ESTIMATION OF HYBRID MODELS FOR REAL-TIME CRASH RISK ASSESSMENT ON FREEWAYS

by

ANURAG PANDE
B.Tech. Indian Institute of Technology Bombay, 2002
M.S. University of Central Florida, 2003

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Civil and Environmental Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2005

Major Professor
Dr. Mohamed Abdel-Aty

ABSTRACT

Relevance of reactive traffic management strategies such as freeway incident detection has been diminishing with advancements in mobile phone usage and video surveillance technology. On the other hand, capacity to collect, store, and analyze traffic data from underground loop detectors has witnessed enormous growth in the recent past. These two facts together provide us with motivation as well as the means to shift the focus of freeway traffic management toward proactive strategies that would involve anticipating incidents such as crashes.

The primary element of proactive traffic management strategy would be model(s) that can separate 'crash prone' conditions from 'normal' traffic conditions in real-time. The aim in this research is to establish relationship(s) between historical crashes of specific types and corresponding loop detector data, which may be used as the basis for classifying real-time traffic conditions into 'normal' or 'crash prone' in the future. In this regard traffic data in this study were also collected for cases which did not lead to crashes (non-crash cases) so that the problem may be set up as a binary classification.

A thorough review of the literature suggested that existing real-time crash 'prediction' models (classification or otherwise) are generic in nature, i.e., a single model has been used to identify all crashes (such as rear-end, sideswipe, or angle), even though traffic conditions preceding crashes are known to differ by type of crash. Moreover, a generic model would yield no information about the collision most likely to occur.

To be able to analyze different groups of crashes independently, a large database of crashes reported during the 5-year period from 1999 through 2003 on Interstate-4 corridor in Orlando were collected. The 36.25-mile instrumented corridor is equipped with 69 dual loop detector stations in each direction (eastbound and westbound) located approximately every ½ mile. These stations report speed, volume, and occupancy data every 30-seconds from the three through lanes of the corridor. Geometric design parameters for the freeway were also collected and collated with historical crash and corresponding loop detector data.

The first group of crashes to be analyzed were the rear-end crashes, which account to about 51% of the total crashes. Based on preliminary explorations of average traffic speeds; rear-end crashes were grouped into two mutually exclusive groups. First, those occurring under extended congestion (referred to as regime 1 traffic conditions) and the other which occurred with relatively free-flow conditions (referred to as regime 2 traffic conditions) prevailing 5-10 minutes before the crash. Simple rules to separate these two groups of rear-end crashes were formulated based on the classification tree methodology. It was found that the first group of rear-end crashes can be attributed to parameters measurable through loop detectors such as the coefficient of variation in speed and average occupancy at stations in the vicinity of crash location. For the second group of rear-end crashes (referred to as regime 2) traffic parameters such as average speed and occupancy at stations downstream of the crash location were significant along with off-line factors such as the time of day and presence of an on-ramp in the downstream direction. It was found that regime 1 traffic conditions make up only about 6% of the

traffic conditions on the freeway. Almost half of rear-end crashes occurred under regime 1 traffic regime even with such little exposure. This observation led to the conclusion that freeway locations operating under regime 1 traffic may be flagged for (rear-end) crashes without any further investigation. MLP (multilayer perceptron) and NRBF (normalized radial basis function) neural network architecture were explored to identify regime 2 rear-end crashes. The performance of individual neural network models was improved by hybridizing their outputs. Individual and hybrid PNN (probabilistic neural network) models were also explored along with matched case control logistic regression. The stepwise selection procedure yielded the matched logistic regression model indicating the difference between average speeds upstream and downstream as significant. Even though the model provided good interpretation, its classification accuracy over the validation dataset was far inferior to the hybrid MLP/NRBF and PNN models. Hybrid neural network models along with classification tree model (developed to identify the traffic regimes) were able to identify about 60% of the regime 2 rear-end crashes in addition to all regime 1 rear-end crashes with a reasonable number of positive decisions (warnings). It translates into identification of more than ¾ (77%) of all rear-end crashes.

Classification models were then developed for the next most frequent type, i.e., lane change related crashes. Based on preliminary analysis, it was concluded that the location specific characteristics, such as presence of ramps, mile-post location, etc. were not significantly associated with these crashes. Average difference between occupancies of adjacent lanes and average speeds upstream and downstream of the crash location were found significant. The significant variables were then subjected as inputs to MLP and

NRBF based classifiers. The best models in each category were hybridized by averaging their respective outputs. The hybrid model significantly improved on the crash identification achieved through individual models and 57% of the crashes in the validation dataset could be identified with 30% warnings. Although the hybrid models in this research were developed with corresponding data for rear-end and lane-change related crashes only, it was observed that about 60% of the historical single vehicle crashes (other than rollovers) could also be identified using these models. The majority of the identified single vehicle crashes, according to the crash reports, were caused due to evasive actions by the drivers in order to avoid another vehicle in front or in the other lane. Vehicle rollover crashes were found to be associated with speeding and curvature of the freeway section; the established relationship, however, was not sufficient to identify occurrence of these crashes in real-time.

Based on the results from modeling procedure, a framework for parallel real-time application of these two sets of models (rear-end and lane-change) in the form of a system was proposed. To identify rear-end crashes, the data are first subjected to classification tree based rules to identify traffic regimes. If traffic patterns belong to regime 1, a rear-end crash warning is issued for the location. If the patterns are identified to be regime 2, then they are subjected to hybrid MLP/NRBF model employing traffic data from five surrounding traffic stations. If the model identifies the patterns as crash prone then the location may be flagged for rear-end crash, otherwise final check for a regime 2 rear-end crash is applied on the data through the hybrid PNN model. If data from five stations are not available due to intermittent loop failures, the system is

provided with the flexibility to switch to models with more tolerant data requirements (i.e., model using traffic data from only one station or three stations). To assess the risk of a lane-change related crash, if all three lanes at the immediate upstream station are functioning, the hybrid of the two of the best individual neural network models (NRBF with three hidden neurons and MLP with four hidden neurons) is applied to the input data. A warning for a lane-change related crash may be issued based on its output. The proposed strategy is demonstrated over a complete day of loop data in a virtual real-time application. It was shown that the system of models may be used to continuously assess and update the risk for rear-end and lane-change related crashes.

The system developed in this research should be perceived as the primary component of proactive traffic management strategy. Output of the system along with the knowledge of variables critically associated with specific types of crashes identified in this research can be used to formulate ways for avoiding impending crashes. However, specific crash prevention strategies e.g., variable speed limit and warnings to the commuters demand separate attention and should be addressed through thorough future research.

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my advisor Dr. Mohamed Abdel-Aty. Without his valuable guidance and constant support this dissertation would not have taken its present shape. I would also like to acknowledge the support of my committee members, Dr's Essam Radwan, Haitham Al-Deek, Nizam Uddin and Chris Lee. I would like to thank Dr's Nizam Uddin and Chris Lee again for being part of the research group and helping me with several relevant issues. Dr. Fathy Abdalla, Dr. Jeong Yu, Raja, Albinder, and Jeremy also deserve thanks for their help as members of my research group. I would like to thank Dr. Haitham Al-Deek again for providing access to the data used here in. Ravi, Patrick and Yueliang also deserve mention here, for their expertise in coding made my task easier. I would like to thank Nezamuddin, Albinder and to an extent Ravi for being 'come with' guys especially for coffee breaks in the middle of several nights. Last but not the least I would like to acknowledge the blessings of Lord Ganesha, my parents, elder sister Deepti and brother-in-law Vivek back in India. Their best wishes helped me sail through my toughest times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Research Motivation

Crashes are incidents involving collision among vehicles or between vehicles and other fixed/moving objects on or off the roadway. Traffic crashes claim more human years than any other incident or disease. They also result in tremendous property losses. The motivation for this research mainly stems from concern to save human lives associated with freeway crash occurrences. According to Traffic Safety Facts (2002), a quarter of all urban fatal crashes occurred on uninterrupted flow facilities. However, the losses from freeway crashes go beyond what is exemplified through the aforementioned statistic. Even the least severe of crashes impact traffic operation in a big way and turn freeways into virtual parking lots. In fact, most of the congestion on freeways may be attributed to incidents consisting mostly of crashes. These facts signify that preventing freeway crashes is not only important from a traffic safety stand point but from an operation point of view as well.

Having identified the extent of the problem, the first step towards an effective solution is to identify the primary causes. The 'nut behind the wheel' is usually perceived as the cause of most traffic crashes and it is not hard to find statistics to support this claim. This perception might lead to the assumption that countermeasures must concentrate on changes in the attitude and behavior of the drivers. Henderson (1971) compared this notion about traffic crashes to the way people used to think about Cholera. Since Cholera

mostly struck poor people it was believed that Cholera would be obliterated only if poor people would change their unhygienic ways of life. It was pointed out that the control of the environment (i.e., purified water and construction of sewage systems), and not changes in human behavior, brought the disease under control. It was concluded that focusing too much on the driver behavior as the cause (even if it is an important one), and therefore the solution to crashes, masks our ability to see other efforts that could reduce traffic crashes. It is important to understand that crashes are caused due to bad decisions made by the driver in an environment resulting from surrounding traffic conditions and geometric design created by the engineer. The influence of geometric design on the likelihood of a driver making bad decision has been well understood since, but the attention given to the surrounding traffic conditions immediately preceding crash occurrence has almost been non-existent. Limited progress of the real-time traffic surveillance technology might be one of the reasons behind it.

1.2 Problem Statement

In the recent past tremendous growth has been observed in traffic management and information systems. The capability of storing and analyzing data has grown manifolds and considerable amounts of data are being collected and stored for ITS applications. These data include speed, vehicle counts and occupancy archived from each lane every 30 seconds by a series of loop detectors installed beneath the freeway pavement. While this growth has been around for some time now, research efforts directed toward the application of freeway loop detector data for traffic safety have gained momentum only recently. Since traffic flow would be measured in terms of loop data collected in real-

time, the approach diverges from traditional safety studies aimed at estimation of crash frequency or rate on freeway sections through aggregate measures of traffic flow (e.g., AADT or hourly volumes).

Application of loop detector data for traffic management has been limited to incident detection algorithms. These algorithms are developed by analyzing historical post-incident loop data and attempt to detect incidents as quickly as possible. It essentially is a reactive strategy which is being rendered irrelevant with widespread use of mobile phones and surveillance cameras. This research is an effort in the direction of proactive traffic management that would involve anticipating impending incidents such as crashes. Crashes are arguably the most critical and 'predictable' types of incidents. In some of our initial efforts (Pande (2003), Abdel-Aty et al. (2004) and Abdel-Aty and Pande (2005)) as part of the ongoing research at University of Central Florida, the concept of relating archived ITS-data with crash occurrences has been shown to work. But these studies, as well as some of the related work done elsewhere (e.g., Lee et al. 2002, 2003), are limited in their scope due to generic nature of the models. The term generic implies that a single model is recommended to anticipate all crashes. Traffic conditions following the crashes are somewhat similar in nature and are not a function of the type of crash. Therefore, the incident detection algorithms can rely on generic models. This "one size fits all" approach, however, is not sufficient for anticipating crashes because crashes initiated by different harmful events are known to be associated with distinct traffic characteristics. For example, rear-end crashes are expected to occur in congested traffic regimes where

drivers have to slow down and speed up quite often; on the other hand, sideswipe crashes might result from excessive lane changing in an 'at capacity' regime.

To develop separate models for groups of crashes one needs to segregate the crash data into smaller groups. To ensure sufficient sample size in resulting categories crash data needs to be adequately large. A large initial sample of crashes will also ascertain that, after accounting for intermittent failures of loop detectors, enough crashes with corresponding loop data available would be left for analysis. The problem of discriminating crash prone conditions from normal freeway traffic is set up as a classification problem in this research. To this end we also need to collect "non-crash" loop detector data representing 'normal' traffic. Traffic data corresponding to crash (and non-crash) cases need to be augmented with geometric design parameters of the freeway and some measures of driver characteristics to examine their impact on crash occurrence. These three groups of parameters (traffic, geometric, and driver characteristics) should then be used simultaneously as inputs to the binary (crash vs. non-crash) classification models.

Based on a thorough literature review, it may be argued that real-time identification of freeway "black-spots" in a traffic management framework is in its primitive phase. To provide reliable information to traffic management authorities about the likelihood of crashes in real-time, a disaggregate analysis of historical crash data is required. The disaggregate approach demands a large crash database to be assembled and systematically correlated with traffic, geometric and driver characteristics. This research

effort is aimed at development of a system that involves simultaneous application of multiple models and can be used to reliably flag freeway sections for an impending crash.

1.3 Research Objectives

The main objectives of this research are the following:

1. Critically review the studies applying ITS archived data for traffic safety analysis and the studies employing relevant methodologies in other transportation engineering applications.
2. Emphasize the need to develop separate models to identify crashes initiated by distinct harmful events (e.g., rear-ends, sideswipes etc.) from 'normal' traffic conditions.
3. Assemble a database with crash and corresponding traffic surveillance data sufficient for disaggregate analyses of crashes by type from the 36.25-mile instrumented corridor of Interstate-4.
4. Augment the database with geometric design features of the freeway at crash location and devise ways to incorporate the measure of driver population composition in the analysis.
5. Use the database to explore several modeling techniques such as logistic regression, classification tree and neural networks etc. and develop multiple models for separating crashes of different types from normal freeway operation. Examine hybrids of these models to improve on the classification performance of individual models.

6. Demonstrate the application of multiple models in the framework of a crash prediction system.

1.4 Organization of the Dissertation

Following this introductory chapter, a detailed review of the relevant literature is provided. The review, of course, includes studies examining the relationship between crashes and freeway loop detector data. In addition, applications of neural networks and other data mining techniques in transportation engineering have also been reviewed. The third chapter presents theoretical overview of the data analysis techniques used in this study. The next chapter explains the data preparation effort for this research. Five years of crash data have been assembled from the 36.25-mile instrumented Interstate-4 corridor along with corresponding loop detector data and geometric/driver population composition characteristics of the freeway corridor under consideration. The purpose of extensive data collection is to have sufficient data available for segregating the crashes into smaller groups and analyze them separately. In the following two chapters, detailed analysis of rear-end crash data is presented. The rear-end crashes are first separated by the traffic regime prevailing prior to their occurrence and then multiple modeling methodologies are employed to devise a strategy for separating the resulting groups of crashes from normal conditions. A similar approach is proposed for lane change related crashes in the next chapter with necessary modifications for data requirements and availability. After the modeling procedure, a chapter is dedicated to simultaneous implementation of the models for different types of crashes. In the chapter a brief analysis of single vehicle crashes is also provided which happen to be the next most frequent type of crashes after rear-end

and lane-change related crashes. The application of these models is demonstrated on data collected from all stations for a whole day on the Interstate-4. The final chapter consists of summary and conclusions from this research and provides insight into how the findings from this research may be used in future to prevent crashes. The chapter also acknowledges the fact that although we have established a reliable way to separate crash prone conditions from 'normal' freeway traffic; further investigations are needed to develop strategies for avoiding potential crashes identified based on the findings from this research.

CHAPTER 2

LITERATURE REVIEW

2.1 General

This chapter reviews previous studies from the literature relevant to this research. The literature review is divided into two sections. Traffic safety studies with real-time identification of crash prone conditions on the freeway as their objective are summarized first. All of these studies are very recent; indicating that the idea of using loop detector data for traffic safety applications is in its nascent stages. These studies are categorized into two groups: a) the exploratory studies and b) studies establishing statistical links. The review of safety applications of the ITS-archived data is followed by the summary of data mining applications in the areas of incident detection and crash analysis.

2.2 Safety Applications of ITS-archived Data

Golob et al. (2004, a) categorized traffic safety related studies into two groups; First, the aggregate studies, in which units of analysis represent counts of crashes or crash rates for specific time periods (typically months or years) and locations (specific roads or networks). The traffic flow in these studies is represented by the parameters of statistical distributions of traffic (e.g., AADT) for similar time and location (e.g., Zhou and Sisiopiku, 1997). The second group of studies consist of disaggregate analysis, in which the units of analysis are the crashes themselves and traffic flow is represented by parameters of traffic flow at the time and location of each crash.

While determination of freeway crash patterns has been the stated focus of traffic safety literature, most of the studies belong to the former group. Disaggregate studies are relatively new, and are made possible by the recent enhancements in capabilities to collect, store and analyze real-time traffic data thorough intelligent transportation system (ITS) applications. In this section such previous studies are summarized and critically reviewed since our research falls in the group of disaggregate studies.

2.2.1 Exploratory studies

Hughes and Council (1999) were one of the first authors to explore relationship between freeway safety and peak period operations using loop detector data. They concluded that the macroscopic measures, such as AADT and even hourly volume, in fact, correlate poorly to real-time system performance. Their work mostly relied upon the data coming from a single milepost location during the peak periods of the day, on which they tried to overlay the crash time at that particular location to infer about the changes in system performance as it approaches the time of the crash. The changes in the performance were also examined from "snapshots" provided by cameras installed on the freeway.

One of their most important observations was that as "design inconsistency" has been identified as a factor of crash causation; future research should consider "traffic flow consistency" as perceived by the driver as an important variable from a human-factor standpoint. They also expressed a need for determining the exact time of the crash to avoid "cause and effect" fallacy.

2.2.2 Studies establishing statistical links

Madanat and Liu (1995) came up with an incident likelihood prediction model using loop data as input. The focus of their research was to enhance existing incident detection algorithms with likelihood of incidents. They actually considered two types of incidents *a)* crashes and *b)* overheating vehicles. Binary logit was the methodology used for analysis. They concluded that merging section, visibility and rain are statistically the most significant factors for crash likelihood prediction.

Lee et al. (2002) introduced the concept of "crash precursors" and hypothesized that the likelihood of crash occurrence is significantly affected by short-term turbulence of traffic flow. They came up with factors such as speed variation along the length of the roadway (i.e. the difference between the speeds upstream and downstream of the crash location) and also across the three lanes at the crash location. Another important factor identified by them was traffic density at the instant of the crash. Weather, road geometry and time of the day were used as external controls. With these variables, a crash prediction model was developed using log-linear analysis. According to the authors the log-linear model was chosen so that the exposure can be easily determined, which would have been difficult, if instead a logit model was used. In order to test the goodness of fit for the model, Pearson chi-square test was performed. The test measured how close the expected frequencies are to the observed frequencies for any combination of crash precursors and control factors. At 95 % confidence level the model yielded a good fit.

In a later study (Lee et al., 2003), they continued their work along the same lines and modified the aforementioned model. They incorporated an algorithm to get a better estimate of time of the crash and the length of time slice (prior to the crash) duration to be examined. They concluded that variation of speed has relatively longer term effect on crash potential rather than density and average speed difference between upstream and downstream ends of roadway sections. It was also observed that the average variation of speed difference across adjacent lanes doesn't have direct impact on crashes and hence was eliminated from the model.

The prediction models in both studies relied upon the log-linear models developed in the past to estimate crash frequencies on freeways using the aggregate measures of traffic flow variables. The main difference being that they determined the crash precursors included in the model in an objective manner and not based on their subjective categorization. In one of their most recent related studies Lee et al. (2004) proposed the application of these models and estimated real-time crash potential. The main focus of this study was to reduce the crash potential obtained from the model through different control strategies of variable speed limits (*VSL*). To mimic responses from the drivers to changes in speed limits, microscopic simulation tool *PARAMICS* was used. At least on the simulated data the VSL showed significant safety benefits measured in terms of reduction in crash potential estimated from their model.

Oh et al. (2001) showed that five minutes standard deviation of 30-second speed measurements was the best indicator of "disruptive" traffic flow leading to a crash as

opposed to “normal” traffic flow. They used the Bayesian classifier to categorize the two possible traffic flow conditions. Since Bayesian classifier requires probability distribution function for each competing class, the standard deviations of speed over crash and non-crash cases were used to fit non-parametric distribution functions using Kernel smoothing techniques. The potential application of the model in real-time was also demonstrated.

A more detailed analysis of patterns in crash characteristics as a function of real-time traffic flow was done by Golob and Recker (2001, 2004). The methodology used was non-linear (nonparametric) canonical correlation analysis (NLCCA) with three sets of variables. The first set comprised a seven-category segmentation variable defining lighting and weather conditions; the second set was made up of crash characteristics (collision type, location and severity); and the third set consisted of real-time traffic flow variables. Since NLCAA requires reducing collinearity in the data, principal component analysis (PCA) was performed to identify relatively independent measurements of traffic flow conditions. The results of the PCA are shown in Table 2-1.

Table 2-1: Interpretation of principal components and variable selection (Golob and Recker, 2001)

Factor	Interpretation	Represented by
1	Central tendency of speed	Median volume/occupancy interior lane
2	Central tendency of volume	Mean volume left lane
3	Temporal variation in volume—Left and interior lanes	Variation in volume for left lane
4	Temporal variation in speed—Left and interior lanes	Variation in volume/occupancy interior lane
5	Temporal variation in speed—Right lane	Variation in volume/occupancy right lane
6	Temporal variation in volume—Right lane	Variation in volume right lane

It was concluded that the collision type is the best-explained characteristic and is related to the median speed, and to left-lane and interior lane variations in speed. Moreover the severity of the crash tracks the inverse of the traffic volume, and is influenced more by volume than the speed.

Based on these results, in one of their later study (Golob et al., 2004 (b)) used data for more than 1000 crashes over six major freeways in Orange County , California and developed a software tool *FITS* (Flow Impacts on Traffic Safety) to forecast type of crashes that are most likely to occur for the flow conditions being monitored. A case study application of this tool on a section of *SR 55* was also demonstrated.

Golob et al. (2004, a) also showed that certain traffic flow regimes are more conducive to traffic crashes than the others. Of the eight traffic flow regimes found to exist on the six freeways in Orange County (California), the study found that nearly 76% of all crashes occurred in the four traffic regimes that represent flow nearing or at congestion. This displays a correlation between the types of flow and crashes and indicates that understanding the patterns in real-time traffic flow might be the key to 'predict' crashes on urban freeways. It should be noted that none of the studies by Golob et al. (2004 (a), 2004 (b)) included non-crash loop data as a measure of 'normal' traffic conditions.

This link between traffic congestion and freeway crashes was also noted by Zhang et al. (2005) in a study that explored the relationship between crashes, weather conditions, and traffic congestion. The study showed that the relationship between the "Relative Risk Ratio" (a measure of crash probability) resembles a U-shaped curve with a peak value during moderate congestion and low points at free flow and heavy congestion.

Park and Ritchie (2004) showed that the lane-changing behavior and presence of long vehicles with-in a freeway section has significant impact on section speed variability. The section speed variance rather than the point speed variance was used to demonstrate the traffic changes more efficiently. The traffic data for their study were not obtained from more conventional single or dual loop detectors. A state-of-the-art vehicle-signature based traffic monitoring technology providing individual vehicle trajectories as well as accurate vehicle classification was used, instead.

While almost all studies have indicated a relationship between crash occurrence and speed variability, a recent study by Kockelman and Ma (2004) found no evidence to the fact that speeds measured as 30-second time series or their variations trigger crashes. The study was conducted for the same area as Golob and Recker (2004). Their sample size was limited to 55 severe crashes that occurred during January 1998 and with such a small sample their conclusions remain suspect. Similarly, Ishak and Alecsandru (2005) were unable to separate pre-incident, post-incident, and non-incident traffic regimes from each other and it was indicated that conditions before a crash might not be discernible in real-time. The study was performed using part of the ITS-archived data from the Interstate 4 in Orlando, Florida that has been used in this research as well. However, data for only 116 crashes were used which raises concerns about the validity of the findings from this research.

Our group at University of Central Florida (UCF) has also been actively involved in research linking crash patterns with loop detector data. Various modeling methodologies have been explored e.g., Probabilistic neural network (PNN) (Abdel-Aty and Pande, 2005), matched case-control Logistic Regression (Abdel-Aty et al. 2004), multi-layer perceptron (MLP)/radial basis function (RBF) neural network architectures (Pande, 2003) and Generalized Estimation Equation (Abdel-Aty and Abdalla, 2004). The data for these studies were collected from 13.2-mile central corridor of Interstate-4 in Orlando. All these studies made significant contributions towards enriching the literature. However, as explained later in this chapter, it must be acknowledged that there remains sufficient scope of improvement.

2.2.3 Critical review

It is evident that the idea of exploring the loop data in traffic safety research is still in its preliminary stages. Some of the aforementioned studies do have a potential application in the field of real-time proactive traffic management, but they have not fully analyzed the "recipe" of crashes. This is besides the fact that the statistical analysis in some cases isn't really sound from a theoretical point of view.

The research conducted in Canada (Lee et al. 2002, 2003) has the advantage over other research groups with dual loops placed closer to each other (38 loops on 10 km stretch of the freeway). Their analysis is based on a log-linear crash frequency model, which is not best suited for real-time classification of the loop data patterns.

Golob and Recker (2001) have established sound statistical links between environmental factors, traffic flow as obtained from loop data, and crash occurrence but their findings are limited by the fact that the traffic data is obtained from single loop detectors and speed has to be estimated using a proportional variable (volume/occupancy). The *FITS* tool developed by Golob et al. (2004, b) has limited application due to a systematic pattern of missing values within the data used for development of this tool. Also, the geometric characteristics of the freeways in the study area are not considered by this tool.

The classification model developed by Oh et al. (2001) seems to have most promising online application, also demonstrated in the paper, but due to lack of crash data (only 52 crashes) their model remains far from being implemented in the field. The only factor

used for classification is the 5-minute standard deviation of speed, other significant factors such as geometry and other traffic flow variables were not considered. It is also to be understood that if a crash prediction model has to be useful we need to classify the data much ahead of the crash occurrence time and not just 5-minutes prior to the crash so that Regional Transportation Management Center (*RTMC*) has some time for analysis, prediction and dissemination of the information.

Lack of crash and traffic data is what causes concerns about the findings by Ishak and Alecsandru (2005) as well. In the study pre-incident, post-incident, and non-incident traffic flow regimes were described by 30-second average speed and its variation depicted through spatio-temporal contour charts. Using second-order statistical analyses, the charts were measured for smoothness, homogeneity, and randomness. No consistent pattern for any of the statistical measures was found within three different categories of traffic regimes (i.e., the pre-incident, post-incident, and non-incident). Therefore, it was concluded that conditions belonging to these regimes could not be differentiated from each other based on loop data. However, only 116 crashes were used in the analysis with speed and its variation as the only independent parameters. It is likely that more crash and non-crash data along with different flow parameters from a range of stations located around crash locations would have yielded better results towards separating these three distinct traffic regimes. The findings from some of the previous studies by Abdel-Aty et al. (2004; differentiating pre-crash from non-crash) and Ishak and Al-Deek. (1999; separating post-incident data from non-incident) that used the loop data from same corridor make this postulation all the more plausible.

Besides the lack of data, there are certain key issues, which are either have been completely overlooked or proper attention has not been given to them. One of them is the determination of exact time of the crash. Except for Lee et al. (2003) all the studies have either relied on the police reports or at the most visual inspection of the loop data plots. Even the algorithm developed by Lee et al. (2003) has errors associated with shock-wave progression speed. Also, any freeway crash predictive model cannot be implemented without inclusion of the geometry of freeway section. None of the studies reviewed, except for Abdel-Aty et al. (2004) has accounted for horizontal curvature on freeways in their analysis even though the influence of curvature on freeway crashes is well understood and documented. None of the studies except for those done at UCF have analyzed data from series of loop detectors in order to examine progression of crash prone conditions on freeways.

Park and Ritchie (2004) proved the effect of lane-changing on speed variation (and thereby on crash occurrence) using data obtained through a sophisticated traffic surveillance technology. Since this technology is not going to be available on most freeways in near future, the association of lane-changing related variables derived from conventional dual loop detector data with specific type of crash occurrences would be worth examining.

Neither of the studies has incorporated driver population characteristics in a crash 'prediction' framework. Another point worth mentioning is that most of the studies

including our previous research work have focused on development of a generic crash identification model. Although we achieved satisfactory classification accuracy for 13.2-mile dense urban section of I-4, a careful analysis of results from those models clearly showed the relationship between classification accuracy and type/time (of day) of the crashes. The need for crash-type specific models should also be seen in the background of findings from the study by Golob et al. (2004, a) which concluded that certain crash types are more likely under certain traffic flow conditions.

The critical review presented here shows a sufficient scope of improvement not only in terms of data analysis but in data assembly related issues (e.g., incorporation of driver characteristics, etc.) as well. In this study we attempt to address these issues by analyzing freeway traffic and geometric design data by segregating them by crash-type and examine them against a sample of non-crash data representing "normal" conditions on the freeway. The data preparation and related issues are discussed in detail in Chapter 4.

2.3 Applications of Data Mining/Neural Networks in Transportation

Data mining is defined as the process of extracting valid, previously unknown and ultimately comprehensive information from large databases (Hand et al., 2001). Over the years data mining has emerged as a powerful new instrument offering value across a broad spectrum of information intensive industries involving huge amounts of data including banking, logistics, etc. The potential of various data mining techniques in the field of transportation engineering, however, remains under utilized with the exception of neural network applications for incident detection.

The neural networks attempt to achieve good performance through dense interconnection of simple computational elements (i.e., neurons). Of all neural network applications in transportation engineering, the "incident detection" algorithms are the most relevant to our research problem, since detecting an incident also involves classification of traffic flow patterns emanating from loop detectors. The critical distinction being that while we are interested in 'pre-crash' data; detection algorithms involve analysis of 'post-incident' loop data. In the following section neural network based incident detection algorithms are reviewed.

2.3.1 Incident detection: neural networks based algorithms

Cheu and Ritchie (1995) developed three types of neural network models, namely multi-layer feed forward (MLF), the self-organizing feature map (SOFM) and adaptive resonance theory 2 (ART2) to classify traffic data obtained from the loop detectors with the objective of using the classified output to detect lane-blocking freeway incidents.

The ANNs (Artificial neural network models) were designed to classify the input data into one of the two states, an incident or incident-free condition. ANN models were trained using post-incident loop detector data generated from INTRAS, a microscopic traffic simulation model, as according to the authors would have been impractical to put extensive effort in collecting real life data. INTRAS initially generated the incident and incident free input vectors in a ratio of 1:4. The incident input vectors were later replicated to make the number of state 1 and state 2 vectors equal in the training data set. The input vectors used were 16-dimensional, consisting of upstream and downstream

detectors' volume and occupancy at 30 seconds slices after the time of the incident. Based on the performance of these networks on field evaluation data they reported that MLP (multi-layer perceptron) neural networks always produced consistently better results than the other two networks and that these results were also better than the traditional detection algorithms.

Abdulhai and Ritchie (1999) tried to identify the requirements of a successful detection framework and found that inability to address the issues of predicted probability of incident occurrence is one of the major shortcomings of detection algorithms. They proposed the concept of statistical distance and a modified probabilistic neural network model (PNN2) in addition to Bayesian based traditional probabilistic neural network (PNN) model to detect the patterns in the loop data. They also reported that these two models were competitive with the more frequently used MLP neural networks for incident detection.

A study, which did not use simulation data and training and testing of the neural network models for incident detection was done through real-life loop data only, was conducted by Ishak and Al-Deek (1999). The data used by Ishak and Al-Deek (1999) were collected from part of the same Interstate-4 corridor for which the crash prediction models are being developed in this study. Input patterns of various dimensions were attempted and the network size was changed accordingly in order to achieve better performances. One of their interesting finding was that while using the MLF neural network the incidents might be detected better with the speed patterns alone rather than the using occupancy

patterns or a combination of speed-occupancy patterns. This observation is utilized later in Chapter 4 to examine the proximity of the reported time of crash with its actual time.

2.3.2 Data mining/neural network applications in traffic safety

A comparison between the fuzzy K-nearest neighbor algorithm and MLP neural network to identify crash-prone locations was made by Sayed and Abdelwahab (1998). Results showed that MLP produced slightly more accurate results and achieved higher computational efficiency than fuzzy classification.

Awad and Janson (1998) applied an MLP to model truck crashes at interchanges in Washington State. Results of the neural network were compared with a linear regression model. Comparison was based on the root mean square of error (RMSE). The trained neural network showed a better fit when the training data is presented. However, the ability of the trained ANN to predict "unseen" test data was unsatisfactory.

Mussone et al. (1999) adopted an MLP approach to analyze traffic crashes that occurred at intersections in Milan, Italy. Results showed that the neural network models could extract information, such as factors explaining crashes and contributing to a higher degree of danger.

