police officer investigating the collision. The contributing factor cannot provide complete and detailed information on exactly how each crash occurred. However, finding a strong connection between contributing factors and road features might offer more reliable information about how to reduce collisions, as we can learn under what road conditions those contributing factors tend to be observed. Knowing both the relevant contributing factor and the apparently conspiring road characteristics offers much clearer direction into how to improve the road to reduce the incidence of the indicated collisions.

Studies have shown that the relationship between crash rate and traffic flow related variables varies by collision type and that to consider this variation in crash modeling improves the prediction results. The traditional ways of categorizing collision types have concentrated on differentiating in terms of number of involved vehicles, direction of travel, and the combination of these factors. Lord *et al.* (2005) considered traffic density and volume to capacity (V/C) ratio for crash risk prediction, and they asserted the necessity of separating single and multi-vehicle crashes for obtaining better model prediction. Zhou and Sisiopiku (1997) also found that a U-shaped pattern explains the relationship between V/C value and crash rate for multi-vehicle collisions, while a decreasing trend with V/C value is observed for single-vehicle collisions. Qin *et al.* (2004) disaggregated crashes into four collision types: single-vehicle, multi-vehicle same direction, multi-vehicle opposite direction and multi-vehicle intersecting direction, and showed that the relationship between crashes and traffic flow is different for each of these collision types.

By focusing on relating traffic flow to crash rate for showing differences among collision types, these studies did not consider an analytical methodology for defining crash categories. Hauer *et al.* (1988) re-categorized intersection collision types according to the involved traffic flows, and they argued that to define crash categories just by collision type would lead to the loss of cause-and-effect relationship with traffic flow. Their study took a step further at considering the involved traffic flow of the crashes, yet, there is still more to deliberate to relate the crash categorization with the actual causalities.

To develop a more analytical strategy for categorizing collision types, the differentiation among collision types ought to be captured by examining a wider range of factors leading to the particular collision type. These factors include crash causality or contributing factor for its direct relationship with the crash occurrences. Also, similar patterns observed in severity distributions also indicate that the proper categorization will further improve the crash severity prediction. Consequently, the crash prediction models estimated by the redefined crash category can then potentially generate better results showing the relationship between crash risk and the explanatory variables.

3.3 Study Data

We randomly selected 655 segments with uniform length of 1-km from 50 two-lane rural roads in Connecticut. Only minor intersections and driveways are included within the segments to limit the complexity due to the traffic operation and control. Of the 655 two-lane road segments, lane width, shoulder width, and speed limit values range from 8 ft (2.5 m) to 14 ft (4.5 m), 1.5 ft (0.5 m) to 6.5 ft (2 m), and 25 mph (40 km/h) to 50 mph (80 km/h), respectively. Along with the segment geometric variables, crash data were collected from the Connecticut Collision Analysis System developed by Connecticut Department of Transportation, which provides a complete crash database of information on Connecticut motor vehicle collisions, and that of the involved vehicles and persons. After confirming the consistency of segment characteristics over the study period, 6-years of crash data from 1996 to 2001 were collected for the selected segments.

The crash data were gathered from reports written by the police officers investigating the crashes (as is standard for road safety research). It is understandable that the contributing factors represent the subjective opinion of the reporting officers, and that in many cases where multiple contributing factors may apply, only one is recorded for the collision. However, the dataset used in this study is large enough so that the observed distribution of reported contributing factors for each collision type should be representative of the distribution expected in the entire population of contributing factors assigned. For each crash we extracted the collision type, the contributing factor, severity and some driver characteristics. In the Connecticut crash database, there is only one contributing factor assigned to each collision; therefore, this study was conducted based on the collision dataset with regard to the collision type and contributing factor of each particular case. Tables 3.1 and 3.2 list the possible values for collision type and contributing factor respectively. Table 3.3 presents the number of crashes observed for each combination of collision type and contributing factor, revealing some notable patterns. Figure 3.1 presents these relationships graphically. First, there is a great diversity in contributing factor even for a single collision type, so it is necessary to examine how each contributing factor relates to each type of collision. For example, fixed object crashes are most often associated with the contributing factor "speed too fast for conditions" but were also associated with "defective equipment". For the latter situation, if the equipment had not failed at that instant that fixed object collision likely would not have occurred. Obviously, no improvements to the road or the traffic control can prevent equipment failure on an individual vehicle, so it is not necessarily helpful for our purpose to use that contributing factor for classifying crashes. Another interesting pattern was that several collision types were associated with multiple contributing factors that actually all imply a similar influence from the road conditions. For example, the most common contributing factors for head-on collisions are "speed too fast for conditions" and "driver lost control", both of which likely imply restricted sight distance along with a roadway design that encourages speeding. Finally, the same combinations of contributing factor seem to be associated with multiple collision types. For instance, 39 percent of head-on and 42 percent of fixed object collisions are associated with "speed too fast for conditions" and 16.9 percent of head-on and 27.3 percent of fixed object collisions with "driver lost control", respectively.

Table 3.1 Collision Type Definitions

Crash Type	Description	Abbreviation
1	Turning - Same Direction	TSD
2	Turning – Opposite Direction	TOD
3	Turning - Intersecting Paths	TIP
4	Sideswipe - Same Direction	SSD
5	Sideswipe - Opposite Directions	SOD
6	Miscellaneous Non-Collision	
7	Overturn	OVT
8	Angle	ANG
9	Rear-end	RE
10	Head-on	HO
11	Backing	
12	Parking	
13	Pedestrian	PED
14	Jackknife	
15	Fixed Object	FO
16	Moving Object	
17	Unknown	

Table 3.2 Contributing Factor Definitions

Contributing Factor	Description	Abbreviation
1	Driving on Wrong Side of Road	DOW
2	Speed Too Fast For Conditions	STF
3	Violated Traffic Control	VTC
4	Under the Influence	UTI
5	Failed to Grant Right of Way	FGR
6	Improper Passing Maneuver	IPM
7	Improper Lane Change	ILC
8	Following Too Closely	FTC
9	Slippery Surface	SS
10	Driver Lost Control	DLC
11	Animal or Foreign Object In Road	AIR
12	Fell Asleep	FA
13	Defective Equipment	DE
14	Driver Illness	DI
15	Driver's View Obstructed	DVO
16	Unsafe Tires	UT
17	Unsafe Use of Highway by Pedestrian	UBP
18	Unsafe Right Turn on Red	UOR
19	Driverless Vehicle	DV
20	Insufficient Vertical Clearance	IVC
21	Proper Turn Signal Not Displayed	PND
22	Disabled or Illegally Parked Vehicle	DPV
23	Abnormal Road Condition	ARC
24	Vehicle Without Lights	VWL
25	Traffic Signal Not Operating	TNO
26	Vehicle Involved in Emergency	VIE
27	Entered Roadway in Wrong Direction	EIW
28	Roadway Width Restricted	RWR
29	Unknown	UKN
30	Unsafe Backing	UB
31	Improper Turning Maneuver	ITM

Table 3.3 Crash Counts by Collision Type and Contributing Factor

	Same-direction (SD)			*Intersecting (INT)*			*Opposite-direction (OD)*		*Single-vehicle (SV)*		
	TSD	**SSD**	**RE**	**TOD**	**TIP**	**ANG**	**SOD**	**HO**	**OTN**	**FO**	**Total**
STF	10	26	182	4	20	19	152	67	81	1488	2051
FTC	51	28	1480				2	2	2	24	1589
DLC	1	21	34		4	4	79	29	51	967	1199
FGR	9	3	2	271	430	99		1	4	26	850
SS	1	16	34	2	4	8	34	7	23	337	466
FA		1	4				17	9	8	265	304
IPM	137	86	18	1	4	1	9	3	1	18	278
UTI	1	8	13	1	7	1	22	6	9	200	272
DOW				1	5		119	48		29	202
VTC		2	1	5	59	100	1			8	177
AIR			5						3	159	167
ITM	67	2	1	13	31					20	134
DI			4				5	3		67	79
UKN		1	3	2	1	1	17	2		37	67
DVO		2	13	1	13	1	1			21	53
DE		1	6		1		6		1	25	40
UT							5		2	19	26
ILC		16								1	17
ARC							2	1	2	11	16
RWR		2					5			6	15
DV			1		1	2	1	1		4	11
DPV		3	5		1			2			11
PND	10										10
IVC										9	9
UB			3							5	9
UOR					2						2
VWL	1		1	1						1	4
TNO						1					1
VIE										4	4
EIW				1	1	1	2	1			6
UBP											0
Total	288	218	1810	303	584	238	479	182	187	3751	8104

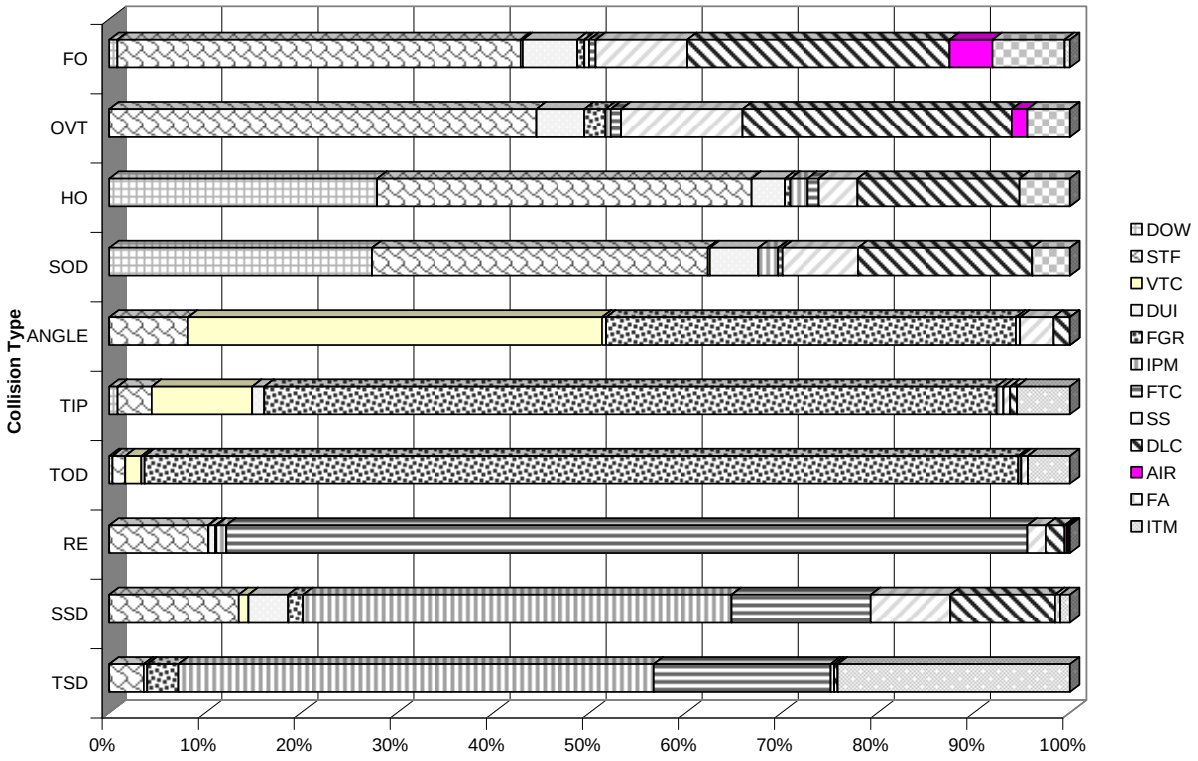


Figure 3.1 Contributing Factor Distributions by Collision Type

As a result, to simply aggregate collision types to obtain the crash count for the model prediction without looking at the underlying contributing factors severely restricts the ability of the resulting models to tell us anything about why the crashes occurred. Instead, if we define collision type categories in terms of contributing factors, the crash risk of a certain collision type could then be an indication of a strong correlation between a set of contributing factors and the explanatory variables. In addition, to relate the specific contributing factors to different collision types can help to interpret crash model results by indicating which segment descriptor variables relate to specific contributing factors, supporting consideration of the role of driver behavior. Therefore, we analyzed the distribution of contributing factors for each collision type and their patterns across different collision types.

We also extracted crash severity indications from these data, to investigate whether or not this classification of collision types helps to directly predict the expected severity distribution. First we examined the distribution of crash severity for each collision type, rated on a KABCO scale, where the crash severity level depends on the most severely injured person in each crash. Since the severity distribution might vary from collision type to collision type due to differences in relative speeds and colliding directions, we expected it may be necessary to categorize the collision types further to uniquely predict the severity distribution for each category. For collision types within the same category, the fatality (K) and severe injury (A) rate or total injury (K, A, B, and C) rate can vary greatly, which might be due to slight difference in their contributing factor distributions or the combination of contributing factors and some other explanatory variables that need to be further addressed. By categorizing collision type according to severity distribution, we hope to achieve new collision categories with more homogeneous severity distributions and being capable of more accurately predicting crash severity.

3.4 Collision Type and Contributing Factor

The 17 pre-defined collision types and 31 contributing factors in the Connecticut crash database were examined to identify those that are not related to the road characteristics but rather due to unusual traffic maneuvers or mechanical failure. Among the 17 collision types, 6 were omitted due to the inestimable nature of the incidents, such as aberrant driver behavior or an unclear crash situation. These 6 collision types include the crashes involving backing, improper parking, moving object (often used for collisions with animals), those due to unknown factors and miscellaneous non-collisions, and jackknifes. Crash counts for the remaining 10 collision types (excluding collisions involving pedestrians) were then matched to the 31 different contributing factors. Note that 19 contributing factors were observed to have occurred in less than 100 collisions during the 6-year study period, and among them 7 were observed to have occurred in fewer than 10 collisions. Most of these appeared to have occurred due to vehicle malfunction or due to very unusual driving maneuvers such as backing up on the roadway. Finally, we chose the most commonly observed 12 contributing factors for analysis.

In Table 3.4, the cell contents are crash percentages under each of the 12 contributing factors for each collision type (Table 3.3 shows the actual crash counts for the crash types caused by each contributing factor). The pattern of values greater than 10 percent (those marked in bold face) shows that these predominant contributing factor/collision type pairings form three distinct clusters. For the collision types involving vehicles traveling in the same-direction, "improper passing maneuver", "following too closely", "driver lost control", "speed too fast for conditions", and "improper turning maneuver" are the main contributing factors. Turning-intersecting crashes were more likely to occur due to a driver "failing to grant right of way" or "violating traffic control". Opposite-direction collisions, including sideswipe opposite direction and head-on, and single vehicle collisions, including overturn and fixed object, seem to share the same main contributing factors such as "speed too fast for conditions" and "driver lost control", except that the distribution of these contributing factors varies slightly among these collision types. These same patterns are also easily seen in Figure 3.1.

Table 3.4 Proportion of Crashes by Contributing Factor for Each Collision Type

	TSD	SSD	RE	TOD	TIP	ANG	SOD	HO	OVT	FO
FA	0	0.005	0.002	0	0	0	0.039	0.052	0.044	0.075
SS	0.004	0.083	0.019	0.007	0.007	0.034	0.078	0.041	**0.126**	0.095
DOW	0	0	0	0.003	0.009	0	**0.274**	**0.279**	0	0.008
DLC	0.004	**0.109**	0.019	0	0.007	0.017	**0.182**	**0.169**	**0.280**	**0.273**
STF	0.036	**0.135**	**0.103**	0.013	0.035	0.082	**0.349**	**0.390**	**0.445**	**0.420**
FTC	**0.184**	**0.145**	**0.834**	0	0	0	0.005	0.012	0.011	0.007
IPM	**0.495**	**0.446**	0.010	0.003	0.007	0.004	0.021	0.017	0.005	0.005
ITM	**0.242**	0.010	0.001	0.044	0.055	0	0	0	0	0.006
FGR	0.032	0.016	0.001	**0.909**	**0.762**	**0.427**	0	0.006	0.022	0.007
VTC	0	0.010	0.001	0.017	**0.105**	**0.431**	0.002	0	0	0.002
UTI	0.004	0.041	0.007	0.003	0.012	0.004	0.051	0.035	0.049	0.056
AIR	0	0	0.003	0	0	0	0	0	0.016	0.045

(Values in Bold exceed 0.10)

3.5 Methodology

K-means cluster analysis is conducted to group the 10 collision types using the contributing factors as reference variables according to the Euclidean distance between the descriptor variables (contributing factors). In the clustering procedure, the 10 collision types were partitioned into four groups such that a pair of elements in the same cluster tends to be more similar than a pair of elements belonging to different clusters. The aim of the K-means cluster algorithm is to classify a set of data points into K categories through the K clusters defined a priori (1967). The main idea is to define K initial cluster

centers $C^o = \{c_1^o, c_2^o, ..., c_K^o\}$, one for each cluster, and then to relate the rest of the data points to these centers depending on the closest Euclidean distance,

$$\boldsymbol{d(x_i, x_j) = \left[\sum_{v=1}^{V} (x_{i,v} - x_{j,v})^2 \right]^{1/2}} \tag{3.1}$$

where $d(x_i, x_j)$ is the Euclidean distance between point *i* and *j*, and *V* is the dimension of the reference variable vector. Since each distinct data cluster should include data points relatively near one another, the initial cluster centers need to be as far away from each other as possible. From the three natural groupings shown in bold in Table 3.4, we chose "turning same-direction", "turning intersecting-path", and "overturn" collision types to serve as the initial cluster centers after observing that these three types have virtually no overlapping contributing factors and represent the largest number of collisions in each cluster. Noticing that over 80 percent of the rear-end collisions are attributed to "following too close" compared to less than 20 percent for turning same-direction and sideswipe same-direction collisions, we finally chose four initial centers including the rear-end collision type as well.

The program then applies the following two steps iteratively. First, it assigns each data point to the nearest cluster center according to the Euclidean distance using equation (3.2).

$$\begin{aligned} \boldsymbol{c_j^{i+1}} &= \boldsymbol{\{x \mid d(x, \mu_j^i) \le d(x, \mu_{j'}^i), 1 \le j, j' \le k \text{ and } j' \ne j\},} \\ \boldsymbol{C_j^{i+1}} &= \boldsymbol{\{c_1^{i+1}, c_2^{i+1}, ..., c_k^{i+1}\}} \end{aligned} \tag{3.2}$$

Then when all points have been assigned to a cluster, the cluster centers are recalculated based on their assigned data points using equation (3.3),

$$\boldsymbol{\mu_j^i = \frac{1}{| c_j^i |} \sum_{x \in c_j^i} x} \tag{3.3}$$

where *j* refers to the clusters, *i* is the iteration counter, $| c_j^i |$ is the number of points of cluster *j* after the *i*th iteration, and μ_j^i is the center of cluster *j* after the *i*th iteration. The procedure using equations (3.2) and (3.3) is repeated until the cluster centers do not change any more (that is, they converge), and then the K clusters each with a subset of data points become the final categories.

The K-means cluster analysis was applied to the contributing factors percentages (Table 3.4) using SPSS software (2004). We used the percentage of crashes identified with each contributing factor as the dimension for determining the final center of each of four clusters of collision types. Once the final clusters are determined, the distance from each observation (in this case, the collision types) to the center of the cluster to which it was assigned is calculated. Comparing these distances tells how the observations within the same cluster are ranked based on their closeness to the center. Also, the distances between the cluster centers demonstrate how far apart the clusters are from each other. The final cluster centers can also be obtained to show the average value of the dimension variables in that cluster. For this case, since the contributing factor percentages were used in clustering, the larger value of the dimension variables at the cluster center show a higher influence of that contributing factor in grouping the collision types in that cluster.

3.6 Results and Discussion

The twelve contributing factors shown in Table 3.4 were used as dimension variables to categorize the 10 collision types; the clustering results are reported in Tables 3.5 and 3.6. As seen in Table 3.5, rear-end collisions formed a cluster on their own, and Table 3.6 shows that "following too closely" was the major contributing factor differentiating rear-end crashes from the other crash categories. Though it has been common practice to place rear-end collisions in the same category with sideswipe same-direction collisions, these clustering results suggest that rear-end collisions occur mostly due to different causal effects compared to the other same-direction collision types. Also observed in Table 3.5, turning same-direction collisions were clustered together with sideswipe same-direction collisions to form the "same-direction" cluster; the results in Table 3.6 show that "improper passing maneuver" and "following too closely" are the major contributing factors differentiating same-direction collisions from the rest of the

collision types. Turning opposite-direction, turning intersecting-path, and angle collisions were clustered together as the "intersecting-direction" cluster, with "fail to grant right of way" and "violate traffic control" as the two main contributing factors. Again Table 3.5 shows that sideswipe opposite-direction, head-on, overturn and fixed object collisions formed the "segment-crash" cluster, and Table 3.6 illustrates that "speed too fast for conditions", "driver lost control", and "drive on wrong side of road" discriminated these four collision types from the others.

Table 3.5 Final Cluster Memberships by Collision Types

Collision Type	Cluster	Intra-cluster Distance
Turning Same Direction (TSD)	1 (Same-direction Cluster)	0.147
Sideswipe Same Direction (SSD)	1 (Same-direction Cluster)	0.147
Turning Opposite Direction (TOD)	2 (Intersecting-direction Cluster)	0.271
Turning Intersecting Path (TIP)	2 (Intersecting-direction Cluster)	0.105
Angle (ANG)	2 (Intersecting-direction Cluster)	0.372
Rear End (RE)	3 (Rear-end Cluster)	0.000
Sideswipe Opposite Direction (SOD)	4 (Segment-crash Cluster)	0.152
Overturn (OVT)	4 (Segment-crash Cluster)	0.163
Head On (HO)	4 (Segment-crash Cluster)	0.158
Fixed Object (FO)	4 (Segment-crash Cluster)	0.147

Table 3.6 Final Cluster Centers for Each Crash Category

Contributing Factor	Cluster Centers			
	1	2	3	4
Drive on Wrong Side of Road (DOW)	0.000	0.004	0.000	**0.140**
Speed Too Fast (STF)	0.085	0.044	**0.103**	**0.401**
Violate Traffic Control (VTC)	0.005	**0.184**	0.001	0.001
Drive Under Influence (DUI)	0.023	0.007	0.007	0.048
Fail to Grand Right of Way (FGR)	0.024	**0.700**	0.001	0.009
Improper Passing Maneuver (IPM)	**0.470**	0.005	0.010	0.012
Following Too Close (FTC)	**0.165**	0.000	**0.834**	0.008
Slippery Surface (SS)	0.043	0.016	0.019	0.085
Driver Lost Control (DLC)	0.056	0.008	0.019	**0.226**
Animal or Foreign Object in Road (AIR)	0.000	0.000	0.003	0.015
Fell Asleep (FA)	0.003	0.000	0.002	0.053
Improper Turning Maneuver (ITM)	**0.126**	0.033	0.001	0.001

The above findings imply interesting conclusions regarding the correlation between contributing factors and road characteristics. First, rear-end collisions seem to stand out instead of being categorized together with the other same-direction collisions., This is due to the fact that more than 80 percent of rear-end collisions were caused by a vehicle following another too closely. This obviously occurs when the lead vehicle is traveling slower than the following vehicle, and implies either that the two drivers have different perceptions about the appropriate speed on the road, or that the lead vehicle is slowing to make a turn or due to other circumstances (such as sudden poor sight distance). When drivers select different speeds, this suggests ambiguous cues from the prevailing road environment, including road features such as the speed limit, the roadway width, roadside land development and the alignment.

Second, as expected, sideswipe same-direction and turning same-direction collisions form the same-direction collision cluster. The main contributing factors–"improper passing maneuver", "following

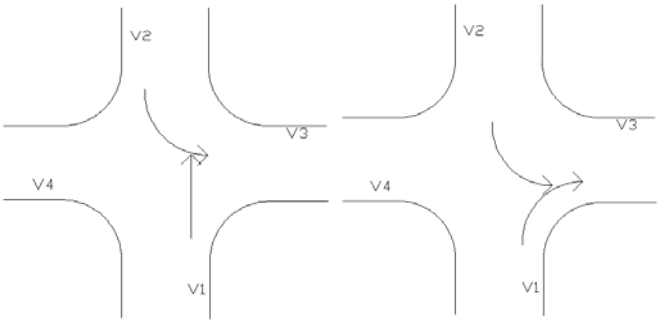
too closely", and "improper turning maneuver"–imply either that the following vehicle left an insufficient time or distance gap, as is the case with rear end crashes, or attempted to execute a passing or turning maneuver in a reckless manner. These crashes therefore appear to be related to ambiguous road design as for rear end crashes, but also require the presence of either a passing opportunity or a turning vehicle. Because police reports might not always account for a vehicle slowing to turn as a turning collision, instead of classifying it as a rear-end collision if the lead vehicle was not actually turning when the crash occurred, the difference between these crashes and rear-end collisions may be somewhat blurred, and these similar contributing factors confirm this.

Third, it is also not surprising to notice that the category of intersecting collisions, including turning opposite-direction, turning intersecting-path, and angle collisions, has "fail to grant right of way" and "violate traffic control" as its main contributing factors. Figure 3.2 illustrates the possible collision patterns for this cluster. From the clustering result, we notice that drivers' judgments on the traffic situation made before traversing the junction play an important role in determining whether a turning-involved crash could be avoided or not. However, we need to consider whether these collisions occur more often at some particular locations, and find out which aspects of the traffic control and intersection design at these locations seem to contribute to such driver behaviors.

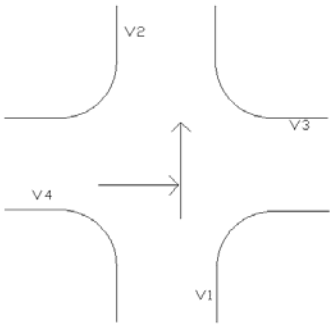
Another interesting finding is that opposite-direction collisions are categorized together with single vehicle collisions, with "speed too fast for conditions", "driver lost control", and "drive on wrong side of road" being the important contributing factors. This finding suggests that opposite-direction and single-vehicle collisions are more or less equivalent in the sense that they both should be treated as segment crashes and appear to result from the same contributing factors. Both "driver lost control" and "drive on wrong side of road" could be caused by driver's inattention and being distracted. "Speed too fast for conditions" could be due to drivers over-estimating the safe speed for the road, but it could also occur because of the restricted sight distance or the lead vehicle traveling at a much slower speed. Thus, all four collision types under such contributing factors can occur to through traffic at any point along the road. For instance, when a driver loses control, the vehicle collides into either another vehicle on the road at the same time or a roadside object, such as a guide rail, if there is no another vehicle on the road at that time. Other variables might come into play, such as vehicle speed, driver reaction time, lane width, shoulder width, and roadway alignment, to determine whether the collision can be avoided or not. Also, we need to find out whether such driver actions occur more frequently at some locations and how the frequency of such driver behavior correlates with the explanatory variables of the locations. Therefore, modeling opposite-direction and single-vehicle collisions together appears to be reasonable because they occur due to the same contributing factors and differentiating between them depends on knowledge about minute to minute traffic conditions on the road.

Based on the collision type categorization according to contributing factors, we further examined the distribution of severity by collision type. In Figure 3.3, the collision types are arranged in the following order: rear-end collisions, same-direction collisions, intersecting collisions, and segment collisions, and the distribution of severities were plotted for the collision types of each category. Within each category, the member collision types seem to have relatively similar distributions of severity. Especially for the distribution of fatality (K), severe injury (A), and observable injury (B), turning same-direction and sideswipe same-direction collisions have very similar patterns, and rear-end collisions are comparable to both of them. For the intersecting collision category, turning opposite-direction and angle collisions have very similar severity distributions, but turning intersecting-path collisions have lower injury rates at all four levels than the other two. For the segment collision category, head-on collisions have the highest injury rate, which is also the highest among all the collision types; and fixed object collisions seem to have the lowest injury rate of these four.

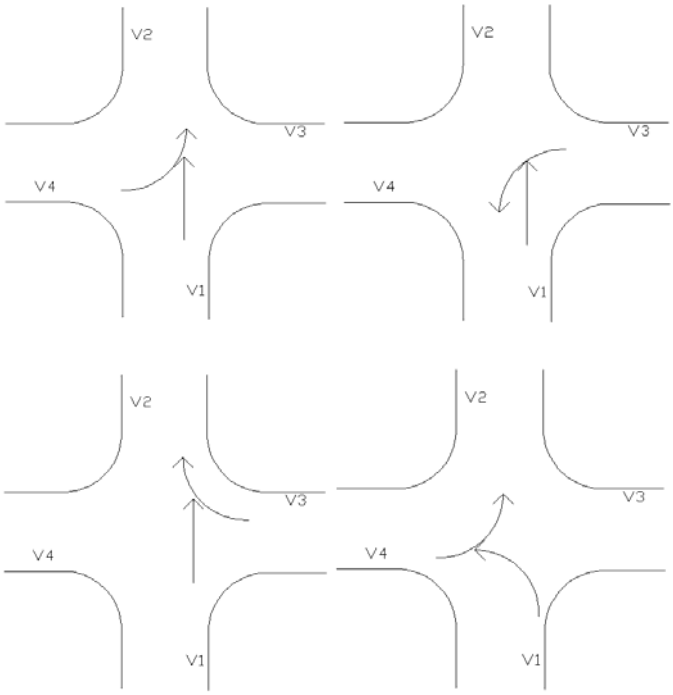
The above results suggest that in order for the crash categories to have relatively homogeneous injury severity distributions, some of the groups work well, but others must be divided further. For example, in the same-direction group, turning same-direction and sideswipe same-direction collisions have quite similar severity distributions, possibly due to lower relative colliding speeds and the same direction of travel for both collision types. Similarly, vehicles in rear-end collisions also have lower relative speed, and the slight difference in the injury rate between rear-end and the same-direction collisions could be due to a slightly different relative colliding direction of the vehicles. Intersecting traffic-related collision types, as illustrated in Figure 3.4, all have somewhat higher injury rates than rear-end and same-direction collisions, but themselves do not all have the same injury distributions: turning opposite-direction and angle collisions resulted in more and severer injuries than turning intersecting-path collisions.



(a) Turning Opposite-direction Collisions



(b) Angle Collisions



(c) Turning Intersecting-path Collisions

Figure 3.2 Illustration of Possibilities for Intersecting Traffic Collisions

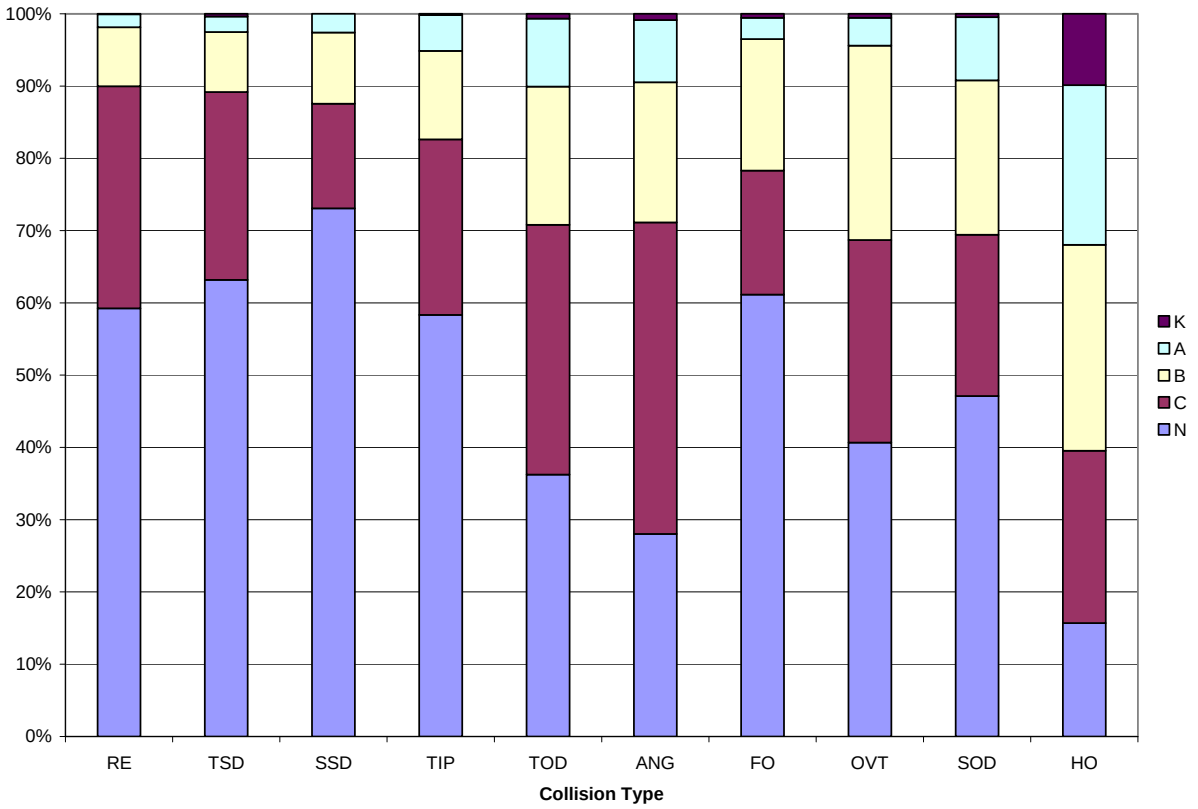


Figure 3.3 Severity Distributions by Collision Type

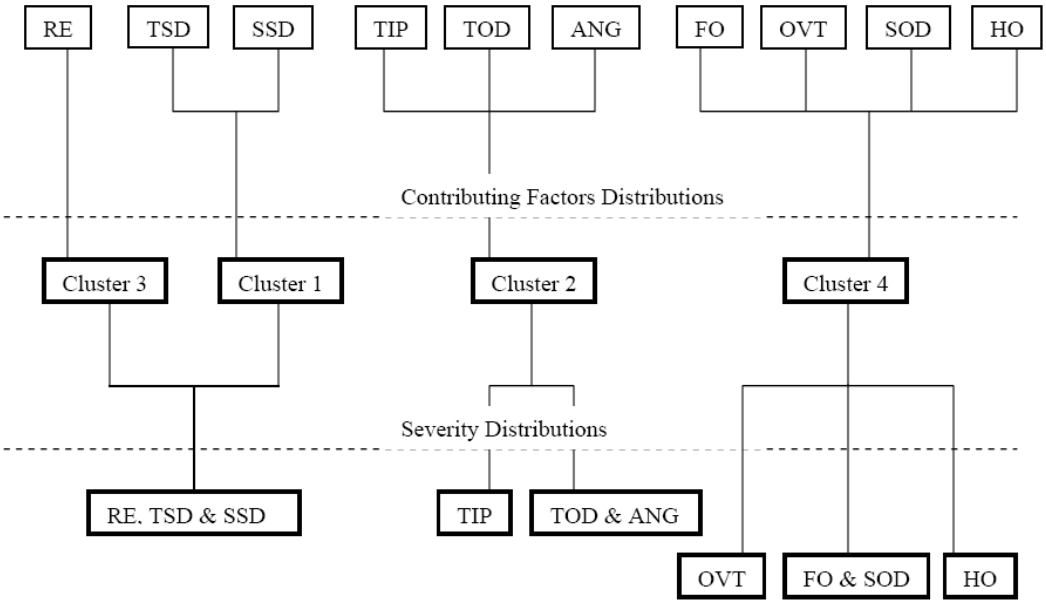


Figure 3.4 Categorization of Collision Types Based on Contributing Factors and Severity Distributions