Through a sequential review of literature it was observed that only neural network architecture explored for traffic safety analysis was MLP, until Abdelwahab and Abdel-Aty (2001) developed Fuzzy ART neural networks to predict driver injury severity in

traffic crashes at signalized intersections. These models were compared with the MLP architecture and it was concluded that MLP models were superior tools compared to ordered logit model and Fuzzy ART. In a later work by same authors (Abdelwahab and Abdel-Aty, 2002) ANN models were used for traffic safety analysis of toll plazas. Driver injury severity (no injury, possible injury, evident injury, severe injury/fatal crashes) and location of the crash (before plaza, at the plaza and after the plaza) were analyzed using MLP as well as radial basis function (RBF) neural network. They reported that for analyzing crash location the nested logit model was the best, while RBF neural network was the best model for driver injury severity analysis.

In the recent past data mining techniques other than neural networks have also figured in traffic safety literature. Vorko and Jovic (2000) used multiple attribute entropy models to classify school-age injuries. Shon and Shin (2001) employed neural networks and decision tree algorithms to develop classification models for road traffic crash severity (bodily injury or property damage) as a function of potentially correlated categorical factors. It was noticed that classification accuracy of the individual models from both algorithms was relatively low. It was noticed that the use of data fusion or ensemble algorithms were able to increase the classification accuracy. Data fusion techniques try to combine classification results obtained from several individual classifiers and are known to improve the classification accuracy when some results of relatively uncorrelated classifiers are combined. The resulting performance is usually more stable than that of a single classifier.

2.4 Conclusions from Literature Review

An extensive review of relevant literature conducted in this chapter demonstrates the applications, albeit limited so far, of ITS archived data and/or data mining techniques in the field of traffic safety.

The issues not addressed adequately by studies using real-time loop detector data for 'predicting' crashes, referred to by Golob et al. (2004, a) as disaggregate studies, have been thoroughly discussed in section 2.2.3. These issues include accuracy of the time of crash, no accounting for the geometric design parameters etc. and leave enough scope for continuing research in the area. The reason why these studies have overlooked various factors seems to be the effort involved in collection and fusion of the data for crash, traffic, geometric and driver population characteristics. The time and effort involved in gathering of sufficient crash data also happens to be the reason behind the fact that none of the models developed so far, to separate crashes from normal (non-crash) conditions, differentiate crashes by type. It is a major shortcoming considering that the traffic conditions preceding crashes are likely to differ by type of crash, e.g., the rear-end crashes might be expected to occur under congested traffic regime where the drivers have to slow down and speed up quite often, on the other hand the single vehicle crashes might result from excessive speeds on a curved freeway section. Although studies by Golob et al. (2004 (a), 2004 (b)) dwelled into this issue but without loop data representing non-crash traffic conditions their findings are inadequate to develop a system that can reliably separate crash prone traffic conditions from 'normal' ones.

This research is aimed at developing such a system while addressing the problems identified in the literature at the modeling and data preparation stage. A sufficiently large database with crash and non-crash data over five year period from 1999 through 2003 has been assembled for this study from the 36.25 instrumented Interstate-4 corridor in Orlando metropolitan area. Besides, geometric design parameters and factors to represent measures of the driver population composition are also included in the database. This extensive data is analyzed for different types of collisions through multiple classification algorithms belonging to the realm of data mining and/or traditional statistical models. Moreover, results from various modeling techniques are combined with each other in order to make the performance of the proposed system more reliable. In the next chapter these modeling techniques are described in the context of the present research problem.

CHAPTER 3

METHODOLOGY AND MODELING TECHNIQUES

3.1 General

Reliable models that can separate crash prone conditions from "normal" freeway traffic in real-time are the focus of this research. These models are envisioned to be the primary component of a proactive traffic management system. Factors critically associated with different types of crashes would be identified and used as inputs to the classification models in this research. An insight into these factors would also help in devising remedial measures for crash prone conditions. As mentioned earlier, loop detector data from stations surrounding the crash location are used here as a surrogate measure of traffic flow. These data are continuously collected and archived for various *ITS* (Intelligent Transportation Systems) applications such as travel time prediction etc. The archived data represents an "observational" dataset for this study. An "observational" dataset essentially means that the objectives of analysis have no bearing on the data collection strategy which happens to the case here with ITS related archived data. A data mining approach is usually recommended for such datasets and is adopted here too. Data mining is analysis of large "observational" datasets to find unsuspected relationships that might be useful to the data owner (Hand et al., 2001).

It essentially means that at various stages of this research tools from a range of fields such as machine learning (e.g., clustering algorithms), statistics (e.g., classification tree), and artificial intelligence (e.g., neural networks) have been used in a step by step manner.

Classification tree would be used as the data preparation tool for identification of critical variables which would be used as inputs to the neural network models in subsequent steps. The neural network based algorithms explored in this study include multi-layer perceptron (MLP), normalized radial basis function (NRBF) and probabilistic neural networks (PNN). In addition, with-in stratum matched case control logistic regression, a statistical technique borrowed from epidemiological studies, has also been explored for classification of real-time loop detector data patterns. In this chapter theoretical overview of these data mining and statistical techniques is provided in context of their application to this research.

3.2 Data Mining Methodologies

3.2.1 Decision tree based classification and its application for variable selection

A classification tree represents segmentation of data created by applying a series of simple rules. Each rule assigns an observation to a group based on the value of an input. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final or terminal nodes are called leaves. For each leaf, a decision is made and applied to all observations in that leaf. Decision trees are the most widely utilized tools in data mining applications besides neural networks and may be used for classification of binary variables as well as for continuous target. The later application, of course, is not relevant here. The advantage of classification tree over other modeling tools, such as neural networks, is that it produces a model that may represent interpretable English rules or logic statements. The other advantage associated with trees is that no assumptions are necessary about the data and

the model form which makes them an excellent data exploration tool. Classification trees can be used to automatically rank the input variables based on the strength of their contribution to the tree. This ranking may be used for variable selection in subsequent modeling procedures such as the neural networks. In this section theoretical details of the classification trees are described along with the variable selection procedure. Since we would invariably deal with binary target variable in this study the details of the methodology are provided in the context of a binary target.

3.2.1.1 Decision tree methodology for binary classification

The basic element in classification tree construction is to split each (non-terminal) node such that the descendent nodes are 'purer' than the parent node. To achieve this, a set of candidate split rules is created, which consists of all possible splits for all variables included in the analysis. For example, for a dataset with 215 observations and 19 input variables there would be 215*19=4085 splits available at the root node. These splits are then evaluated based on a criterion to choose amongst various available splits at every non-terminal node (including the root node). There are three measures (i.e., 'purity' functions) which may be used to rank candidate splits for a binary target variable:

1. Chi-square test – This criterion uses *–log (p-value)* measure with the p-value corresponding to the Pearson contingency table chi-sq. test of independence between the binary target and the ordinal variable resulting from the split. Of course, the best split is the one with smallest p-value. A significance threshold may be specified so as to compare only a limited number of splits for computational efficiency.

2. Entropy reduction -- The entropy measure of node impurity as a split criterion for classification tree was first proposed by Quinlan (1993). The entropy function is zero for a split resulting in 'pure' child node, i.e., a node that only consists of observations belonging to one particular class. For a given node, the predictor and splits are chosen, from all predictors and all admissible splits, that maximize the impurity reduction between the parent node and its descendents.
3. Gini reduction -- The application of reduction in Gini index, which essentially is a measure of variability in categorical data, as a measure of split criteria was proposed by Brienamn et al. (1984).

One of these criteria is applied recursively to the descendents, which become the parents to successive splits, and so on. The splitting process is continued until the criteria of minimum reduction in impurity and/or minimum size of a node are satisfied. To stop the splitting process one may also choose the classification accuracy over the validation dataset (i.e., the dataset not used for estimating the splits) as the criterion. The classification accuracy may be assessed after every split and the process may be terminated if the classification accuracy declines after a particular split.

In this study all three measures of assessing candidate splits would be employed one at a time and compared with each other. It should be noted that the classification tree would not be used to develop final models but for data exploration and variable selection for

other modeling tools. Hence, the criterion resulting in most comprehensive list of variables may be selected for application.

3.2.1.2 Application of classification trees for variable selection

Brieman et al. (1984) devised a variable importance measure (VIM) for trees. This measure may be applied as a criterion to select a promising subset of variables for other modeling tools, especially for other flexible modeling tools such as neural network. Let $s(x_j, k)$ be the split at the k^{th} internal node using the variable x_j. The variable importance measure for variable x_j is the weighted average of the reduction in the Gini impurity for all splits using variable x_j across all internal nodes of the tree and the weight is the node size. The formula for the importance for variable x_j may be given by the following:

$$VIM\left(x_j\right) = \sum_{t=1}^{T} \frac{n_t}{N} \Delta Gini\left(s\left(x_j, t\right)\right) \qquad (1)$$

where T is the total number of nodes in the tree and N is the total size of the training sample. The formula above depicts the variable importance measure in its raw form as proposed by Brieman et al. (1984). In this study, however, the VIM used has been scaled by maximum importance for the tree so that it lies between 0 and 1. One may conveniently use a threshold of 0.05 on VIM to separate variables critically associated with the target from variables that are not. These critical variables can then be used as inputs to the classification models in subsequent steps. Moreover, a closer examination of the rules, based on which the VIM is calculated, also provides insight into crash precursors and their association with crash occurrence of a specific type.

3.2.2 MLP neural network architecture

A neural network may be defined as a massively parallel-distributed processor made up of simple processing units having natural propensity for storing experimental knowledge and making it available to use. The ability to learn and generalize provides neural networks with the computing power it possesses. Generalization refers to the ability of a "trained" network to provide satisfactory responses even for the inputs that it has not seen during the training process. Neural network models may usually be specified by three entities, namely; model of processing elements themselves, model of interconnections and structures (i.e. network topology), and the learning rules. In this section we describe multi-layer perceptron (*MLP*) that is one of the most commonly used neural network architectures.

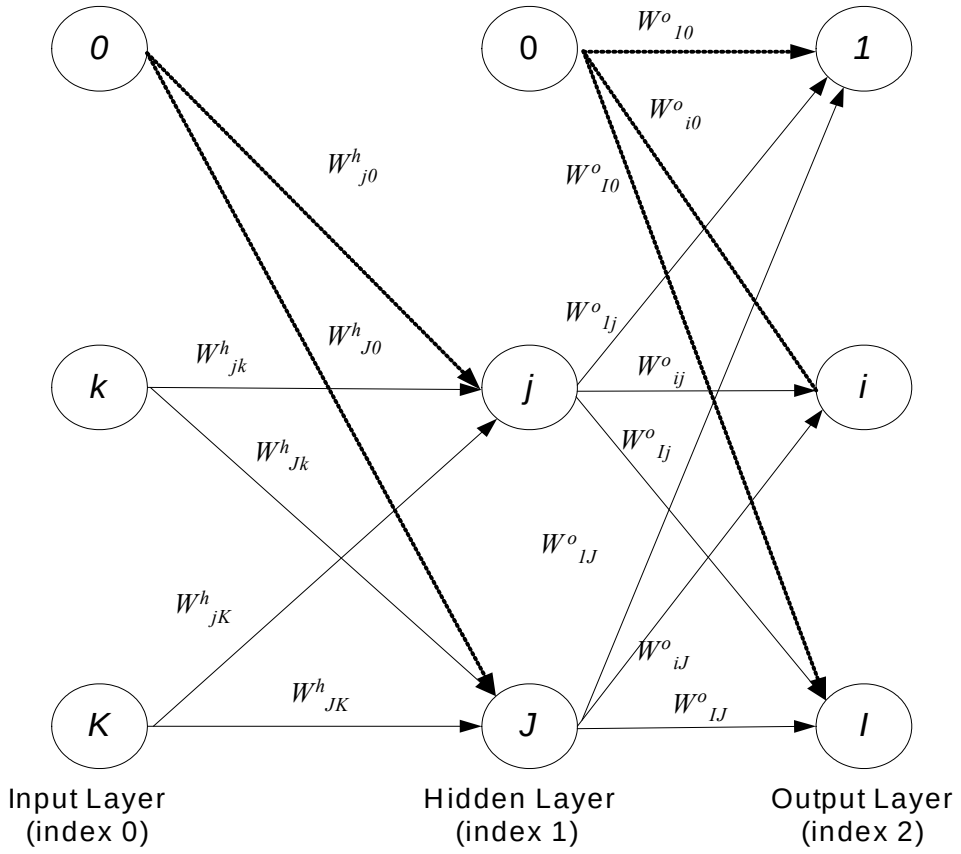

Figure 3-1: MLP neural network architecture

An *MLP* neural network shown in Figure 3-1 has input layer of size *K*, a hidden layer of size *J* and output layer of size I along with input and output bias. In the *MLP* architecture shown here the connections are of feed-forward type; it means that the only connections allowed between nodes are from a layer of a certain index to layers of higher index. The net input to hidden layer neurons is determined through inner product between the vector of connection weights and the inputs. The activation function is applied to this net input of hidden neurons and the weights from hidden to output layer are then used to get the output of the network. These weights are the parameter estimated during the supervised training process and are then used to 'score' unseen observations. The activation function of hidden neurons is non-linear in nature and is critical in the functioning of the neural network for it allows the network to 'learn' any underlying relationship of interest between inputs and outputs. Of course the procedure adopted for training is also crucial in performance of a neural network.

3.2.2.1 Training of MLP-NN: *Levenberg-Maraquardt (LM) algorithm*

Training the neural network is essentially an exercise in numerical optimization of a non-linear function. Error Back-propagation (EBP) algorithm proposed by Rumelhart et al. (1986) has been a significant milestone in neural network literature and still remains the most widely used supervised training algorithm. It however has been known to have a poor convergence rate for more complex problems (Wilamowski et al., 2001). A significant improvement on realization performance may be achieved by using second order approaches such as the Levenberg-Maraquardt (LM) optimization technique. The LM algorithm is widely accepted as the most efficient algorithm in terms of realization

accuracy as it provides a good balance between speed of Newton algorithm and stability of EBP algorithm (Hagan and Mehnaj, 1994). The only problem with this training algorithm, as pointed out by Wilamowski et al., (2001), is that with increase in number of independent variables the computational complexity grows exponentially. The details of the algorithm would make this point clearer and are provided below. For LM algorithm the objective function takes the following form:

$$F(w) = \sum_{p=1}^{P}\left[\sum_{k=1}^{K}(d_{kp} - o_{kp})^2\right] \qquad (2)$$

where $\boldsymbol{w} = [w_1 \;\; w_2 \;\; \ldots \;\; w_N]^T$ consists of the interconnection weights in the network, d_{kp} and o_{kp} are the desired and actual values, respectively, for k_{th} output and p_{th} pattern. N is the total number of weights, P is the number of patterns, and K is the number of network outputs. The above equation may be rewritten as

$$F(w) = E^T E \qquad (3)$$

$\mathbf{E} = [e_{11} \ldots e_{K1}\, e_{12} \ldots e_{K2} \ldots e_{1P} \ldots e_{KP}]^T, \qquad e_{kp}=d_{kp}-o_{kp} \qquad k=1, \ldots, K, \qquad p=1, \ldots, P$

where **E** is the cumulative error vector (for all patterns). Based on Equation 3 the Jacobian matrix is defined as

$$J = \begin{bmatrix} \frac{\partial e_{11}}{\partial w_1} & \frac{\partial e_{11}}{\partial w_2} & \cdots & \frac{\partial e_{11}}{\partial w_N} \\ \frac{\partial e_{21}}{\partial w_1} & \frac{\partial e_{21}}{\partial w_2} & \cdots & \frac{\partial e_{21}}{\partial w_N} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial e_{KP}}{\partial w_1} & \frac{\partial e_{KP}}{\partial w_2} & \cdots & \frac{\partial e_{KP}}{\partial w_N} \end{bmatrix} \qquad (4)$$

and the weights are adjusted using the following equation

$$w_{t+1} = w_t - \left(J_t^T J_t - \lambda_t I\right)^{-1} J_t^T E_t \qquad (5)$$

where $\boldsymbol{I}$ is the identity unit matrix, λ is the learning parameter and $\boldsymbol{J}$ is the Jacobian of m output errors with respect to the N weights of the neural network. It should be noted that if $\lambda=0$ then the above equation becomes the Gaussian-Newton method while for very large λ algorithm is equivalent to the EBP. The learning parameter is automatically adjusted at every iteration in order to secure convergence. The algorithm requires computation of Jacobian matrix and inversion of the $\boldsymbol{J^T J}$ matrix at each iteration step. Since the dimension of the matrix to be inverted is $\boldsymbol{N}$x$\boldsymbol{N}$, for large size neural networks the LM algorithm is not practical. In this regard Wilamowski et al. (2001) proposed modifications to the LM algorithm to avoid these impracticalities and make it more stable.

It can be argued, however, that in this research we have a reliable classification tree based variable selection algorithm due to which only the significant variables would be used as inputs to the neural networks. It would control the size of the network and hence in this study we could work with the original form of the LM algorithm to train the networks.

3.2.3 Radial basis function (RBF) neural network

3.2.3.1 Architectural issues

In feed forward neural network architectures the activation function of hidden neurons is applied to a net single value that is obtained by combining input vectors with the vector of connection weights between input layer to hidden layer. The function that combines

the inputs with the weights may be referred to as the ‘combination function’. In the MLP neural network architecture the combination function was simply the inner product of the inputs and weights. A radial basis function (RBF) network is a feed forward network with a single hidden layer for which the ‘combination function’ is more complex and is based on a distance function (referred to as width) between the input and the weight vector. Ordinary RBF (ORBF) networks using radial combination function and exponential activation function are universal approximators in theory (Powell, 1987), but in practice they are often ineffective estimators of the multivariate function. The individual basis functions have a local effect around their center while for the MLP neural networks the effect is distributed across the input space. Due to the localized effect the ORBF neural networks often require an enormous number of hidden units to avoid an unnecessarily bumpy fit.

To avoid the pitfalls of ORBF networks, softmax activation function may be used. It essentially normalizes the exponential activations of all hidden units to sum to one. This type of network is called a "normalized RBF" or NRBF network. Note that the output bias has no role in an NRBF network since the constant bias term would be linearly dependent on the constant sum of the hidden units due to the softmax activation. The distinction and advantages of NRBF networks over the ORBFs are discussed in detail by Tao (1993). It was argued by Tao (1993) that the normalization not only is a desirable option but in fact is imperative.

In NRBF networks one may add another term to the Gaussian combination function referred to as the 'altitude' which determines the maximum height of the Gaussian curve over the horizontal axis. Based on the two parameters (width and height) defining the shape of combination function the NRBF networks may be categorized into five different types:

1 NRBFUN: Normalized RBF network with unequal widths and heights

2 NRBFEV: Normalized RBF network with equal volumes $(a_i=w_i)$

3 NRBFEH: Normalized RBF network with equal heights (and unequal widths) $(a_i=a_j)$

4 NRBFEW: Normalized RBF network with equal widths (and unequal heights) $(w_i=w_j)$

5 NRBFEQ: Normalized RBF network with equal widths and heights $(a_i=a_j)$ and $(w_i=w_j)$

where w_i *and* a_i represent the widths and heights, respectively, of the neurons in the hidden layer. Note that the last four categories of networks are special cases of the first and are more parsimonious in nature. It essentially means that with certain assumptions about the shape of the combination functions they reduce the number of parameters to be estimated.

In this research the networks from the first category would be used. Even though these networks need to calibrate more parameters (the connection weights as well as height and altitude) through the training process; they are preferred over other architectures since no assumptions are needed. Note that the discussion so far has been on the architectural

issues pertaining to NRBF networks. While comparing various architectures training issues must be separated from architectural issues to avoid the most common sources of confusion in the understanding of neural networks. In the next section a discussion on training process used for NRBF networks is provided.

3.2.3.2 Training procedure for NRBF networks

The NRBF networks may be trained by "hybrid" methods, in which the hidden weights (centers) are first obtained by unsupervised learning and then the output weights are obtained by supervised learning. In an unsupervised method for choosing the center a random subset of training cases are first selected to serve as centers. The training cases may then be clustered based on the value of input variables and the mean of cluster centers can be used as the center. Heuristic methods proposed by Moody and Darken (1989) can be employed to estimate the widths of these RBF centers. Once the centers and widths are estimated, the output weights can be learned very efficiently, since the computation reduces to a linear or generalized linear model. The hybrid training approach can thus be much faster than the nonlinear optimization (e.g., LM algorithm described in the previous section) that would be required for supervised training of all of the weights in the network. However, note that the supervised training algorithm would optimize the locations of the centers, while hybrid training wouldn't. Hence the Hybrid training will usually require more hidden units than supervised training. For a given number of hidden units supervised training will provide a better approximation for the underlying function to be learned. Thus, the supervised training will often let one use fewer hidden units (with a fewer training cases) for a given accuracy of approximation than the hybrid training

(Tarassenko and Roberts, 1994). Moreover, the number of hidden units required by hybrid methods becomes an increasingly serious problem as the number of inputs increase. To escape from this 'curse of dimensionality' fully supervised training methods are adopted for the NRBF networks as well. Note that for the MLP networks the unsupervised (or hybrid training) was not an option. Supervised training for RBF networks can be is accomplished using Levenberg-Marquardt algorithm which would also be used for the MLP neural network architecture.

3.2.4 Theoretical background of the PNN

The Probabilistic Neural Network (*PNN*) is a neural network implementation of the well-established multivariate Bayesian classifier, using Parzen estimators to construct the probability density functions of different classes (Specht, 1996). One can think of PNN as an RBF network in which there is a hidden unit centered at every training case.

3.2.4.1 Parzen estimator

Parzen estimator uses the weight function *W(d)* (frequently referred to as potential function or a kernel) having largest value at *d=0* and it decreases rapidly as the absolute value of *"d"* increases. The weight functions are centered at each training sample point with the value of each sample's function at a given abscissa is being determined by the distance *"d"* between *x* and that sample point. The *PDF* estimator is the scaled sum of that function for all the sample cases. The method can be stated mathematically using the following equation:

$$g(x) = \frac{1}{n\sigma} \sum_{i=1}^{n} W\left(\frac{x - x_i}{\sigma}\right) \qquad (6)$$

The scaling parameter σ defines the width of the bell curve that surrounds each sample point. This parameter has a profound influence on performance of a *PNN*. While the too small values will cause individual training cases to have too much of an influence, losing the benefit of aggregate information, the large values will cause so much blurring that the details of density will be lost (Masters, 1995).

3.2.4.2 Multivariate Bayesian discrimination and PNN

The objective of the *PNN* is to separate classes of objects, i.e. define the boundaries between the existing classes and classify new objects to one of the existing classes. A vector in a *p*-dimensional input space, where *p* is the number of features or variables, defines an object. In this section the mathematics for the case of two competing classes is explained.

Let $f_1(x)$ and $f_2(x)$ be the probability density functions (*PDFs*) associated with the p-dimensional input vector *X* for the populations' p_1 and p_2, respectively. A reasonable classification rule that minimizes the expected cost of misclassification (*ECM*) is to assign a new vector to either class π_1 or class π_2 based on the density ratio, the misclassification cost ratio and the prior probability ratio as follows:

X belongs to:

π_1 if $f_1(x)/f_2(x) \geq \{[C(1|2)/C(2|1)]*[P_2/P_1]\}$

π_2 otherwise, (7)

Where:

C (i|j) is the cost of misclassifying an object as belonging to population π_i while it belongs to population π_j

P_i is the prior probability of occurrence of population π_i.

The key for using the above classification rule is the ability to estimate the *PDFs* based on training patterns. Typically, the a priori probability can be estimated, and the cost ratio can be estimated either subjectively or objectively.

The accuracy of the decision boundaries' estimation and the subsequent classification depends on the accuracy with which the underlying *PDFs* are estimated. A nice feature of this approach and the related *PNN* implementation is estimation consistency. Consistency implies that the error in estimating the *PDF* from a limited sample gets smaller as the sample size increases. The estimated *PDF* (the class estimator) collapses on the unknown true *PDF* as more patterns in the sample become available.

An example of the Parzen estimation of the *PDFs* (described in the preceding section) is given below for the special case that the multivariate kernel is a product of the univariate kernels. In the case of the Gaussian kernel, the multivariate estimates can be expressed as:

$$f_k(X) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{m}\sum_{i=1}^{m}\exp\left[\frac{-(X-X_{ki})^T(X-X_{ki})}{2\sigma^2}\right] \quad (8)$$

where k is the class or category; i the pattern number; m the total number of training patterns; X_{ki} the ith training pattern from category or population π_k; σ the smoothing parameter and p the dimensionality of feature (input) space. Note that the estimated *PDF* for a given class, say $f_1(x)$, is the sum of small multivariate Gaussian distributions centered at each training sample. However, the sum is not necessarily Gaussian. It can, in fact, approximate any smooth density function. The smoothing factor σ can alter the resulting *PDF*. Larger values of σ causes a vector X to have about the same probability of occurrence as the nearest training vector. The optimal σ can be easily found experimentally.

An interesting feature of the *PNN* approach is that the estimated *PDFs* can be used not only for classification but also to estimate the posterior probability that a vector X belongs to class π_i. If the classes are mutually exclusive, we have from Bayes theorem:

$$P[\pi_1 \mid X] = \frac{P_1 f_1(X)}{P_1 f_1(X) + P_2 f_2(X)} \tag{9}$$

Also the maximum of $f_1(x)$ and $f_2(x)$ is a measure of the density of the training samples in the vicinity of X which can be used to indicate the reliability of the classification.

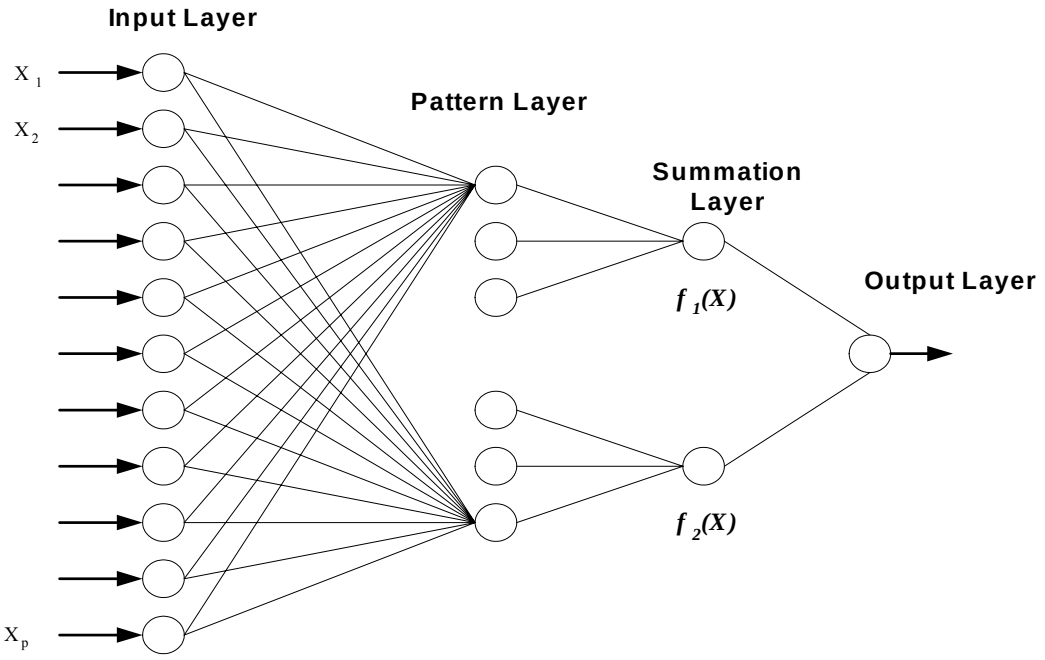

Figure 3-2: The traditional PNN architecture for a two-class classification problem

The original neural network implementation of the above theory (Specht, 1996) is shown in Figure 3-2 for a two-class classification problem. The input units are merely distribution units that supply the same input values to all of the pattern units. Each pattern unit forms a dot product of the new incoming input pattern vector X with one exemplar pattern i stored as a weight vector W_i such that $Z_i = X.W_i$, and then performs a nonlinear operation on Z_i before outputting its activation level to the summation unit. Instead of the sigmoid activation function commonly used for the *MLP* neural networks the nonlinear operation used here is $exp\ [(Z_i-1)/\ \sigma^2]$. Assuming that both X and W_i are normalized to a unit length, this is equivalent to using $exp\ [-(W_i-X)^T\ (W_i-X)/2\sigma^2]$ and since all inputs to the classifier have norm 1, both the dot products X^TX and W^TW equal unity and the exponential term reduces to $exp\ [(X^TW_i-1)/\ \sigma^2]$. Each summation unit sums the outputs

from the pattern units that correspond to one of the classes. The output layer in the traditional PNN architecture act as the threshold discriminator and the test case is assigned the class corresponding to which the output of the summation layer is maximum (Abdulhai and Ritchie, 1999).

3.3 Methodology from Statistical Background

3.3.1 Simple logistic regression and hazard ratio

In a logistic regression setting the function of dependent variables yielding a linear function of the independent variables would be the logit transformation.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \qquad (10)$$

Where $\pi(x) = E(Y|x)$ *is* the conditional mean of *Y* (dummy variable representing crash occurrence in our case) given *x* when the logistic distribution is used. Under the assumption that the logit is linear in continuous covariate *x,* the equation for the logit would be $g(x) = \beta_0 + \beta_1(x)$. It follows that the slope coefficient, β_1, gives the change in the log odds for an increase of 1 unit in *x*, i.e. $\beta_1 = g(x+1) - g(x)$ for any value of *x*. Hazard ratio is defined as the exponential of this coefficient, in other words it represents how much more likely (or unlikely) it is for the outcome to be present for an increase of "1" unit in *x* (Agresti, 2002).

3.3.1.1 Within stratum matched case-control sampling

The matched case-control sampling technique has been adopted from epidemiological studies. The purpose of the matched case-control analysis is to explore the effects of

independent variables of interest on the binary outcome while controlling for other confounding variables through design of the study. In this section application of a multivariate logistic regression model in a within stratum matched sampling framework has been described in the context of present research problem.

If there are N strata with 1 case and m controls in stratum j, $j = 1, 2, \ldots N$. The conditional likelihood for the j^{th} stratum is the probability of the observed data given the total number of observations and the number of crashes observed in the stratum. Let $p_j(x_{ij})$ be the probability that the i^{th} observation in the j^{th} stratum is a crash where $x_{ij} = (x_{1ij}, x_{2ij}, \ldots x_{kij})$ is the vector of k traffic flow variables $x_1, x_2, \ldots x_k$; $i = 0, 1, 2, \ldots m$; and $j = 1, 2, \ldots N$. This crash probability $p_j(x_{ij})$ may be modeled using a linear logistic model as follows:

$$logit\ (p_j(x_{ij})) = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots\ldots + \beta_k x_{kij} \qquad (11)$$

Note, that the intercept term α is different for different strata. It summarizes the effect of variables used to form strata on the probability of crash. In order to take account of the stratification in the analysis of the observed data, one constructs a conditional likelihood. This conditional likelihood function is the product of N terms, each of which is the conditional probability that the crash in a particular strata, say the j^{th} strata, is the one with explanatory variables x_{0j}, conditional on $x_{0j}, x_{1j}, \ldots x_{mj}$ being the vectors of explanatory variables in the j^{th} stratum. The mathematical derivation of the relevant likelihood function is quite complex and is omitted here. The reader may consult Collett (1991) for full derivation of the conditional likelihood function that can be expressed as:

$$L(\beta) = \prod_{j=1}^{N} [1 + \sum_{i=1}^{m} \exp\{\sum_{u=1}^{k} \beta_u (x_{uij} - x_{u0j})\}]^{-1} \quad (12)$$

where, β*'s* are the same as in Equation 11. The likelihood function *L* (β) is independent of the intercept terms $\alpha_1, \alpha_2, \ldots\ldots \alpha_N$. So the effects of matching variables cannot be estimated and hence Equation 11 cannot be used to estimate crash probabilities. However, note that the values of the β parameters that maximize the likelihood function given by Equation 12 are also estimates of β coefficients in Equation 11. These estimates are log odds ratios and can be used to approximate the relative risk of a crash.