This difference could also be due to the different colliding direction of vehicles: vehicles involved in turning opposite-direction and angle collisions always have higher relative speeds and mostly result in a broadside impact, but for vehicles involved in turning intersecting-path collisions, as seen in Figure 3.4, some such collisions can also have lower relative speed and tend to have a colliding result that is more similar to turning same-direction collisions. For the segment collision category, overturn collisions stand out to have the second highest injury rate, probably because the vehicle turning over is more violent than other collision types. Moreover, vehicle type and weather/surface conditions are probably more critical factors correlated with overturn collisions.

3.7 Summary and Conclusions

The major objective of this study was to examine the collision type categorization based on contributing factors and severity distributions. The crash data reported by on-duty police officers were used for collecting collision information including collision type, contributing factor, and severity level. Based on 12 contributing factors, using K-means cluster analysis methodology, 10 collision types were categorized into 4 different categories, namely, rear-end, same-direction (sideswipe and turning), intersecting (turning intersecting-path, turning opposite-direction and angle), and segment (fixed object, overturn, head-on, and sideswipe opposite-direction) collisions. The cluster results also show the common influential contributing factors for each of the categories.

From the analysis results, we found that the contributing factors are beneficial for classifying collision types into different categories with similar severity distributions. Furthermore, the variation in severity distribution within each category can be explained by further analyzing the occurrence details of the particular collision types. As shown in Figure 3.4, considering the distribution of both contributing factors and severities, we recommend to group together rear-end, sideswipe same-direction, and turning same-direction collisions due to potentially ambiguous distinction (by police officers) between their contributing factors and the similar severity distributions. However, turning intersecting-path collisions need to be separated from the category of intersecting traffic-related collisions due to the great discrepancy in their severity distributions. Turning opposite-direction and angle collisions not only have the same major contributing factors, but also share the very similar severity distribution. For the segment collision category, head-on collisions need to be in a separate group due to their extremely high injury rate at all levels, and the fixed object and sideswipe opposite-direction collisions can be grouped together due to the similar reason as for turning opposite-direction and angle collisions. The overturn collision type stands out alone because of the unique nature of its occurrence.

Compared to the collision type categories used in previous crash analysis studies, these categories do not exclusively classify collision types by the number and travel direction of involved vehicles. Moreover, this categorization methodology helps to move crash prediction analysis closer towards crash causality study by identifying the actual driver behavior (mistakes) that are correlated with road characteristics and provides information about what is useful to explain the correlation between crash risk and causalities. Eventually, such contributing factor prediction models will be able to tell traffic engineers more clearly how to improve roads to reduce collisions.

Future research in this vein could explore the possible factors attributed to the differences in severities of same category collision types (e.g., turning intersecting direction, turning opposite direction and angle), such as the actual direction of travel of the vehicles involved, which is not always perfectly clear in all crash reports. The more stable severity distributions of the collision type categories proposed here also are likely to lend themselves to better fitting crash severity models, which account for factors such as occupant age, seating position and weather and surface conditions. In any case, as more and more crash prediction modeling segregates models by collision type (as is occurring in current research for the Highway Safety Manual), it is becoming more important to standardize the categorization of crash types; this research offers a rigorously defended suggestion for such categorization.

4 Same-direction Crash Prediction Models Using Opportunities as Exposure

4.1 Introduction

Collisions involving vehicles traveling in the same direction are very common on road networks. These are typically defined in police accident reports as rear-end, sideswipe same-direction, or turning same-direction collisions. Rear-end collisions occur when the front of one vehicle strikes the rear of the vehicle traveling in front of it; typically such collisions occur due to drivers keeping close time/distance headways between one another. Sideswipe and turning same-direction collisions can occur under different scenarios, such as improper passing or turning maneuvers and following other vehicles too closely. In addition, turning same-direction collisions can also be caused by improper turning maneuvers. Although on average, same-direction collisions, especially rear-end collisions, result in lower severities compared to intersecting-direction or opposite direction collisions, their high frequency of occurrence can make their cost to society just as important. In 2004, nearly one million rear-end collisions occurred nationwide, 30 percent of the total number of collisions in that year. In the same year, head-on crashes were only 1.9 percent of total crashes (TSF 2004, NCSA).

4.2 Background

The previous chapter describes a crash type categorization method was proposed involving contributing factors, and broad collision types were defined to categorize collisions into four types with common contributing factors. The K-means clustering algorithm was implemented to complete this categorization, in which rear-end collisions formed a category of their own, because more than 80 percent had "following too closely" as the contributing factor, while same-direction sideswipe and turning collisions were grouped together in a second category, with a substantially different distribution of contributing factors.

Kim *et al.* (2007) modeled crash probabilities at rural intersections in the State of Georgia, USA. They found that treating collision counts by crash type as dependent variables along with other crash-related and intersection-related explanatory variables would result in more accurate explanation of crash risk variation. In their study, they defined five crash types at intersections: angle, head-on, rear-end, sideswipe same-direction, and sideswipe opposite-direction, and showed that comparing rear-end and sideswipe same-direction crash models, intersection-related variables seem to explain more variation in crash risk for the former. This finding further supports the notion of modeling crashes by type and also pointed out the underlying difference between rear-end and sideswipe same-direction collisions.

As mentioned earlier, for the three types of same-direction collisions to occur, there must be vehicles following one another in the same direction. It is therefore reasonable to assume that the higher the traffic volume is the more chances vehicles in the traffic stream have to accidentally collide with the ones traveling before them. The usual exposure measurement for segments, vehicle kilometers traveled (VKT), is defined as the average daily traffic (ADT) multiplied by the segment length and the number of days in the analysis time period. Wang and Abdel-Aty (2006) used Generalized Estimating Equations (GEE) models to estimate the effects of geometric design features on rear-end collisions at intersections in Florida and found the ADT on both major and minor roadways to be significant in explaining the rear-end crash risk.

As an exposure definition, VKT certainly helps to account for the effect of increasing traffic volume on increases in crash incidence. However by not taking into consideration the directional split of the traffic volume, it ignores potential differences from one segment to another in the operation of the traffic stream with regard to vehicles following and passing one another, even when the two-way volume is constant. For example, when the traffic volume is heavier in one direction than the other (e.g. a 70/30 directional split), vehicles traveling in the direction with heavier traffic might find more need to make passing maneuvers than when the directional traffic is more balanced (*i.e.* 50/50). On the other hand, when vehicles on a two-lane road find few opportunities to pass due to heavier traffic in the oncoming direction (e.g., the 30 percent direction), they are forced to follow slower vehicles, increasing the possibility of rear-end collisions.

In a study using inductive loop detector data to identify rear-end collision risk on freeways, Oh *et al.* (2006) defined a stopping distance index (SDI) which counts the number of rear-end conflict events by computing and comparing the safe stopping distance of a lead vehicle and the following vehicle. The safe stopping distance of a lead vehicle should always be greater than that of the following vehicle in avoiding a rear-end collision. Dividing this SDI by the number of vehicles passing a point over a certain time interval, they obtained a rear-end collision risk index (RCRI) and clustered the traffic situations regarding rear-end collision risk level based on vehicle categories and RCRI values. However, for two-lane road segments, it is difficult to have an exposure measurement based on the SDI since such a point-based index loses its meaning when trying to describe the traffic situation over a length of segment during a time interval.

Considering all of this past research, the purpose of this study is to evaluate a new definition of exposure to better account for the effects of traffic volume on the incidence of same-direction collisions on two-lane rural roads. This new exposure measure for same-direction crashes on two-lane highways is intended not only to represent the level of traffic intensity, but also the traffic flow state in a way that more accurately accounts for the occurrence of opportunities for same-direction collisions. Generalized linear modeling is applied on same-direction crash sets to estimate prediction models. The two distinctive same-direction crash sets are (1) rear-end collisions and (2) same-direction sideswipe and turning collisions, as categorized in Chapter 3. Through the model results, this study also shows that the new definition of same-direction crash exposure works better for collisions caused by the same contributing factors. In other words, for same-direction collisions resulting from similar occurrence mechanisms, this new exposure better accounts for traffic flow effects on crash risk because it is found to be linearly related to crash count.

4.3 Methodologies

4.3.1 Same-direction Crash Opportunities: Vehicle-Time-Spent-Following

Percent-time-spent-following (PTSF) was selected for counting the opportunities for vehicles to experience same-direction crashes on roadway segments. According to the Highway Capacity Manual (TRB 2005), "PTSF represents the freedom to maneuver and the comfort and convenience of travel." It is defined as the average percentage of travel time vehicles spend in platoons behind slower vehicles due to the inability to pass. It is difficult to measure PTSF in the field on a specific roadway segment, but the HCM offers a procedure to estimate this value. The idea of using PTSF in estimating same-direction crash opportunities relies on the concept that the risk of having same-direction collisions would increase when a higher percentage of vehicles are observed following other vehicles closely in the same direction of traffic. In Chapter 2, opportunities were computed for opposite-direction collisions by time of day, as early morning (2:00 am to 5:59 am), morning peak (6:00 am to 9:59 am), midday (10:00 am to 1:59 pm), afternoon peak (2:00 pm to 5:59 pm), evening (6:00 pm to 9:59 pm), and overnight (10:00 pm to 1:59 am). This time of day framework is necessary because on a daily basis the directional distribution tends to be aggregated to 50/50 for most roads. Based on the same time of day framework, the proposed equation for same-direction crash opportunities is

$$VTSF_{it} = \sum_{h \in H_t} \frac{V_{ih} L_i}{\overline{u}_{ih}} \times PTSF_{ih} \tag{4.1}$$

where $VTSF_{it}$ is the vehicle time spent following on segment i in time period t, H_t is the set of hours during time of day t, V_{ih} is the flow rate on segment i in hour h, L_i is the length of segment i, $\overline{u}_{ih}$ is the average traffic speed on segment i in hour h, and $PTSF_{ih}$ is the computed percent-time-spent-following on segment i in hour h (given the hourly traffic flow and speed condition on the segment, also taking into consideration the distribution of passing and no-passing zones). According to the HCM procedure, one *PTSF* value is computed for each two-lane highway segment by hour. Due to the difficulty of gathering average traffic speed along the study segments, $\overline{u}_{ih}$ is represented by the posted speed limit. Although the number of access points (driveways and minor intersections) might affect average vehicle headways and speeds, it is difficult to include this within the calculation of $VTSF_{it}$. Therefore, the number of access points enters the prediction models as a covariate to explain the crash risk, instead of contributing to the exposure measurement.

The HCM procedure for two-lane highways is used to compute *PTSF* in equation (4.1) (TRB 2000). We used the HCM procedure because of its widespread use for analyzing the operation of two-lane roads and its ease of calculation. According to *HCM*,

$$PTSF = BPTSF + f_{d/np} \tag{4.2}$$

where *BPTSF* is the base percent-time-spent-following, and $f_{d/np}$ is the adjustment factor accounting for the combined effect on *PTSF* of the directional distribution of traffic flow and the percentage of no-passing zones. For the segments used for this study, both the directional hourly traffic volumes and indication for passing zone information are known, so $f_{d/np}$ can be determined via *HCM* Exhibit 20-12 as shown in Table 4.1. The HCM also provides the equation for *BPTSF as*

$$\boldsymbol{BPTSF} = 100[1 - \exp(-0.00087 v_p)] \tag{4.3}$$

where v_p is the two-way demand flow rate.

Ninety-five one-km long state-maintained rural two-lane highway segments in Connecticut were randomly selected for analysis (same data set as used in Chapters 2 and 3). During this segment selection, intersections with signal or stop control on the major approaches and higher intensity of land development and driveway access, such as in a thickly settled village area, were avoided. The geometric information for each segment was collected using ConnDOT Photolog software, which provides images of state-maintained highways every 0.01 kilometer in both directions. The photolog was used to acquire roadway width (including lane width and shoulder width), posted speed limit, centerline information and number of driveways and minor intersections along the selected segments. Directional hourly traffic flow is needed for computing the same-direction crash opportunities. For each segment, the Traffic Management System for ADT (TMSADT) program developed by ConnDOT provides directional hourly traffic counts measured by loop detector in the field at a regular weekday (non-holiday) for a single 24-hour period. After applying the adjustment factors computed by ConnDOT to account for monthly and day-of-week effects on the 24-hour directional traffic counts, we obtained the estimated yearly average hourly traffic volumes for all 95 segments. $VTSF_{it}$ was then calculated using equation (4.1) based on the gathered segment data for each segment. Another exposure measurement, vehicle kilometers traveled in period (*VKTP*), which is equivalent to VKT but for just the specified time of day, was computed and used instead of VKT in the prediction models to be compared with the models using opportunities. Accident records were extracted from the Connecticut Collision Analysis System compiled by ConnDOT from police reports.

4.3.2 Statistical Methodologies

This section describes the fitting of Poisson and negative binomial Generalized Linear Models (GLIMs) for predicting crash counts including an explanation of choosing the best model using two scale criteria deviance and Pearson chi-squared statistic. Diagnostics checks on cumulative residuals are introduced to evaluate and compare the validity of the format of exposure measurement.

GLIM has been widely used in traffic crash prediction research since it allows an assumption of a Poisson or a negative binomial distribution to model the dependent variable of crash counts. Many studies have applied such GLIM's to crash data, but few have extensively investigated model choice, viz., the difference in model goodness of fit for different distributional assumptions, especially the scaled Poisson distribution (used in the context of over-dispersion). Here it is elaborated how the difference in distribution assumption affects the model outcome.

Both Poisson and negative binomial distributions belong to the exponential family of distributions (McGullagh and Nelder 1989), which takes the form

$$f_Y(y;\theta,\varphi) = \exp\{(y\theta - b(\theta))/a(\varphi) + c(y,\varphi)\} \tag{4.4}$$

where y is the dependent variable, θ is the parameter or vector of parameters of interest, and $a(\varphi)$ accounts for the dispersion in the distribution. The log-likelihood function of the exponential distribution can be written as $l(\theta,\varphi;y) = \log f_Y(y;\theta,\varphi)$, and has the following properties:

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0, \tag{4.5}$$

Table 4.1 HCM Table for Adjustment Factor $f_{d/np}$ (TRB 2005)

Two-Way Flow Rate, v_p (pc/h)	Increase in Percent Time-Spent-Following (%)					
	No-Passing Zones (%)					
	0	20	40	60	80	100
Directional Split = 50/50						
≤ 200	0.0	10.1	17.2	20.2	21.0	21.8
400	0.0	12.4	19.0	22.7	23.8	24.8
600	0.0	11.2	16.0	18.7	19.7	20.5
800	0.0	9.0	12.3	14.1	14.5	15.4
1400	0.0	3.6	5.5	6.7	7.3	7.9
2000	0.0	1.8	2.9	3.7	4.1	4.4
2600	0.0	1.1	1.6	2.0	2.3	2.4
3200	0.0	0.7	0.9	1.1	1.2	1.4
Directional Split = 60/40						
≤ 200	1.6	11.8	17.2	22.5	23.1	23.7
400	0.5	11.7	16.2	20.7	21.5	22.2
600	0.0	11.5	15.2	18.9	19.8	20.7
800	0.0	7.6	10.3	13.0	13.7	14.4
1400	0.0	3.7	5.4	7.1	7.6	8.1
2000	0.0	2.3	3.4	3.6	4.0	4.3
≥ 2600	0.0	0.9	1.4	1.9	2.1	2.2
Directional Split = 70/30						
≤ 200	2.8	13.4	19.1	24.8	25.2	25.5
400	1.1	12.5	17.3	22.0	22.6	23.2
600	0.0	11.6	15.4	19.1	20.0	20.9
800	0.0	7.7	10.5	13.3	14.0	14.6
1400	0.0	3.8	5.6	7.4	7.9	8.3
≥ 2000	0.0	1.4	4.9	3.5	3.9	4.2
Directional Split = 80/20						
≤ 200	5.1	17.5	24.3	31.0	31.3	31.6
400	2.5	15.8	21.5	27.1	27.6	28.0
600	0.0	14.0	18.6	23.2	23.9	24.5
800	0.0	9.3	12.7	16.0	16.5	17.0
1400	0.0	4.6	6.7	8.7	9.1	9.5
≥ 2000	0.0	2.4	3.4	4.5	4.7	4.9
Directional Split = 90/10						
≤ 200	5.6	21.6	29.4	37.2	37.4	37.6
400	2.4	19.0	25.6	32.2	32.5	32.8
600	0.0	16.3	21.8	27.2	27.6	28.0
800	0.0	10.9	14.8	18.6	19.0	19.4
≥1400	0.0	5.5	7.8	10.0	10.4	10.7

$$\boldsymbol{E}\left(\frac{\partial^2 \boldsymbol{l}}{\partial^2 \theta}\right)+\boldsymbol{E}\left(\frac{\partial \boldsymbol{l}}{\partial \theta}\right)^2=0\,, \tag{4.6}$$

from which the mean and variance of *y* are derived to be $\boldsymbol{E(y)}=\mu=\boldsymbol{b'(\theta)}$ (where μ is the expectation of y, and $\boldsymbol{b'(\theta)}$ is the first derivative of $\boldsymbol{b(\theta)}$), and $\boldsymbol{Var(y)}=\boldsymbol{b''(\theta)a(\varphi)}$ (where $\boldsymbol{a(\varphi)}$ is a function of the known parameter φ for exponential family model with canonical parameter θ, and $\boldsymbol{b''(\theta)}$ is the second derivative of $\boldsymbol{b(\theta)}$).

Transformed into the exponential distribution family format, the Poisson probability distribution (pmf)

$$\boldsymbol{f(y)}=\frac{\boldsymbol{e}^{-\mu}\mu^{\boldsymbol{y}}}{\boldsymbol{y!}} \tag{4.7}$$

becomes

$$\boldsymbol{f(y)}=\boldsymbol{\exp\{y\log(\mu)-\mu-\log(y!)\}} \tag{4.8}$$

and has

$$\boldsymbol{b(\theta)}\text{ as }\boldsymbol{\exp(\theta)}\text{ since }\theta=\boldsymbol{\log(\mu)} \tag{4.9}.$$

This leads to the finding that both the first and second derivative of $\boldsymbol{b(\theta)}$ equal μ, demonstrating the Poisson distribution's equal mean and variance. Similarly, the negative binomial distribution

$$\boldsymbol{f(y)}=\frac{\Gamma(\boldsymbol{y}+1/\boldsymbol{k})}{\Gamma(\boldsymbol{y}+1)\Gamma(1/\boldsymbol{k})}\frac{(\boldsymbol{k}\mu)^{\boldsymbol{y}}}{(1+\boldsymbol{k}\mu)^{\boldsymbol{y}+1/\boldsymbol{k}}} \tag{4.10}$$

can be written in exponential form as

$$\boldsymbol{f(y)}=\boldsymbol{y\log(k\mu)-(y+1/k)\log(1+k\mu)+\log}\left[\frac{\Gamma(\boldsymbol{y}+1/\boldsymbol{k})}{\Gamma(\boldsymbol{y}+1)\Gamma(1/\boldsymbol{k})}\right] \tag{4.11}$$

Since not one, but two parameters, μ and *k*, need to be estimated by maximizing the log-likelihood, $\boldsymbol{b(\theta)}$ takes the form (Ulsson 2002)

$$\boldsymbol{b(\theta)}=1/\boldsymbol{k}[\theta-\boldsymbol{\log}(1-\boldsymbol{e}^{\theta})] \tag{4.12}$$

where $\boldsymbol{\theta}$ on the left-hand side is a vector of parameters. Although $\boldsymbol{a(\varphi)}$ is still 1 in this case, the additional parameter *k* is able to take care of the over-dispersion in the data as shown in the variance function

$$\boldsymbol{Var(y)}=\mu+\boldsymbol{k}\mu^2 \tag{4.13}$$

Numerical methods are usually implemented to maximize the log-likelihood in estimating the parameters. The most popular methods are Newton-Raphson method and Fisher scoring method (Ulsson 2002). The Newton-Raphson method is adopted in the SAS Genmod procedure, a popular choice for crash prediction. If the first derivative of $\boldsymbol{l(\theta,\varphi;y)}$ with respect to θ is defined to be *g(θ)* and the second derivative *H(θ)*, usually called the Hessian matrix, the Newton-Raphson method works by a Taylor series expansion of *g(θ)* around $\hat{\theta}$

$$\boldsymbol{g(\hat{\theta})}=\boldsymbol{g(\theta_0)}+\boldsymbol{(\hat{\theta}-\theta_0)H(\theta_0)} \tag{4.14}$$

where θ_0 is an initial estimate of theta and $\hat{\theta}$ is the estimate of θ. Setting $\boldsymbol{g(\hat{\theta})}=0$, $\hat{\theta}$ can be solved as

$$\hat{\theta}=\theta_0-\boldsymbol{g(\theta_0)H^{-1}(\theta_0)} \tag{4.15}.$$

Substituting $\hat{\theta}$ for θ_0 and repeating this step until the value of $\hat{\theta}$ reaches a stable level when the algorithm converges, the maximum likelihood estimate of θ is obtained.

When there is lack of nominal dispersion in the data, that is, the variance of the Poisson distribution appears to be greater than its mean, or the mean of the dependent variable is less than 1, there are two ways to address this issue. One is to use the negative binomial distribution (Lord 2006), which has a variance function allowing for heterogeneity; the other one is to adjust the Poisson distribution variance function by multiplying by a scale parameter φ, which can be roughly estimated by the deviance divided by

degree of freedom (*d. f.*) (Scaled deviance) or Pearson chi-squared statistic divided by *d. f.* (Pearson scale). In models that do not explicitly incorporate any scale parameter, an estimated deviance or Pearson scale larger than 1.0 in the output indicates the data may be over-dispersed. For the Poisson distribution, this scale parameter is used to adjust the variance to be $\boldsymbol{Var(y) = \varphi\mu}$ where φ is the scale parameter and μ is the Poisson mean. The function obtained by dividing the maximum log-likelihood by the scale parameter is an example of the quasi maximum log-likelihood function and in the case for the Poisson distribution, standard errors of covariate parameters and likelihood ratio statistics are computed based on this quasi log-likelihood function (SAS 1999).

Residual analysis has long been used as a method to assess the adequacy of regression models. However, conventional residuals computed as the absolute difference between the actual observation and the predicted value do not suit GLIM. For example, in crash prediction models the dependent variable, crash count, is assumed to follow a Poisson distribution, but the predicted value for the dependent variable is the estimated Poisson mean for each observation, which is usually a real number. As a result, residuals measuring the disparity between the prediction model and observations should be computed as the difference between the actual crash count and a count value that follows the Poisson distribution with the estimated mean at a certain probability level. Such count values and probability levels are difficult to determine; therefore, conventional residuals analysis is usually not considered in this type of study, although theoretically it can be conducted.

Lin *et al.* (2002) proposed a model checking framework based on cumulative residuals that still can make valid use of the conventional residuals. For GLIM, the residuals are defined as

$$\boldsymbol{e_i = Y_i - v(\hat{\beta}' X_i), (i = 1,...,n)} \tag{4.16}$$

where Y_i is the i^{th} observed value of the dependent variable and $\boldsymbol{v(\hat{\beta}' X_i)}$ is the predicted dependent variable value. Then the cumulative sum of $\boldsymbol{e_i}$ over X_{ji} is defined as

$$\boldsymbol{W_j(x) = n^{-1/2} \sum_{i=1}^{n} I(X_{ji} \le x) e_i} \tag{4.17}$$

where X_{ji} is the jth component of covariate vector X_i, n is the sample size, and $I(.)$ is the indication function, which generates a value of 1 when the specified situation is satisfied and 0 when the situation is not. The primary objective of this cumulative residual analysis is to graphically examine model assumptions, such as the functional form of covariates. Lin *et al.* also showed that the null distribution of $\boldsymbol{W_j(x)}$, also known as a distribution of random error to be compared with the actual distribution of $\boldsymbol{W_j(x)}$, can be approximated by simulating the corresponding zero-mean Gaussian process whose realizations can be easily generated by computer process. So if the P-value of this null distribution test for $\boldsymbol{W_j(x)}$ is greater than a set threshold value, we can accept the functional form of the covariate as appropriate. In this study, cumulative residual analysis is used for assessing the functional form of the exposure measurements, VKTP and VTSF, and evaluating how well they fit in the models by comparing the null distribution test P-values that are generated through the simulation process.

4.4 Results and Discussion

This section presents the model results using vehicle kilometers traveled in period (*VKTP*) and Vehicle Time Spent Following (*VTSF*) as crash exposure, respectively, for same-direction crashes. The disparity between different Poisson and negative binomial models regarding goodness of fit and parameter estimates is discussed, and further, the application of the two types of scale estimates in adjusting the variance in Poisson distribution is evaluated. In an attempt to acquire best performance models, full models were estimated first, each including the exposure measurement as well as all other roadway geometric variables, such as roadway width, speed limit and access point density. Models with only the significant covariates are then used for comparing VTSF and VTKP. An assessment of comparison between VKTP and VTSF as exposure measurements using graphical illustration of actual crash count plotted against VKTP or VTSF and cumulative residual analysis is also conducted and described. In the end, the meaning of the exposure parameter estimates is elaborated upon, relating to the functionality of exposure itself.

4.4.1 Model Selection

Crash prediction models are estimated for three different sets of same-direction crashes: (1) all three types of same direction collisions; (2) rear-end collisions only; and (3) sideswipe and turning same direction collisions. In modeling rear-end collisions separately from the other two, we intend to show the effects on prediction models of the newly defined collision type categories according to contributing factors, as discussed in the previous chapter. Three different roadway design and land use variables are considered in this study: directional roadway width, defined to be narrow (3, 3.5 and 4 m), medium (4.3 and 4.5 m), and wide (>4.5 m), speed limit, categorized as low speed limit (< 72 km/hr or 45 mi/hr) and high speed limit (>= 72 km/hr or 45 mi/hr), and the number of driveways or minor intersections along the segment. These categories were defined so as to roughly equalize the number of observations in each category, and to account for differences in operation from one category to another. Time period is also included as a covariate in the models aiming to assess the possible difference in driver behavior or other effects at different times of day. For each crash set, the model with the highest log-likelihood is selected as best. AIC (defined as negative of twice the maximized log likelihood plus twice the number of covariates) is not used in this step, because the number of covariates is the same in all the models.

In total, twenty four full models are estimated for the three crash data sets, assuming four different crash count distributions, and using two distinctly defined exposure factors. The four distribution assumptions are (1) unscaled (nominal) Poisson distribution, (2) deviance-scaled-Poisson distribution, (3) Pearson-scaled-Poisson distribution, and (4) the negative binomial distribution. Tables 4.2 to 4.4 show the estimated parameter coefficients and fit diagnostics of all the full models for the three collision types, respectively. In the full models, all the aforementioned variables are included. Not all are found to be significantly correlated with crash risk.

According to the maximized log likelihood values, the total same-direction crashes seem to fit better with the Pearson-scaled-Poisson distribution. For VKTP models, after adjusting for over-dispersion in the data by a Pearson scale parameter of 1.104, the maximized log likelihood reaches -184.653, which is much higher than the other models. Similarly for VTSF models, with a Pearson scale parameter of 1.105 a log likelihood of -188.328 is achieved. Rear-end crashes fit better with the same Poisson distribution assumption. For VKTP models, the Pearson-scaled Poisson model has a maximized log likelihood value of 181.571, and for VTSF models, the maximized log likelihood value is -185.004. Negative binomial models are found unable to converge for same-direction sideswipe and turning crashes, which could be a result of the log likelihood value being unable to reach optimality through the iterative numerical process. Both Pearson and deviance scale parameters are less than 1.0, indicating under-dispersion in the data. This lack of over-dispersion might be due to the relatively rare occurrence of turning and sideswipe collisions, as the largest number of turning or sideswipe collisions on any segment did not exceed 2.0. In comparison, the number of rear-end collisions per segment varies from 0 to 15. So we choose non-scaled Poisson for the models.

Table 4.5 shows the parameters and statistics for the reduced models that contain only the significant covariates, identified in the full model results, and use VKTP and VTSF as exposure, respectively. For the total same-direction collisions, the Pearson-scaled Poisson distribution models have the highest maximized log-likelihood values, -180.020 for VKTP as exposure and -194.236 for VTSF as exposure. The significant covariates are VKTP or VTSF, roadway width categories, and number of access points. The same distribution approach is also best for rear-end crashes, for the model including VKTP the log likelihood is -175.881 and for the model including VTSF the maximized log likelihood is -186.543. The significant variables are the same too. For same-direction sideswipe and turning crashes, non-scaled Poisson distribution models have the highest maximized log-likelihood values, and only access point density is significant in addition to VKTP or VTSF in the models.

In these models, the maximized log likelihood of VKTP models is slightly higher than for the corresponding VTSF models. This may be caused by potential inaccuracy in the estimated VTSF values, which are derived from PTSF, which is quite sensitive to the accuracy of the traffic flow data. For example, a small difference in hourly traffic flow or directional traffic split might result in a 2 percent to 6 percent difference in $f_{d/np}$, and hence more variation in PTSF (VTSF). Our aggregated data do not reflect traffic variation from day to day and therefore might cause the inaccuracy in the aggregated VTSF to have a relatively large impact on model result. Perhaps more problematic is our need to substitute speed limit for the actual average speed, and the attendant effect on the estimation of VTSF.