The log odds ratios can also be used for crash prediction under this matched case-control framework. Consider two observation vectors $x_{1j} = (x_{11j}, x_{21j}, \ldots\ldots, x_{k1j})$ and $x_{2j} = (x_{12j}, x_{22j}, \ldots\ldots, x_{k2j})$ from the j^{th} strata on the k traffic flow variables. Using Equation 11, one may verify that the log odds ratio of crash occurrence due to traffic flow vector x_{1j} relative to vector x_{2j}

$$\log\left\{\frac{p(x_{1j})/[1-p(x_{1j})]}{p(x_{2j})/[1-p(x_{2j})]}\right\} = \beta_1(x_{11j} - x_{12j}) + \beta_2(x_{21j} - x_{22j}) + \ldots\ldots\ldots + \beta_k(x_{k1j} - x_{k2j}) \quad (13)$$

The right hand side the equation above is independent of α_j and can be estimated using the β coefficients estimated through multivariate logistic regression. We may utilize the above relative log odds ratio for predicting crashes by replacing x_{2j} by the vector of values of the traffic flow variables in the j^{th} stratum under normal traffic conditions. One may conveniently use simple average of all control observations within that stratum for each variable. If $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j}, \ldots, \bar{x}_{k2j})$ denotes the vector of means for k variables over

control cases from j^{th} stratum, then the log odds of this case being a crash relative to the controls may be approximated by:

$$\log\left\{\frac{p(x_{1j})/[1-p(x_{1j})]}{p(\bar{x}_{2j})/[1-p(\bar{x}_{2j})]}\right\}=\beta_1(x_{11j}-\bar{x}_{12j})+\beta_2(x_{21j}-\bar{x}_{22j})+.....+\beta_p(x_{k1j}-\bar{x}_{k2j}) \qquad (14)$$

The above log odds ratio can then be used to 'predict' crashes by establishing a threshold value that yields that desirable crash classification accuracy (Abdel-Aty et. al, 2004). Note that the variables used in the logistic regression model would *not* be selected through the classification tree based variable selection procedure described earlier in the chapter. Standard stepwise variable selection method will be used instead. The details of stepwise variable selection procedure for logistic regression may be found in Hosner and Lemeshow (1989).

3.4 Summary

In this chapter theoretical details of the methodologies used to develop the models constituting the proposed real-time crash prediction system are provided. Among the data mining based techniques classification tree was discussed in the context of its application to the present work. Three categories of the neural network architectures, namely, the *MLP, NRBF* and *PNN* were discussed following the classification tree based variable selection algorithm. In addition to the tools belonging to data mining family, an epidemiological approach of within stratum sampling based logistic regression to classify loop data patterns into crash (case) and non-crash (controls) was described. In the next chapter an introduction to study area is provided along with the data preparation effort.

CHAPTER 4

DATA PREPARATION AND RELATED ISSUES

4.1 General

The final goal of this research is to develop a predictive system for crash occurrence on Interstate-4 corridor equipped with underground loop detectors. To achieve this objective we need to systematically correlate between the crash characteristics and the loop data (representing ambient traffic flow configuration). Moreover it has to be collated with the geometric design of the freeway at the location of the crash and the environmental conditions at the time of the crash. The system needs to recognize the patterns not leading to crash occurrence as well; hence traffic characteristics corresponding to selected "non-crash" cases or "normal" freeway operation must be a part of the database. Drivers belonging to certain groups are known to have high likelihood of being involved in crashes; therefore, a measure for driver population by freeway mile-post location, time of day and day of week should also be included in the database.

The traffic parameters in this study would be measured in terms of 30-seconds time series obtained from inductive loop detectors in the vicinity of the crash location for a period leading up to the crash. It is not difficult to realize the importance of properly fusing the loop detector data with crash data and geometric/environmental/driver related factors that might affect the probability of crash occurrence.

4.2 Introduction to the Study Area

The study is being conducted using data from Interstate-4 (I-4) corridor in Orlando. The corridor is considered to be an integral part of Central Florida's transportation system. It carries greater number of people and vehicles than any other facility in the region and serves many of the area's primary activity centers. Though originally designed to serve long distance travelers, the I-4 corridor now has evolved to one serving many shorter trips. No wonder a significant amount of growth in the region is occurring within close proximity to I-4. In recent years, congestion on I-4 has extended well beyond normal peak hours and major crashes have closed the freeway, subsequently resulting in traffic congestion throughout the Orlando metropolitan area. Hence, congestion and delays blended with very frequent crashes are the major transportation problems facing the freeway.

The freeway section under consideration is 36.25 miles long and is spread over three counties, namely Osceola, Orange and Seminole. It has a total of 69 loop detector stations, spaced out at nearly half a mile. Each of these stations consists of dual loops in each direction and measures average speed, occupancy and volume over 30 seconds period on each of the through travel lane. The loop detector data are continuously transmitted to the Regional Traffic Management Center (*RTMC*). The source of crash and geometric characteristics data for the freeway is *FDOT* (Florida Department of Transportation) intranet server.

4.3 Crash Data Collection

The first step was to collect crash data for the instrumented freeway corridor over a period of time. Since the loop detectors are known to suffer from intermittent failures it was likely that some of the crashes may not have corresponding loop data available. To ensure that loop data for sufficient number of crashes are available to establish reliable link between crash and traffic characteristics represented by loop data it was decided to be on the conservative side and collect crash data for a period of five years ranging from 1999 through 2003.

There were *4189* crashes reported in all during the five year period, while we expected some of them to have corresponding loop detector data missing, it was believed that we will be left with a sample large enough for analysis purposes. However, the information extracted for each crash case to create a complete crash database for is shown in Table 4-1. Part of the information in Table 4-1, i.e., the "Time of crash" and "Direction (EB or WB)" was not readily available in the FDOT database. To extract these fields every crash report was physically examine one by one.

Table 4-1 Crash characteristics

Crash Number	Crash report number	Direction (EB or WB)	Mile post	Date and Time of crash	First harmful event	Lane of the crash	Visibility on the roadway	Pavement Condition (Wet, slippery or dry)	Number of fatalities	Number of injuries
1	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx
2	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx
|	|	|	|	|	|	|	|	|	|	|
4189	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx

The table shown above provides information about each crash; the field "first harmful event" represents *type of the crash.* All other fields are self explanatory. The mile-post location for crashes in the FDOT database is actually the distance in miles from the beginning of a section/subsection within a county to the location of the crash. Hence, these mileposts start from zero at the border of a new county. Hence, all mileposts were transformed into a variable named "base_milepot" starting from zero at the first loop detector station. Now, the variable representing the mile-post location of the crash was monotonic increasing in the eastbound direction and decreasing in the westbound direction without any discontinuity at the county boundaries. The "base_milepost" derived from the crash characteristics table (Table 4-1) was used to determine loop detector station nearest to location of each crash. This station was referred to as the station of the crash.

4.4 Reported Time of Historical Crashes: How Accurate is it?

4.4.1 Background

This study relies on linking pre-crash loop detector data patterns with crash characteristics and therefore time of historical crashes used in the analysis becomes critical. The reason being that if the reported time of the crash is for example, 10 minutes later than the actual time of crash occurrence it would lead to a "cause and effect" fallacy as pointed out by Hughes and Council (1999). Fortunately for us, there is an automated system is in place in Florida that records the exact time when a crash is reported to the Police. With the wide spread use of the mobile phones the difference between times of occurrence and reporting of a crash is usually minimal. Moreover, surveillance cameras

are located approximately every mile on the Interstate-4 corridor. The range of each camera overlaps with adjacent cameras in upstream and downstream direction. The whole corridor is therefore visible to the RTMC operators. According to these operators the crashes are reported 'as soon as' they occur. These feedbacks from the Florida Highway Patrol (FHP) and RTMC officials about accurate reporting of the time of the crash indicated that the reported time might in fact be close to actual time of crash occurrence. However, since it was one of the most critical issues identified in the literature (e.g., Hughes and Council, 1999), concurrence of reported time of crash with the actual time needed to be verified before proceeding further.

In this regard a rule based shockwave methodology, presented later in this chapter, was developed. The methodology is based on estimation of speed of the individual shockwave resulting from each crash. A modified time obtained through the application of this methodology was estimated for crashes that fulfilled somewhat expansive data requirements of the methodology. The accuracy of the aforementioned methodology depends upon the severity of the drop in speed observed at the loop detector(s) preceding the location of the crash in upstream direction. For some crashes, however, no drop in speed is observed at any of the upstream stations because of very low existing demand levels. Due to this limitation along with extensive data requirements (loop data from each of the three lanes) the methodology could only be successfully applied to a small proportion of crashes. Again, fortunately for us the results from this reduced sample of crashes provided us ample proof of concurrence between reported and actual time of crashes and thereby validating the contention of FHP officials. The details of the

methodology and procedure to validate the concurrence are provided in the following sections of this chapter.

4.4.2 Loop data used to estimate time of historical crashes

To get an estimate of time of the crash, loop detector data from the station of the crash, four upstream stations and two downstream stations were collected for a period of 90 minutes around the reported time (one hour prior and half an hour later) of every crash. Note that for estimating the time of crashes, the loop data in their raw form, as time series with 30-seconds interval, were used. Note, that since it was one of the preliminary steps in analysis and at the time we only had crash data available through the year 2002, this methodology was developed using crashes belonging to the four year period (from 1999 through 2002). Out of the 3755 crashes belonging to the four year period 1705 crashes did not have any loop detector data available, i.e., none of the seven detectors from which data were sought were functioning on day of these crashes. The remaining crashes had at least partial data available but there was no guarantee that detectors from all three lanes at these stations were reporting data. Besides, the loop detectors are known to suffer from intermittent hardware problems that result in unreasonable values of speed, volume and occupancy. Values that include Occupancy>100, speed=0 or >100, flow>25, and flow =0 with speed>0, were removed from the raw 30-second data (Chandra and Al-Deek, 2004).

4.4.3 Impact of crashes on traffic flow

Crashes are a specific type of incident and generally have more profound impact on freeway operation. The effects of a crash on traffic flow patterns develop over time both

upstream and downstream of the crash. However, the changes in traffic flow characteristics are distinct on loop detectors located upstream and downstream directions. On the upstream direction, a queue is observed to form, resulting in significant reduction and increase in lane speed and occupancy, respectively. On the other hand, decrease in lane flow and occupancy is observed downstream. The critical aspect for determining the time of crash is the time elapsed in the progression of the shockwave from the crash location to the upstream loop detector station. In general this duration (i.e., the shockwave speed) and changes observed in the loop data are affected by the severity of that crash, roadway geometry, presence of on- and off-ramps, the distance between loop detector stations, and prevailing traffic flow conditions (Adeli and Karim, 2000, Al-Deek et al. 1995 and Al-Deek et al., 1996).

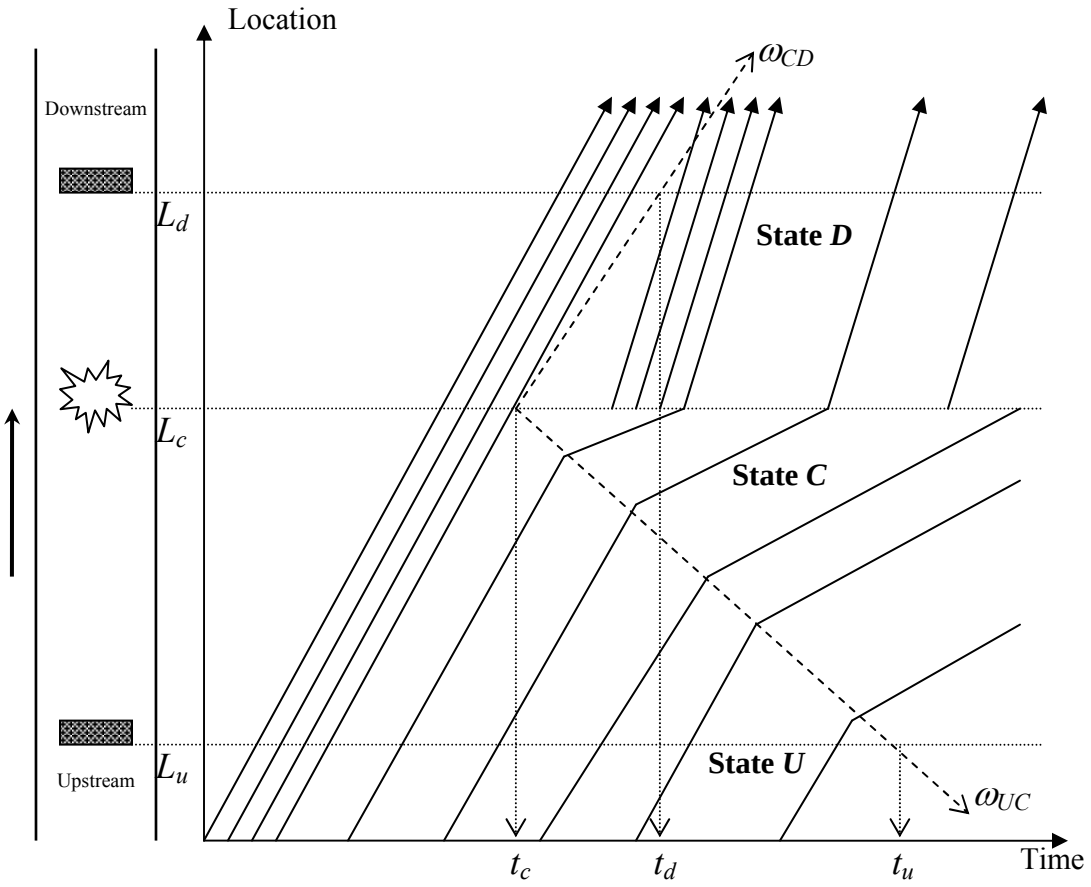

Figure 4-1 Time-space diagram in the presence of a crash (Lee et al. 2002)

The impact of a crash under the assumption of a constant shockwave speed may be shown by a time-space diagram (Figure 4-1). L_d and L_u represent the location of detector stations downstream and upstream of the crash site, respectively. The time t_c, t_d and t_u are time of the crash and time of shockwave arriving at downstream and upstream stations, respectively. It is clear from the figure that if the speed of backward forming shockwave is known then the time of the crash could be estimated. The times of shockwave hitting two adjacent upstream stations may be determined by observing when the drops in speed profiles of the two stations occur. The gap between the two arrival times is the time that

backward forming shockwave takes to travel from first upstream station to the next upstream station.

4.4.4 Time of the crash: estimation and validation

First step in estimating the time of the crash was to estimate the speed of the backward forming shockwave resulting from the crash. The difference between times of shockwave arrival at the two adjacent stations located immediately upstream of the crash location was used. Since the mileposts of all loop detectors on I-4 were known accurately, distance between the two detectors could be used to get the shockwave speed. Once the shockwave speed is known it is not difficult to determine t_c, using the milepost of crash location (also known with certain precision from the FDOT crash database). The following equation may be used for the estimation:

$$t_u - t_c = \frac{(L_u - L_c)}{\omega_{UC}}$$

All the variables in the above equation have the notation used in Figure 4-1. Due to the underlying assumption made here, that shockwave speed remains constant while it hits the first and second stations in the upstream direction, it was mandatory to validate the results. The critical issue in the validation was that there is no way to know the actual time of the crash (true value) to compare the shockwave model estimates with. The model was validated using the traffic simulation package *PARAMICS*. A small freeway section on Interstate-4 was simulated and three traffic flow statistics (speed, volume and density) were obtained from locations separated half mile apart on the section just as the loop data is archived for Interstate-4. Crashes were configured to occur at various locations between a set of two detectors (e.g., very near to upstream or downstream loop,

exactly midway between the loops, etc.). The simulation experiment showed that the time of these "artificial" crashes could be accurately estimated using the shockwave method under various scenarios.

4.4.4.1 Aggregation across lanes vs. using lane of the crash

After the methodology was developed and validated as explained above, it could either be applied by aggregating the data across three lanes or by using data from the specific lane on which the crash was known to have occurred through the *FDOT* database. The advantage of using the aggregated data was that the time of the crash could be estimated for a large sample of crashes, since the data for at least one of the lanes would obviously be available for more crashes than the data for a specific lane. On the other hand since the algorithm relies on the impact of shockwave hitting at successive upstream stations, sometimes the aggregated data (averaged over three lanes) might dampen this impact and the drop in speed or rise in occupancy may not be significant enough to be detected by the algorithm as a shock-wave hit. Since the methodology may not be applied to more than a fraction of crashes anyways and our main purpose was to validate or disapprove the claim of *FHP* officials regarding accuracy of the reported time of crash it was decided to apply the algorithm using data from the specific lane of crash for each case. Although the methodology may only be applied to even smaller sample of crashes yet one may be more confident of the results obtained.

4.4.4.2 Examination of traffic speed profiles upstream of crash location

Although results of the algorithm were validated on the simulation data it was necessary to understand the complexities involved in the real data, for example for the crashes that occur on the median it is almost impossible to detect any impact on upstream loop detectors. Even the "rubber neck" effect dies down before being felt at the station immediately preceding the crash location. Hence when the algorithm was applied to estimate time of real crashes the results were carefully examined for differences between reported and estimated time of crash. It was noticed that the differences between reported and estimated time ranged between 0 to 28 minutes; while for majority of crashes the difference was within three minutes there were some crashes for which it was more than 15 minutes. The methodology is expected to work best (i.e., result in more accurate time of crash) when a severe drop is observed at least two upstream stations. In absence of any significant drop in speed the methodology would still pick the maximum difference between two consecutive speed values as the time of shockwave hit at that station but that would be based on noise in the data rather than any drop resulting due to the crash.

To validate the concurrence of reported and estimated time the speed profile at the station located immediately upstream of crash location were examined for several crashes. The difference between the estimated and reported time was extracted from the results of the algorithm. It was observed that crashes in which the difference between reported and estimated time (through the methodology) of crash was zero or very close to zero the speed profiles at the upstream loop detector showed a clear drop in speed. From

PARAMICS validation as well as the details of the methodology it is clear that the methodology would work best in estimating the time of the crash when a clear drop in speed is observed at the upstream station. It provides heuristic validation for the fact that reported time is in fact very close to the estimated time of crash. On the other hand when the speed profiles of upstream stations are examined for the cases where the methodology estimated a time far off from the reported time; there was no pattern of drop in the raw speed time series. Essentially the drop picked up as the indication of a shockwave hit was part of the noise in the speed data.

We now present typical speed profiles from stations located upstream of the location of crash for a couple of crashes to clarify the observations made above (Figure 4-2). The time series shown in the figures has readings obtained from three freeway lanes for a period of 90 minutes (an hour prior and half an hour later to the reported time of each crash). Out of the 180 readings, 120^{th} reading is reported time of the crash. Along with speed profiles the difference between the estimated and reported time is also provided in the following figures. It should be noted that only speed time series was chosen for examination based on the findings by Ishak and Al-Deek (1999).

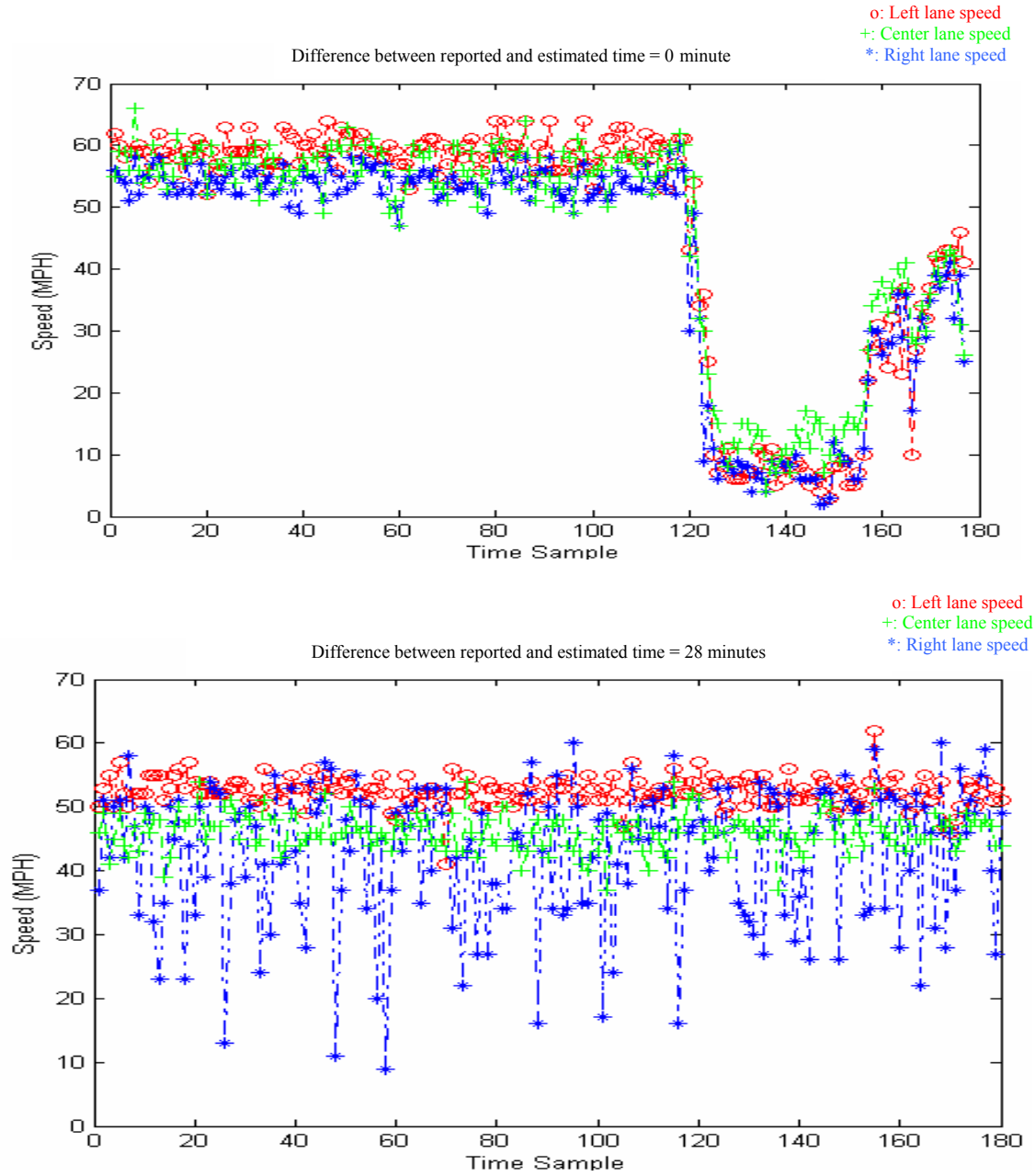

Figure 4-2 Speed profiles from station located upstream of crash location along with the difference in reported and estimated time for two separate crash cases

Based on results from shock-wave and rule based algorithm along with these observations it can be argued that the reported time is in fact very close to the actual time of crash occurrence. It also validates the claim of *FHP* and *RTMC* officials about accurate reporting of the time of the crash due to increasing cell phone usage and the automated crash reporting system. It gives reason to believe that the reported time is in fact close to actual time of crash occurrence and could be used with a certain amount of confidence.

4.5 Loop Data Collection

An essential part of the data used in this study is the loop detector data corresponding to crashes and non-crash cases. For the five-year period 1065 crashes had no loop detector data available at all. Hence, the loop data were collected for the remaining 3124 crashes. The format of the data extracted from the loop detector database largely depends upon the methodology used. Past experience of the research group (e.g., Pande, 2003, Abdel-Aty et al. 2004, Abdel-Aty and Abdalla, 2004) with data from 7-month period of the year 1999 was very beneficial in this regard. Three separate databases consisting of loop detector data have been assembled for this study. The first database was used in the previous section to verify the concurrence of actual and reported time of the crash while the other two were assembled for modeling purposes.

4.5.1 Data for matched case-control analysis

The matched case-control methodology was identified as an effective tool for modeling the binary outcome: crash or non-crash. To compare traffic characteristics (measured during time prior to crash occurrence from locations surrounding the crash location) that

lead to a crash with corresponding normal traffic conditions that did not lead to a crash, traffic data were extracted is a specific matched format.

Loop data were extracted for the day of crash and on all corresponding (non-crash) days to the day of every crash. The correspondence here means that, for example, if a crash occurred on April 12, 1999 (Monday) 6:00 PM, I-4 Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, three loops upstream and three loops downstream of station 30 for half an hour period prior to the estimated time of the crash for all the Mondays of the same season[1] in that year at the same time. This matched sample design controls for all the factors affecting crash occurrence such as driver population, season, day of week, location on the freeway, etc (thus implicitly accounting for all these factors). Hence, this case will have loop data table consisting of the speed, volume and occupancy values for all three lanes from the loop stations 27-33 (on eastbound direction) from 5:30 PM to 6:00 PM for all the Mondays of the year 1999, with one of them being the day of crash (crash case). More details of this sampling technique and application of this methodology may be found in one of the papers by our research group (Abdel-Aty et al., 2004). The format of data tables for this hypothetical crash is shown in Table 4-2.

[1] Summer season includes months May through August, while non-summer season includes other months of the year

Table 4-2 Format of the matched data extracted from the I-4 loop detector database for a hypothetical crash case

Day	Station	Y	Time	ELS	ECS	ERS	ELV	ECV	ERV	ELO	ECO	ERO
04/05/99	27	0	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	27	0	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	|	0	|	|	|	|	|	|	|	|	|	|
04/05/99	|	0	|	|	|	|	|	|	|	|	|	|
04/05/99	33	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	33	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	27	1	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	27	1	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	|	1	|	|	|	|	|	|	|	|	|	|
04/12/99	|	1	|	|	|	|	|	|	|	|	|	|
04/12/99	33	1	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	33	1	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	27	0	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	27	0	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	|	0	|	|	|	|	|	|	|	|	|	|
04/19/99	|	0	|	|	|	|	|	|	|	|	|	|
04/19/99	33	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	33	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
|	|	0	|	|	|	|	|	|	|	|	|	|
|	|	0	|	|	|	|	|	|	|	|	|	|
12/27/99	33	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
12/27/99	33	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx

Y: Binary variable representing crash (Y=1) and non-crash (Y=0) cases
ELS, ECS, ERS: Eastbound left, center and right lane speeds respectively
ELV, ECV, ERV: Eastbound left, center and right lane volume respectively
ELO, ECO, ERO: Eastbound left, center and right lane occupancy respectively

The field *Y* in the table above represents whether the data row corresponds for the crash case or a matched non-crash case. Such tables were extracted for all 3124 crashes with any loop data available. Note that number of observations in these tables for different

crashes was different due to random failures of the loops. Also, the cleaning mechanism suggested for these raw 30-second data by Chandra and Al-Deek (2004) was again adopted to clean the raw data.

4.5.2 Extraction of random non-crash cases

The database described above was prepared with a with-in stratum logistic regression analysis in perspective. However, there are other modeling techniques for binary target (crash vs. non-crash) which require random sampling of events from the two classes. To ensure that at any stage data requirements do not limit the scope of this study it was decided to collect loop detector data for random non-crash cases as well. Since crashes are rare events it was imperative to include all crashes with loop detector data available in this database. The sample may still be argued to be a random one since crashes are random events and loop detectors also tend to fail randomly. To generate random non-crash sample, the 5-year period was divided in to one minute periods (60minutes *24hours*1826 days over five years). 2629440 (one-minute periods) is the number of options available to choose the "time of non-crash". Similarly we have 138 stations (69 stations in two directions; EB and WB) to choose as "station of non-crash". In all, we can choose from 362862720 (2629440 one-minute periods* 2 directions* 69 stations) options to draw a random combination of time, station and direction to assign as random non-crash case. 150000 such combinations were selected randomly with corresponding station and direction equivalent to station and direction of crash, respectively. Similarly random time period from the combination corresponded to the time of crash and corresponding station to be the station of the crash. This random combination was used to extract sets of

35-minute data (30-minute prior and 5-minute later to the assumed time of the non-crash from 3 stations upstream and 3 stations downstream of the assumed station of non-crash) to create a random non-crash sample. Out of these 150000 random available non-crash cases, a non-crash sample may be drawn depending on the data requirements of the methodology used for analysis.

4.6 Geometric Design Parameters

Although the main purpose of this study is to establish link between real-time traffic characteristics (measured through loop detectors) and crash occurrences, it is extremely important to consider geometric characteristics on the freeway with respect to the crash characteristics. For example, the traffic characteristics leading to a crash on a curved section might be different from those leading to crash on a straight section. To obtain the geometric design of the I-4 corridor the Roadway Characteristics Inventory (*RCI*) database available on *FDOT* Intranet server was used. Geometric design features were extracted for the location of each loop detector station since it was the common link between crash and loop detector database. The structure of this database is shown in Table 4-3. Geometric design of the freeway might differ from one direction to the other, hence the dataset has 138 (*69*2=138*) observations.

Table 4-3 Geometric design of the freeway at loop detector station locations

Observation	Loop	Direction	Mile post	Radius (ft)	# of Lanes	Median type and width
1	2	E	xxx	xxx	xxx xxx	xx xx
2	2	W	xxx	xxx	xxx xxx	xx xx
|	|		|	|	|	| |
137	71	E	xxx	xxx	xxx xxx	xx xx
138	71	W	xxx	xxx	xxx xxx	xx xx

Another critical geometric design parameter was the location of ramps. The impact of ramps on freeway crash frequency is well documented. The location of ramps was determined with respect to crash location for each crash rather than the location of the station of crash. The milepost of each ramp on both direction of Interstate was collected from the geometric design database. This along with the variable "base_milepost" for each crash were used to determine the distance of nearest on and off ramp from the crash location in both upstream and downstream direction. Essentially we created four more variables, namely, "upstreamon", "upstreamoff", "downstreamon", and "downstreamoff" for each crash case. The modified structure of the crash database is shown in Table 4-4.

Table 4-4 Ramp location with respect to crash location

Crash Number	Crash report number	Direction (EB or WB)	Base _mile post	Upstreamon (Distance to nearest upstream on ramp)	Upstreamoff (Distance to nearest upstream off ramp)	Downstreamon (Distance to nearest down stream on ramp)	Downstreamoff (Distance to nearest down stream off ramp)
1	xx	xx	xx	xxx	xxx	Xxx	xxx
2	xx	xx	xx	xxx	xxx	Xxx	xxx
|	|	|	|	|	|	|	|
4189	xx	xx	xx	xxx	xxx	Xxx	xxx

For random non-crash cases that were created based on the "station of the crash" we do not have the variable "base_milepost". It would not be appropriate to assign the milepost of the loop detector corresponding to station of crash as the "base_milepost" since all non-crash cases would then be limited to just 69 (# of loop detector stations; as the milepost of loop detectors is identical for both directions) distinct values. It would then be difficult to analyze the impact of the location of the freeway on crash occurrence. It was decided to assign each random non-crash case a mile post location. Since the station of crash was fixed for each random non-crash case, the milepost assigned to it was a random milepost generated from within the influence area of station of crash. The influence area of a station of crash is defined as the section of the freeway, crash on which will be assigned that station as the station of crash. The definition is made clearer in Figure 4-3.

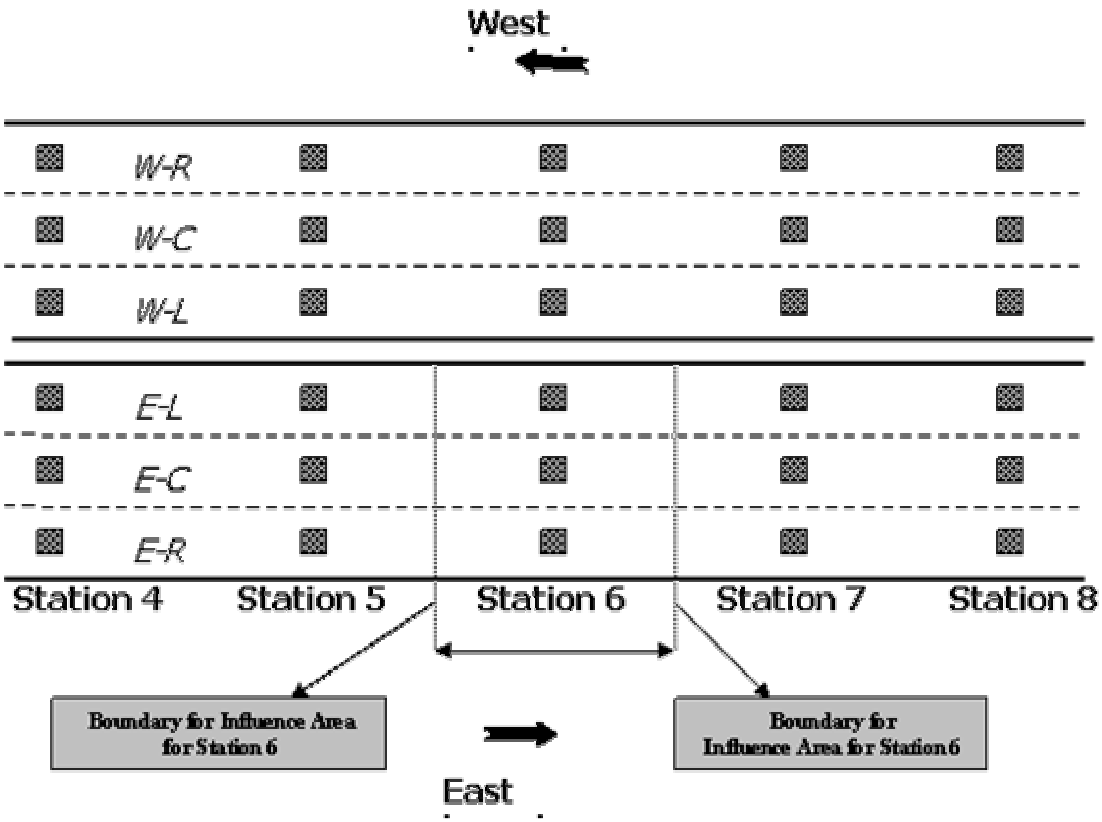

Figure 4-3: Influence area for loop detector stations

The figure shows series of loop detectors on a freeway along with the influence area for station 6 in both east and west directions. Station 6 is the closest station for each point in the section within the boundaries shown in the Figure 4-3. Hence, if a crash occurs within the boundaries shown, the station of the crash would be station 6. To assign mile post to random non-crash cases the mileposts corresponding to these boundaries were estimated for every loop detector station. These mile-posts were merged with each non-crash case based on the station of the crash associated with it. A random number was then chosen between the milepost of these boundaries and assigned as "base_milepost" for that non-crash case. For non-crash cases the distances of closest ramps in upstream and downstream directions were determined with respect to the "based_milepost" assigned to

it. With the variable "base_milepost" available for crash cases (based on the actual mile post location from the FDOT database) and random non-crash cases (assigned using the procedure described above) one can even analyze it as an independent variable. In the analysis presented in later chapters this "base_milepost" variable would be transformed into ordinal variables based on its relationship with crash occurrence.