Table 4.2 Full Model Results for Total Same-direction Crashes

	VKTP				**VTSF**			
Distribution	Non-scaled Poisson	**Pearson-scaled Poisson**	Deviance-scaled Poisson	NB	Non-scaled Poisson	**Pearson-scaled Poisson**	Deviance-scaled Poisson	NB
Maximized Log-likelihood	-225.191	**-184.653**	-220.534	-206.738	-229.866	**-188.328**	-221.484	-208.844
Scale/ Dispersion	1.0	**1.104**	1.010	0.393	1.0	**1.105**	1.109	0.426
VKTP	1.416** (0.095)	**1.416** (0.105)**	1.416** (0.090)	1.378** (0.115)	---	---	---	---
VTSF	---	---	---	---	0.906** (0.063)	**0.906** (0.069)**	0.906** (0.064)	0.887** (0.076)
Intercept	-21.595** (1.426)	**-21.595 (1.575)**	-21.595 (1.441)	-21.000** (1.693)	-13.529** (0.911)	**-13.529** (1.006)**	-13.529** (0.928)	-13.256** (1.079)
2 – 6 AM	-1.127 (0.631)	**-1.127 (0.697)**	-1.127 (0.638)	-1.196 (0.649)	-0.214 (0.634)	**-0.214 (0.700)**	-0.214 (0.646)	-0.266 (0.655)
6 – 10 AM	-0.426 (0.300)	**-0.426 (0.328)**	-0.426 (0.300)	-0.486 (0.343)	0.358 (0.283)	**0.358 (0.313)**	0.358 (0.288)	0.298 (0.318)
10 AM – 2 PM	-0.157 (0.290)	**-0.157 (0.321)**	-0.157 (0.293)	-0.191 (0.332)	0.500 (0.281)	**0.500 (0.311)**	0.500 (0.286)	0.483 (0.313)
2 – 6 PM	-0.302 (0.300)	**-0.302 (0.332)**	-0.302 (0.304)	-0.332 (0.348)	0.517 (0.284)	**0.517 (0.314)**	0.517 (0.290)	0.502 (0.318)
6 – 10 PM	-0.468 (0.300)	**-0.468 (0.328)**	-0.468 (0.300)	-0.462 (0.335)	-0.313 (0.300)	**-0.313 (0.330)**	-0.313 (0.304)	-0.297 (0.332)
Pavement Width < 13 ft	0.137 (0.122)	**0.137 (0.134)**	0.137 (0.123)	0.166 (0.154)	-0.004 (0.119)	**-0.004 (0.132)**	-0.004 (0.121)	0.004 (0.154)
Pavement Width > 15 ft	-0.549** (0.130)	**-0.549** (0.144)**	-0.549** (0.131)	-0.529** (0.164)	-0.568** (0.129)	**-0.568** (0.143)**	-0.568** (0.132)	-0.545** (0.166)
Posted Speed < 45 mph	0.110 (0.124)	**0.110 (0.136)**	0.110 (0.125)	0.210 (0.160)	-0.279** (0.116)	**-0.279** (0.128)**	-0.279** (0.118)	-0.159 (0.151)
Number of Access Points	0.050** (0.008)	**0.050** (0.009)**	0.050** (0.008)	0.046** (0.011)	0.055** (0.008)	**0.055** (0.009)**	0.055** (0.009)	0.052** (0.011)

Number in parentheses is the standard error of the estimated value. ** Significant at 95 percent confidence level

Table 4.3 Full Model Results for Rear-end Only Crashes

	VKTP				**VTSF**			
Distribution	Non-scaled Poisson	**Pearson-scaled Poisson**	Deviance-scaled Poisson	NB	Non-scaled Poisson	**Pearson-scaled Poisson**	Deviance-scaled Poisson	NB
Maximized Log-likelihood	-219.383	**-181.571**	-231.86	-200.765	-225.728	**-185.004**	-232.976	-203.907
Scale/Dispersion	1.0	**1.099**	0.973	0.463	1.0	**1.105**	0.984	0.509
VKTP	1.584** (0.108)	**1.584** (0.119)**	1.584** (0.105)	1.544** (0.133)	---	---	---	---
VTSF	---	---	---	---	1.003** (0.071)	**1.003** (0.079)**	1.003** (0.07)	0.985** (0.088)
Intercept	-24.427** (1.645)	**-24.427** (1.808)**	-24.427** (1.60)	-23.785** (1.972)	-15.233** (1.050)	**-15.233** (1.159)**	-15.233** (1.033)	-14.947** (1.257)
2 – 6 AM	-0.674 (0.654)	**-0.674 (0.719)**	-0.674 (0.636)	-0.740 (0.679)	0.277 (0.657)	**0.277 (0.725)**	0.277 (0.646)	0.240 (0.686)
6 – 10 AM	-0.445 (0.348)	**-0.445 (0.382)**	-0.445 (0.338)	-0.519 (0.401)	0.417 (0.335)	**0.417 (0.370)**	0.417 (0.330)	0.348 (0.373)
10 AM – 2 PM	-0.137 (0.341)	**-0.137 (0.375)**	-0.137 (0.332)	-0.151 (0.388)	0.580 (0.333)	**0.580 (0.368)**	0.580 (0.328)	0.589 (0.367)
2 – 6 PM	-0.304 (0.351)	**-0.304 (0.386)**	-0.304 (0.342)	-0.332 (0.404)	0.598 (0.336)	**0.598 (0.368)**	0.598 (0.331)	0.593 (0.372)
6 – 10 PM	-0.488 (0.350)	**-0.488 (0.385)**	-0.488 (0.341)	-0.498 (0.394)	-0.326 (0.355)	**-0.326 (0.390)**	-0.326 (0.347)	-0.324 (0.390)
Pavement Width < 13 ft	0.182 (0.134)	**0.182 (0.147)**	0.182 (0.130)	0.213 (0.172)	0.024 (0.131)	**0.024 (0.144)**	0.024 (0.129)	0.022 (0.171)
Pavement Width > 15 ft	-0.550** (0.142)	**-0.550** (0.156)**	-0.550** (0.138)	-0.505** (0.182)	-0.577** (0.141)	**-0.577** (0.156)**	-0.577** (0.139)	-0.531** (0.185)
Posted Speed < 45 mph	0.142 (0.137)	**0.142 (0.150)**	0.142 (0.133)	0.252 (0.180)	-0.294** (0.128)	**-0.294** (0.141)**	-0.294** (0.126)	-0.163 (0.169)
Number of Access Points	0.052** (0.009)	**0.052** (0.010)**	0.052** (0.009)	0.048** (0.012)	0.059** (0.009)	**0.059** (0.010)**	0.059** (0.009)	0.054** (0.012)

Number in parentheses is the standard error of the estimated value. ** Significant at 95 percent confidence level

Table 4.4 Full Model Results for Same-direction Sideswipe and Turning Crashes

	VKTP				VTSF			
Distribution	**Non-scaled Poisson**	Pearson-scaled Poisson	Deviance-scaled Poisson	Negative Binomial	**Non-scaled Poisson**	Pearson-scaled Poisson	Deviance-scaled Poisson	Negative Binomial
Maximized Log-likelihood	**-186.900**	-213.546	-403.344	Algorithm unable to converge	**-186.547**	-216.502	-403.682	Algorithm unable to converge
Scale/Dispersion	**1.0**	0.936	0.681		**1.0**	0.928	0.680	
VKTP	**0.700** (0.203)**	0.700** (0.190)	0.700** (0.138)		**---**	---	---	
VTSF	**---**	---	---		**0.473** (0.134)**	0.473** (0.125)	0.473** (0.091)	
Intercept	**-12.676** (2.887)**	-12.676** (2.701)	-12.676** (1.965)		**-9.066** (1.832)**	-9.066** (1.701)	-9.066** (1.246)	
2 – 6 AM	**-23.442 (>10,000)**	-23.442 (>10,000)	-23.442 (>10,000)		**-22.820 (>10,000)**	-22.820 (>10,000)	-22.820 (>10,000)	
6 – 10 AM	**-0.014 (0.609)**	-0.014 (0.570)	-0.014 (0.415)		**0.374 (0.556)**	0.374 (0.516)	0.374 (0.378)	
10 AM – 2 PM	**0.069 (0.591)**	0.069 (0.553)	0.069 (0.402)		**0.403 (0.551)**	0.403 (0.511)	0.403 (0.374)	
2 – 6 PM	**-0.062 (0.630)**	-0.062 (0.589)	-0.062 (0.429)		**0.339 (0.568)**	0.339 (0.528)	0.339 (0.386)	
6 – 10 PM	**-0.055 (0.587)**	-0.055 (0.549)	-0.055 (0.400)		**0.017 (0.579)**	0.017 (0.537)	0.017 (0.393)	
Pavement Width < 13 ft	**0.036 (0.298)**	0.036 (0.278)	0.036 (0.203)		**-0.041 (0.296)**	-0.041 (0.275)	-0.041 (0.201)	
Pavement Width > 15 ft	**-0.438 (0.328)**	-0.438 (0.307)	-0.438** (0.224)		**-0.439 (0.326)**	-0.439 (0.303)	-0.439 (0.222)	
Posted Speed < 45 mph	**0.044 (0.292)**	0.044 (0.273)	0.044 (0.199)		**-0.143 (0.279)**	-0.143 (0.259)	-0.143 (0.190)	
Number of Access Points	**0.041* (0.021)**	0.041 (0.020)	0.041****** (0.014)		**0.045** (0.021)**	0.045** (0.020)	0.045** (0.014)	

Number in parentheses is the standard error of the estimated value. * Significant at 90 percent confidence level; ** Significant at 95 percent confidence level

Table 4.5 Reduced Model Results for the Three Crash Data Sets

	VKTP			VTSF		
	All Same-direction	Rear-end	Same-direction sideswipe and turning	All Same-direction	Rear-end	Same-direction sideswipe and turning
Best Distribution Assumption	Scaled Poisson (Pearson)	Scaled Poisson (Pearson)	Non-scaled Poisson	Scaled Poisson (Pearson)	Scaled Poisson (Deviance)	Non-scaled Poisson
Scale/Dispersion	1.131	1.127	-404.460	1.138	1.151	-190.430
Maximized Log likelihood	-180.020	-175.881	0.689	-194.236	-186.543	1.0
Intercept	-21.684** (1.285)	-24.123** (1.478)	-14.301** (1.477)	-14.668** (0.958)	-16.327** (1.121)	-10.801** (1.570)
VKTP	1.402** (0.080)	1.546** (0.092)	0.793** (0.096)	---	---	---
VKTP 95% CI Limits	1.245, 1.560	1.366, 1.726	0.604, 0.982	---	---	---
VTSF	---	---	---	0.998** (0.062)	1.097** (0.072)	0.596** (0.108)
VTSF 95% CI Limits	---	---	---	0.876, 1.120	0.956, 1.239	0.385, 0.807
Narrow vs. Medium Width	0.147 (0.135)	0.182 (0.148)	---	-0.008 (0.135)	0.023 (0.150)	---
Wide vs. Medium Width	-0.573** (0.140)	-0.583** (0.152)	---	-0.456** (0.141)	-0.459** (0.156)	---
No. of Access Points	0.050** (0.140)	0.053** (0.010)	0.046** (0.021)	0.054** (0.010)	0.057** (0.011)	0.049** (0.021)

Number in parentheses is the standard error of the estimated value.
** Significant at 95 percent confidence level

The Pearson scale parameter being greater than 1.0 suggests over-dispersion in the crash data under the Poisson distribution assumption. Ulsson (2002) pointed out that the main reason for over-dispersion is the lack of homogeneity in the data. This lack of homogeneity might exist between groups of cases, between cases and within cases (Ulsson 2002). In reality, such lack of homogeneity can also be caused by modeling using insufficient variable predictors. It is not an easy task to include sufficient variables in model prediction. For example, driver behavior-related variables are both difficult to collect and to apply, but they provide the most accurate explanation of the crash count. In Ulsson's argument, cases correspond to the roadway segments for which the crash prediction models are estimated. According to the categorical predictors used for modeling, the segments can be separated into several groups. Therefore, first, it is possible that the between-group variation in crashes might exceed what the Poisson distribution variance can account for; second, the segments which fall into the same group still might not share the exact same crash risk; and finally, among individual segments, there might still be factors, related to driver behavior and all other conditions that are not considered in the study, that help to explain the crash risk.

4.4.2 Comparison between VKTP and VTSF

We just showed that the models using VKTP have slightly better goodness of fit than those using VTSF. Furthermore, the significant roadway geometric characteristic variables and their parameter estimates (indicated by ** in Table 4.5) are not much different between the models using VKTP and VTSF. However, the parameter estimates for VKTP and VTSF vary widely. The VTSF parameters for both total same-direction and rear-end collisions are not significantly different from 1.0, and their 95 percent

confidence intervals are not very wide, all overlapping 1.0. This is a strong indication of a linear relationship between crash counts and VTSF, indicating that VTSF can be used in the denominator to calculate a normalized crash rate (crashes per VTSF) that would be constant for all levels of VTSF. Such a relationship is not indicated for VKTP, for which the estimated parameter value ranges from about 0.8 to over 1.5, with confidence intervals not including 1.0.

Plots of rear-end crashes versus the two different exposure measures are used to demonstrate the linear relationship between crash counts and VTSF as shown in Figures 4.1 through 4.6. Each pair of plots illustrates rear-end crash counts versus VKTP and VTSF, in units of vehicle-kilometers ($\times 10^6$) and vehicle-minutes-spent-following ($\times 10^6$), respectively, for a different combination of roadway width category and access point density level, such as "narrow width and low access-point density" or "narrow width and high access-point density". Access point counts less than or equal to 10 correspond to the low level of access point density, and counts greater than 10 correspond to the high level of access points density. The scattered dots in Figure 4.1a, especially for the 0 crash counts, seem to gather closer than in Figure 4.1b, and the dots at higher crash counts also show a less scattered trend. Very similarly, in Figures 4.2 through 4.6, the plotted observations for VTSF all more or less illustrate a stronger tendency centering on a straight line passing through the coordinate origin.

In Figures 4.7a and 4.7b, the dark lines are the cumulative residuals of the models using VKTP and VTSF, respectively. The P-value in the lower right corner tells the significance level of the null distribution test, which can be used to evaluate model goodness of fit given the functional form of the exposure measurement. As seen in the cumulative residual plots, the p-value for the model using VKTP is 0.088 and for the model using VTSF it is 0.1250. Both values show that the zero-mean Gaussian distribution assumption cannot be rejected at a 95 percent confidence level, therefore the exposure functional forms are proper. Moreover, the p-value for using VTSF in the model is apparently higher than that for using VKTP in the model, which implies a better fit using VTSF as exposure.

4.4.3 Implication of Parameter Estimates of VTSF

In finding VTSF to be a consistent exposure format we:

- represent traffic flow so as to reflect in prediction models a logical relationship between crashes and traffic flow conditions;
- reduce confounding issues caused by estimating parameters representing traffic flow and other variables correlated with traffic flow in the models simultaneously;
- solve the practical problem of defining a meaningful crash rate that is constant at all levels of the denominator; and
- further confirm the value of predicting crashes by collision type and suggesting an alternative categorization technique that begins to account for crash causality through reported contributing factors.

Estimating a parameter for VTSF that is not significantly different from 1.0 implies a constant crash risk for any value of this exposure measure when the predictor variables are held constant. Previous crash prediction models have found that crash count for different types of collisions follows different nonlinear relationships with traffic flow. For example, for multi-vehicle collisions, crash rate tends to be lower at low traffic flow and higher at high traffic flow. In other words, multi-vehicle crashes in the models have an increasing nonlinear relationship with traffic flow, which results in an exponent larger than 1.0 for the traffic flow variable. The opposite has typically been found for single-vehicle crashes, that is, traffic flow takes an exponent less than 1.0 in prediction models. Both of these cases indicate that the crash rate (crashes divided by exposure) is not constant with traffic flow as exposure. So crash rates cannot be compared from one road to another when their traffic volumes differ substantially. The nonlinear relationship between crash count and traffic flow also creates a problem in defining crash rate in words, as it is difficult to explain to the general public what "rear-end crash per traffic flow raised to the power of 1.55" means, for example.

Therefore, the linear relationship identified between same-direction crashes and VTSF defines VTSF to be a consistent exposure measurement that results in the crash rate being independent of the traffic volume. This makes it possible to more consistently identify road segment features that result in higher crash risk without confounding their effects with traffic volume. Table 4.6 shows the results of estimating models assuming the Pearson-scaled Poisson distribution and including VKTP or VTSF as an offset (the parameter estimate is set to be 1.0) in rear-end crash models. The covariate parameter estimates in VTSF

model are almost the same as those in Table 4.5. However, due to the confounding issue between traffic flow and other covariates, the corresponding estimates in the VKTP model are substantially different from the results in Table 4.5. Also, importantly, the VTSF model turns out to have a higher log likelihood than the VKTP model.

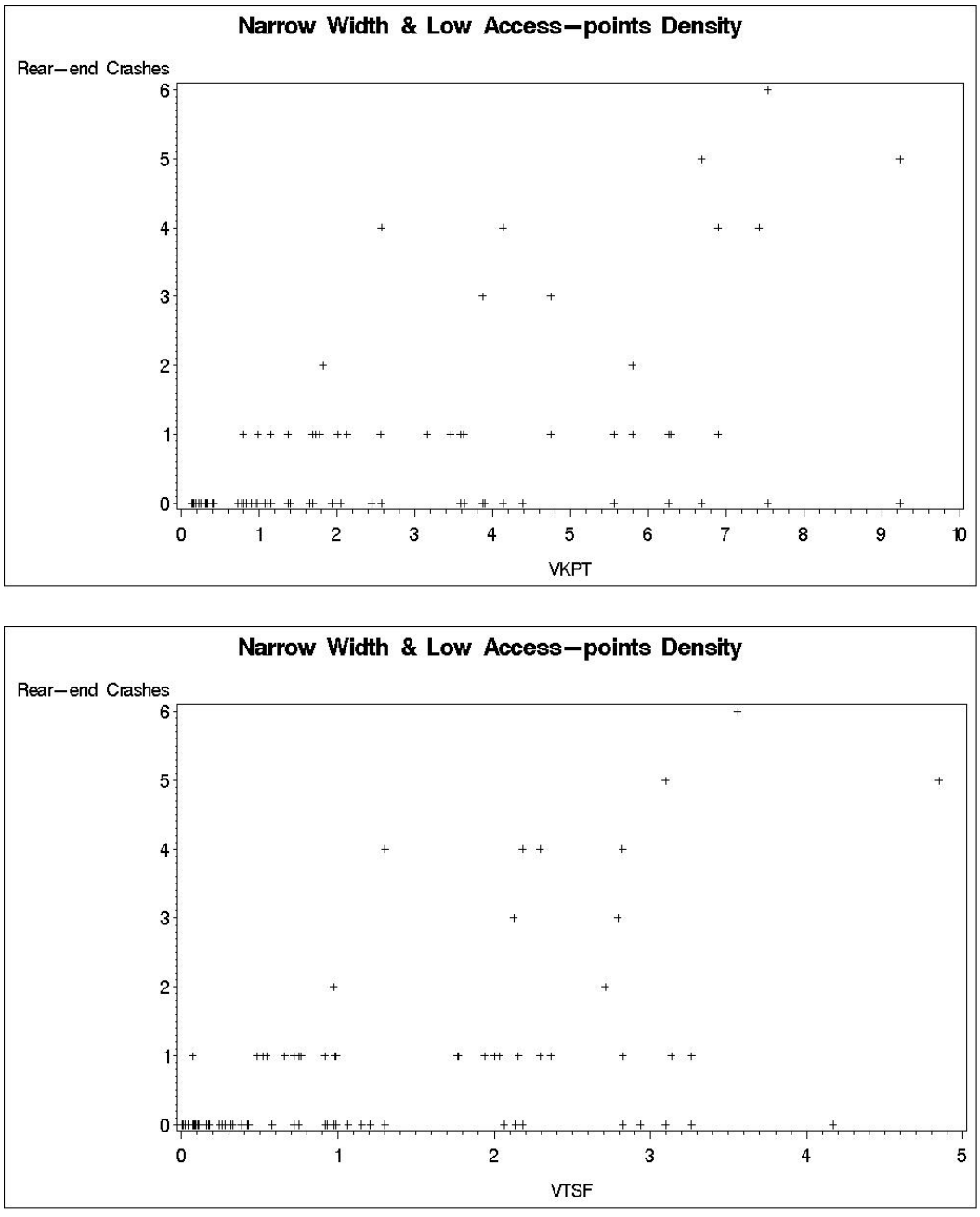
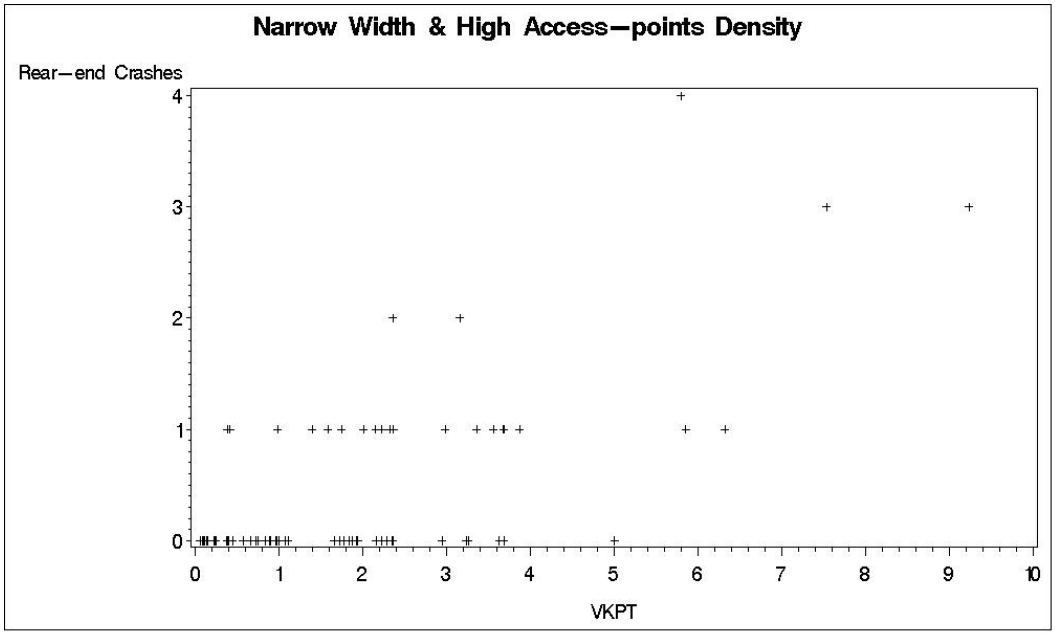


Figure 4.1 Rear End Collisions Versus VKTP and VTSF: Segments with Narrow Width and Low Access-Point Density



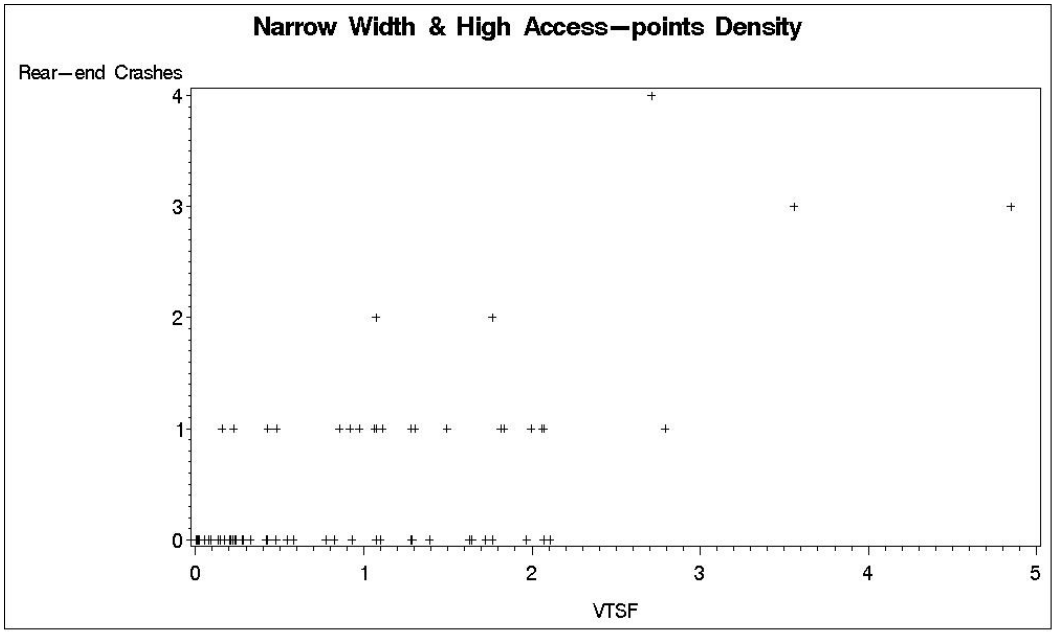
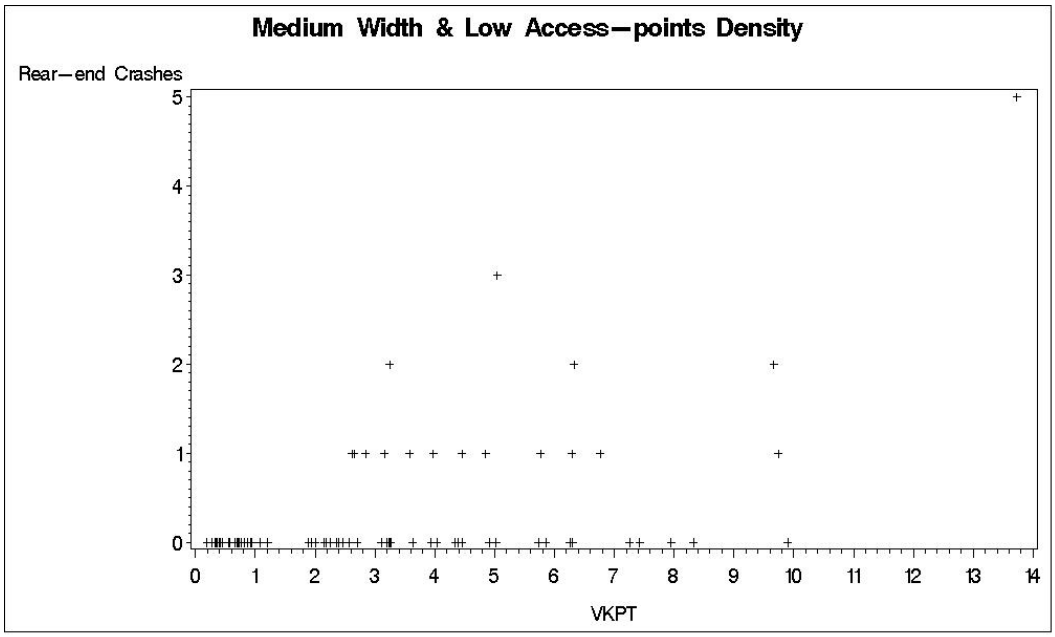


Figure 4.2 Rear End Collisions Versus VKTP and VTSF: Segments with Narrow Width and High Access-Point Density



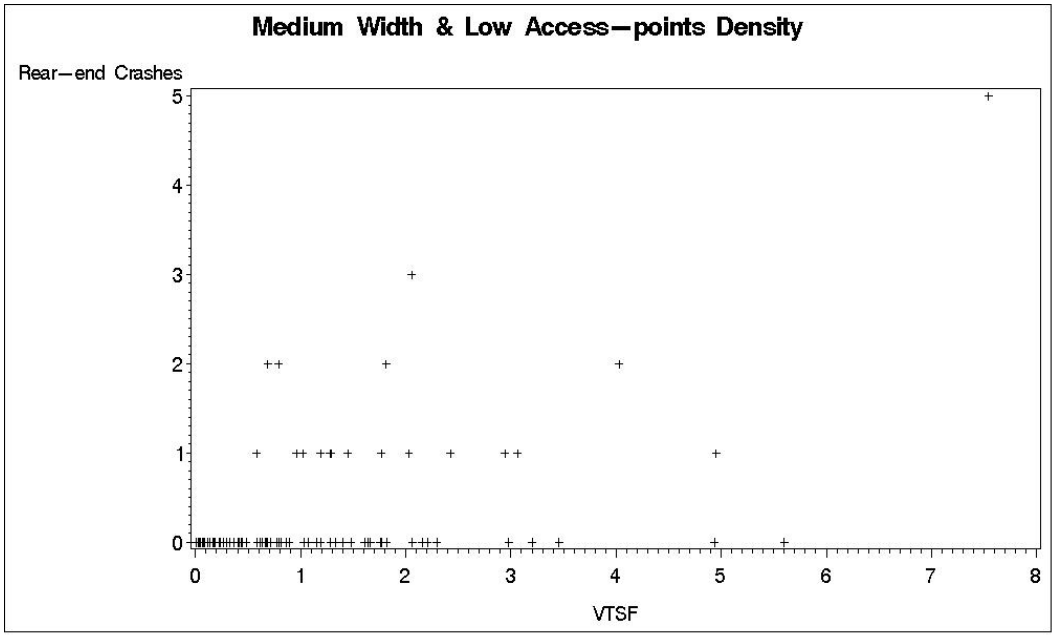
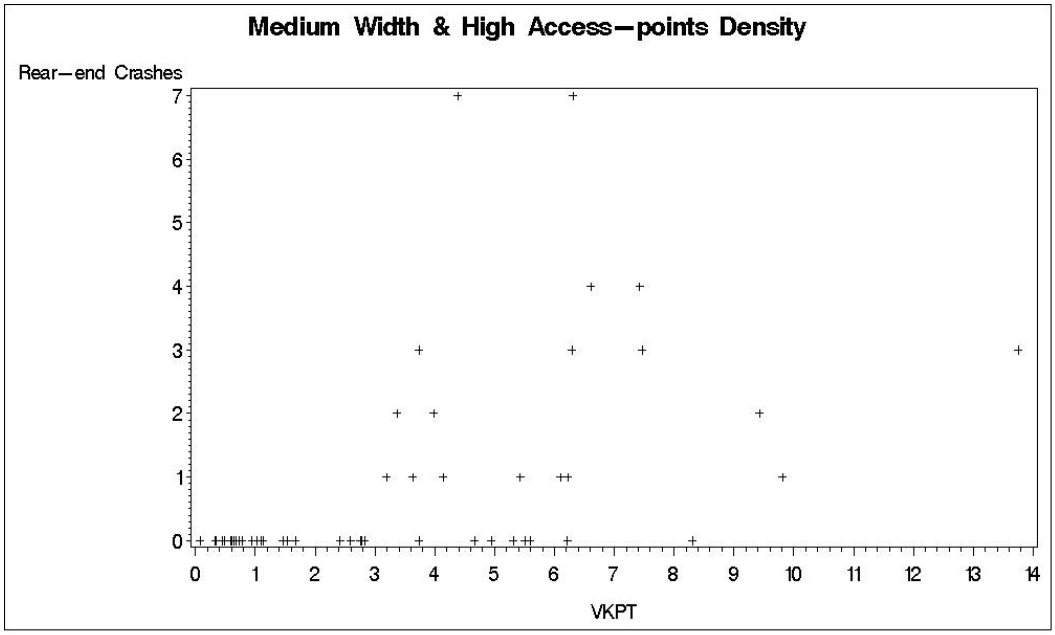


Figure 4.3 Rear End Collisions Versus VKTP and VTSF: Segments with Medium Width and Low Access-Point Density



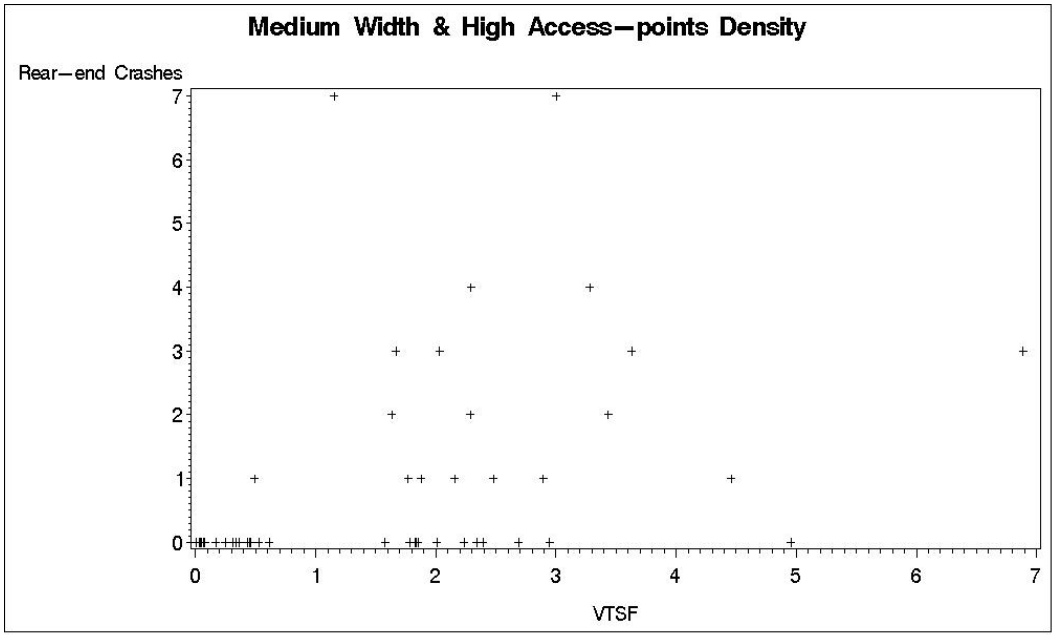
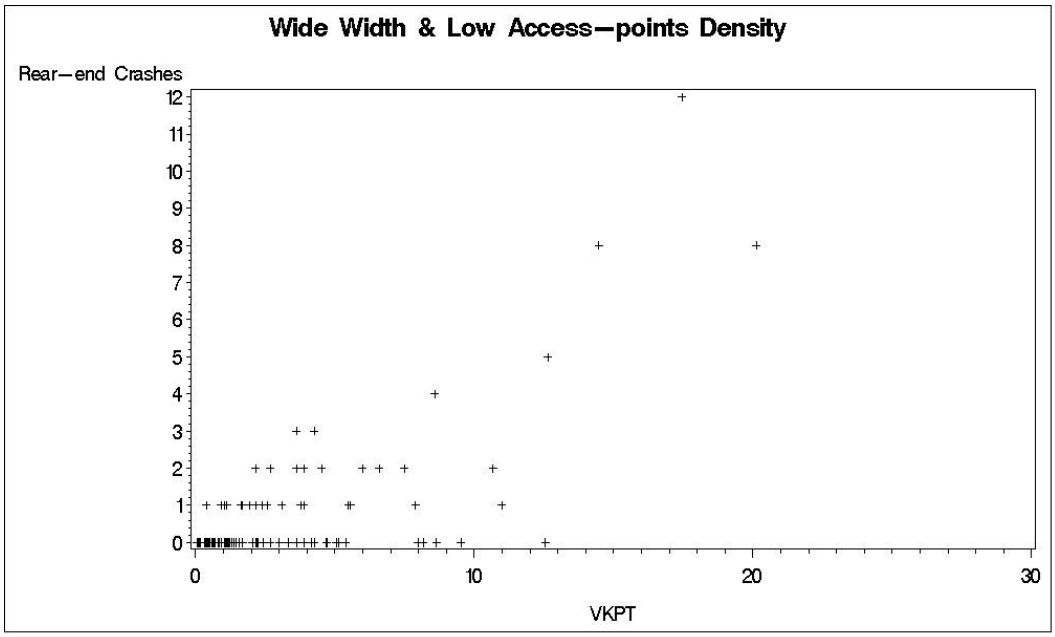