4.7 Driver Population Characteristics

4.7.1 Conceptual background

While crash involvement of drivers belonging to certain age group or gender etc. has been a major area of research in traffic safety; none of these factors have been incorporated into real-time crash prediction models developed so far. A possible way to incorporate driver characteristics would be to identify the composition of driver population on the corridor at different times of the day, e.g. morning peak will consist of mostly middle aged commuters while a Friday evening will have more of younger drivers. Results from the past studies (e.g., Stamatiadis and Deacon, 1997) about the variation in risk of crash involvement amongst distinct groups of drivers (categorized based on their age, gender, etc.) combined with information on road user population based on time of the day, day of the week etc. may be used as inputs in real-time crash prediction models.

The information about composition of driver population on Interstate-4 was deduced using the induced exposure concept proposed by Stamatiadis and Deacon (1997). The induced exposure method uses not-at-fault driver involved in crashes as a measure of

exposure. The at-fault drivers generally belong to a certain group prone to commit driving errors but it can be assumed that "not-at-fault" or "victim" drivers represent a random sample of the road user driver population. Essentially, we intend to derive odds of finding driver belonging to a certain group by time of day, day of week and segment of the freeway. These odds associated with various types of drivers may then be used as independent variables in the models developed in later chapters.

4.7.2 Database properties

The driver information for *4189* crashes that occurred in the study area (from 1999 through 2003) was extracted from the Drivers table of the *DHSMV* crash database. The factors such as the age and gender of the driver(s) involved in each crash are part of this database. From the database we have information on 8761 drivers involved. While 4186 *(≈ 48%)* of them were cited for some form of traffic violation; 4575 *(≈ 52%)* was the size of sample of not-at-fault drivers which is expected to be representative of the overall population of drivers on the freeway.

The driver characteristics examined here include; race, gender, age and residency status of the driver. Out of these factors, race (White, Black, Hispanic and Others), gender (male and female), and residency status (County of crash, Elsewhere in state, Non-resident of state, and Foreign) were measured at nominal scales while age of the driver(s) involved was a continuous variable. The driver-age was transformed into an ordinal variable with five levels. The drivers were categorized into five levels, namely, Very young (Less than 20 years of age), Young (between 20 and 25 years of age), Middle aged

(between 25 to 45 years), Old (between 45 to 60 years), and Very old (More than 60 years of age) (Abdel-Aty et al., 1998). In the next section relevant distributions are explored to make comparisons between the populations of the drivers that were not-at-fault with at-fault drivers so as to determine which driver groups may be considered significantly more risky or safe.

4.7.3 Distribution of driver population

We intend to deduce factors that would appropriately represent the composition of the driver population in terms of its impact on overall odds of crash occurrence. In this regard one of the first tasks was to examine which of these factors related to the drivers would significantly alter their crash involvement. In this section distribution of driver population belonging to certain groups is presented among at-fault and not-at fault drivers. Based on these distributions we can assess that drivers belonging to which group are more likely to be at-fault and cause crash. Figures 4-4 through 4-7 show the percentage of drivers by age group, race, residency status, and gender, respectively.

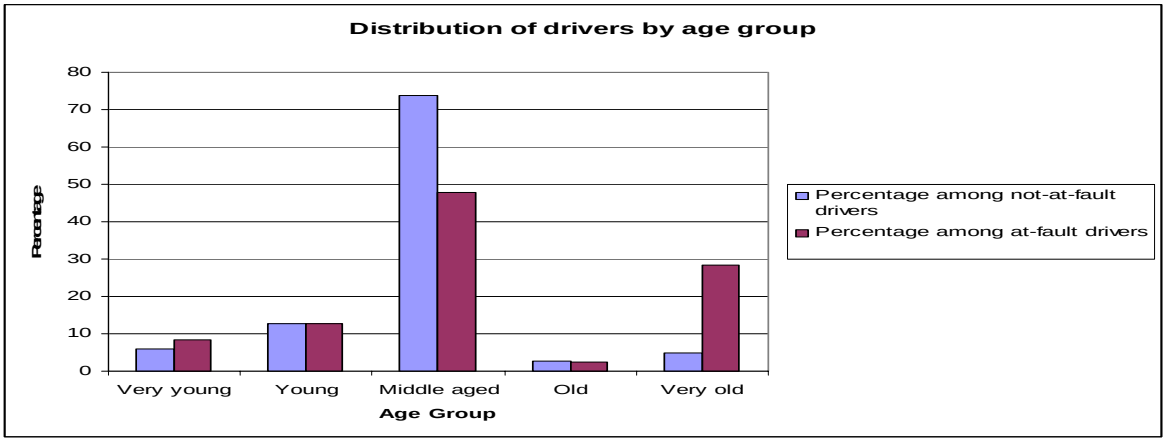

Figure 4-4: Percentage of driver by age group in at-fault and not-at-fault driver samples

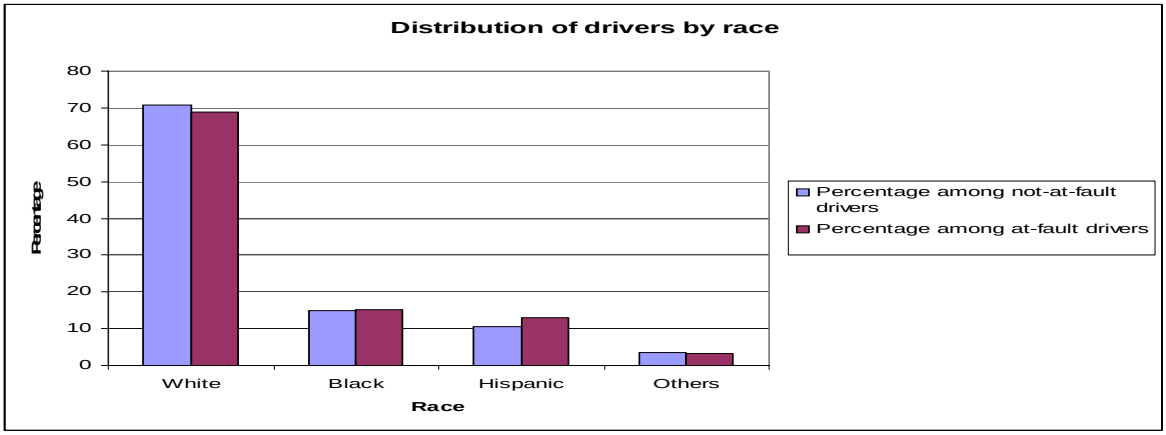

Figure 4-5: Percentage of driver by race in at-fault and not-at-fault driver samples

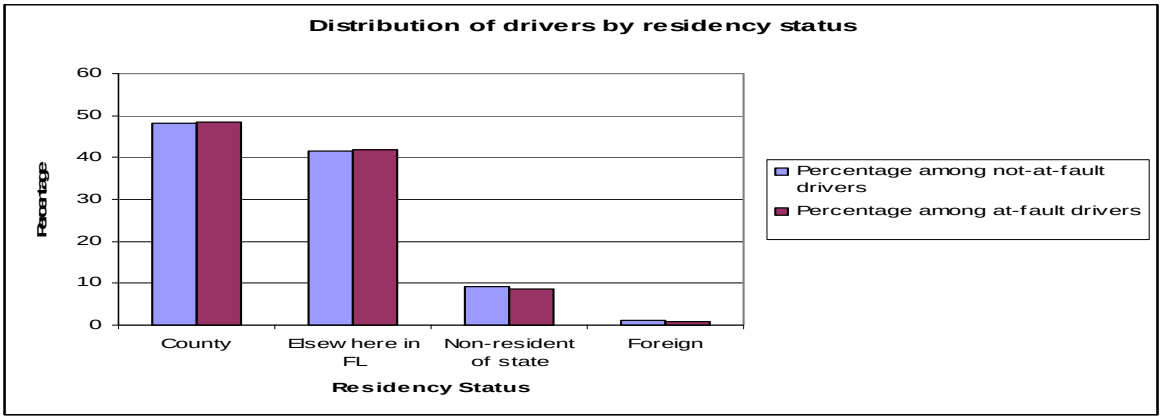

Figure 4-6: Percentage of driver by residency status in at-fault and not-at-fault driver samples

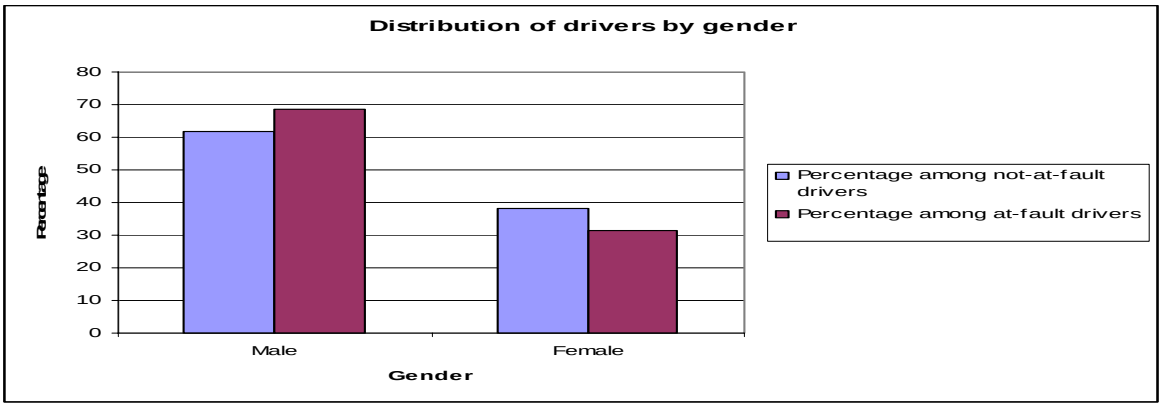

Figure 4-7: Percentage of driver by gender in at-fault and not-at-fault driver samples

It is clear from the figures that while for different categories of variable "race" (Figure 4-5) the proportions of drivers among the not-at-fault and at-fault sample is almost the same. It indicates that drivers belonging to any particular race are not prone toward crash involvement. A similar trend or lack of it is observed with respect to the variable "residency status" (Figure 4-6). However, proportion of male drivers is slightly more in the sample of guilty drivers than their proportion in the innocent drivers. It indicates that male drivers are slightly more "unsafe" on the Interstate than the female drivers (Figure 4-7).

The starkest contrast in terms of difference in proportion among at-fault and not-at-fault drivers is depicted by different categories of age-group. We may see (Figure 4-4) that the drivers in the category young and old have the almost same percentage in the sample of guilty and innocent drivers. For Very young and Very old drivers, however, the proportion among guilty drivers is significantly more than their proportion among the innocent drivers. In contrast the proportion of Middle-aged drivers is significantly lower

(compared to their proportions in not-at-fault) in at-fault drivers. If we combine these observations with our premise that the composition of not-at-fault drivers' sample represents the overall population of drivers on the freeway corridor; one may infer that while majority of drivers on the roads are middle-aged they are reasonably safe drivers. Very old drivers tend be less exposed (with very less percentage among the not-at-fault drivers) but are highly crash prone. In the next section odds of drivers belonging to various age groups on the Interstate segments by time of day and day of week would be calculated as potential input variables to real-time models.

4.7.4 Odds of drivers from certain age-groups: Factors representing driver population composition

The 36.25-mile Interstate-4 corridor under consideration was initially divided into 10 segments of equal length, with Disney area being first segment and Lake Mary area being the tenth segment. Downtown Orlando area is located around fifth and sixth segments. The time of day was categorized into rush hours (morning and afternoon peak hours; 6:00 - 9:00 AM and 4:00 -7:00 PM), mid-day off-peak (between 9:00 AM to 4:00 PM) and night off-peak (After 7:00 PM up to 6:00 AM in the morning). Days of week were categorized into weekend (Saturday and Sunday) and weekday (Monday through Friday). These classifications were made based on broad understanding of driver population composition.

Odds of finding drivers belonging to certain age group would depend upon their proportion among not-at-fault driver sample by segments of the Interstate. Note that there are some limitations to the methodology adopted here such as the sample size that is used

as the estimate of driver population would be different for each interstate segment. Moreover, sample size would be biased towards locations/time period with high crash frequency (more drivers sample for locations with high crash frequency). More discussion on these issues may be found in the relevant reference (Stamatiadis and Deacon, 1997). Since estimating driver population is not the main focus of this research and just a measure arguably representing the driver composition by age-group is needed, we can work with the limitations of this approach. However, the dependence of sample size (available to estimate driver population composition) on the crash frequency raises another concern. During periods and locations with smaller crash frequency (e.g., Lake Mary area during night off-peak on a Sunday night) one usually does not find a significant number of drivers belonging to rare age groups (such as very old drivers) and confidence of our estimates suffers. One way to deal with this problem would be to collect more crash data. While collecting more crash data would involve significant effort it may not be entirely useful since historical driver population obtained from beyond a certain period in the past might not be relevant any more to deduce current driver population estimates. Therefore, we examined distribution of drivers of different age-groups at these segments by time of day and day of week to see if some of them may be combined in order to boost the sample size. These distributions are shown in Figure 4-8 and 4-9.

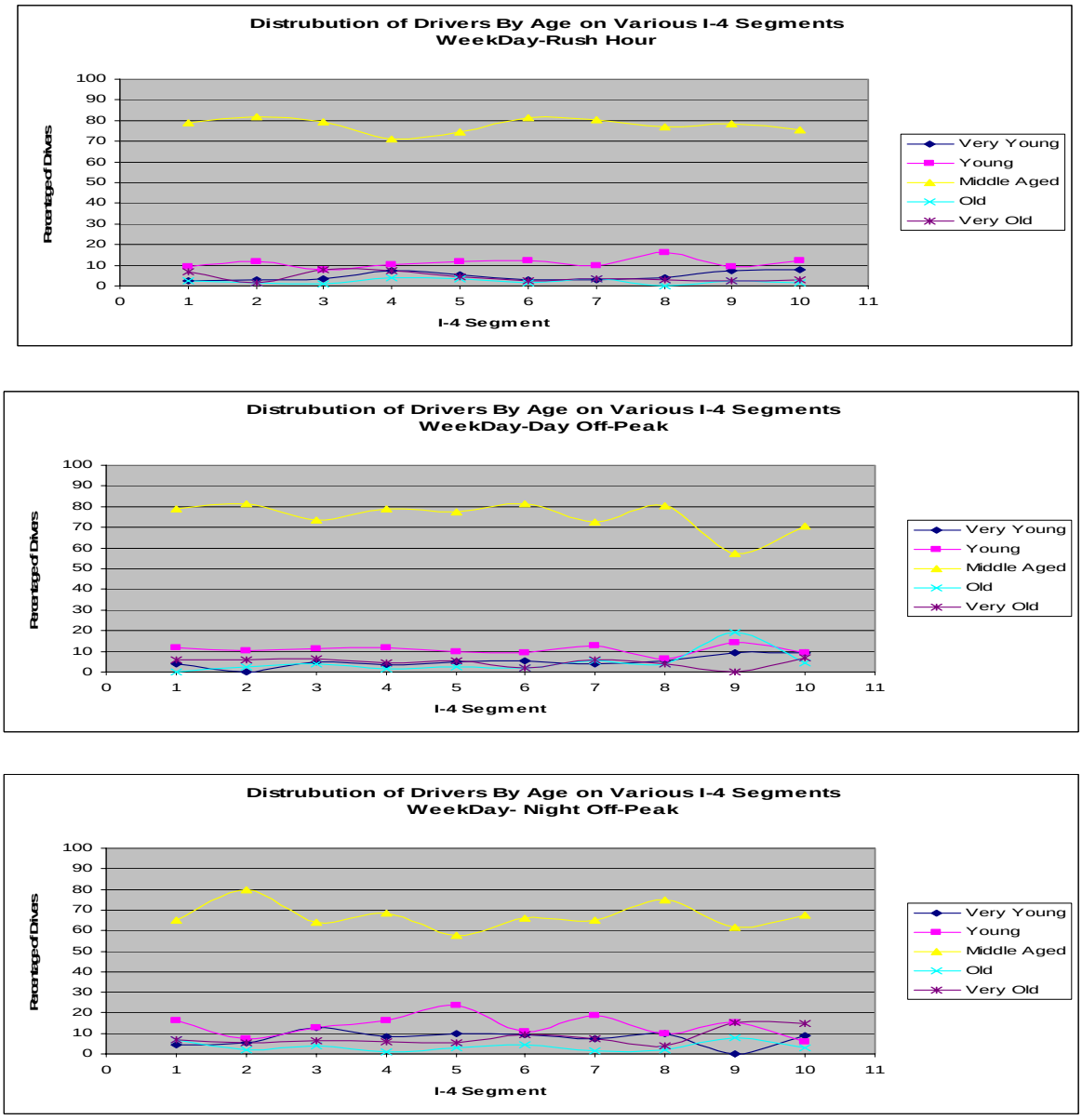

Figure 0-1: Distribution of drivers of different age-group on weekdays by time of day at different I-4 locations

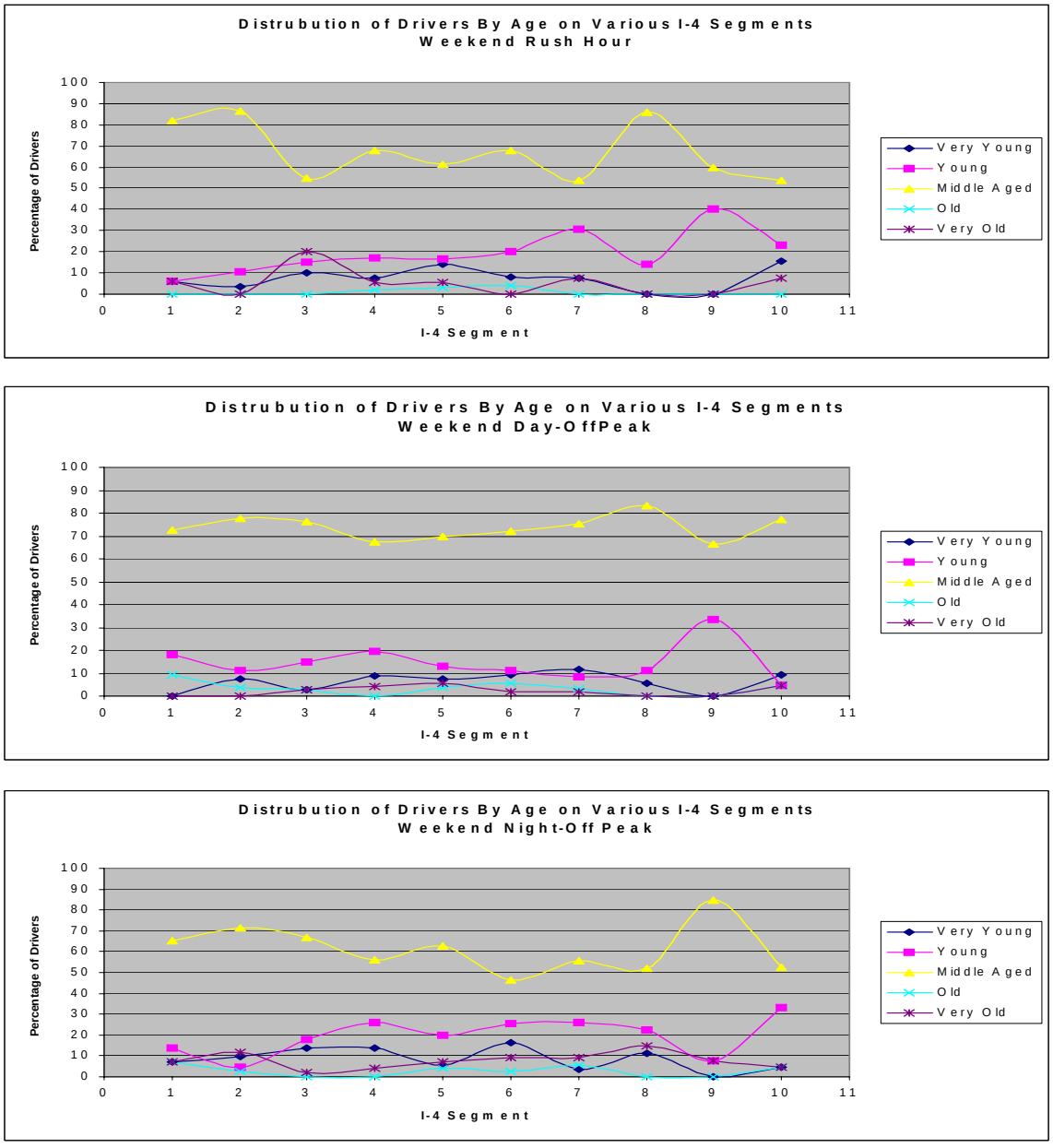

Figure 0-2: Distribution of drivers of different age-group on weekends by time of day at different I-4 locations

Based on the figures (Figure 4-8 and Figure 4-9) shown above it may be said that the variation of driver population composition is not very significant among segments 3 through 7 and therefore these segments may be combined together. Similarly we can combine segments 1 and 2 and segments beyond segment 7. These observations along with the demographic and commercial characteristics of the I-4 segments led to the conclusion that initially proposed ten segments should be reduced to the following three segments: One Disney to Universal Studios, the attractions, second the downtown Orlando area and third the segment east of downtown Orlando (beyond loop detector station 55).

Table 4-5 provides the proportions of drivers of the five age-groups on these newly defined segments of freeway by time of day and day of week. Note that these estimates are based on a sample of 4575 drivers that were involved in a crash within the five year period but were not found at-fault by the police officer on the scene.

Table 4-5 Proportion of drivers belonging to different age groups by Interstate segment, time of day and day of week

			Proportion Estimate by Age-group				
Segment	**Time of Day**	**Day of Week**	**Very young**	**Young**	**Middle aged**	**Old**	**Very old**
Attractions	rush hours	weekdays	0.049	0.098	0.777	0.016	0.059
Downtown Orlando	rush hours	weekdays	0.037	0.117	0.783	0.028	0.035
East of downtown	rush hours	weekdays	0.084	0.122	0.756	0.015	0.023
Attractions	mid-day off-peak	weekdays	0.031	0.120	0.768	0.021	0.061
Downtown Orlando	mid-day off-peak	weekdays	0.046	0.102	0.778	0.032	0.042
East of downtown	mid-day off-peak	weekdays	0.080	0.100	0.710	0.070	0.040
Attractions	night off-peak	weekdays	0.074	0.140	0.696	0.029	0.062
Downtown Orlando	night off-peak	weekdays	0.096	0.165	0.638	0.033	0.069
East of downtown	night off-peak	weekdays	0.072	0.096	0.675	0.036	0.121
Attractions	rush hours	week-end	0.073	0.138	0.732	0.000	0.057
Downtown Orlando	rush hours	week-end	0.091	0.171	0.648	0.034	0.057
East of downtown	rush hours	week-end	0.087	0.261	0.609	0.000	0.044
Attractions	mid-day off-peak	week-end	0.040	0.158	0.753	0.030	0.020
Downtown Orlando	mid-day off-peak	week-end	0.103	0.113	0.718	0.036	0.031
East of downtown	mid-day off-peak	week-end	0.056	0.111	0.778	0.028	0.028
Attractions	night off-peak	week-end	0.108	0.140	0.675	0.019	0.057
Downtown Orlando	night off-peak	week-end	0.085	0.244	0.542	0.035	0.095
East of downtown	night off-peak	week-end	0.071	0.238	0.619	0.024	0.048

To calculate the odds (defined as *p/ (1-p)*) the proportions shown in the above table may be used. For example, the odds of very old drivers during week day peak hours on Interstate-4 near Attractions (Disney/Universal) would be equal to *0.059/ (1-0.059)* (*=0.063*). Note that 0.059 is the proportion of these drivers highlighted in Table 4-5. Table 4-6 shows odds corresponding to the proportions shown in Table 4-5.

Table 4-6 Odds of drivers belonging to different age groups by Interstate segment, time of day and day of week

			Odds Estimate by Age-group				
Segment	Time of Day	Day of Week	Very young	Young	Middle aged	Old	Very old
Attractions	rush hours	weekdays	0.052	0.109	3.484	0.017	0.063
Downtown Orlando	rush hours	weekdays	0.038	0.132	3.615	0.029	0.036
East of downtown	rush hours	weekdays	0.092	0.139	3.093	0.016	0.023
Attractions	mid-day off-peak	weekdays	0.032	0.136	3.308	0.021	0.065
Downtown Orlando	mid-day off-peak	weekdays	0.048	0.113	3.507	0.033	0.044
East of downtown	mid-day off-peak	weekdays	0.087	0.111	2.448	0.075	0.042
Attractions	night off-peak	weekdays	0.080	0.163	2.284	0.030	0.066
Downtown Orlando	night off-peak	weekdays	0.106	0.197	1.760	0.034	0.074
East of downtown	night off-peak	weekdays	0.078	0.107	2.074	0.037	0.137
Attractions	rush hours	week-end	0.079	0.160	2.727	0.000	0.060
Downtown Orlando	rush hours	week-end	0.100	0.206	1.838	0.035	0.060
East of downtown	rush hours	week-end	0.095	0.353	1.556	0.000	0.045
Attractions	mid-day off-peak	week-end	0.041	0.188	3.040	0.031	0.020
Downtown Orlando	mid-day off-peak	week-end	0.114	0.127	2.545	0.037	0.032
East of downtown	mid-day off-peak	week-end	0.059	0.125	3.500	0.029	0.029
Attractions	night off-peak	week-end	0.121	0.163	2.079	0.019	0.061
Downtown Orlando	night off-peak	week-end	0.092	0.322	1.185	0.036	0.104
East of downtown	night off-peak	week-end	0.077	0.313	1.625	0.024	0.050

These odds depict prevailing driver population composition by age-group on the freeway at different times of day. For example, one might expect that at a time of day if the odds of very old drivers are high then it should have a positive impact of overall chances of crash occurrence.

4.7.5 Odds of drivers with certain residency status: Factors representing driver population composition

Second categorization of drivers considered was based on their residency status. Note that according to the Figure 4-6 driver groups with different residency status do not show differences among their proportions in the sample of innocent and guilty drivers and their crash involvement is in accordance to their exposure. The Figure was based on the overall distribution of the drivers by residency status over the whole I-4 corridor. However, if we examine the data disaggregated by 'time of day', 'day of week' and freeway segment (as we did with driver age) the possibility of varying population of tourist drivers (identified by residency status 'out of state' or 'foreign') having an impact on odds of crash can not be out rightly rejected. It is certainly possible that these drivers drive as 'safely' as the commuters in the downtown region where they need not worry about missing the desired exit but as they approach popular tourist destinations (e.g., near Disney area attractions) they might do some maneuvers in their lookout for directions that involve more risk, thereby resulting in increased odds of crash occurrence. Since we have no reason to believe that disaggregate analysis with the parameters 'race' or 'gender' would bring a similar scenario into picture; the odds of drivers belonging to different races are not estimated here.

The 36.25-mile Interstate-4 corridor under consideration was divided into three segments: First, Disney to Universal Studios, the attractions, second, downtown Orlando area and third the segment east of downtown Orlando (beyond loop detector station 55) for the disaggregate analysis. The time of day was separated into rush hours (morning and afternoon peak hours; 6:00 - 9:00 AM and 4:00 -7:00 PM), mid-day off-peak (between 9:00 AM to 4:00 PM) and night off-peak (After 7:00 PM up to 6:00 AM in the morning). Days of week were categorized into weekend (Saturday and Sunday) and weekday (Monday through Friday). These classifications were made based on broad understanding of driver population composition in the previous section itself.

Table 4-7 Proportion of drivers with different residency status by Interstate segment, time of day and day of week

			Proportion Estimate by Residency Status			
Segment	**Time of Day**	**Days of Week**	**County**	**Elsewhere in FL**	**Non-resident of state**	**Foreign**
Attractions	rush hours	weekdays	0.449	0.398	0.1293	0.0238
Downtown Orlando	rush hours	weekdays	0.5465	0.3884	0.0643	0.0009
East of downtown	rush hours	weekdays	0.4063	0.5469	0.0391	0.0078
Attractions	mid-day off-peak	weekdays	0.3865	0.4078	0.1667	0.039
Downtown Orlando	mid-day off-peak	weekdays	0.5292	0.3861	0.0792	0.0055
East of downtown	mid-day off-peak	weekdays	0.3776	0.5306	0.0816	0.0102
Attractions	night off-peak	weekdays	0.4249	0.3777	0.1674	0.03
Downtown Orlando	night off-peak	weekdays	0.4968	0.4146	0.0791	0.0095
East of downtown	night off-peak	weekdays	0.2568	0.5811	0.1351	0.027
Attractions	rush hours	week-end	0.388	0.491	0.112	0.009
Downtown Orlando	rush hours	week-end	0.581	0.372	0.047	0.000
East of downtown	rush hours	week-end	0.500	0.409	0.091	0.000
Attractions	mid-day off-peak	week-end	0.3366	0.5644	0.0792	0.0198
Downtown Orlando	mid-day off-peak	week-end	0.4663	0.4456	0.0777	0.0104
East of downtown	mid-day off-peak	week-end	0.3056	0.4444	0.1944	0.0556
Attractions	night off-peak	week-end	0.4094	0.396	0.1409	0.0537
Downtown Orlando	night off-peak	week-end	0.5104	0.4219	0.0677	0
East of downtown	night off-peak	week-end	0.3902	0.5366	0.0732	0

Table 4-7 provides the proportions of drivers of the four residency groups on three segments of freeway by time of day and day of week. Note that these estimates are based on a sample of 4575 drivers that were involved in a crash within the five year period but were not found at-fault by the police officer on the scene.