Figure 4.4 Rear End Collisions Versus VKTP and VTSF: Segments with Medium Width and High Access-Point Density



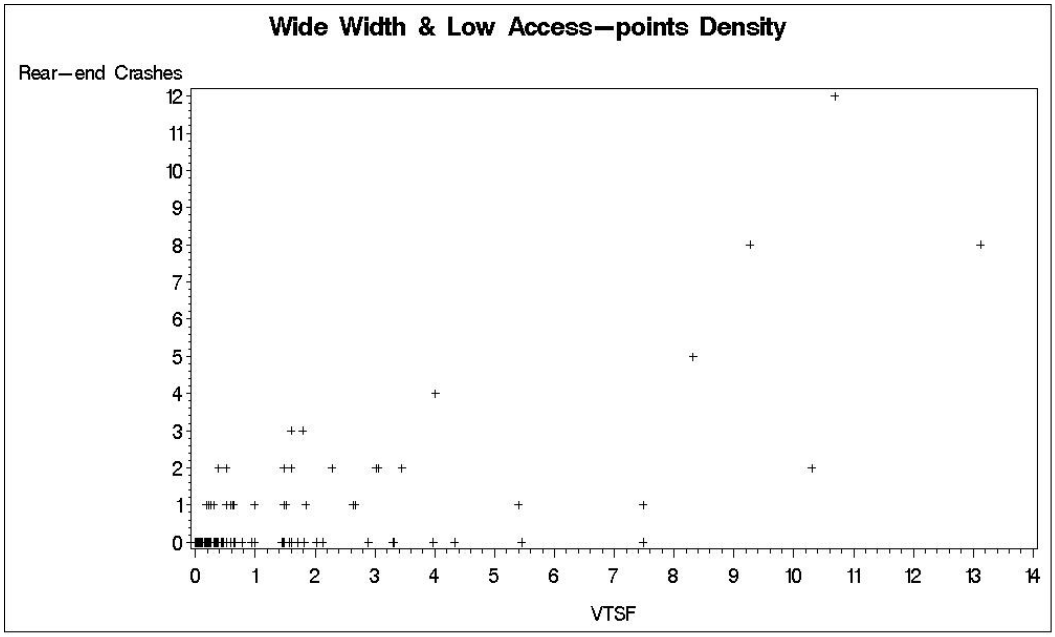
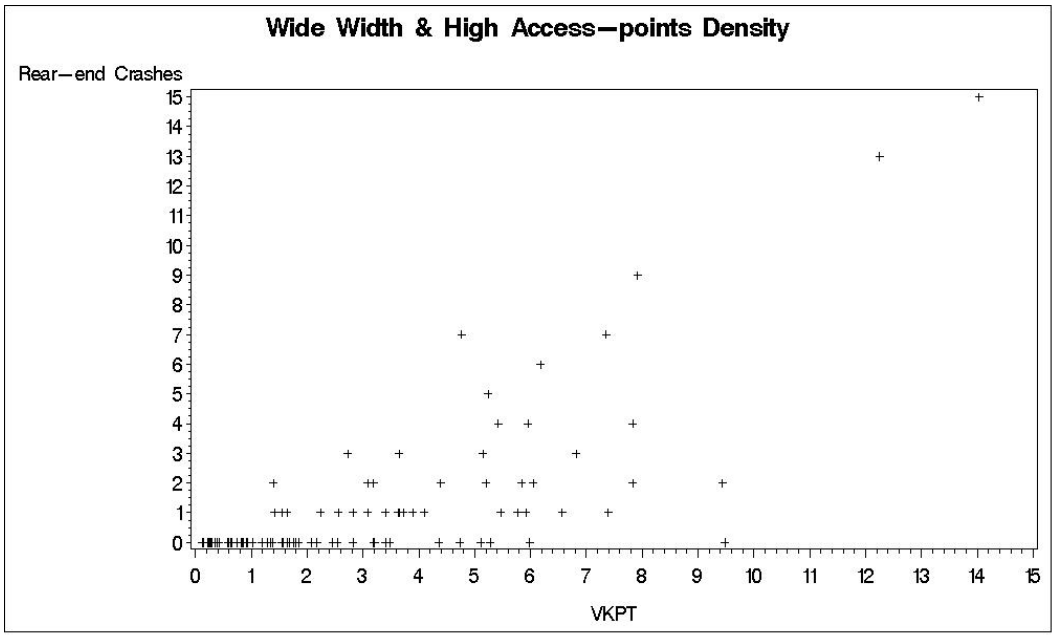


Figure 4.5 Rear End Collisions Versus VKTP and VTSF: Segments with Wide Width and Low Access-Point Density



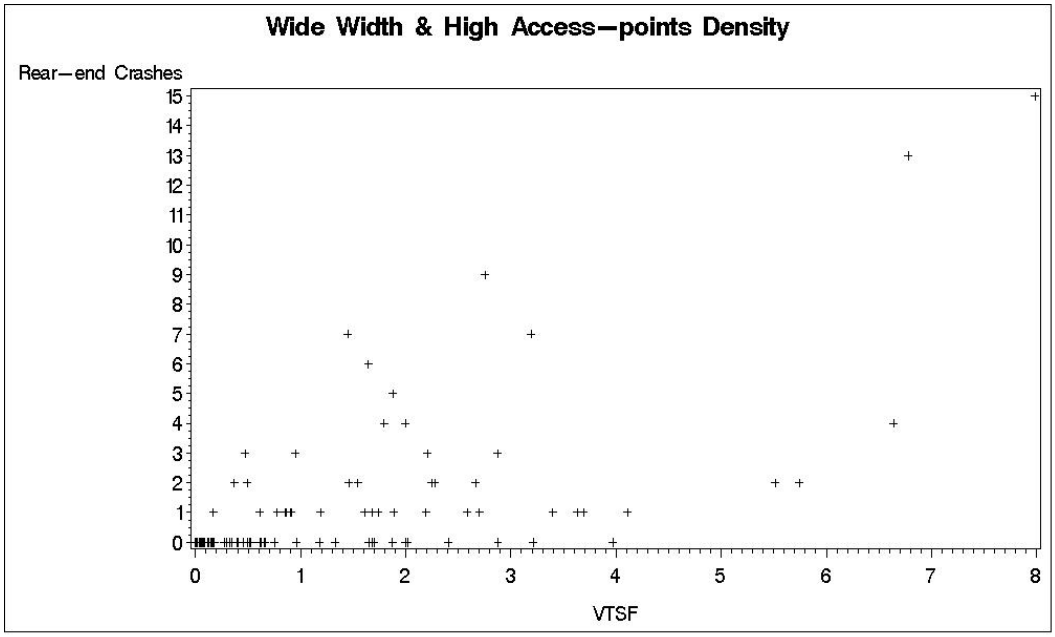
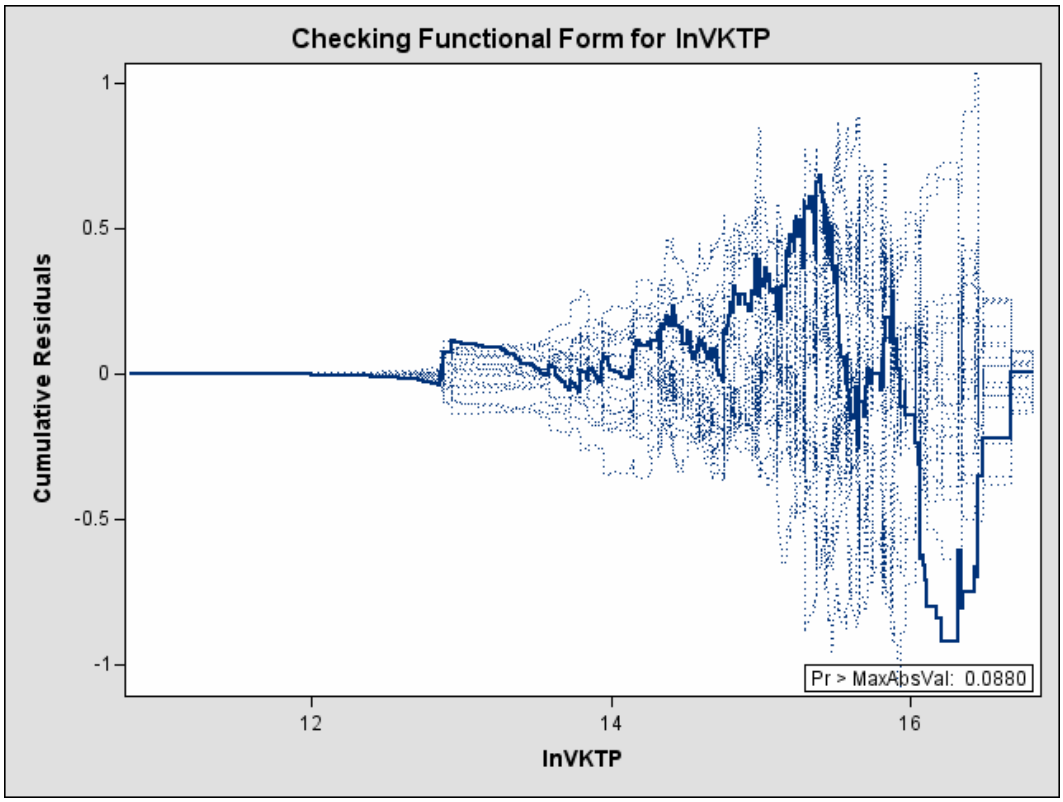


Figure 4.6 Rear End Collisions Versus VKTP and VTSF: Segments with Wide Width and High Access-Point Density



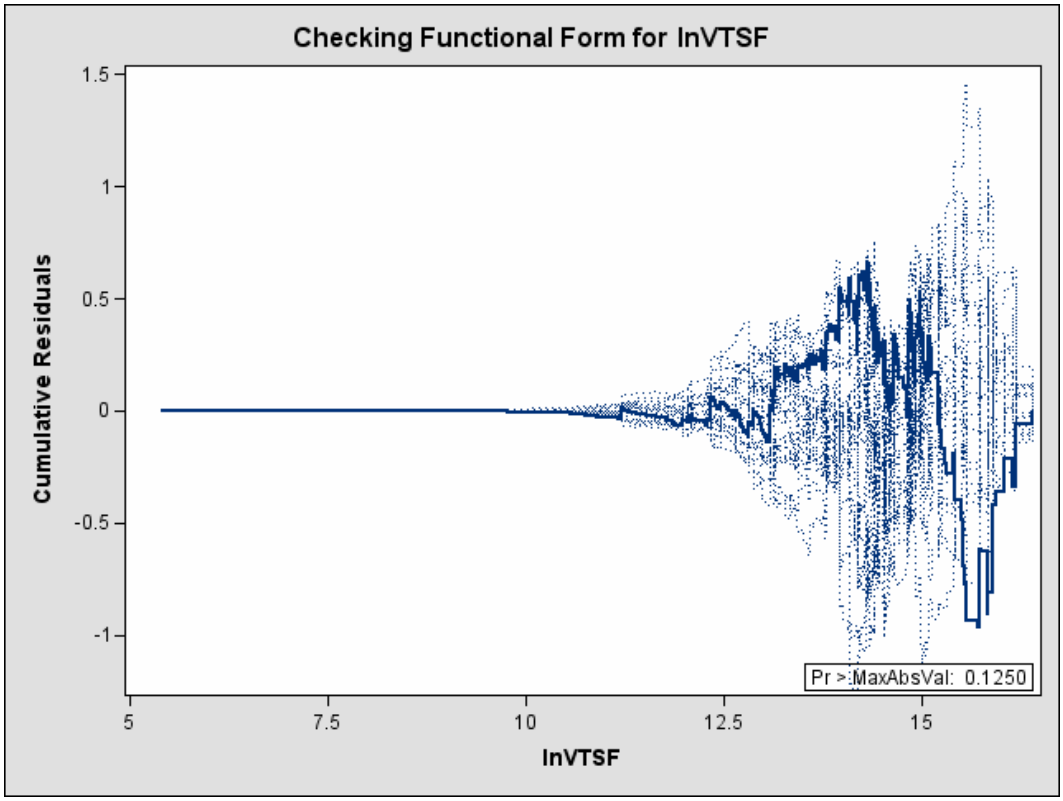


Figure 4.7 Cumulative Residuals Plot for Checking VKTP and VTSF Functional Forms

Table 4.6 Full Rear-end Crash Model Results Using VKTP and VTSF as Offset

	VKTP	VTSF
Distribution	Pearson-scaled Poisson	Pearson-scaled Poisson
Log-likelihood	-203.434	-185.601
Scale/Dispersion	1.076	1.103
Intercept	-15.884** (0.351)	-15.187** (0.357)
2 – 6 AM	-0.936 (0.701)	0.273 (0.718)
6 – 10 AM	0.369 (0.344)	0.422 (0.352)
10 AM – 2 PM	0.572 (0.344)	0.584* (0.352)
2 – 6 PM	0.635* (0.337)	0.604* (0.345)
6 – 10 PM	0.082* (0.364)	-0.321 (0.373)
Pavement Width < 13 ft	-0.007 (0.136)	0.022 (0.140)
Pavement Width > 15 ft	-0.679** (0.150)	-0.578** (0.153)
Posted Speed < 45 mph	-0.194 (0.129)	-0.296** (0.135)
Number of Access Points	0.041** (0.010)	0.059** (0.010)

* Significant at 90 percent confidence level

** Significant at 95 percent confidence level

As Table 4.6 shows, in the model for all same-direction crashes the parameter estimate of VTSF has a 95 percent confidence interval (0.891, 1.106) that includes 1.0. However, when we separate rear-end collisions from the other same-direction collisions, only the VTSF parameter estimate in the rear-end collision models still has a value close to 1.0 with a 95 percent confidence interval of (0.974, 1.220). The parameter estimate of about 0.5 in the model for sideswipe and turning same-direction collisions shows that vehicle time spent following for these two types of collisions does not account for exposure as much as it does for rear-end collisions. The contributing factor-based collision type categorization from the previous study showed that "following too closely" is responsible for almost 90 percent of rear-end crashes but only about 40 percent of turning and sideswipe collisions. This implies VTSF is a more suitable exposure measurement for collisions involving vehicles following each other closely. The reason behind this result might be that percent-time-spent-following is an accurate representation of car following conditions. Thus, VTSF, as a linear function of percent-time-spent-following, then successfully explains in the models the traffic flow condition for rear-end collisions that are due to following too closely. This finding supports the argument that perhaps using contributing factors alone to categorize collisions will yield better prediction models than considering collision types. For the same concept, for same-direction sideswipe and turning crashes, "number of vehicles passing" or "number of vehicles turning" might be good exposure candidates that can address the collision causes due to vehicle passing or vehicle turning. These measures would likely involve number and volume of traffic in and out of access points along the road segment.

For future studies, this is also the first step towards estimation of a crash prediction model that identifies causal relationships between predictor variables and crash incidence. It is also crucial to examine significant covariates in the models other than the exposure measurement. By categorizing the collision types using contributing factors, we intend to bridge the gap between crash prediction models and the collision causality studies, *i.e.* to possibly associate the predictor variables with collision causes. In this case, we find "following too closely" contributes to most rear-end collisions and thus the significant predictor variables could then help to pinpoint the geometric design features which have influence on such contributing factors.

In the models estimated with VKTP and VTSF included as exposure variables, roadway width categories and the number of access points are found to be significant. The covariate parameters estimated in VKTP and VTSF models are not substantially different. The reduced rear-end Pearson-scaled Poisson model results show narrow roadway (6-8 m) tends to be correlated with higher crash risk but does not correspond to a significantly higher rear-end crash count than medium roadway (8.6-9 m). However, compared to medium roadway, wide roadway (> 9 m) decreases rear-end crashes marginally by nearly 40 percent given all other variables the same. Drivers having plenty of space on a wide roadway to move away from a vehicle they are about to rear-end might be the reason for this result. Access point density also has a

significant effect on increasing rear-end crashes by 6 percent per access point. Since vehicles making a turn have much slower speed than the regular traveling speed and minor intersections and driveways have vehicles turning in and out to merge onto the major roadway traffic, chance of rear-ending apparently is higher at roadway segment where the access point density is higher.