Table 4-8 Odds of drivers with different residency status by Interstate segment, time of day and day of week

			Odds Estimate by Residency Status			
Segment	**Time of Day**	**Days of week**	**County**	**Elsewhere in FL**	**Non-resident of state**	**Foreign**
Attractions	rush hours	weekdays	0.815	0.661	0.149	0.024
Downtown Orlando	rush hours	weekdays	1.205	0.635	0.069	0.001
East of downtown	rush hours	weekdays	0.684	1.207	0.041	0.008
Attractions	mid-day off-peak	weekdays	0.630	0.689	0.200	0.041
Downtown Orlando	mid-day off-peak	weekdays	1.124	0.629	0.086	0.006
East of downtown	mid-day off-peak	weekdays	0.607	1.130	0.089	0.010
Attractions	night off-peak	weekdays	0.739	0.607	0.201	0.031
Downtown Orlando	night off-peak	weekdays	0.987	0.708	0.086	0.010
East of downtown	night off-peak	weekdays	0.346	1.387	0.156	0.028
Attractions	rush hours	week-end	0.634	0.966	0.126	0.009
Downtown Orlando	rush hours	week-end	1.389	0.593	0.049	0.000
East of downtown	rush hours	week-end	1.000	0.692	0.100	0.000
Attractions	mid-day off-peak	week-end	0.507	1.296	0.086	0.020
Downtown Orlando	mid-day off-peak	week-end	0.874	0.804	0.084	0.011
East of downtown	mid-day off-peak	week-end	0.440	0.800	0.241	0.059
Attractions	night off-peak	week-end	0.693	0.656	0.164	0.057
Downtown Orlando	night off-peak	week-end	1.042	0.730	0.073	0.000
East of downtown	night off-peak	week-end	0.640	1.158	0.079	0.000

To calculate the odds (defined as *p/ (1-p)*) the proportions shown in Table 4-7 may be used. Note that the procedure to estimate the odds from the available proportion is similar to the one used in previous section. These odds depict prevailing driver population composition by residency status on the freeway at different times of day. Note that the only category where the odds of observing an 'out of state' drivers near attractions are less than 0.1 is mid-day off-peak on the weekend, when most of the tourist may be expected to be actually enjoying the attractions. Another segment with high odds of observing 'out of state' tourist drivers is the section East of downtown Orlando during mid-day off-peak. Possible reason for that might be the beach traffic from or towards Daytona.

These odds or combination of them obtained for the variable driver-age and residency status may now be used as inputs to the real-time models based on the time-of-day, day of week and freeway location. It should be understood that overall composition of driving population will be used as a surrogate for the individual drivers on the freeway. Of course the driving error(s) will be committed by individual driver(s) but if the freeway has considerable fraction of "un-safe" drivers it would increase the overall chances of having a crash and is expected to reflect in real-time odds of crash occurrence. Hence, incorporating the odds of certain category of drivers on the freeway based on time of day, day of week and roadway segment in a real-time prediction system should improve the explanatory power and prediction accuracy of the models developed.

4.8 Concluding Remarks

This Chapter describes the data gathering and preparation effort for this study. The data have been prepared keeping in mind the requirements of methodologies to be used later in the analysis. Significant amount of time and effort has been devoted to collection and assembling of data. First, it was established through a detailed shockwave speed based algorithm that actual time of historical crashes is in fact very close to the reported time of crash. Therefore, it was decided that reported time of crash would be used for collection of loop detector data corresponding to crashes. The crash data was combined with non-crash data collected in two separate formats. First, the non-crash database created through with-in stratum matching and second, the randomly selected non-crash data extracted from the loop database. Induced exposure analysis was used to deduce odds of drivers belonging to certain age-groups by time of day, day of week and segment location on the freeway. These odds will be used later in the analysis as input to the models separating crash prone conditions from normal conditions.

With five years of crash and non-crash data, the database created here are by far the most comprehensive database created for a real-time crash prediction study. The information about driver population composition makes the assembled database even more valuable. In the coming chapters these databases would be combined and used to estimate models separating distinct types of crash occurrences such as rear-end, side swipe from normal conditions on the freeway.

CHAPTER 5

DATA MINING ANALYSIS OF REAR-END CRASHES

5.1 General

The essential idea of a fully functional proactive traffic management system would involve anticipating incidents, such as crashes, prior to their occurrence and then intervene in a certain manner to reduce their likelihood. The goal of this research is to develop a system of models that would efficiently identify the conditions prone to crash occurrences on freeways. Most of the existing real-time crash prediction models are generic in nature, i.e., one model has been used to predict different types (such as rear-end, sideswipe, or angle) of crashes. This "one size fits all" approach is not sufficient because different types of crashes have been known to be related to distinct traffic flow characteristics.

While the traffic conditions following crashes of different types (such as rear-end, sideswipe or angle crashes) are similar in nature; the conditions preceding them are likely to differ from type to type. E.g., the rear-end crashes might be expected to occur under congested traffic regime where the drivers have to slow down and speed up quite often, on the other hand the single vehicle crashes might result from excessive speeds on a curved freeway section. Therefore, while generic models may be used to separate post-incident traffic surveillance data from a non-incident scenario; the approach for proactive traffic management should be type (of crash) specific in nature. Such specific models would also be useful in devising remedial measures to improve the safety situation on the

freeway which would differ for each type of crash, e. g, the variable speed limits for rear-end crashes or a temporary "no lane-changing" sign to avoid an impending sideswipe crash.

In this chapter a data mining approach is presented to separate rear-end crashes from non-crash conditions based on the freeway traffic data collected through the loop detector stations surrounding the location of historical crashes. The formation and structure of the dataset used for the analysis were discussed in detail in one of the earlier chapters (Chapter 4). However, it should be noted at this point that the non-crash data used in the analysis were drawn as a random sample from the loop detector dataset. Although logistic regression technique with a matched study design was previously employed successfully to develop generic crash prediction models, for this part of the research random sampling of non-crash cases was used since it enables us to explore the impact of offline factors such as the ramp locations, curvature etc.; along with their possible interactions with real-time traffic parameters, on the occurrence of a rear-end crash. In the matched design these factors were implicitly controlled for and were assumed to be included in the intercept term of the logistic regression model. As per the model structure, the estimate for the intercept was neither available nor required for prediction. In this analysis off-line factors are also considered to examine if these characteristics impact the occurrence of rear-end crashes and improve on the performance that was achieved through the models developed from matched study design. Studying the impacts of these "off-line" characteristics on crash occurrence is also critical since it might come into play while

devising measures for crash avoidance, for example, different approach might be required to calm crash prone conditions on a curved freeway segment than on a straight one.

This chapter is divided into five sections. The next section deals with preliminary explorations and identification of clusters in the rear-end crash data based on the prevailing traffic speed configurations on the freeway section around the crash locations. Characteristics of individual groups (clusters) of rear-end crashes such as their distribution over time of the day and locations are also discussed in this section. Based on the frequency of traffic conditions belonging to groups/clusters of rear-end crashes in the freeway loop detector data, two strategies of data analysis are proposed. In the subsequent sections 'prediction' models developed for individual groups of rear-end crashes using the crash and random non-crash data are presented. Modeling issues such as proportions of the crash vs. non-crash cases in the samples used to develop the models and choice of modeling tools have also been addressed. The final section summarizes modeling results and application strategies to identify two groups of rear-end crashes. The proposed application strategy would be incorporated in the detailed implementation plan presented later in the dissertation.

5.2 Loop Data Aggregation

As explained in the previous chapter loop data were then extracted for every crash in a specific format, for example, if a crash occurred on April 12, 1999 (Monday) 6:00 PM, I-4 Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, three loops upstream and three loops downstream of station 30 for 30-minute

period prior to the reported time of the crash. Hence, this crash case will have loop data table consisting of the 30-seconds averages of speed, volume and occupancy for all three lanes from the stations 27 through 33 (on eastbound direction) from 5:30 PM to 6:00 PM for April 12, 1999. The analysis however was limited to five stations and 20 minutes for rear-end crashes based on results from our previous studies (Abdel-Aty et al. 2004, 2005).

The raw 30-second data have random noise and are difficult to work with in a modeling framework. Therefore, the 30-second raw data were combined into 5-minute level in order to obtain averages and standard deviations. For 5-minute aggregation 20-minute period was divided into four time slices. The stations were named as "*D*" to "*H*", with "*D*" being farthest station upstream and so on. It should be noted that "*F*" is the station closest to the location of the crash with "*G*" and "*H*" being the stations downstream of the crash location. Similarly the 5-minute intervals were also given "IDs" from 1 to 4. The interval between time of the crash and 5 minutes prior to the crash was named as time-slice 1, interval between 5 to 10 minutes prior to the crash as time-slice 2 and so on. The parameters were further aggregated across the three lanes and the averages (and standard deviations) for speed, volume and lane-occupancy at 5-minute level were calculated based on 30 (10*3 lanes) observations. Therefore, even if at a location the loop detector from a certain lane was not reporting data, there were observations available to get a measure of traffic flow at that location. The format of the traffic data collected with respect to time and location of crashes is provided in Figure 5-1.

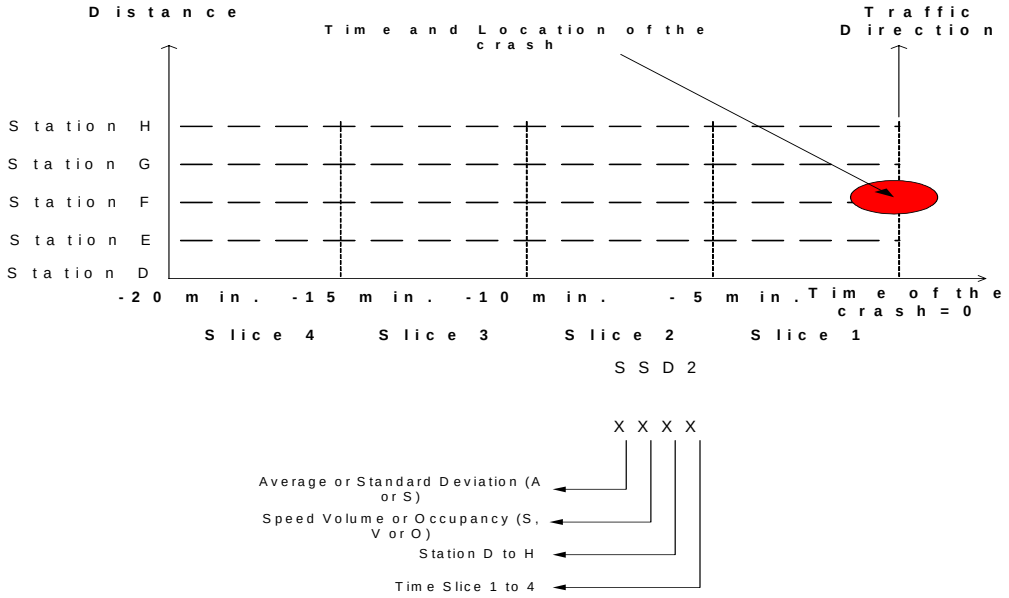

Figure 5-1: Traffic data collection in a time-space framework and nomenclature of independent variables with respect to time and location of the crash

The figure also shows the description of variable nomenclature. The variable *"SSD2"* shown for example represents the standard deviation of 30 speed observations during the 5-minute period of 5-10 minutes prior to a crash at station "D" which is the farthest upstream station. Another variable "y" was created with its value as 1 for all crash cases. It would later be used as the binary target variable in the analysis. Note that due to random intermittent failure of loops traffic data were not available for all crashes. The analysis presented in this chapter is based on 1620 rear-end crashes which had the corresponding loop data available.

5.3 Rear-end Crashes: Preliminary Explorations

As part of preliminary analysis, distributions of average speeds were explored at various loop detector locations surrounding the crash location. Figures 5-2, 5-3, and 5-4 show the histogram distributions for the variables *ASD1, ASF1*, and *ASH1* respectively, under crash and non-crash scenario. "*ASF1*" is the average of speeds measured from the three lanes at the station closest to the crash location (*Station F*) during the 5-minute period leading to the crash (*Slice 1*), while *ASD1* and ASH1 are the same parameters measured at station D (located about 1-mile upstream of the crash location) and station H (located about 1-mile downstream of the crash location), respectively. The histogram distributions for these parameters under rear-end crash scenario appear to have the shape of two adjacent approximately mound-shaped distributions. The overlapping frequencies are observed over average speed values ranging between 35 to 45 mph (Figure 5-2, 5-3 and 5-4). In contrast the average distributions under non-crash scenarios appear to have a single distribution. The two relative peaks in the frequency distributions of average traffic speeds under the rear-end crash scenario histograms suggest that the crashes belonging to each peak needs to be analyzed separately. In one of our previous studies (Abdel-Aty et al., 2005) crashes were separated by simply splitting the crash data based on the average speeds just before the crash (time slice 1, 0-5 minutes before the crash) at station F.

In this analysis, the idea of separating crashes by prevailing conditions only at station of the crash (station F) is refined. It is imperative because specifically rear-end crashes are being analyzed here. The rear-end crashes at freeway locations are expected to be affected not only by the prevailing regimes at that location but complex interaction

between traffic regimes at the locations upstream and/or downstream of it. For example, low speeds downstream of a site accompanied by high speeds on the upstream of the location would be more likely to result in a rear-end crash. To reflect this fact, it was decided to divide the rear-end crashes into two clusters/groups, not only based on just *ASF1* but on the three parameters *ASD1, ASF1,* and *ASH1*. Note that these parameters are derived from three separate stations, with station *D* and *H* being the stations located approximately one mile upstream and downstream of the crash site, respectively. They essentially represent traffic speeds measured at the extremities of a 2-mile stretch around the crash location.

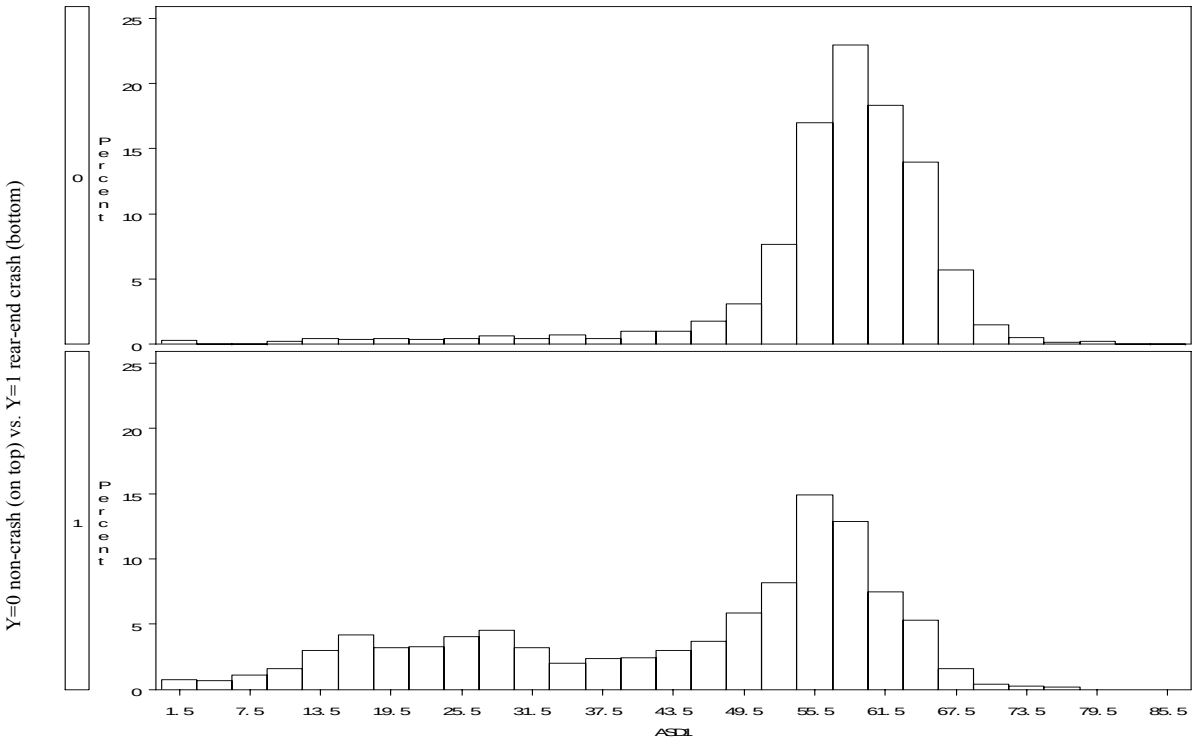

Figure 5-2: Histogram distribution of ASD1 for non-crash (on top) and rear-end crashes (bottom)

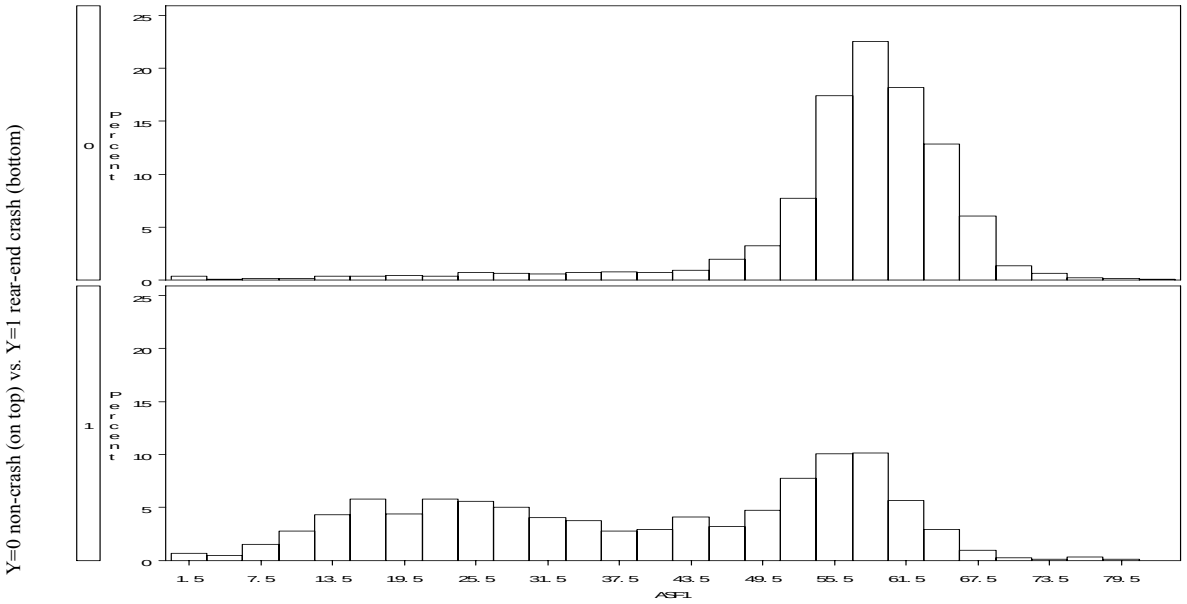

Figure 5-3: Histogram distribution of ASF1 for non-crash (on top) and rear-end crashes (bottom)

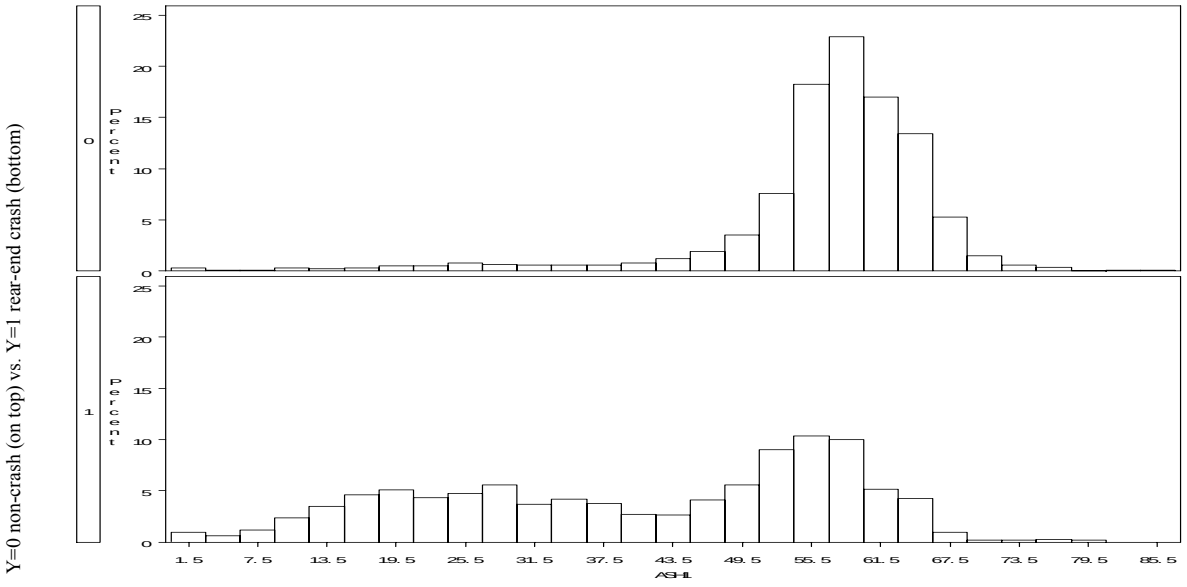

Figure 5-4: Histogram distribution of ASH1 for non-crash (on top) and rear-end crashes (bottom)

5.3.1 Clustering of rear-end crashes based on prevailing speed configurations

Clustering places objects into groups or clusters suggested by the data. The objects in each cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. Since the first objective in this analysis is to separate crashes based on different configurations of prevailing traffic speeds at upstream and/or downstream freeway locations with respect to the crash location, it was decided to divide the crash data in two clusters. Using the output from cluster analysis these two groups of crashes could be analyzed separately along with a non-crash sample. Note that there might be more than two natural clusters in the rear-end crash data based on the prevailing traffic speed regimes at the three locations (*station D, F* and *H*). The number of clusters, however, was forcibly limited to two because with more clusters, some of them might end up with smaller (and hence insufficient) sample size for further analysis of crash prone conditions within each cluster.

Kohonen vector quantization (KVQ) training technique was used to cluster the crash data into two groups based on the value of the three average speed parameters(*ASD1, ASF1,* and *ASH1*). SOM/Kohonen node of the SAS Enterprise Miner (SAS Institute, 2001) was used to employ this technique. Vector quantization networks are competitive networks that can be viewed as unsupervised density estimators or auto-associators (Kohonen, 1988). Each competitive unit corresponds to a cluster, the center of which is called a codebook vector or cluster seed. Kohonen's learning law is an online algorithm that finds the cluster seed closest to each training case and moves that "winning" seed closer to the training case. KVQ may also be used for offline learning (as is the case here), in which

case the training data is stored and Kohonen's learning law is applied to each case in turn, cycling over the data set many times (incremental training). Convergence to a local optimum can be obtained as the training time goes to infinity if the learning rate is reduced in a suitable manner. In KVQ training method one may specify the number of clusters to be created. The initialization seeds may be chosen randomly or through some preliminary analysis. In this analysis random initial seeds were chosen; hence we started with a high learning rate, of 0.5. Note that if initial seeds are obtained through some preliminary analysis, then the initial learning rate should be much lower. The theoretical and application details of the algorithm may be found in Kohonen (1988) and SAS Institute (2001), respectively.

The dataset consisting of 1620 rear-end crashes was subjected to the SOM/Kohonen node of the SAS Enterprise Miner, to divide all the crashes into two clusters. Only the three average speed variables, namely, *ASD1, ASF1,* and *ASH1* were used as input to the vector quantization technique. Of course, as any clustering algorithm, vector quantization networks are example of unsupervised learning, hence the learning is done through the input data and no target variable is specified. The output dataset from the Kohonen node, in addition to all the existing variables, consisted of a newly created binary variable named, *_segmnt_* for each crash. This variable represented the cluster assigned to each observation (i.e., a rear-end crash) in the dataset. The cluster, to which each crash belongs, based on traffic speeds at *ASD1, ASF1,* and *ASH1*, was now known. Out of 1620 rear-end crashes in the sample 47.2 % were grouped in cluster 1 while 52.8% were grouped in cluster 2.

5.3.2 Classification tree model for identification of clusters

It is intended that separate models would be developed and applied to predict the two groups (clusters) of rear-end crashes identified by the KVQ learning algorithm. From an application perspective, one must be able to identify the cluster with which an incoming real-time data under consideration belongs so that appropriate model(s) may be applied to assess whether or not it is a crash prone pattern. It can not be achieved through an unsupervised clustering technique such as the KVQ method. Therefore, a set of classification rules need to be formulated so that real-time data patterns may be assigned to one of the two clusters. These rules can also be used to identify these clusters in randomly selected loop data to examine how frequently traffic conditions belonging to the two groups of rear-end crashes occur under "normal" traffic.

Classification tree was selected as the tool to classify data into either of the two groups. A classification tree represents a segmentation of data created by applying a series of simple rules. Each rule assigns an observation to a group based on the value of one input. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The original group that contains the entire data set is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final or terminal nodes are called leaves. For each leaf, a decision is made and applied to all observations in that leaf. The choice of tree as the classification model was based on the fact that it would provide simple interpretable rules for identification of the two clusters in crash and non-crash

data. Simple rule based approach to separate clusters would add to the understanding of traffic conditions that constitute the two clusters/groups of rear-end crashes.

As explained in the previous section the clusters of crashes were obtained based on traffic speeds prevailing right before time of the crash occurrence (0-5 minutes; time-slice 1). Ideally, identification of these clusters would be best achieved if the same parameters (i.e., *ASD1, ASF1,* and *ASH1*) are now used as input to the tree model. However, since this tree model is intended to be used on real-time loop detector data as part of the stepwise procedure of freeway crash risk assessment it would be more appropriate if parameters from time-slice 2 (i.e., *ASD2, ASF2,* and *ASH2*) are used to identify the cluster each data point belongs to. It was found that the three parameters from time-slice 2 also followed similar distribution over crash cases as their slice 1 counterparts. The distributions for these three slice 2 parameters over all rear-end crashes are shown in Figure 5-5 (a, b, and c).

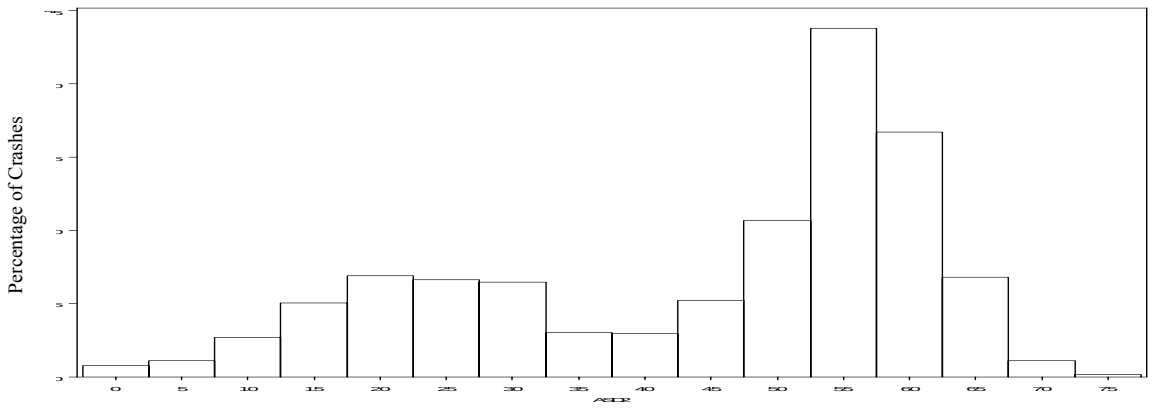

Figure 5-5 (a): Histogram distribution of ASD2 for all rear-end crashes

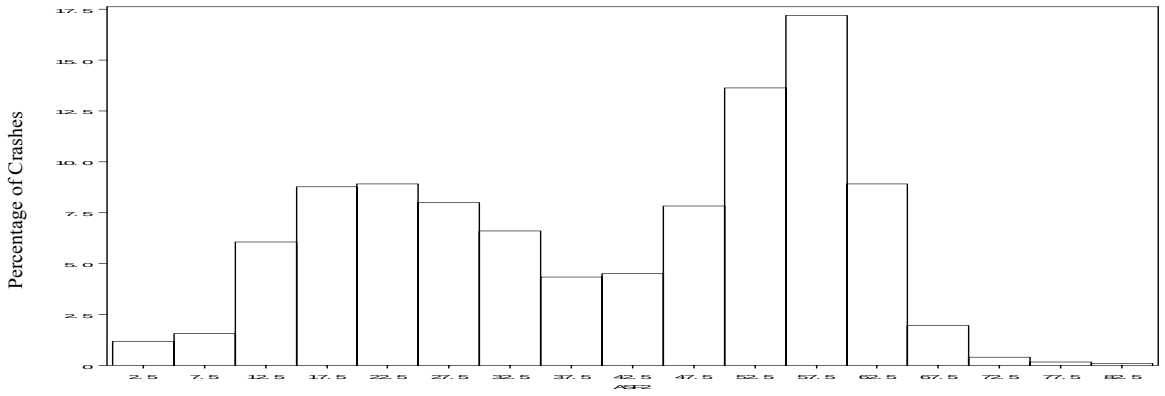

Figure 5-5 (b): Histogram distribution of ASF2 for all rear-end crashes

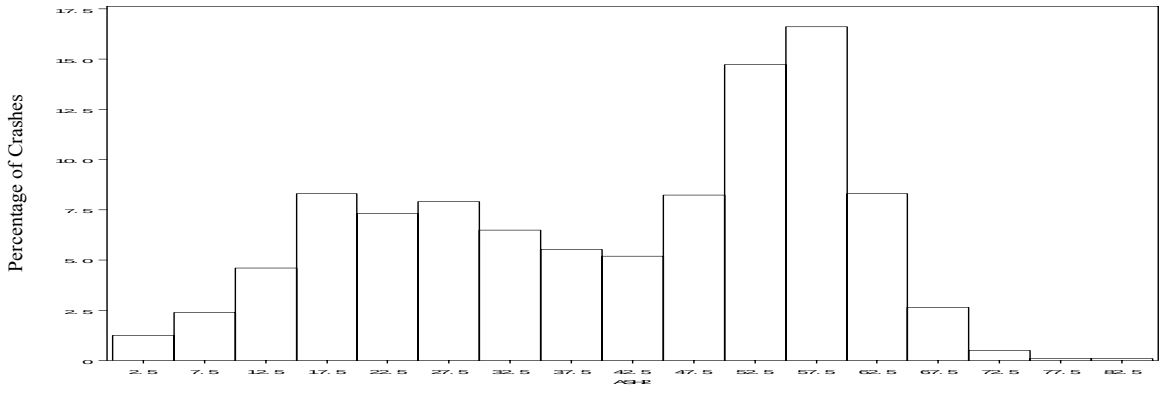

Figure 5-5 (c): Histogram distribution of ASH2 for all rear-end crashes

Figure 5-5: Histogram distributions of ASD2, ASF2, and ASH2 for all rear-end crashes

To ensure that the use of parameters from slice 2 would lead to sufficiently accurate identification of the group, to which any crash belongs, we calibrated two separate tree models; one model with input parameters ASD1, ASF1, and ASH1 (average speeds 0-5 minutes before the crash) while the other with same parameters from time slice 2 (5-10 minutes before the crash). The tree models were calibrated using the output dataset from SOM/Kohonen clustering algorithm. The dataset has loop data corresponding to 1620 rear-end crashes along with the cluster/group assigned to it by the clustering algorithm. It was decided to use 70% (i.e., 1073 crashes) data for calibration and 30% (547 crashes) data for validation of these tree models. The enterprise Miner data mining flow diagram for the process is provided in Figure 5-6.

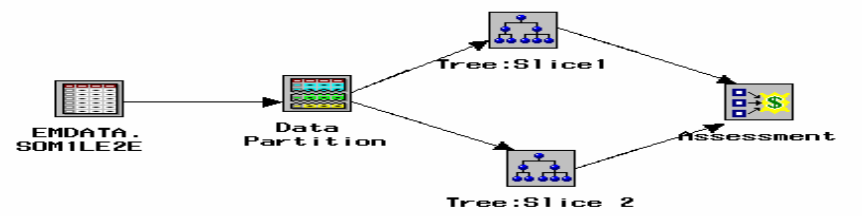

Figure 5-6: Data mining process flow diagram to develop and evaluate classification tree models for binary target variable _segmnt_ (i.e., the cluster to which each crash belongs)

In this analysis the best split among available set of candidate splits was determined using the *chi-sq.* test with *p-value=0.2* as the criterion. The results of the tree model were assessed using a lift chart. The lift chart displays the cumulative percentage response rate for the predictive models developed. The performance of a model may be measured by determining what percentage of the target event has been captured by the model at various percentiles. Figure 5-7 shows the cumulative captured response lift plots for both

classification tree models developed to identify the groups (clusters). The performance of the two models (Models to identify the cluster with traffic speed inputs from time-slice 1 and time-slice 2, respectively) is mostly comparable with the curves from both models running close to each other for the most part. The predicted output (indicated by the variable "*i__segmnt_*") from the two classification tree models for all crashes was also subjected to formal chi-sq. test and Fisher's exact test. Both tests indicated that the two outputs are closely associated. It was concluded that either of the two models (with input parameters from slice 1 or slice 2) may be used to classify the real-time loop data into two clusters (groups) identified for rear-end crashes. Since using time-slice 2 parameters is more suitable form a real-time application perspective it was decided that the tree model with slice 2 parameters would be used for segmentation of the crash data. The rules formulated by this tree model are used to separate rear-end crashes belonging to one cluster from the other. Therefore, more than the model's classification performance we are interested in the actual rules formulated by the tree. Using these rules we can score any dataset of interest and estimate the cluster (group) to which any data point would belong.

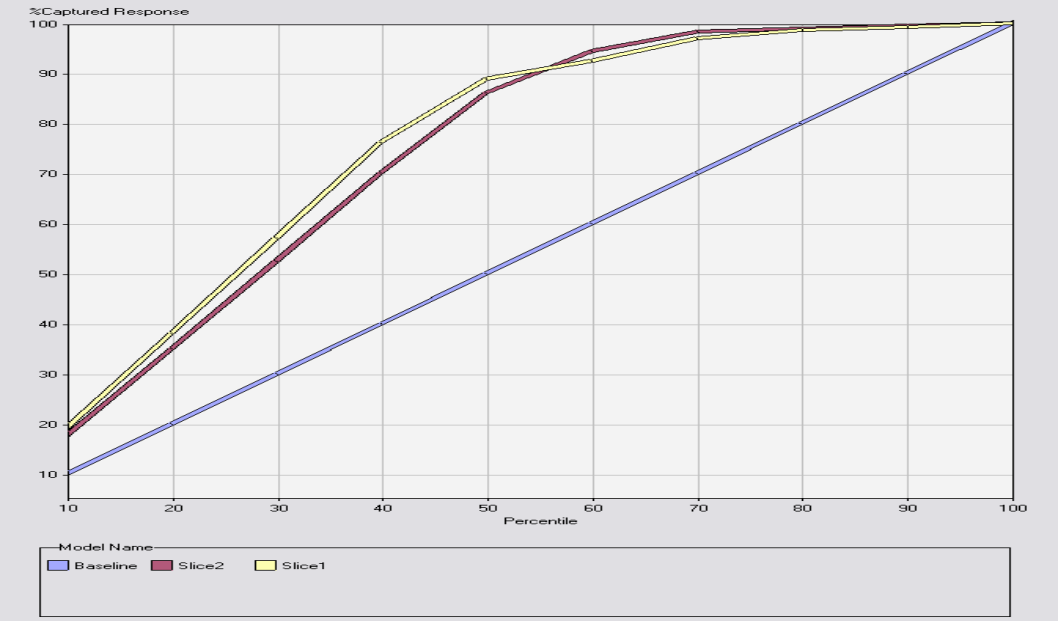

Figure 5-7: Lift Chart showing the performance of the two classification tree models on the validation dataset

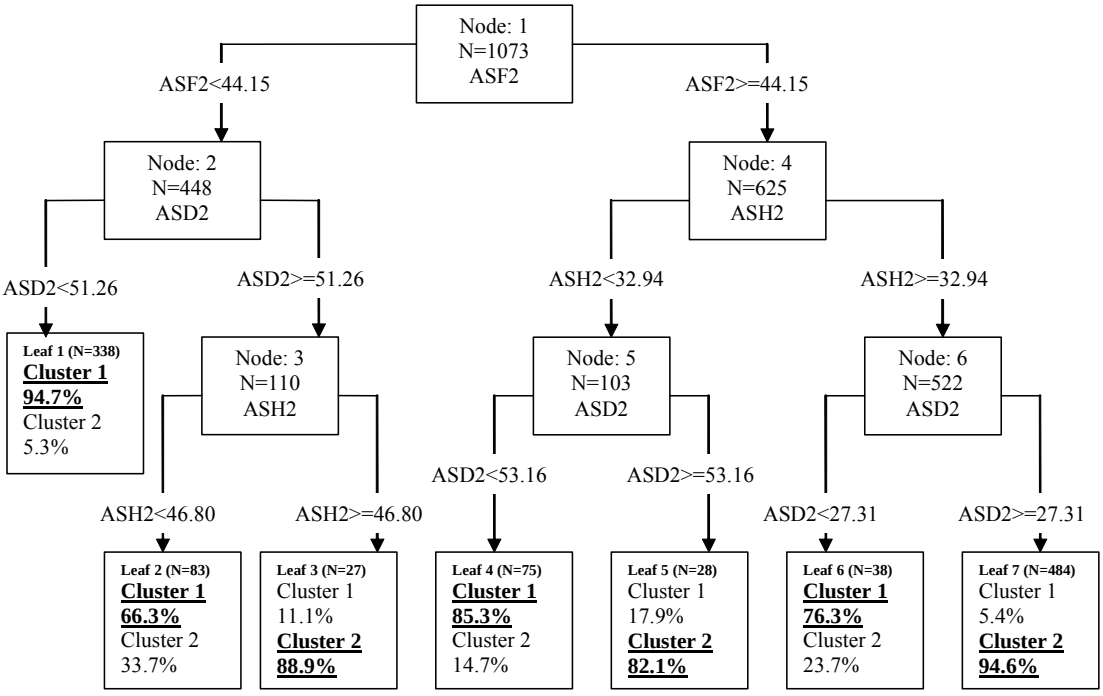

Figure 5-8: The structure of the decision tree with inputs from time-slice 2 for target variable _segmnt_

Figure 5-8 demonstrates the hierarchical structure of the classification tree produced by SAS Enterprise Miner Tree node with *ASD2, ASF2* and *ASH2* as inputs. The tree has seven terminal nodes or leaves. The initial split at the root node is based on the variable *ASF2* (average speed 5-10 minutes before the crash at the loop detector located nearest to crash location). The tree then directs observations with *ASF2* less than *44.15 MPH* to the left forming node 2; and the rest of observations form node 4. Node 2 is further split based on the speed at station located approximately 1-mile upstream of crash location (*ASD2*) and leaf 1 (terminal node, where decision is made) is created on the left with observations having *ASD2* less than *51.26 MPH.* For each leaf or terminal node in a tree structure, a decision is made and applied to all observations in that leaf. In Figure 5-8 the classification (cluster 1 or 2) assigned to the observations belonging to a particular leaf is underlined along with the posterior probability of that cluster within each leaf. As indicated by leaf 1, if *ASF2* less than *44.15* and *ASD2* less than *51.26;* the tree predicts the crash to be a cluster 1 crash. On the right of node 2, node 3 is formed which would be further split at the next level. Tree algorithm splits node 3 based on the variable *ASH2* producing leaves 2 and 3. At leaf 2, if *ASH2 < 46.80* then the rear-end crash would belong to cluster 1 (posterior probability *67%*). Leaf 3 indicates that if *ASH2 >= 46.80* then the crash would be cluster 2 (posterior probability *88.9%*).

Now the tree branches to the right of root node, i.e., nodes 4 to 6 and leaves 4 to 7 are explained. Node 4 is split based on the speed at station located approximately 1-mile downstream of crash location. Node 5 has observations having *ASH2 < 32.94 MPH* and node 6 has the observations with *ASH2>= 32.94 MPH*. Further split in node 5 creates

terminal nodes (leaves) 4 and 5. As indicated by leaf 4, if *ASD2 < 53.16* then the tree predicts the crash to be part of cluster 1. In leaf 5 where *ASD2 >= 53.16* the crash is most likely be cluster 2. Leaf 6 consist of the observations with *ASH2 >= 32.94* and *ASD2 < 27.31* and crashes in this leaf are identified as cluster 1. Last leaf has rear-end crashes with *ASH2 >= 32.94* and *ASD2 >= 27.31*. The posterior probability of these crashes belonging to cluster 2 is 94.6%.

Table 5-1 summarizes the series of rules leading to each leaf of the classification tree. Out of seven leaves four favor cluster 1 while the other three favor cluster 2. In the last two columns the table also shows the number and percentage of training dataset observations ending up in that leaf.

Table 5-1: The series of rules formulated by the classification tree model to identify clusters in rear-end crash data

Leaf	Conditions (Series of Rules)	Cluster Assigned	Number of observations	Percentage of observations (training dataset)
1	ASF2<44.146 and ASD2<51.26	cluster 1	338	31.50
2	ASF2<44.146 and ASD2>=51.26 and ASH2 <46.8	cluster 1	83	7.74
4	ASF2>=44.146 and ASH2<32.941 and ASD2<53.165	cluster 1	75	6.99
6	ASF2>44.146 and ASH2>=32.941 and ASD2<27.30	cluster 1	38	3.54
3	ASF2<44.146 and ASD2>=51.26 and ASH2 >=46.8	cluster 2	27	2.52
5	ASF2>44.146 and ASH2<32.941 and ASD2>=53.165	cluster 2	28	2.61
7	ASF2>44.146 and ASH2>=32.941 and ASD2>=27.30	cluster 2	484	45.11

From the general structure of the tree, it may be inferred that the cluster 1 rear-end crashes generally belong to low speed traffic regime, while those in cluster 2 belong to medium to high speed traffic regime. The conditions near the crash location (Station F) and upstream of it (Station D) are both somewhat congested 5-10 minutes before a cluster 1 crash as indicated by leaf 1. Leaf 2, which also belongs to cluster 1 rear-end crashes indicates although the speeds upstream of the station of the crash are high (ASD2>=51.26); 5-minute average speeds at station of crash as well as downstream of it are on the lower side. Note that barring leaf 6 (which only has 38 observations) all leaves

classifying crashes into cluster 1 have the condition of lower speeds at two out of three stations. It indicates that cluster 1 rear-end crashes may be identified with persisting congested conditions over longer periods (at least 10 minutes) and extended segments (approximately one mile indicated by two stations). The leaf or terminal node with maximum (almost 90%) observations resulting in cluster 2 classification is terminal node 7. These observations generally have higher speeds at all three stations. Hence, most of cluster 2 crashes occur under relatively free flow conditions that commonly prevail on freeways. Based on these observations one could infer that cluster 1 rear-end crashes occur during congested conditions that prevail on the freeway for small part of the day and have very low exposure. Cluster 2 rear-end crashes mostly occur under traffic speed conditions with higher exposure.

Based on the discussion provided above, the traffic speed conditions which result in cluster 1 classification (Rows 1 through 4 of Table 5-1) are referred to as traffic regime 1 and the traffic speed conditions which result in cluster 2 classification (Rows 5 through 7 of Table 5-1) are referred to as traffic regime 2. The crashes that occur in these two traffic regimes are referred as regime 1 and regime 2 rear-end crashes, respectively. In the next section properties of these two groups of rear-end crashes are explored. We would also be applying the tree model (shown in Figure 5-8) to score a randomly selected sample of loop detector data to verify the inference about the exposure of the conditions belonging to the two traffic regimes.

5.3.3 Properties of rear-end crashes belonging to the two regimes

It was apparent from the model performance evaluations that the classification tree model with parameters from time-slice 2 was able to identify, with sufficient accuracy, the cluster to which any rear-end crash belonged. Therefore, as the next step in the analysis we scored the dataset consisting all rear-end crashes with the tree model developed in the previous section and obtained a variable named *i__segmnt_* depicting the classification assigned to it. Samples from this dataset were used earlier to train and validate the classification tree model. Note that the rules formulated by the classification tree model may now be used to score any sample of traffic data. A sample of the available random non-crash cases, to be used later in the chapter for crash-non-crash modeling, was also scored using the slice 2 tree model to examine the frequency of the two regimes under randomly sampled traffic data. It would let us verify the postulation that the exposure of traffic speed conditions belonging to regime 1 is much less than those belonging to regime 2.

The frequency of each cluster as identified by the tree model in the rear-end crash dataset is shown in Table 5-2. In the output dataset obtained from the SOM/Kohonen clustering procedure (used to calibrate the tree model) 47.2% crashes were grouped into cluster 1 and 52.8% were grouped into cluster 2. Since Table 5-2 is based on the prediction obtained from the classification tree it slightly differs from the output of clustering procedure. We could have alternatively used that output from SOM/Kohonen node for exploratory analysis of the rear-end crashes in the two regimes. However, since the frequency of the two clusters on randomly selected non-crash cases may only be obtained

by scoring it using the tree model it was decided to be consistent and use the datasets scored by the tree model for both crash as well as non-crash cases. On a random non-crash dataset (with 7030 observations) scored with the tree model, the frequency of the two clusters is provided in Table 5-3.

Table 5-2: Frequency table of clusters idetified by the tree model for all rear-end crashes

Classification assigned by the Tree model	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cluster 1	742	45.8	742	45.8
Cluster 2	878	54.2	1620	100

Table 5-3: Frequency table of clusters idetified by the tree model for a sample of random non-crash cases

Classification assigned by the Tree model	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cluster 1	441	6.27	441	6.27
Cluster 2	6589	93.73	7030	100

Note that regime 1 makeup 45.8% of the crash dataset but it only makes 6.27% of the random non-crash sample. It indicates that the crashes belonging to regime 1 may be 'predicted' (or anticipated) using the tree model calibrated in the previous section. If we predict all traffic patterns belonging to regime 1 as rear-end crashes, we would be able to identify about 46% of rear-end crashes by issuing warnings just over 6% of the times. This by itself is a significant finding and it would be recalled later to formulate an online application strategy. It also verifies the postulation about very less exposure for the speed conditions belonging to traffic regime 1.

While analyzing the average speed distributions under rear-end crash scenario it was observed that the two mound shape distributions for average speeds were reduced to a

single mound shape distribution within each regime. Before proceeding with complex prediction models crash and non-crash frequency histogram distributions of some of the variables known to be related to freeway crashes were examined for the two clusters of rear-end crashes. In the following figures, the top pair of histograms belong to regime 1 rear-end crashes while the bottom pair belongs to regime 2. On the left distribution over crash cases is shown while on the right the distribution over randomly selected non-crash cases.

Figures 5-9 and 5-10 depict these histograms for *CVSF2* (coefficient of variation in speed at station of crash) and *AOG2* (average occupancy downstream of the crash site). These two real-time traffic parameters have been known to be critically associated with freeway crash occurrence in some of the generic models developed earlier (Abdel-Aty et al., 2004, 2005). These parameters are considered critical here; even while separately analyzing rear-end crashes, because generic models are expected to be biased towards rear-end crashes due to their relatively high frequency on freeways. It may be inferred from the figures (Figures 5-9 and 5-10) that while *CVSF2* would be an important predictor in the case of only regime 1 rear-end crashes (*CVSF2* has identical distribution over crash and non-crash cases for regime 2 rear-end crashes), *AOG2* might be critically associated with rear-end crashes from both regimes.

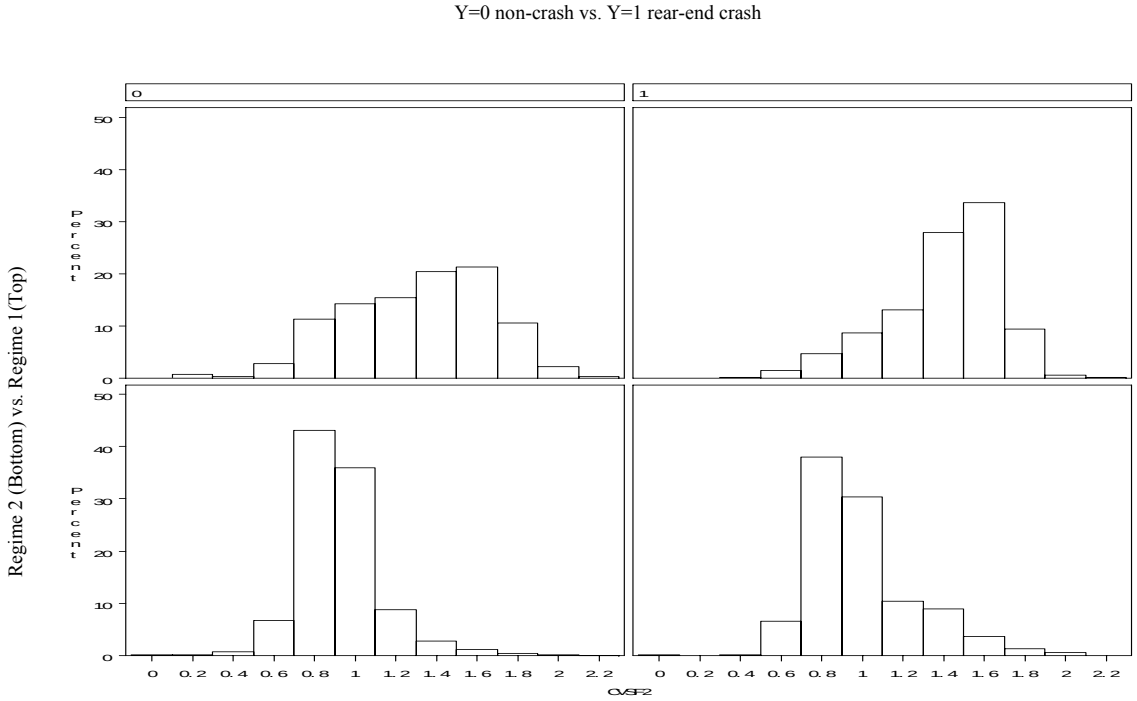

Figure 5-9: Histogram distribution of CVSF2 wrt binary variables _ segmnt _ and Y (crash vs. non-crash)

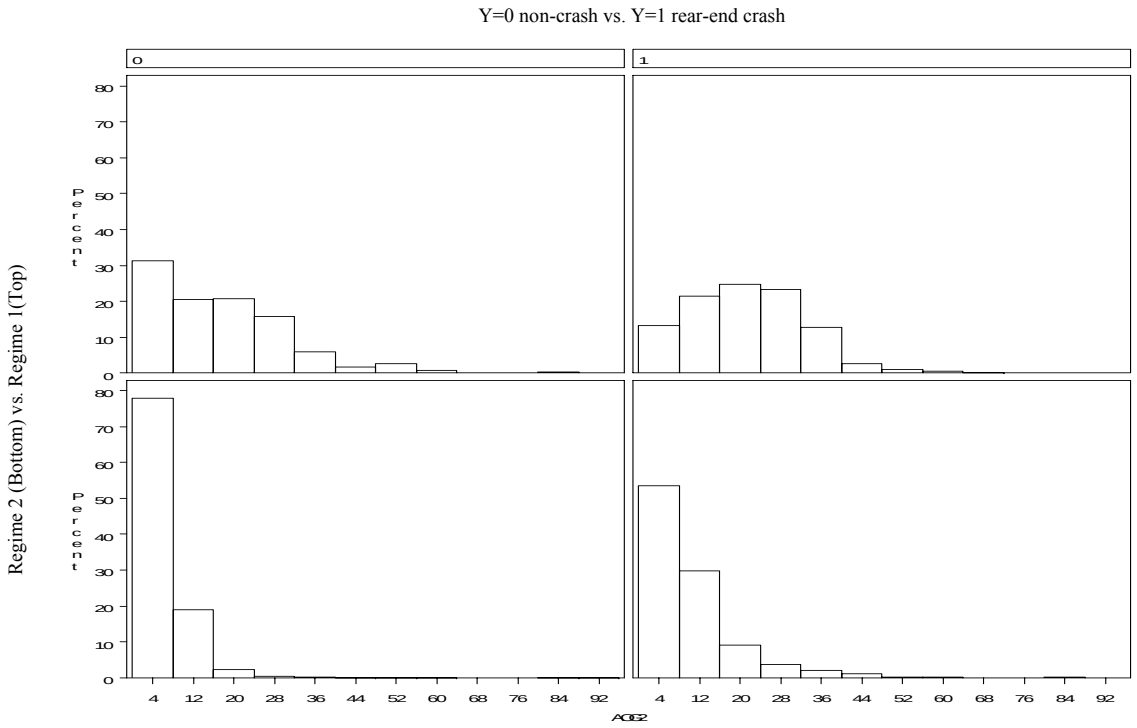

Figure 5-10: Histogram distribution of AOG2 wrt binary variables _ segmnt _ and Y (crash vs. non-crash)

The distributions of these two groups of rear-end crashes by time of the day, day of the week and mile-post location on the 36-mile corridor under consideration was analyzed next. As mentioned earlier, in this analysis randomly sampled non-crash cases have been used rather than the matched sampling design; hence the effects of offline (i.e., static and location specific) factors on rear-end crashes can be examined. In the matched case-control analysis these factors were implicitly controlled for by the study design. Figures 5-11 through 5-13 show histogram distributions of these off-line factors for the two groups of rear-end crashes.

In Figure 5-11 it may be seen that crashes belonging to both regimes are almost equally frequent from Monday through Thursday. Regime 1 crashes (identified by leaves 1, 2, 4 and 6 of the classification tree shown in Figure 5-8) are less frequent on weekends (Saturday and Sunday); which is expected since they are mostly related to more congested traffic conditions (See Table 5-1). On Fridays, however, both types of rear-end crashes are more frequent compared to other weekdays (Monday through Thursday). It indicates that on Friday crash prone conditions on the freeway might be more prevalent along the freeway corridor. From an field application perspective it might mean more warnings on Fridays.

Frequency distribution of these two groups of rear-end crashes is shown with respect to time of the day (expressed in terms of seconds past midnight) in Figure 5-12. While frequency of regime 1 crashes peaks during morning (7:30 to 8:30 AM) and afternoon peak period (3:45 to 5:15 PM); for regime 2 crashes it is the maximum just before the

afternoon peak sets in. It may be inferred that the conditions prone to regime 2 rear-end crashes generally prevail in the off-peak period or just prior to the beginning of peak traffic on the freeway.

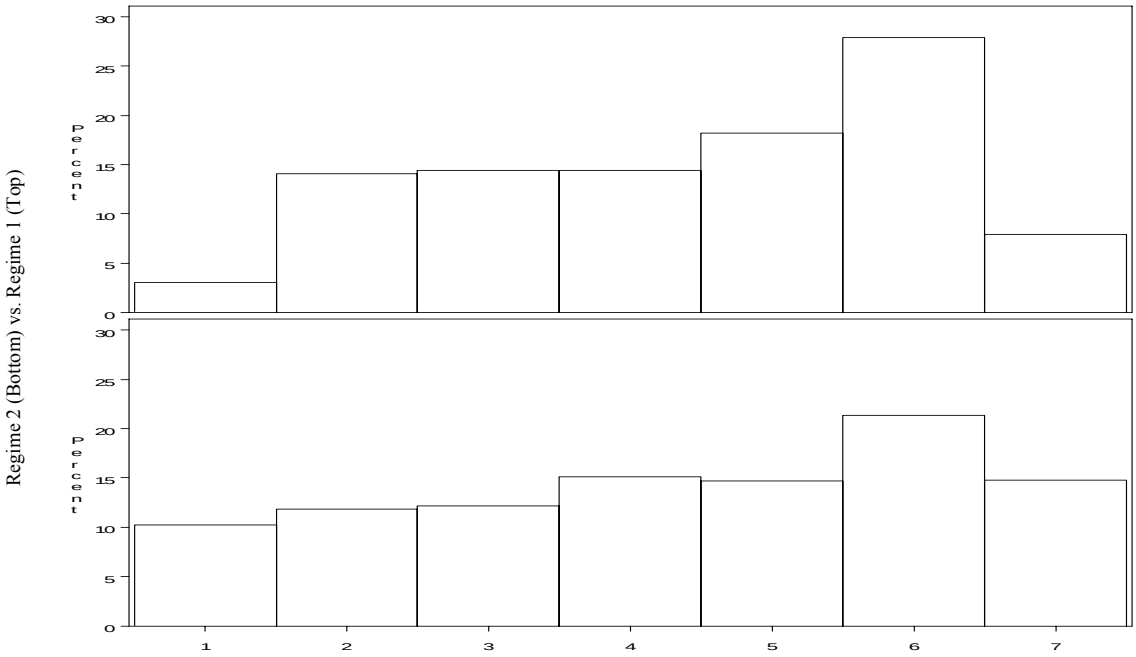

Figure 5-11: Histogram distribution of crashes from two regimes over day of the week (1: Sunday to 7: Saturday)

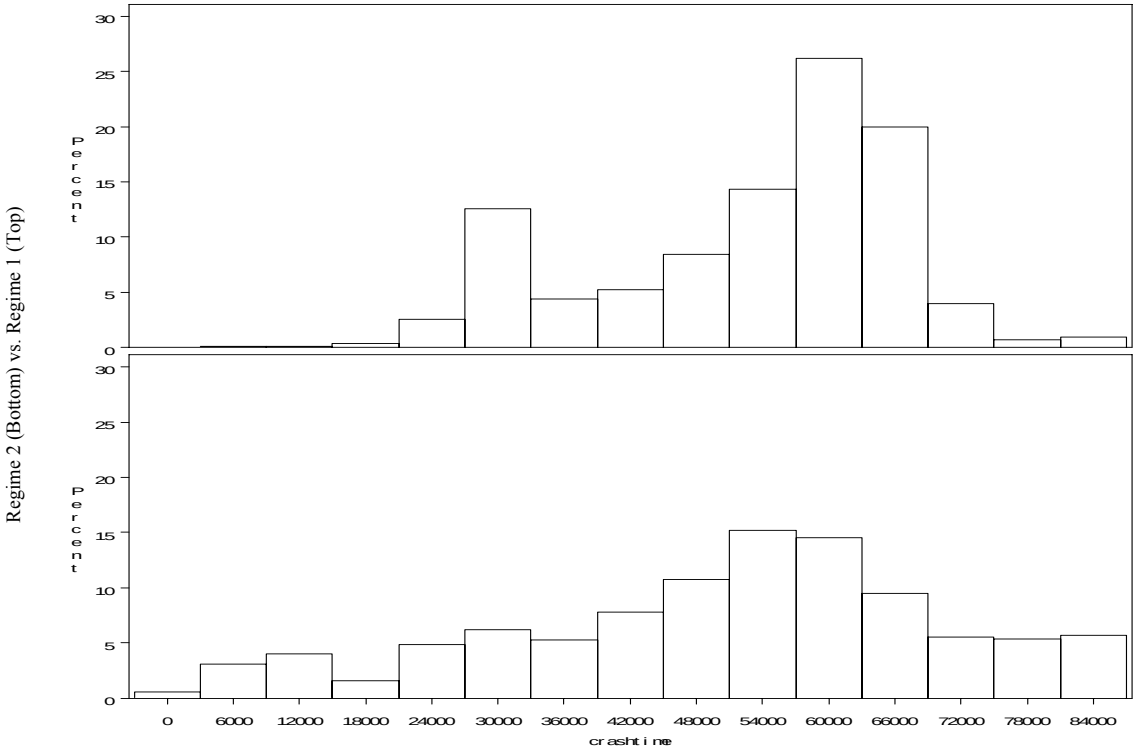

Figure 5-12: Distribution of rear-end crashes belonging to two regimes over time of the day expressed in terms of seconds past midnight

From the frequency distribution of crashes over the corridor by milepost (Figure 5-13), it may be inferred that regime 1 rear-end crashes are prevalent in the stretch located in downtown Orlando area (milepost location 18<*base_milepost*<27) of the corridor. Regime 2 crashes, although peak in the same stretch of the freeway, show lower but relatively significant frequency in and around Disney area (*base_milepost*>33) and at the beginning of Orlando city limits (12<*base_milepost*<15). These plots indicate that these off-line factors might be critical in identifying recurring crash prone conditions, especially for regime 2 rear-end crashes.

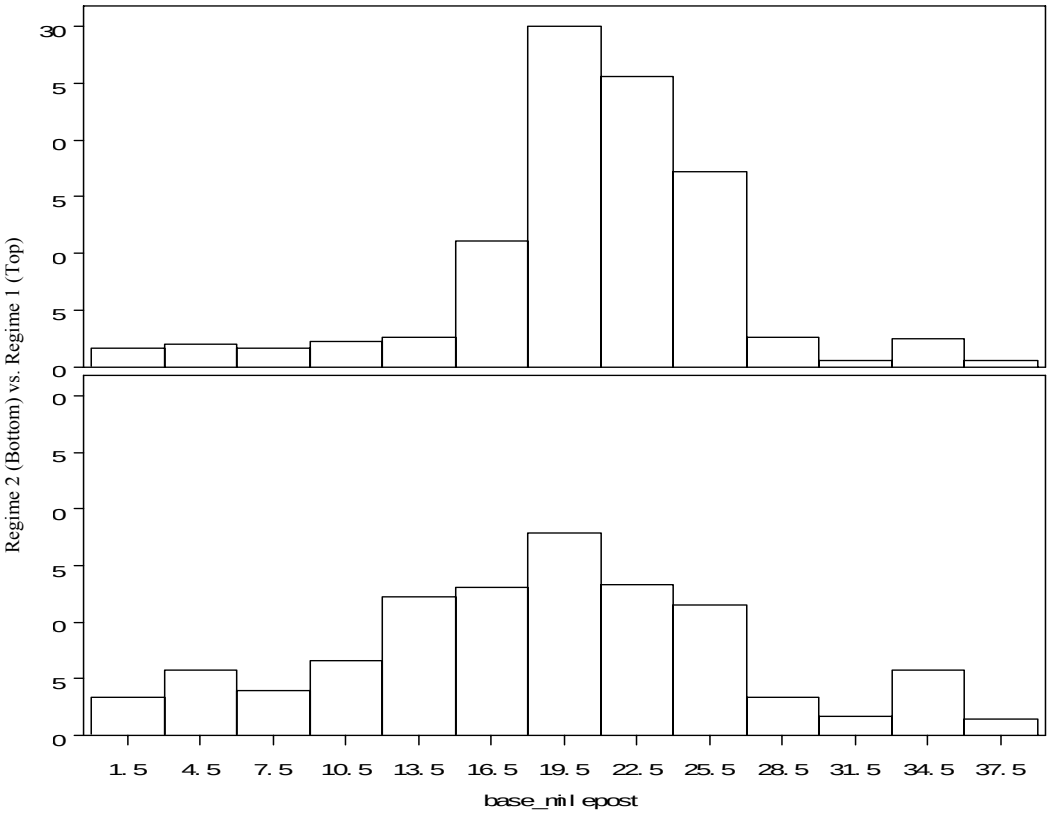

Figure 5-13: Distribution of mile post location of the rear-end crashes belonging to two regimes

We next examined the frequency distributions of mile-post location over both crash and random non-crash data. Random non-crash database with 7030 observations was scored with the classification tree model to identify the cluster to which these data point belong (i.e., traffic regime). The resulting frequency distribution of these two regimes in the non-crash dataset was shown in Table 5-3 earlier. The distribution of milepost locations is shown for both crash (with 1620 observations) and non-crash data (having 7030 observations) by regime in Figure 5-14. It was noticed that while non-crash data belonging to regime 2 were almost uniformly distributed over the freeway corridor, it was not the case with the non-crash data belonging to regime 1. For regime 1 non-crash cases

frequency distribution was somewhat similar to the distribution for crashes belonging to this regime. Note that the four distributions shown in Figure 5-14 are all based on different number of observations. While the rear-end crash location distribution is based on 742 and 878 crashes for regime 1 and 2 respectively; the non-crash location distributions are based on 441 and 6589 observations for regime 1 and regime 2, respectively.

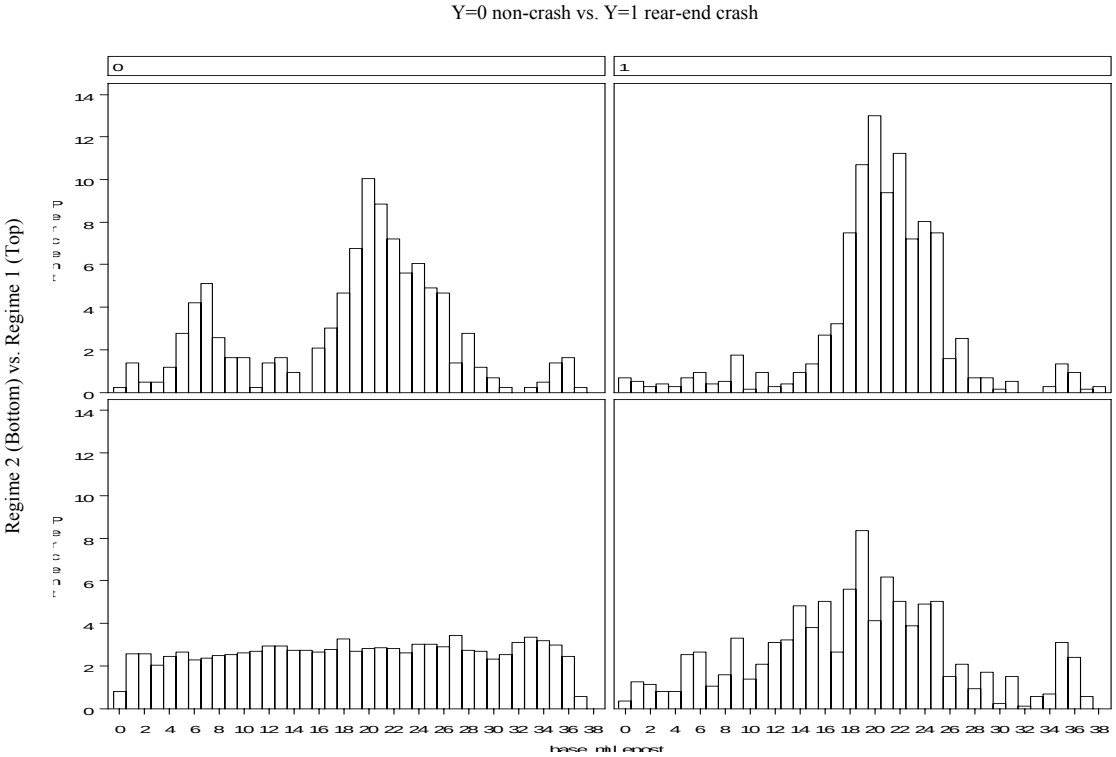

Figure 5-14: Histogram distribution of "base_milepost" wrt binary variables _ segmnt _ and Y (crash vs. non-crash)

The exploratory analyses of these distributions provide critical inferences about association of some of the real-time and static variables with rear-end crash occurrence. In the subsequent sections detailed classification analysis has been carried out to separate

rear-end crashes belonging to each of the two regimes from randomly selected non-crash cases.