4.5 Summary and Conclusions

This study is focused on evaluating prediction models for same-direction crashes using a newly defined exposure measurement (VTSF) in comparison with models using VKTP, a similar form to the commonly accepted exposure VKT except with regard to accounting for time of day. Three different same-direction crash data sets are considered for the modeling as a continuation from the studies described in Chapters 2 and 3. The new exposure measurement is investigated together with the contributing factor-based collision type categorization. The model goodness of fit statistics show that the over-dispersion in all same-direction and rear-end crash data sets can be best compensated for using the Pearson-scaled Poisson model. Models using VTSF as exposure show no improvement by log-likelihood compared to those using VKTP as exposure. However, when VTSF and VKTP are taken as offset in the models, the VTSF model has much higher log likelihood and therefore out performs VKTP model.

The parameter estimates for VTSF in the models for all same-direction and rear-end crash data sets are found not significantly different from 1.0, while those for VKTP are significantly higher than 1.0. This finding shows that VTSF could be used as the appropriate functional form of exposure that has linear relationship with same-direction, especially rear-end crash counts. Furthermore, it suggests the key in investigating same-direction crashes is to associate the traffic flow condition to collision cause, as the cause of collisions might be more correlated with exposure values specifically defined for the type of driver error involved. The traffic condition, itself, of vehicles following others closely implies crash risk of cars rear-ending. PTSF then well formulates this traffic situation and so VTSF as a linear function of PTSF can represent both the traffic flow intensity and state in crash prediction models. In showing some statistical support, the cumulative residual analysis also indicates that the functional form of VTSF fits better in the model than that of VKTP as shown in Figure 4.7.

This finding about VTSF can possibly extend our use of same-direction, especially rear-end crash forecasting models. In the previous studies, the prediction models were only used in interpreting crash risk for the data from one area or region. Usually the parameter of exposure is estimated differently from model to model, thus it is impossible to deliver universally applicable Safety Performance Functions (SPF) concerning the effects on safety of the roadway design features. The linear relationship found between VTSF and rear-end crashes is a first step over this barrier. We hope by properly representing traffic flow intensity and state using VTSF, there can eventually be consistent SPF's estimated for evaluating road safety performance from different locations.

5 Conclusions

5.1 Summary of Results

The three steps proposed in Section 1.4 are conducted as described in Chapters 2 through 4. For step 1, Chapter 2 defines the opportunities for opposite-direction crashes and compares the model results using opportunities as exposure with those using VKT as exposure. Chapter 3 carries out the step 2 assignment and defines collision type categories based on contributing factors. Similar to Chapter 2 but also taking the results of Chapter 3 into consideration, Chapter 4 illustrates results that are obtained not only by evaluating the same-direction crash opportunities as exposure in crash prediction models compared to VKTP but also by applying the categorizing results from Chapter 3 in the prediction models and comparing the results with those assuming the commonly defined categories. Together, the studies find the concept of crash opportunities shows promise in same-direction crash models but not much in opposite-direction and single vehicle crash models, at least as defined here. Furthermore, collision type categorization based on contributing factors is also backed by model results showing new categories help to reveal the actual relationship between crashes and exposure. Below is a summary of the findings and conclusions of each study as well as the contributions of the research to crash prediction modeling.

For opposite-direction crashes, including sideswipe opposite-direction and head-on collisions, models using opportunities as exposure do not show improvement in goodness-of-fit compared to those using VKT as exposure. However, AADT added as a covariate in the models seems to help to increase log-likelihood of the models using opportunities as exposure more than that of the models using VKT. In addition to the explanations listed in Chapter 2, constricting the exposure measurements to have parameter estimate 1.0 in the models could be another reason to cause the low performance of opportunities. As seen in equation (2.5), besides segment length and AADT, the definition of opportunities involves more variables that VKT does not consider, thus more variables having parameters to be forced to be 1.0. Statistically, this constriction directly leads to lower log-likelihood.

The contributing factor-based collision type categorization regroups ten collision types into four categories. One important finding is that rear-end crashes fall into one single category, while sideswipe and turning same-direction crashes belong to a different category. This separation is in contrast to the usual practice of grouping all three types of crashes into one individual collision category. Here they are separated due to the overrepresentation of “following too close” as contributing factor for rear-end collisions. “Following too close” is coded as contributing factor for nearly 85 percent rear-end collisions, while only for less than 20 percent of the other two same-direction collision types. Another important finding is that opposite-direction collision types including sideswipe and head-on are grouped into the same category together with single vehicle collision types including fixed object and overturn. The reason behind this aggregation is that all four collision types share a very similar pattern in contributing factors, especially “driver lost control” and “speed too fast”, which are coded for nearly 60-70 percent collisions of each of the four types. Severity is also considered for categorization in the study, but only the contributing factor-based results are considered for further crash prediction models.

The results of same-direction crash models find a linear relationship between the new exposure VTSF and the total same-direction crashes as well as rear-end crashes. This linearity implies a well defined traffic flow intensity and state by the new exposure, and provides for the possibility of a constant crash rate by dividing the number of rear-end crashes by this exposure measurement. Moreover, finding VTSF being linearly related to the rear-end crash count, which is mainly caused by contributing factor following too closely, further suggests categorizing crashes by crash causality rather than the nature of the collision and defining the exposure measurement based on the contributing factor can lead to more accurate and explainable prediction models.

5.2 Application of Results

This report contributes to current traffic safety research by redefining exposure for crash prediction models to represent traffic flow intensity as well as flow state. The study about crash opportunities described in this report emphasizes the variation among collision types caused by different contributing factors and flow characteristics under which different type of collisions occurs relating to those contributing factors. As seen in Chapter 2, for a head-on or opposite-direction sideswipe collision to

occur, vehicles traveling in opposite directions have to meet. Thus, the definition for opposite-direction crash opportunities calculates the number of times vehicles have to meet at the selected road segments during the study time period. Similarly, Chapter 4 defines crash opportunities for same-direction collisions, which occur only if vehicles traveling in same-direction follow one another closely. Percent-time-spent-following (PTSF) described in HCM 2000 can well define this flow situation, and so the crash opportunities, vehicle-time-spent-following (VTSF), computed linearly to PTSF is able to represent the traffic flow state that is necessary for a same-direction collision to take place.

By including the newly defined exposure measurements in the prediction models, the results show that VTSF, when used as an offset, not only corresponds to higher model log likelihood but increases/decreases linearly with rear-end crash counts. This finding directly relates to the outcome of the collision categorization based on contributing factors. Due to a rather high percentage of rear-end collisions caused by contributing factor "following too closely", VTSF apparently explains the traffic situation to which rear-end collisions are closely related. For same-direction sideswipe and turning, since VTSF is not established to measure percent of vehicle turning or percent of vehicle passing, it is therefore hard to justify a linear relationship between the two. For opposite-direction crash models, crash opportunities do not perform as well as the commonly used exposure, Vehicle Kilometers Traveled (VKT). First, this might be due to the statistical constriction caused by setting all the parameter estimate of the variables involved in computing crash opportunities to 1. Second, it is likely because the opposite-direction crash opportunities is not defined to explain the most important contributing factors, driver lost control and speed too fast, for head-on and opposite-direction sideswipe crashes.

The linearity found between VTSF and rear-end crashes implies that VTSF can be used as the appropriate functional form of exposure for rear-end crashes. In addition to suggesting the necessity of defining exposure measurement based on collision causalities, this finding further confirms the collision categorization with regard to crash contributing factors described in Chapter 3. This linearity also enables computing a meaningful Safety Performance Function (SPF) concerning the effects on safety of the roadway design features. Current safety studies found it is difficult to reach a stable SPF results due to the variation in data, especially traffic flow condition, from location to location. It is thus hopeful that the use of VTSF can standardize traffic flow representation for all different locations and achieve a rather uniform measurement of SPF. More generally, to use VTSF as exposure, a crash rate computed by dividing rear-end crash counts by VTSF is more understandable and acceptable by the general public.

To follow up this report study, future research can be conducted to redefine crash exposure measurement for other collision categories, such as intersecting-direction crashes. More than 80 percent of intersection-related collisions are found to be caused by aberrant driver behaviors, including "fail to grant right of way" and "violate traffic control". Therefore, a measurement that is able to account for the traffic conflicts due to such driver misbehavior might be a good candidate for intersecting-direction crash opportunities. Additionally, models using such exposure measurement can also link the opportunity-based crash prediction to predictor variable estimates. In other words, intersection characteristics can be sought to help to identify the most likely responsible cause to the driver misbehaviors and thus the collisions. For rear-end collisions, further modeling process involving larger amount of data for determining the SPF values by different road classifications can also be carried out by using VTSF as exposure.

References

Abdel-Aty, M; and Radwan, E. "Modeling Traffic Accident Occurrence and Involvement." *Accident Analysis and Prevention*. Vol. 32, pp. 633-642, 2000.

Agarwal, M; Maze, T; and Souleyrette, R. "Impacts of Weather on Urban Freeway Traffic Flow Characteristics and Facility Capacity." Proceedings of Mid-Continent Transportation Research Symposium. Ames, Iowa, August 18–19, 2005.

Bruhning, E; and Volker, R. "Accident Risk in Road Traffic – Characteristic Quantities and Their Statistical Treatment." *Accident Analysis and Prevention*. Vol. 14, pp. 65 – 80, 1982.

Davis, G. "Possible aggregation bias in road safety research and a mechanism approach to accident modeling." *Accident Analysis and Prevention*, No. 36, pp. 1119-1127, 2004.

Davis, G; and Swenson, T. "Collective responsibility for freeway rear-ending accidents? An application of probabilistic causal models." *Accident Analysis and Prevention.* No. 38, pp. 728-736. 2006.

Delen, D; Sharda, R; and Bessonov, M. "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks." *Accident Analysis and Prevention,* No. 38, pp. 434-444. 2006.

Elvik, R; Christensen, P; and Amundsen, A. "Speed and Road Accidents – An Evolution of the Power Model." TÖL. Oslo, Norway, 2004.

Finch, D; Kompfner, P; Lockwood, C; and Maycock, G. "Speed, Speed Limit and Accidents." TRL. London, United Kingdom, 1994.

Fridstrom, L; Ifver, J; Ingebrigtsen, S; Kulmala, R; and Thomsen, L "Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to The Variation in Road Accident Counts." *Accident Analysis and Prevention.* Vol. 27, pp. 1-20, 1995.

Hauer, E. "Traffic Conflicts and Exposure." *Accident Analysis and Prevention*. Vol. 14, pp. 359-364, 1982

Hauer, E; Ng, J; and Lovell, J. "Estimation of safety at signalized intersections." *Transportation Research Record*, No. 1185, pp. 48-61, 1988.

Hauer, E. "On Exposure and Accident Rate." *Traffic Engineering and Control.* Vol. 36, pp. 134 – 138, 1995.

Highway Capacity Manual (HCM). Transportation Research Board of National Academies. Washington, D.C., 2000.

Khorashadi, A; Niemeier, D; Shankar, V; and Mannering, F. "Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis." *Accident Analysis and Prevention*, No. 37, pp. 910-921. 2005.

Ivan, J; Pasupathy, R; and Ossenbruggen, P. "Difference in Causality Factors for Single and Multi-vehicle Crashes on Two-lane Roads." *Accident Analysis and Prevention*. Vol. 31, pp. 695-704, 1999.

Ivan, J; Wang, C; and Bernardo, N. "Explaining Two-lane Highway Crash Rates Using Lane Use and Hourly Exposure." *Accident Analysis and Prevention*. Vol. 32, pp. 787-795, 2000.

Jonsson, T. "Predicting Models for Accidents on Urban Links: A Focus on Vulnerable Road Users." Dissertation at Lund Institute of Technology, Lund, Sweden, 2005.

Kim, D and Washington, S. "The Significance of Endogeneity Problems in Crash Models: An Examination of Left-turn Lanes in Intersection Crash Models." *Accident Analysis and Prevention.* Vol. 38, pp. 125-134, 2006.

Kim, D; Lee, Y; Washington, S and Choi, K. "Modeling Crash Outcome Probabilities at Rural Intersections: Application of Hierarchical Binomial Logistic Models." *Accident Analysis and Prevention.* Vol. 39, pp. 125-134, 2007.

Kim, K; Nitz, L; Richardson, J; and Li, L. "Personal and Behavioral Predictors of Automobile Crash and Injury Severity." *Accident Analysis and Prevention*, No. 27, pp. 469-481. 1995.

Lin, D; Wei, L; and Ying, Z. "Model-Checking Techniques Based on Cumulative Residuals." *Biometrics*. Vol. 58, pp. 1-12, March 2002.

Lord, D; Manar, A; and Vizioli, A. "Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments." *Accident Analysis and Prevention.* Vol. 37, pp. 185-199, 2005.

Lord, D. "Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter." *Accident Analysis and Prevention*. Vol. 38, pp. 751-766, 2006.

Maher, M; and Summersgill, I. "A Comprehensive Methodology for the Fitting of Predictive Accident Models." *Accident Analysis and Prevention.* Vol. 28, pp. 281-296, 1996.

McCullagh, P; and Nelder, J. *Generalized Linear Models.* Cambridge University Press, Great Britain, 1989.

McQueen, J. "Some methods for classification and analysis of multivariate observations." 5th Berkeley symposium on mathematical statistics and probability, Berkeley, University of California Press, 1: 281-297, 1967.

Mountain, L; Maher, M; and Fawaz, B. "The Influence of Trend on Estimates of Accidents at Junction." *Accident Analysis and Prevention*. Vol. 30, pp. 641-649, 1998.

National Highway Traffic Safety Administration. "Traffic Safety Facts 2004, A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System." National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, DC, 2004.

Nilsson, G. "Traffic safety dimensions and the power model to describe the effect of speed on safety." Bulletin 211, Institutionen för Teknik och samhälle, Lunds Universitet, Lund, Sweden, 2004.

Oh, C; Park, S; and Ritchie, S. "A Method for Identifying Rear-end Collision Risks Using Inductive Loop Detectors." *Accident Analysis and Prevention.* Vol. 38, pp. 295-301, 2006.

Olsson, U. "Generalized Linear Models, an Applied Approach." Studentlitteratur, Lund, Sweden, 2002.

Ossenbruggen, P; Pendharkar, J; and Ivan, J. "Roadway Safety in Rural and Small Urbanized Areas." *Accident Analysis and Prevention.* Vol. 33, pp. 485-498, 2001.

Papacostas, C; and Prevedouros, P. *Transportation Engineering and Planning*. Prentice-Hall, Inc., Eaglewood Cliffs, New Jersey, 1993.

Persaud, B; and Dzbik, L. "Accident Prediction Models for Freeways." *Transportation Research Record*. Vol. 1401, pp. 55-60, 1993.

Qin, X; Ivan, J; and Ravishanker, N. "Selecting exposure measures in crash rate prediction for two-lane highway segments." *Accident Analysis and Prevention.* Vol. 36, pp. 183-191, 2004.

SAS 8.02, SAS Institute Inc., Cary, NC, USA, 1999.

SPSS 13.0, SPSS Inc., Illinois, 2004.

Shankar, V; Mannering, F; and Barfield, W. "Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies." *Accident Analysis and Prevention*. Vol. 27, pp. 371-389, 1995.

Wang, X and Abdel-Aty, M. "Temporal and Spatial Analyses of Rear-end Crashes at Signalized Intersectional." *Accident Analysis and Prevention.* Vol. 38, pp. 1137-1150, 2006.

Zhang, C; Ivan, J; ElDessouki, W; and Anagnostou, E. "Relative Risk Analysis for Studying the Impact of Adverse Weather Conditions and Congestion on Traffic Accidents." Transportation Research Board 84th Annual Meeting CD ROM, Washington DC, USA, Jan. 2005.

Zhang, C; Ivan, J and Jonsson, T. "Predicting Two-lane Highway Crashes Using Crash Opportunities: A Newly Defined Measure of Exposure." Transportation Research Board 85th Annual Meeting CD ROM, Washington DC, USA, 2006.

Zhang, C; Ivan, J and Jonsson, T. "Collision Type Categorization Based on Crash Causality and Severity Analysis." Transportation Research Board 86th Annual Meeting CD ROM, Washington DC, USA, 2007.

Zhou, M; and Sisiopiku, V. "Relationship between volume-to-capacity ratios and accident rates." *Transportation Research Record*. No. 1581, pp. 47-52, 1997.