5.4 Models for Rear-end Crashes: Procedure and Relevant Issues

In this chapter a data mining approach has been adopted to separate rear-end crashes from non-crash data. The research problem is formulated as a classification problem and the outcome of interest is a rear-end crash vs. no crash (with binary target variable *y=1* for rear-end crash vs. *y=0* for non-crash). The non-crash data were part of random sample drawn from available historical loop detector database. The process of drawing this random sample has been described in the chapter discussing data aggregation issues (Chapter 4).

In the previous section 1620 rear-end crashes (from 1999 through 2003) were divided into two clusters based on prevailing traffic speed configurations prior to their occurrence. There were 742 crashes belonging to regime 1 (identified by leaves 1, 2, 4 and 6 of the classification tree shown in Figure 5-8) while 878 were identified as belonging to regime 2 (identified by leaves 3, 5 and 7 of the classification tree shown in Figure 5-8). It was also observed through some exploratory analysis that crashes from the two regimes show distinct patterns not only in terms of traffic speed configurations upstream and downstream of the crash location but in terms of freeway characteristics at crash location. Hence, it is only logical that separate set of models are developed for the two groups of rear-end crashes.

SAS Institute (2001) defines data mining as the process of **S**electing, **E**xploring, **M**odifying, **M**odeling, and **A**ssessing (SEMMA) large amounts of data to uncover previously unknown patterns that can be utilized for business advantage. Enterprise Miner software from SAS Institute (2001) is used to implement SEMMA data mining process for the research problem at hand. SAS Enterprise Miner contains a collection of sophisticated modeling and data preparation tools with a common user-friendly interface. It may be conveniently used to create and compare multiple models. The modeling tools included in the Miner include decision trees, regression, and neural networks. The theoretical background of these tools is provided in Chapter 3. Note that Miner may also be used to create hybrid or ensemble of multiple models. In this research the terms 'hybrid' and 'ensemble' are used interchangeably.

Among the available modeling tools, the classification trees are considered unstable and are usually recommended for variable selection at the data preparation stage. Brieman et al. (1984) devised a variable importance measure (*VIM*) for trees. *VIM* may be used as a criterion to select promising subset of variables for other flexible modeling tools such as the neural networks (See Chapter 3 for details). As a data preparation tool classification trees offer interpretability, no strict assumptions concerning the functional form of the model and computational efficiency. Two different types of neural network architectures; the multi-layer perceptron (MLP) and the radial basis function (NRBF) were explored as the tools to develop crash vs. non-crash classification models. The theoretical details of these tools may be found in any standard neural network text e.g., Haykin (1999) or Christodoulou and Georgiopoulos (2001). Besides neural networks, logistic regression

models were also estimated for the binary target. Classification tree was not used to select variables for logistic regression. The reason being that unlike neural networks, logistic regression is not a "flexible" modeling technique. Standard variable selection procedures, forward, backward, and step-wise were used, instead. The details of these selection procedures may be found in Collett (1991).

There were some critical issues that needed to be addressed before proceeding with the modeling exercise. First critical question was the proportion of crash and non-crash cases in the dataset used for modeling. The crashes, however frequent on Interstate-4 corridor under consideration, are still rare events. Sampling their actual proportion in the dataset would mean that the sample would be heavily biased towards non-crash cases (crash cases even less than *0.001 %*). Also, even though the stated goal of these models is to predict crashes, actual phenomena of interest are crash prone conditions. It is reasonable to assume that the crash prone conditions, which would be worth issuing warnings, are more frequent than the crashes themselves. For any model intended to be applied in real-time the ideal sample composition for modeling would have proportion of the two competing events same as that in reality. However, there is no way, at this stage anyways, to estimate the proportion of crash prone conditions on the freeway. Also, since the number of warnings beyond a certain point would mean "unreasonable" number of false alarms the decision from the models can not be positive (i.e., a crash) for something like *50%* of the time. Hence, a sample with equal number of crash and non-crash cases would not make an ideal sample. At this point *15%* was deemed to be an appropriate proportion of times at which conditions may be considered crash prone and warning can be issued.

With sample consisting of crash and non-crash data ratio in the same range as their expected proportion in reality or at least the proportion we intend to issue warnings; there would be no need to do prior probability adjustments at the modeling stage. Therefore, in the datasets to be used for modeling (training or validation) the crash non-crash ratio was kept at *15:85*.

Another problem that arises due to imbalance in the proportions of crashes vs. non-crash cases is related to model performance evaluation. Usually the overall classification accuracy of the model on the validation dataset is an appropriate measure to judge the performance of the model and compare it with other models. However, with only *15%* of the crashes in the sample used for modeling; *85%* overall classification accuracy could be achieved by a model that merely classifies every data point as non-crash. Such a model would of course be useless for crash identification. Also, since the classification performance of the models would vary based on the cut-off set on the output from the models (i.e., the posterior probability) even the classification accuracy over each individual class (at a certain cut-off) would not be appropriate to compare performance of competing models. It will only reflect the performance of the model at a predetermined threshold on output posterior probability. Therefore, a continuous measure of performance evaluation was needed instead and it was decided to compare the models using the cumulative percentage of captured response lift plot for validation dataset. Appropriate cut-off on posterior probability for real-time application may be chosen at the application stage based on the performance of the best model on the real-time data.

It was also decided to examine parameters from one time slice in one model. It will not only avoid the autocorrelation problems but would also lead to an easy practical implementation plan. Using data from the same time duration would be easier than to collect data and wait for the model estimation until after the data from next time slice is recorded. The models presented here are based on parameters calculated between 5-10 minutes before the time of crash (i.e., parameters from time slice 2). Note that we did try to use data from other time slices, i.e., slice 3 and 4 (10-15 and 15-20 minutes before the crash occurrence). Those models would provide more time for application and prediction of crashes. However, it was noticed that conditions 10-20 minutes before the crash did not have sufficient discriminatory power to separate crashes from non-crash cases.

As mentioned earlier, with the tree model shown in Figure 5-8 one can identify the two clusters in non-crash data as well. Therefore, last but not the least, critical question related to sampling was whether to use non-crash data only belonging to individual regimes (regime 1 or regime 2) for modeling each group of rear-end crash. An alternate way would be to choose two separate random samples of appropriate size from the loop detector database and use them as non-crash data irrespective of the traffic regime (as identified by the tree model depicted in Figure 5-8) they belong. Models developed with this approach for regime 1 crashes would be able to separate these crashes from normal traffic conditions. Similarly the models developed for regime 2 would separate regime 2 crashes from normal traffic conditions. The advantage of this approach is that we can identify different factors responsible to discriminate these crashes from normal traffic conditions. However, from an application point of view a better approach might be to

declare every case in the random dataset identified as regime 1 to be a crash. It would be appropriate because the regime 1 (extended slow moving traffic conditions in time and space) make about 46% of rear-end crashes while they make up just over 6% of the sample if it is drawn randomly from a uniform distribution. It essentially implies that by issuing warnings 6-7% of the times we would be able to identify almost half of the rear-end crashes. Also, then there is no need to develop separate models for regime 1 rear-end crashes. One could draw a random sample from the loop detector database, score it with the tree model developed above and remove the 6-7 % observations that are classified as regime 1. The remaining observations can be used as the non-crash samples to develop models for regime 2 rear-end crashes. In this chapter both approaches are explored since both of them have their advantages. Same data mining process is used for analysis based on both approaches.

The mining process is initiated by applying necessary transformation to some of the variables. It includes creation of new ordinal variables through *"Optimal Binning for Relationship to Target"* transformation on continuous variables. The aforementioned transformation optimally splits a variable into *n* groups with regards to the binary target. This binning transformation is useful when a nonlinear relationship is suspected between the input variable and target. An ordinal measurement level is assigned to the transformed variable. "*Transform Variables"* node of the SAS Enterprise Miner was used to achieve this transformation (SAS Institute, 2001). To create the *n* optimal groups, the node applies a recursive process of splitting the variable into groups until the association of the resulting ordinal variable with the target is maximized.

Some of the critical off-line factors, such as "*base_milepost*"(representing mile post location of the crash and non-crash cases), distances of the nearest on and off ramp in the upstream and downstream directions from crash location, namely, "*upstreamon*", "*upstreamoff*", "*downstreamon*", and "*downstreamoff*" were transformed along with "*timeofcrash*" using this procedure. In their original continuous form these variables were not suitable for real-time crash prediction system aimed in this research because their value would change continuously through the freeway corridor. Also, it was logical to combine some hours in the day since the traffic conditions remains largely similar during these hours. Hence, these variables were transformed into ordinal variables having maximum association with the binary target variable. The data was then subjected to the tree model to perform variable selection for subsequent neural network models. For logistic regression models the data was subjected to data partition node without being subjected to the "*Variable Selection Tree"* node. For neural network models data was partitioned after variable selection. In both cases standard 70:30 split was used to obtain training and validation dataset, respectively. Note that a stratified random sampling with binary target variable *y* as the stratification variable was used to partition the data, so that 15:85 crash vs. non-crash ratio is maintained in both training and validation datasets. Note that if the input dataset was balanced in terms of the target there was no need to use stratification along target variable at the partitioning phase.

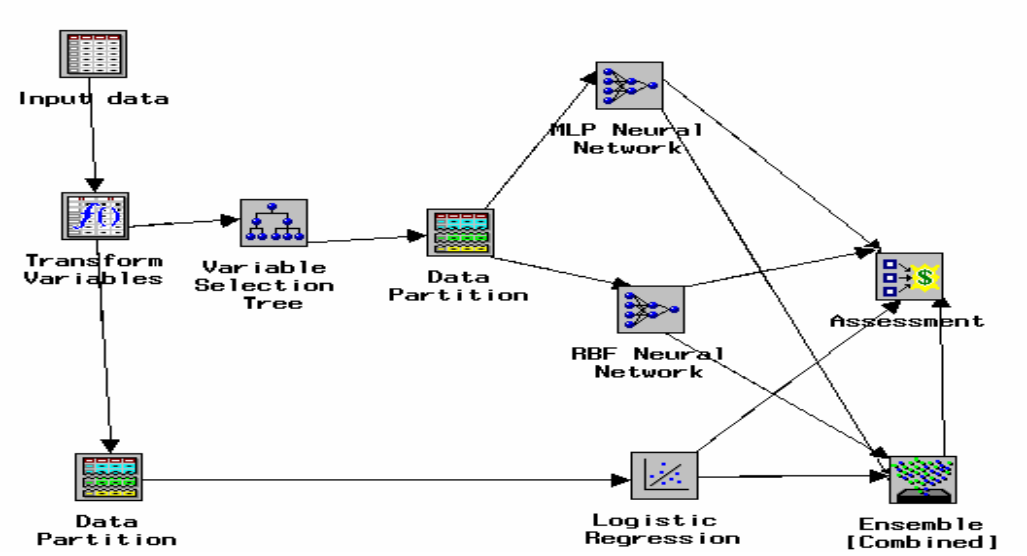

Figure 5-15: Generic data mining process flow diagram

The data partition node following the "*Variable Selection Tree"* node was followed by two sets of neural networks. The first set was used to explore MLP neural network architecture, while the other set was used to explore NRBF neural network architecture. The data partition node, following the *"transform variables"* node was followed by logistic regression models. Performance of different models was assessed using cumulative percentage of captured response plots generated by the "*Assessment"* node of the Enterprise Miner. The generic form of this data mining process is depicted in Figure 5-15.

The first neural network architecture explored for classification is the multi-layer perceptron (MLP) with Levenberg-Marquardt training algorithm. The training procedure starts with an arbitrary randomly chosen set of interconnection weights and then it tries to minimize the difference between network output and the desired outputs for the training

dataset. All runs have been carried out with a maximum number of epochs (a complete list presentation) of 1500, and error goal of 0.01. It has been proven in the literature that an MLP structure with one hidden layer and nonlinear activation functions for the hidden nodes can learn to approximate virtually any function to any degree of accuracy (Cybenko, 1989). The most critical issue then, was to estimate the number of neurons in the hidden layer. The underestimation of hidden neurons leads to a network having an incomplete representation of inputs and by contrast, the over representation reduces the network to a simple look-up table. The methodology adopted for selecting appropriate number of nodes in the hidden layer was to evaluate the performance of the models having hidden nodes varying from 1 through 10. To achieve this; 10 separate *"neural network"* nodes were used in the Enterprise Miner process flow diagram.

For RBF architecture normalized networks were chosen over the ordinary RBFs. The normalized radial basis function (NRBF) networks use the *softmax* activation function applied to radial combination of inputs. The *softmax* constraint causes the basis function to have a distributed effect and makes the network more flexible. Also, of the five varieties available for NRBF (Described in Chapter 3, discussing theoretical details of the methodology) networks, the unconstrained network was chosen since it is the most general form of the network and at this point the underlying relationship of the independent variables with the binary target was not clear. To select appropriate number of nodes in the hidden layer performance of 10 different NRBF networks, with hidden nodes varying from 1 through 10, was examined. Note that the generic data mining process flow diagram (Figure 5-15) shows only one *MLP neural network*, *RBF neural*

network and *logistic regression* node each for demonstration purposes. In reality the process flow diagram consisted of separate neural network nodes to determine the optimal structure (# of hidden nodes) of the hidden layer for each type of neural network. Three separate regression nodes were used also to estimate models utilizing backward, forward and stepwise variable selection procedure. This modeling and assessment procedure would yield the best model for each modeling technique (i.e., NRBF, MLP and Logistic regression). The single best models from each of these techniques were then hybridized using "*Ensemble*" node by averaging the posterior probability from individual models. Performance of the individual best models was compared to that of the combined model through the cumulative percentage of captured response lift plots generated by the "*Assessment"* node. A final model may be arrived at by choosing the best among these four (best NRBF, best MLP, best regression and Combined) models.

As described in the data preparation chapter the loop detector data from 7 stations (3 stations each upstream and downstream along with station of crash; from station C through I) around the crash location were collected. Initially real-time traffic parameters from just the station of crash (Station F) along with the off-line factors were used as potential input parameters to the variable selection procedures. The off-line factors considered were driver population parameters from induced exposure analysis (See Chapter 4 describing data preparation), radius of the horizontal curve, binary variable "*stationf*" (representing the location of station of crash with respect to crash location). Besides, variables created through optimal binning transformation of variables; time of crash, mile-post location and distances of nearest ramps were also included. The best

model with traffic parameters only from station of crash was identified using the procedure described above.

The modeling and evaluation procedure was then repeated by using real-time traffic parameters from three stations (i.e. one station upstream and downstream each besides the station of the crash) along with the aforementioned off-line factors. In the next step traffic parameters from two more stations from either extreme (Station D and Station H) were added as potential independent variables and the modeling procedure was repeated. The choice of estimating models that include parameters from all seven stations (Station C through Station I) was not exercised. The reason being that due to intermittently missing loop detector data from certain stations almost half of the observations would not fulfill the complete case analysis requirements of the methodologies used here for modeling (Abdel-Aty et al., 2005).

Hence, the modeling and performance evaluation exercise was repeated thrice with real-time traffic parameters from 1 (Station F), 3 (Station E, F and G) and 5 (Station D, E, F, G, and H) loop detector stations, respectively. Note that while estimating each of the three sets of models all the off-line factors mentioned above were examined as potential independent variables. It was noticed that there were very few disagreement among models that used real-time traffic parameters from same stations even though they were developed through modeling techniques as diverse as logistic regression and neural networks. Therefore, as expected these models did not improve on the performance of individual models when combined with each other by averaging the output posterior

probability. However, the models developed with loop data from different stations were observed to have different outputs for the same validation cases and were in fact expected to provide better performance than the individual models when hybridized or combined using the "Ensemble" node of the SAS Enterprise Miner. This was the main reason to estimate these three sets of models and then examine the performances of the combined model(s) created from the best model in each set. Note that this procedure was separately applied for rear-end crashes belonging to regime 1 as well as regime 2.

In the following section models to separate regime 1 rear-end crashes from randomly selected non-crash data are presented. In the next section similar models are developed for regime 2 rear-end crashes. Note that the non-crash data used for these models are completely random sample from the loop detector database. In section that follows these two sections we develop another set of models for regime 2 rear-end crashes. In which case, the random non-crash database was first scored using the rules formulated by the classification tree model (See Figure 5-8 and Table 5-1) developed earlier in this chapter. The observations that were classified by the tree model as belonging to regime 1 were removed from the non-crash sample. Conclusions from application of this modeling procedure on three separate datasets are summarized in the last section of this chapter.

5.5 Analysis and Results: Regime 1 Rear-end Crashes vs. Random Non-crash Data

Rules formulated by the tree model shown in Figure 5-8 indicate that low traffic speed conditions prevail around the potential crash location 5-10 minutes before regime 1 rear-

end crashes. Hence, these crashes may generally be associated with high average occupancies and are expected to occur during frequent formation/dissipation of ephemeral queues. Crash types other than rear-end (such as side-swipe or angle) were almost non-existent under such scenario. From one of our previous studies (Abdel-aty et al., 2004) it was inferred that coefficient of variation in speed might be a significant predictor of rear-end crashes under these traffic regimes. Hence, the average and standard deviation of speed were replaced by respective *LogCVS* values defined by Log_{10} (SS/AS*100).

As the first step in the SEMMA data mining process transformations were also applied to the critical off-line factors, such as the milepost location represented by "*base_milepost*", and distance of the nearest on and off ramp in the upstream and downstream directions from crash location, namely, "*upstreamon*", "*upstreamoff*", "*downstreamon*", and "*downstreamoff*". These variables along with "*timeofcrash*" were transformed into ordinal variables using optimal binning with respect to the target *y*. The frequency distributions of the six transformed ordinal variables with respect to the binary target variable (*y=0* for non-crash and *y=1* for regime 1 rear-end crash) are provided in Tables 5-4 through 5-9. Note that along the rows on the first column these tables depict the range of continuous variables that constitute the optimal bins. The two subsequent columns show the frequency (and row percentage) of crash and random non-crash cases, respectively, in the bin represented by corresponding row. Note that the data used here is the complete dataset used to model regime 1 rear-end crashes. In the complete sample there are 742 (≈ 15%) crashes and 4429 (≈ 85%) non-crash cases. Therefore, the bins

with greater than 15 % of crash cases may be considered more crash prone while the bins with less than 15% may be considered relatively safer.

It may be seen from Table 5-4 that based on occurrence of regime 1 rear-end crashes the corridor is divided into four segments with cutoff points located at milepost 13.75, 15.965 and 25.742 miles. Note that regime 1 rear-end crashes have *31%* row frequency (as opposed to varying between *4* to *8%* in other three bins identified through transformation) in the10-mile stretch located in the downtown Orlando area (third bin; mile-post location from 15.965 through 25.74). It indicates that the risk of having a regime 1 rear-end crash is much higher in this region of the study area corridor.

Table 5-4: Frequency table of the variable created through optimal binning transformation of "base_milpost" for crash (regime 1 rear-end crashes) and non-crash cases

Optimal binning of "*base_milepost*" with respect to target variable	**y**		**Total**
	0 (non-crash cases)	**1(crash cases)**	
0 - 13.75	1494 96.08	61 3.92	1555 (100)
13.75 - 15.965	258 92.14	22 7.86	280 (100)
15.965 - 25.74	1343 69.19	598 30.81	1941 (100)
25.742 - 36.25	1334 95.63	61 4.37	1395 (100)
Total	4429 (85)	742 (15)	5171 (100)

Table 5-5 provides similar information for "*timeofcrash*"; four bins (categories) are created with cut-off points at 23652 (midnight to 6:35 AM), 27023 (6:35 AM to 7:31 AM) and 68730 (7:31 AM to 7:06 PM) with period between 7:06 PM to midnight constituting the fourth bin. As expected, row percentage of crash cases is the maximum for the period between 7:31 AM to 7:06 PM, indicating the maximum risk of having a regime 1 rear-end crash during this time period.

Table 5-5: Frequency table of the variable created through optimal binning transformation of "time of crash" for crash (regime 1 rear-end crashes) and non-crash cases

Optimal binning of "*time of crash*" (expressed in terms of seconds past midnight) with respect to target variable	**Y**		**Total**
	0 (non-crash case)	**1(crash case)**	
0 - 23652 (midnight to 6:35 AM)	1147 99.48	6 0.52	1153 (100)
23652- 27023 (6:35 AM to 7:31 AM)	159 91.38	15 8.62	174 (100)
27023- 68730 (7:31 AM to 7:06 PM)	2225 76.67	677 23.33	2902 (100)
68730 - 86400 (7:06 PM to midnight)	898 95.33	44 4.67	942 (100)
Total	4429 (85)	742 (15)	5171 (100)

Next off-line factor to be transformed was the location of ramps. The ramps may be categorized into two types; on-ramp and off-ramp. These ramps are expected to affect the probability of crash occurrence on freeway locations. In this regard their location with respect to the location at which crash risk is being assessed becomes critical. For example, an off-ramp located upstream of a freeway location would effect the odds of crash occurrence in a different way than an on-ramp located downstream. For every crash and non-crash case the distances of nearest on and off ramp in upstream and downstream direction are available from the geometric design database created for this study (See Chapter 4 for details). These continuous variables were named "*downstreamon*", "*downstreamoff*" "*upstreamon*" and "*upstreamoff*". Four ordinal variables with two levels each were created by transforming these variables. These newly created variables are shown in Table 5-6 through 5-9 along with crash and non-crash frequencies in the

resulting categories. The distances of nearest ramps (of both types in both directions) are essentially divided based on a threshold value. This threshold value is obtained with the objective of maximizing the association of the resulting categories of the transformed variable with the target variable. Hence, it is expected that on one side of this threshold the ratio of crash vs. non-crash would be very different from the ratio on the other side of the threshold.

Cut-off for the distances of nearest downstream off-ramp, downstream on-ramp, upstream off-ramp, respectively, are 0.6323, 0.7723, and 0.3196 miles (See column 1 of Tables 5-6 through 5-8). Threshold for the upstream on-ramp is 1.61 miles (Table 5-9) that is much higher than the other three types of ramps. A cut-off value of 1.61 miles would mean that one category of the transformed variable (i.e., upstream on-ramp within 0 to 1.61 miles) would encompass most of the observations. Hence, the location of an on-ramp near or far (up to 1.61 miles upstream) have the same impact on regime 1 rear-end crash occurrence. It is expected because although the presence of an on-ramp would contribute more vehicles on the freeway leading to more congestion; its effect would be independent of the fact whether the ramp is located, for example, 0.1 mile upstream or 1-mile upstream since these many vehicles are going to be on the freeway until an off-ramp is encountered in the downstream direction. The distance between crash (and non-crash) location and the nearest upstream off-ramp has the smallest threshold. It indicates that for a small distance (0.3196 miles according the categorization obtained here) downstream of an off-ramp there is higher probability of a regime 1 rear-end crash. A possible explanation for the same might be that as the vehicles pass besides an off-ramp they

might experience slightly reduced congestion due to some vehicles exiting the freeway. It might prompt some drivers to accelerate, even though the conditions on the freeway are largely unchanged, leading to high speed variance.

The maximum difference in percentage of crash cases in the two categories (bins) is observed in the case of the variable created by transforming the continuous variable "*downstreamon*" (distance of nearest on-ramp in the downstream direction). In the category with observations having nearest downstream on-ramp within 0 through 0.7743 miles there are 21.02% crashes; while in observations with nearest downstream on-ramp greater than 0.7743 miles there are only 4.22% crashes. It indicates that sites within 0.7743 miles upstream of an on-ramp are at considerable higher risk of a regime 1 rear-end crash than other freeway locations. It is somewhat expected since the locations upstream of onramps are the sites of worst recurring congestion and regime 1 rear-end crashes (identified by leaves 1, 2, 4 and 6 of the classification tree shown in Figure 5-8) are indeed associated with lower speeds at extended freeway sections.

The threshold for transforming "*downstreamoff*" is 0.6323 miles and it is the ramp which has the least difference between the two categories in terms of percentage of crashes. In some cases due to queue spillovers from the off-ramps; locations upstream of it might experience the kind of congestion that might cause and be associated with regime 1 rear-end crashes.

Table 5-6: Frequency table of the variable created through optimal binning transformation of "downstreamoff" for crash (regime 1 rear-end crashes) and non-crash cases

Optimal binning of "*downstreamoff*" (distance of nearest downstream off ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0 - 0.6323	2009 81.4	459 18.6	2468 (100)
0.6323 - maximum	2263 89.41	268 10.59	2531 (100)
Total	4429 (85)	742 (15)	5171 (100)

Table 5-7: Frequency table of the variable created through optimal binning transformation of "downstreamon" for crash (regime 1 rear-end crashes) and non-crash cases

Optimal binning of "*downstreamon*" (distance of nearest downstream on ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0 - 0.7743	2439 78.98	649 21.02	3088 (100)
0.7743 - maximum	1906 95.78	84 4.22	1990 (100)
Total	4429 (85)	742 (15)	5171 (100)

Table 5-8: Frequency table of the variable created through optimal binning transformation of "upstreamoff" for crash (regime 1 rear-end crashes) and non-crash cases

Optimal binning of "*upstreamoff*" (distance of nearest upstream off ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0 - 0.3196	1129 75.12	374 24.88	1503 (100)
0.3196 - maximum	3262 89.86	368 10.14	3630 (100)
Total	4429 (85)	742 (15)	5171 (100)

Table 5-9: Frequency table of the variable created through optimal binning transformation of "upstreamon" for crash (regime 1 rear-end crashes) and non-crash cases

Optimal binning of "*upstreamon*" (distance of nearest upstream off ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0 - 1.6107	3653 83.67	713 16.33	4366 (100)
1.6107- maximum	679 95.9	29 4.1	708 (100)
Total	4429 (85)	742 (15)	5171 (100)

It can be argued in general that having a ramp location closer, especially an on-ramp in the downstream direction, would mean increased chances of having a (regime 1 rear-end) crash. Note that we are able to identify these interesting trends because we are comparing randomly selected non-crash data with a sample belonging to a specific type of crash.

Following appropriate transformations modeling procedure described in the previous section was initiated. The first set of models were developed using real-time traffic parameters only from station of the crash along with the offline factors. From the tree model used to perform variable selection (for neural networks) it was found that the tree with entropy maximization criterion resulted in most comprehensive list of variables. The list of variables selected is shown in Table 5-10. It may be seen that none of the factors explicitly related to driver population (measures developed in Chapter 4 such as the odds of observing middle aged drivers or very old drivers by time of day and location along the corridor) had significant *VIM*. However, the binning transformation variable for

"*base_milepost*" (representing the segments along the corridor) was third most significant variable. The effect of driver population on regime 1 rear-end crashes might be implicit in this variable. Similarly binning transformations for distance of nearest downstream on and off ramps and upstream off ramp were found to have significant *VIM*. As expected, the binary transformation of the distance between crash (and non-crash) location and the presence of an on-ramp in the downstream direction is the most significant ramp related variable. Only ramp that does not appear in the list of significant variables is the upstream on-ramp which was expected since most of the observations were concentrated in the first category (0 to 1.61 miles) of the transformed variable created earlier.

Among the real-time traffic variables; average and standard deviation of occupancy, coefficient of variation in speed and standard deviation of volume at nearest loop detector were found to be most associated with binary target *y*. High average occupancy at station F (AOF2) indicates congested traffic regime in which ephemeral queues are being formed and dissipated leading to high variation in speed (CVSF2 in the next most significant variable). Under these driving conditions drivers might have to slow down, stop and speed up again quite often. These conditions are of course prone to rear-end crash. The other two variables SOF2 and SVF2 were also found significant by the tree node used for variable selection. The tree node was followed by 10 parallel MLP and NRBF neural network nodes each in order to estimate neural network models having a range (1 to 10) of hidden nodes. The result from these neural nets showed that the MLP network with 4 hidden nodes and NRBF network with 6 hidden nodes performed best among the models in their respective architectures. The results from the three logistic regression models

employing different selection procedures were almost identical, although backward selection procedure resulted in the model with best performance over the validation dataset.

Table 5-10: Results of variable selection through the classification tree model utilizing Entropy maximization criterion (examined traffic parameters only from Station F)

Name	Variable Importance Measure (VIM)	Variable Description
AOF2	1.0000	Average Occupancy at Station F
CVSF2	0.3529	Coefficient of Variation in Speed at Station F
BASE_MILPOST	0.1576	BASE_MILPOST: Optimal binning for Y =0 if 0<base_milepost<=13.75 =1 if 13.75<base_milepost<=15.965 =2 if 15.965<base_milepost<=25.74 =3 if 25.742<base_milepost<36.25
DOWNSTREAMON	0.1309	DOWNSTREAMON: Optimal binning for Y =0 if nearest downstream on-ramp is located further than 0.7743 miles =1 if nearest downstream on-ramp is located within 0.7743 miles
UPSTREAMOFF	0.1132	UPSTREAMOFF: Optimal binning for Y =0 if nearest upstream off-ramp is located further than 0.3196 miles =1 if nearest upstream off-ramp is located within 0.3196 miles
SOF2	0.1102	Standard Deviation of Occupancy at Station F
SVF2	0.1002	Standard Deviation of Volume at Station F
DOWNSTREAMOFF	0.0765	DOWNSTREAMOFF: Optimal binning for Y =0 if nearest downstream off-ramp is located further than 0.6323 miles =1 if nearest downstream off-ramp is located within 0.6323 miles

As described in the previous section the best model was identified through the lift plot having cumulative percentage of captured response for the validation dataset on the vertical axis. The output of the classification models for any observation is termed as the posterior probability of the event (i.e., a rear-end crash in this case). Posterior probability is a number between 0 and 1. The closer it is to unity the more likely, according to the model, it is for that observation to be a rear-end crash. In a lift chart, the observations in the validation dataset are sorted from left to right by the output posterior probability from each model. The sorted group is lumped into ten deciles[2] (one decile represents 10 percentile) along the horizontal axis. The left-most decile is the *10%* of observations with highest posterior probability i.e., most likely to be a regime 1 rear-end crash. The lift charts used to demonstrate performance of various models in this chapter also display the "*performance*" of a random baseline model which represents the percentage of crashes identified in the validation sample if one randomly assigns observations as crash and non-crash. The performance of each model may be measured by determining how well the models capture the target event across various deciles. Therefore, higher a curve from the baseline curve the better the performance of the corresponding model. From a practical application point of view it must be understood that crashes are rare events and one would need to be parsimonious in issuing warnings for crashes. Therefore, it might be unreasonable to assign more than 20-30% of observations as crashes. Hence, to choose among competing models the position of the curve on first few deciles must be critically examined.

[2] Decile is defined as any of nine points that divided a distribution of ranked scores into equal intervals where each interval contains one-tenth of the scores

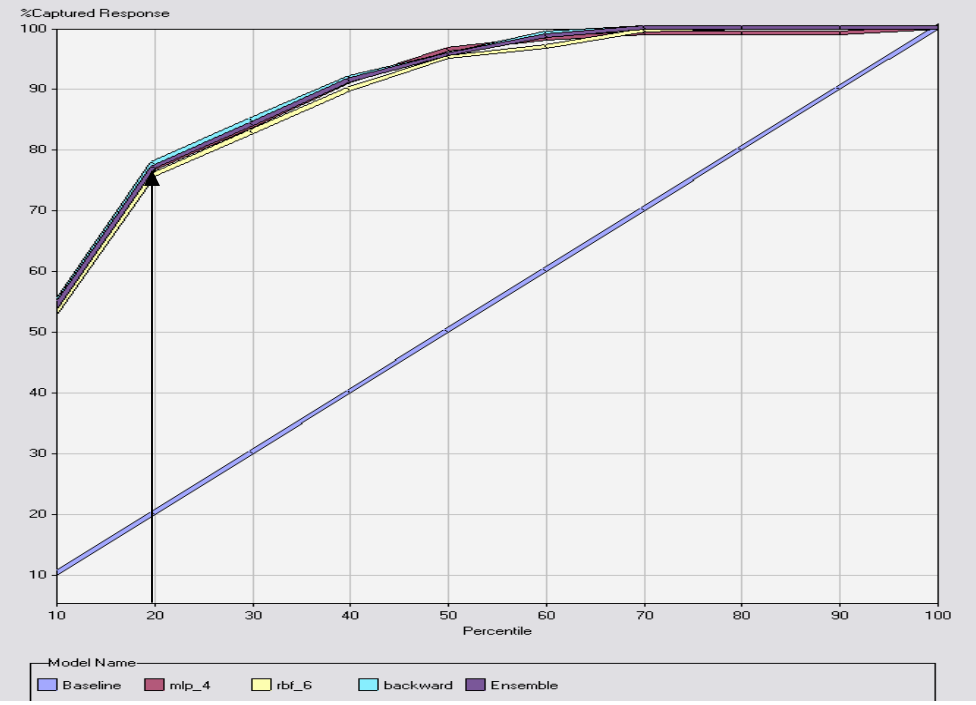

Figure 5-16: Percentage of captured response lift plot for best models belonging to different modeling techniques (input traffic parameters only from Station F)

It may be seen in Figure 5-16 that the performance of the best models from three different modeling techniques is almost identical, and indeed the curve for the ensemble (i.e., hybrid) model created by averaging the posterior probabilities from these three models also all but coincides with these models. The backward logistic regression model depicted by blue curve in the figure had the maximum percentage of response captured in the first two deciles. Almost 78% of crashes (response with $y=1$) from the validation dataset have been identified in the *20%* observations having highest posterior probabilities estimated through this model. Hence, for the models developed with traffic parameters from only station F, backward logistic regression model is recommended as the final model. Note that classification performance of this or any other model for that

matter would vary depending upon the threshold set on the posterior probability. To demonstrate the classification capabilities of this model Table 5-11 shows its performance on validation dataset if the threshold on the output posterior probability to separate crashes from non-crash cases is set at 0.25. 161 out of 224 (71.88%) regime 1 rear-end crashes and 1258 out of 1330 (94.58%) non-crash cases were correctly identified. This performance is much better than any of the generic models developed in our previous studies (Abdel-Aty et al., 2004, 2005)

Table 5-11: Classification performance of the backward regression model on the validation dataset with posterior probability threshold at 0.25

Table of actual by predict			
actual	**predicted**		**Total**
	0 (non-crash)	**1 (crash)**	
0 (non-crash)	1258	72	1330
1 (crash)	63	161	224
Total	1321	233	1554

Note that the model(s) developed above utilized real-time traffic information only from station of the crash (Station F). As the next step in the modeling process potential input variables were increased to include traffic parameters from three stations (Station E, F, and G). The list of variables found significant by the variable selection tree among the traffic parameters and off-line factors is shown in Table 5-12.

It is interesting to note that presence of downstream on-ramp no longer figures in the list of important variables. When traffic parameters from only one station (Station F) were

included it was the most significant variable related to the ramp location (See Table 5-11). Data from series of three stations are now included as the potential independent variables and average occupancy values from all three stations (AOE2, AOF2, and AOG2) are significant. It may be inferred that the congestion effect caused by the on-ramp eventually leading to a regime 1 rear-end crash upstream of the ramp is reflected by these occupancy parameters and which is why the downstream on-ramp location does not have significant *VIM*. Similarly the variable "base_milepost" was also excluded by the variable selection tree model. These variables are replaced by the real-time traffic variables from loop detector station located downstream of the crash site. The only critical traffic variable from Station E (the upstream station) was average occupancy at that station. It is apparent from the list of variables selected (Table 5-12) that traffic conditions measured (in terms of *CVS* and *AO*) at the station nearest to crash location (Station F) and the station downstream of it (Station G); are more critically associated with crash occurrence. The critical traffic related parameters again show that the high occupancy traffic conditions around the crash location (at upstream and downstream stations as well) are causing temporal variation in speed at crash location and downstream of it (CVSF2 and CVSG2 are both significant in that order) that can potentially lead to a regime 1 rear-end crash.

Table 5-12: Results of variable selection through the classification tree model utilizing Entropy maximization criterion (examined traffic parameters from Stations E, F and G)

Name	Variable Importance Measure (VIM)	Variable Description
AOF2	1.0000	Average Occupancy at Station F
AOG2	0.4861	Average Occupancy at Station G
CVSF2	0.3399	Coefficient of Variation in Speed at Station F
AOE2	0.3243	Average Occupancy at Station E
UPSTREAMOFF	0.1520	UPSTREAMOFF: Optimal binning for Y =0 if nearest upstream off-ramp is located further than 0.3196 miles =1 if nearest upstream off-ramp is located within 0.3196 miles
CVSG2	0.1508	Coefficient of Variation in Speed at Station G
DOWNSTREAMOFF	0.1393	DOWNSTREAMOFF: Optimal binning for Y =0 if nearest downstream off-ramp is located further than 0.6323 miles =1 if nearest downstream off-ramp is located within 0.6323 miles
SVF2	0.1387	Standard Deviation of Volume at Station F
AVF2	0.0996	Average Volume at Station F

The significant variables shown in Table 5-12 were included in the neural network models as inputs. The cumulative percentage of captured response lift plots for the best model from each modeling technique (MLP Neural network, NRBF neural network and logistic regression) along with the model created by combining these models (ensemble model) is shown in Figure 5-17. MLP with 6 hidden neurons, NRBF with 8 hidden neurons and backward regression models were found to be the best in their respective categories. It may be seen in Figure 5-17 that the performance of the best models from three different modeling techniques is again almost identical, and indeed the curve for the ensemble model created by averaging the output posterior probabilities from the three

individual models almost coincides with one of these models. Almost *66%* of regime 1 rear-end crashes in the validation dataset were included in top *20%* observations having maximum output posterior probability from the MLP neural network with 6 hidden nodes. This model happened to perform slightly better than the other three (Backward regression, NRBF with 8 hidden neurons, and the ensemble) models.

Hence, the best model with traffic parameters from three stations yields *66%* crash identification in the first two deciles, while the best model developed with data only from station of crash yielded 78% crash identification (See Figure 5-16). It means that the performance of classification models is not positively affected if we include traffic parameters from three stations rather than just one station to identify regime 1 rear-end crashes.

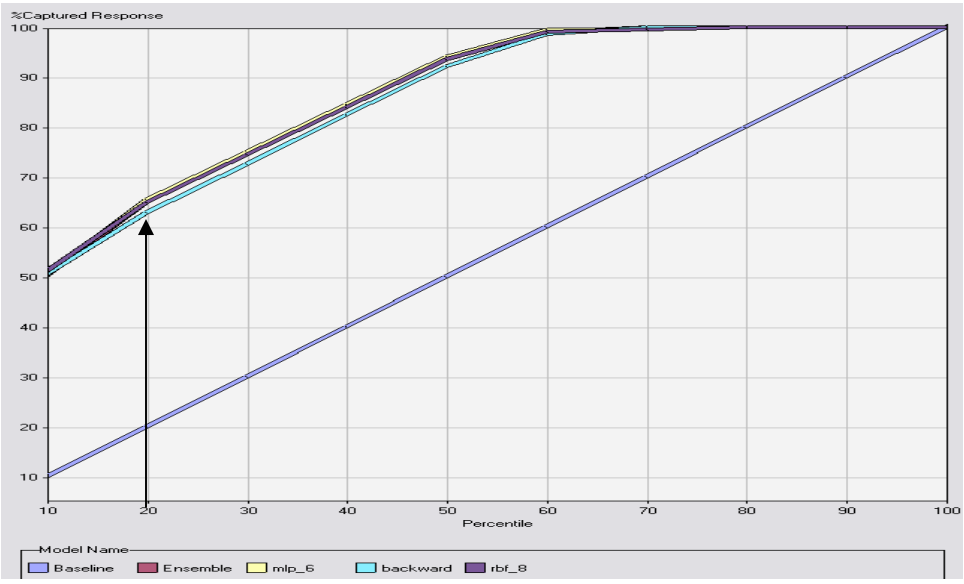

Figure 5-17: Percentage of captured response lift plot for best models belonging to different modeling techniques (input traffic parameters from Station E, F and G)

In the next step traffic parameters from five loop detectors (Station D to H) were included in the analysis as potential independent variables along with selected off-line factors. *AVD2* was the only traffic related parameter from extreme upstream loop detector (Station D) that was found significant (albeit with lowest *VIM* of 0.07; See Table 5-13) by the classification tree model used for variable selection. The complete list of significant variables along with their *VIM*s is presented in Table 5-13. Again, the downstream on-ramp does not figure in the list of critical variables which is because the congestion caused by an on-ramp is well captured by the average occupancy at stations E through H all of which are included in the list of significant variables (Table 5-13). The presence of off-ramps in upstream and downstream direction remains significant even when we include traffic parameters from more stations. The reason could be that the queues spilling over from an off-ramp are generally not very long and any loop detector might not be located in the affected region that can possibly reflect this congestion by means of collected traffic data.

A closer examination of the list of important traffic related variables selected by the tree model suggests that high average occupancy over extended sections of the freeway causes significant temporal variation in speeds and the sites in the vicinity of location experiencing maximum variation have a high probability of having a crash within next 5-10 minutes. This inference is made based on the fact that station F is the station located closest to the crash location and *CVSF2* is the most important variable representing variation in speed. The reason why frequency of regime 1 rear-end crashes peaks at the

middle of the peak period operation (See Figure 5-12) is also clear now. The conditions with high average occupancy at extended segments of the freeway occur during that time.

Table 5-13: Results of variable selection through the classification tree model utilizing Entropy maximization criterion (examined traffic parameters from Stations D through H)

Name	Variable Importance Measure (VIM)	Variable Description
AOF2	1.0000	Average Occupancy at Station F
AOD2	0.4819	Average Occupancy at Station D
AOH2	0.3640	Average Occupancy at Station H
CVSF2	0.2880	Coefficient of Variation in Speed at Station F
SOE2	0.2210	Standard Deviation of Occupancy at Station E
AOE2	0.1731	Average Occupancy at Station E
CVSG2	0.1530	Coefficient of Variation in Speed at Station G
SOH2	0.1476	Standard Deviation of Occupancy at Station H
AOG2	0.1342	Average Occupancy at Station G
SVG2	0.1283	Standard Deviation of Volume at Station G
SOF2	0.1157	Standard Deviation of Occupancy at Station F
CVSE2	0.1098	Coefficient of Variation in Speed at Station E
AVF2	0.0996	Average Volume at Station F
UPSTREAOFF	0.0985	UPSTREAMOFF: Optimal binning for Y =0 if nearest upstream off-ramp is located further than 0.3196 miles =1 if nearest upstream off-ramp is located within 0.3196 miles
SVF2	0.0982	Standard Deviation of Volume at Station F
AVG2	0.0959	Average Volume at Station G
DOWNSTREAMOFF	0.0844	DOWNSTREAMOFF: Optimal binning for Y =0 if nearest downstream off-ramp is located further than 0.6323 miles =1 if nearest downstream off-ramp is located within 0.6323 miles
AVD2	0.0700	Average Volume at Station D

The performance of various classification models in this set is depicted in Figure 5-18. The optimal performance of both MLP and NRBF neural networks over validation dataset was obtained through networks with 8 hidden layer neurons. The best among the four models, MLP neural network with 8 hidden neurons would capture *52.13%* of crashes in the validation dataset in top 20 percentile. It implies that by including data from more stations the capability of the models to correctly identify regime 1 rear-end crashes actually declines (See Figure 5-16 and Figure 5-17).

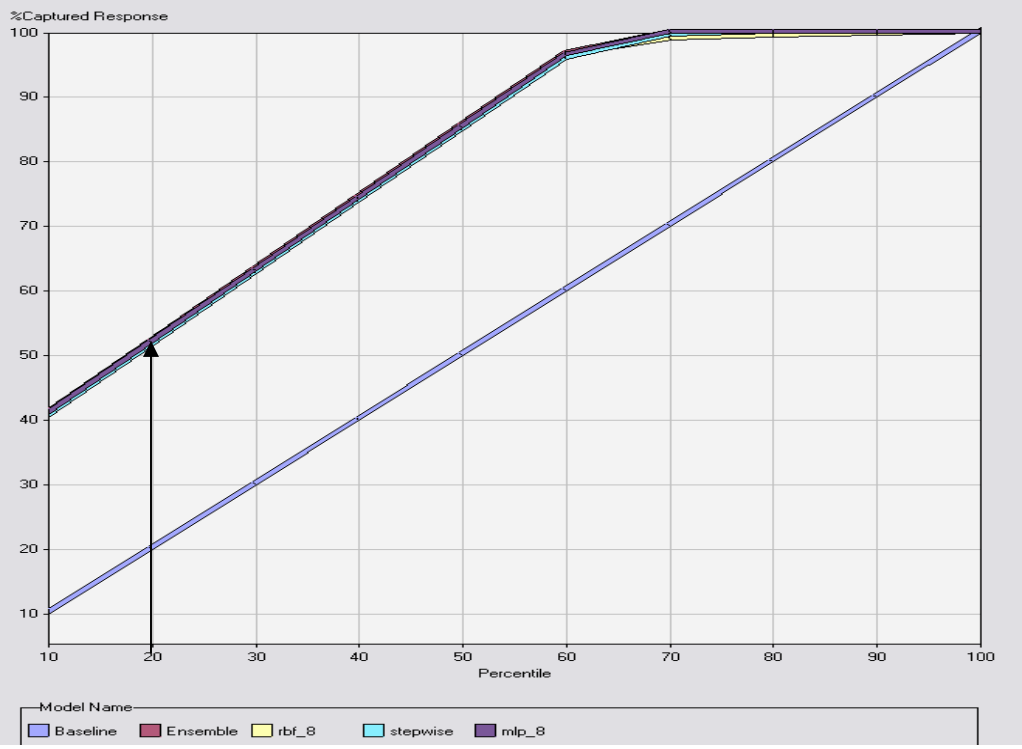

Figure 5-18: Percentage of captured response lift plot for best models belonging to different modeling techniques (input traffic parameters from Station D, E, F, G, and H)

The summary of the performance of models belonging to three sets (utilizing traffic data from 1, 3 or 5 stations) is provided in Table 5-14. It provides the structure (in case of

neural networks) and selection procedure (in the case of logistic regression) along with the percentage of crashes captured within the first two deciles of posterior probability. The "best" model, capturing the highest percentage of crashes within first two deciles is highlighted in the Table.

Table 5-14: Structure and percentage of captured response within the first two deciles for best models estimated for different modeling techniques (Regime 1 rear-end crashes)

		Modeling Technique			
		MLP Neural Network	NRBF Neural Network	Logistic Regression	Ensemble Model
Traffic Parameters from	Station F	76.72% (4 hidden nodes)	75.77% (6 hidden nodes)	77.70% (Backward selection)	77.03%
	Station E, F, and G	65.40% (6 hidden nodes)	64.91% (8 hidden nodes)	62.88% (Backward selection)	64.92%
	Station D, E, F, G, and H	52.13% (8 hidden nodes)	51.85% (8 hidden nodes)	51.54% (Stepwise selection)	51.99%

The performance of the model utilizing data from station of the crash is much better than the models utilizing data from three or five stations. However, it is possible that the three sets of models using data from one, three and five stations respectively are good at identifying distinct sets of crashes and therefore, the performance of individual models may be improved upon by combining the best models in each of the three sets. Hybrids of the highlighted models in each row of Table 5-14 are examined through the Ensemble node of the Enterprise Miner. It was found that the hybrid model does slightly improve upon the performance provided by individual models.

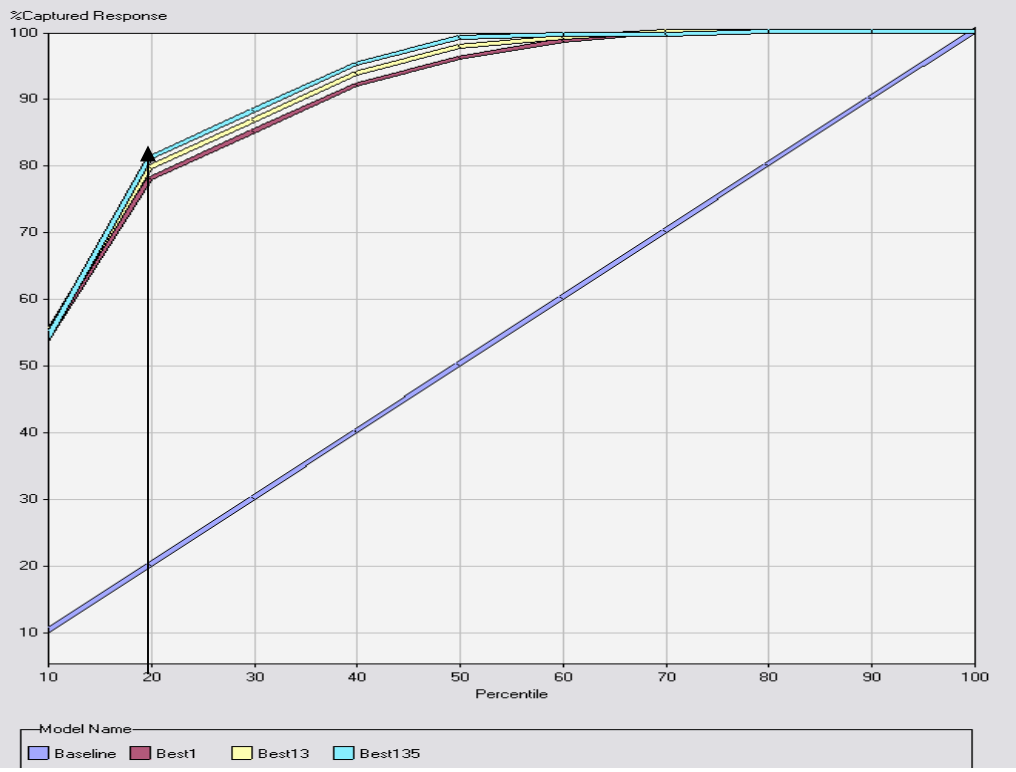

Figure 5-19: Percentage of captured response lift plot for combination of best models for regime 1 rear-end crashes chosen from the three sets

The performance of the two ensemble models (titled *best13* and *best135*) along with the best individual model (backward regression model using traffic data only from station F) is shown in Figure 5-19. The yellow curve represents the performance of ensemble model *best13*, which is the model estimated by combining backward regression model using traffic data only from station F with the MLP with six hidden nodes using traffic data from stations E through G. The blue curve represents the hybrid model titled *best135*, which is the combination of the all three models highlighted in Table 5-14. It may be seen that the blue curve representing combination of the three models is positioned consistently higher than other curves and is therefore recommended for identification of

regime 1 rear-end crashes. With this model 81% of the crashes in the validation dataset can be identified within first two deciles.

It must be said however, that the improvement is marginal compared to the performance of the best individual model (backward regression model represented by purple curve) only utilizing traffic information from station F. Due to intermittent failure of the loops, data from the five stations may not always be simultaneously available to use the recommended hybrid model. In which case, better option would be to use a model that uses data from one station only. Therefore, for real-time implementation of these models their performance should be examined in the context of data requirements.

Since the data used for this analysis was completely random non-crash data the main purpose of this analysis was to provide insight into the variables significantly affecting the probability of regime 1 rear-end crash. The data requirement issues might not be relevant any more because the models predicting regime 1 rear-end crashes would not figure into the application plan proposed in this chapter.

5.6 Analysis and Results: Regime 2 Rear-end Crashes

5.6.1 With completely random non-crash data

From the classification tree developed to separate the two groups of rear-end crashes (belonging to regime 1 and regime 2) it was clear that conditions 5-10 minutes prior to regime 2 rear-end crashes are not congested. Therefore, these crashes do not occur under frequently forming and dissipating queues but are possibly caused by disturbances and

speed differential that is created downstream of a freeway location and to which drivers fail to react. These crashes are generally associated with medium to high speed traffic regimes prevailing just before/after the period of heavy congestion on freeways. In this section models are developed to differentiate regime 2 rear-end crashes from completely random freeway data. The main purpose of this analysis is to make inferences regarding factors responsible for such crashes.

From one of our previous studies (Abdel-Aty et al., 2005) it was inferred that coefficient of variation in speed might not be a significant predictor of rear-end crashes under high speed traffic regimes. Hence, the average and standard deviation of speed were used in their original form rather than the *LogCVS* used earlier for regime 1 rear-end crashes.

First, "*Optimal Binning for Relationship to Target"* transformations, similar to the one used for regime 1 rear-end crash models in the previous section, were applied on critical off-line factors. The "*base_milepost*", "*timeofcrash*", and distance of the nearest on and off ramp in the upstream and downstream directions from crash location, namely, "*upstreamon*", "*upstreamoff*", "*downstreamon*", and "*downstreamoff*" were all transformed and assigned an ordinal measurement level. Tables 5-15 through 5-20 show the frequency of transformed ordinal variables with respect to the target variable for regime 2 crashes. Note that these tables are similar to Tables 5-4 through 5-9 presented in the previous section for regime 1 crashes.

Table 5-15: Frequency table of the variable created through optimal binning transformation of "base_milpost" for crash (regime 2 rear-end crashes) and non-crash cases

Optimal binning of "*base_milepost*" with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1(crash case)**	
0 - 11.93	1454 89.42	172 10.58	1626
11.93 - 25.433	1945 77.4	568 22.6	2513
25.433 - 35.18	1371 93.65	93 6.35	1464
35.181 - 36.25	202 81.78	45 18.22	247
Total	4972 (85)	878 (15)	5850 (100)

It may be seen from Table 15 that regime 2 rear-end crashes have maximum row frequency (22.6%) in the 13-mile segment of the freeway (mile-post location: 11.93 to 25.43). For regime 1 rear-end crashes the segment with maximum row frequency for crashes was the 10-mile stretch of the freeway stating at mile-post location 13.75. Also, note that regime 2 rear-end crashes are more "*uniformly*" distributed over the whole corridor than regime 1 crashes, with second maximum row percentage observed with mile-post location greater than 35.18.

Table 5-16: Frequency table of the variable created through optimal binning transformation of "time of crash" for crash (regime 2 rear-end crashes) and non-crash cases

Optimal binning of "*time of crash*" (expressed in terms of seconds past midnight) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1(crash case)**	
0 - 1577 (midnight to 12:26 AM)	88 100	0 0	88 (100)
1577 - 24326 (12:26 AM to 6:46 AM)	1225 92.45	100 7.55	1325 (100)
24326 - 69868 (6:46 AM to 7:24 PM)	2719 80.8	646 19.2	3365 (100)
69868- 86400 (7:24 PM to midnight)	940 87.69	132 12.31	1072 (100)
Total	4972 (85)	878 (15)	5850 (100)

In a similar table for optimal binning of the continuous variable "*timeofcrash*" it may be seen that bin with highest frequency of regime 2 rear-end crashes (6:46 AM to 7:24 PM) constitutes almost same hours of the day where frequency of regime 1 rear-end crashes was maximum (7:31 AM to 7:06 PM). However, the bin for regime 2 rear-end crashes is a bit wider than that for regime 1 crashes, which indicates that conditions prone to regime 2 rear-end crashes might occur towards beginning and end of congested peak periods on freeway corridor.

In the next step categorization (optimal binning with respect to target) for the distances between crash (and non-crash) locations and nearest on and off ramp in upstream and downstream direction was obtained. Tables 5-17 through 5-21 show the resulting ordinal variables with two levels along with crash and non-crash frequencies for both categories

in the dataset used to analyze regime 2 rear-end crashes. Note that the dataset has 878 *(≈15%)* regime 2 rear-end crashes and 4972 *(≈85%)* randomly selected non-crash cases. As mentioned earlier the distances of nearest ramps (of both types in both directions) are essentially divided based on a threshold value. The threshold value is such that it maximizes the association of resulting categories with the target variable. To explain the threshold values and resulting relative frequencies of crash and non-crash cases in resulting categories we must recall that regime 2 rear-end crashes (identified by leaves 3, 5 and 7 of the classification tree shown in Figure 5-8) occur under relatively free flow locations that possibly precede the congested period on the freeway.

It is observed from Table 5-17 that the threshold for downstream off-ramps is very low and crashes have a high percentage right upstream (only 0.0630 miles upstream) of an off-ramp. It indicates that a rear-end crash associated with somewhat higher speeds might occur if some driver suddenly slows down while approaching an off-ramp. The drivers not familiar with the area might be causing this problem at certain off-ramps.

Even for regime 2 rear-end crashes the most significant type of ramp appears to be downstream on-ramp. An on-ramp located upstream of a freeway location would mean that the drivers get caught unaware of the congested conditions (that are just beginning to expand over the freeway) they are about to experience. For regime 2 rear-end crashes the threshold for upstream on-ramp is 1.91 miles. A cut-off value of 1.91 miles would mean that the one category of the transformed variable (i.e., upstream on-ramp within 0 to 1.91 miles) would encompass most of the observations. The relative frequency of crash cases

in this category is 15.8% which is only slightly higher than 15% that is the overall frequency of rear-end crashes in the dataset. Hence, an upstream on-ramp does not appear to have impact on regime 2 rear-end crashes.

The distance between crash (and non-crash) location and the nearest upstream off-ramp indicates that for a small distance (0.3205 miles according the categorization obtained here) downstream of an off-ramp there is higher probability of a regime 2 rear-end crash. Again a possible explanation for the same might be that as the vehicles pass besides an off-ramp they might experience reduced congestion prompting some drivers to accelerate and run into slower moving vehicles in the downstream direction.

Table 5-17: Frequency table of the variable created through optimal binning transformation of "downstreamoff" for crash (regime 2 rear-end crashes) and non-crash cases

Optimal binning of "downstreamoff" (distance of nearest downstream off ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0 - 0.0638	263 77.13	78 22.87	341 (100)
0.0638 - Maximum	4519 85.34	776 14.66	5295 (100)
Total	4972 (85)	878 (15)	5850 (100)

Table 5-18: Frequency table of the variable created through optimal binning transformation of "downstreamon" for crash (regime 2 rear-end crashes) and non-crash cases

Optimal binning of "downstreamon" (distance of nearest downstream on ramp from crash location) with respect to target variable	y		Total
	0 (non-crash case)	1 (crash case)	
0 - 0.7747	2736 80.57	660 19.43	3396 (100)
0.7747- Maximum	2136 91.01	211 8.99	2347 (100)
Total	4972 (85)	878 (15)	5850 (100)

Table 5-19: Frequency table of the variable created through optimal binning transformation of "upstreamoff" for crash (regime 2 rear-end crashes) and non-crash cases

Optimal binning of "upstreamoff" (distance of nearest upstream off ramp from crash location) with respect to target variable	y		Total
	0 (non-crash case)	1 (crash case)	
0 - 0.3205	1318 77.58	381 22.42	1699
0.3205 - Maximum	3617 88.07	490 11.93	4107
Total	4972 (85)	878 (15)	5850 (100)

Table 5-20: Frequency table of the variable created through optimal binning transformation of "upstreamon" for crash (regime 2 rear-end crashes) and non-crash cases

Optimal binning of "upstreamon" (distance of nearest upstream off ramp from crash location) with respect to target variable	y		Total
	0 (non-crash case)	1 (crash case)	
0 - 1.9117	4417 84.2	829 15.8	5246 (100)
1.9117- Maximum	443 92.87	34 7.13	477 (100)
Total	4972 (85)	878 (15)	5850 (100)

In Table 5-21 frequency distribution of the binary variable "*stationf*" is shown with respect to the target variable. The variable "*stationf*" is defined as 0 if Station F (loop detector station closest to the crash location) is upstream of the crash location and as 1 otherwise. At the modeling stage this variable is found to have a significant VIM. Although the frequency or row percentage do not indicate a large difference between the two levels of this variable it is suspected that relative location of station F with respect to crash location might be a critical in determining whether parameters from station F or parameters from station G would be most significantly associated with crash prone conditions. For example, if station F is downstream of the crash site, it might be significant in predicting a rear-end crash but if its upstream then the next downstream (Station G) might become more significant. Note that even though we did include this variable in potential input variables for regime 1 rear-end crashes it was not found significant at any stage of the modeling process. Therefore, the description of this variable was not provided earlier in the Chapter.

Table 5-21: Frequency table for the variable indicating the location of station of crash (station F) with respect to crash site for crash (regime 2 rear-end crashes) and non-crash cases

(*stationf*) Location of nearest loop detector with respect to crash location	y		Total
	0 (non-crash case)	1 (crash case)	
0 (Loop detector station nearest to crash location is upstream)	2484	347	2831
	87.74	12.26	(100)
1 (Loop detector station nearest to crash location is downstream)	2488	531	3019
	82.41	17.59	(100)
Total	4972	878	5850
	(85)	(15)	(100)

After appropriate transformations modeling procedure was initiated for regime 2 rear-end crashes. Again the first set of models were developed using real-time traffic parameters only from station of the crash along with the offline factors. The list of variables selected by the classification tree model with entropy maximization criterion is shown in Table 5-22. It may be seen that even for regime 2 rear-end crashes none of the factors explicitly related to driver population had significant *VIM*. However, the binning transformation variable for "*base_milepost*" (representing the segments along the corridor) was a significant variable. Similarly binning transformations for distance of nearest downstream on and off ramps were found to have significant *VIM*. Among the real-time traffic variables; average speed and volume at station F were significant. The variable representing the location of station F was also found to be significant. It is remarkable that only two traffic related parameters figure in the list of critical input variables. It indicates that the characteristics of crash location along with traffic conditions downstream of it may be more useful in predicting regime 2 rear-end crashes.

Table 5-22: Results of variable selection through the classification tree model utilizing Entropy maximization criterion (examined traffic parameters only from Station F)

Name	Variable Importance Measure (VIM)	Variable Description
ASF2	1.0000	Average Speed at Station F
AVF2	0.5981	Average Volume at Station F
DOWNSTREAMOFF	0.5156	DOWNSTREAMOFF: Optimal binning for Y =0 if nearest downstream off-ramp is located further than 0.0638 miles =1 if nearest downstream off-ramp is located within 0.0638 miles
DOWNSTREAMON	0.4413	DOWNSTREAMON: Optimal binning for Y =0 if nearest downstream on-ramp is located further than 0.7747 miles =1 if nearest downstream on-ramp is located within 0.7747 miles
BASE_MILPOST	0.3427	BASE_MILPOST: Optimal binning for Y =0 if 0<base_milepost<=11.93 =1 if 11.93<base_milepost<=25.433 =2 if 25.433<base_milepost<=35.18 =3 if 35.181<base_milepost<=36.25
STATIONF	0.3372	Location of Station F relative to crash location =0 if Loop detector station nearest to crash location is located upstream =1 if Loop detector station nearest to crash location is located downstream

The classification tree node for variable selection was followed by 10 parallel MLP and NRBF neural network nodes in order to estimate neural network models having a range (1 to 10) of hidden nodes. The result from these neural nets showed that the MLP network with six hidden nodes and NRBF network also with six hidden nodes performed best among the models in their respective architectures. The result from the three logistic regression models employing different selection procedures showed that stepwise selection procedure resulted in the best model.

As in the case of regime 1 crashes, the best models were identified through the lift plot having cumulative percentage of captured response in the validation dataset on vertical axis. The higher a curve from the baseline curve the better is the performance of the corresponding model. The captured response lift plots for models belonging to regime 2 rear-end crashes are shown in Figure 5-20. It may be noted that corresponding models for regime 1 crashes had their captured response percentage higher than the models shown in Figure 5-20. For example, corresponding models for regime 1 rear-end crashes identified more than 50% of crashes (Figure 5-16) with in the first 10 percentile while regime 2 models shown in Figure 5-20 only identify about 26%. It implies that to identify same percentage of crashes more warnings would have to be issued in the case of regime 2 rear-end crashes. Therefore, while crash identification within first two deciles was used as the evaluation criteria for regime 1 crashes one must increase the number of deciles in case of regime 2 so that reasonable crash identification may be achieved. Crashes being rare events it would be unreasonable to issue warnings more than 20-30% and therefore it was decided to evaluate the model performances with in first three deciles (deciles = 10 percentiles).

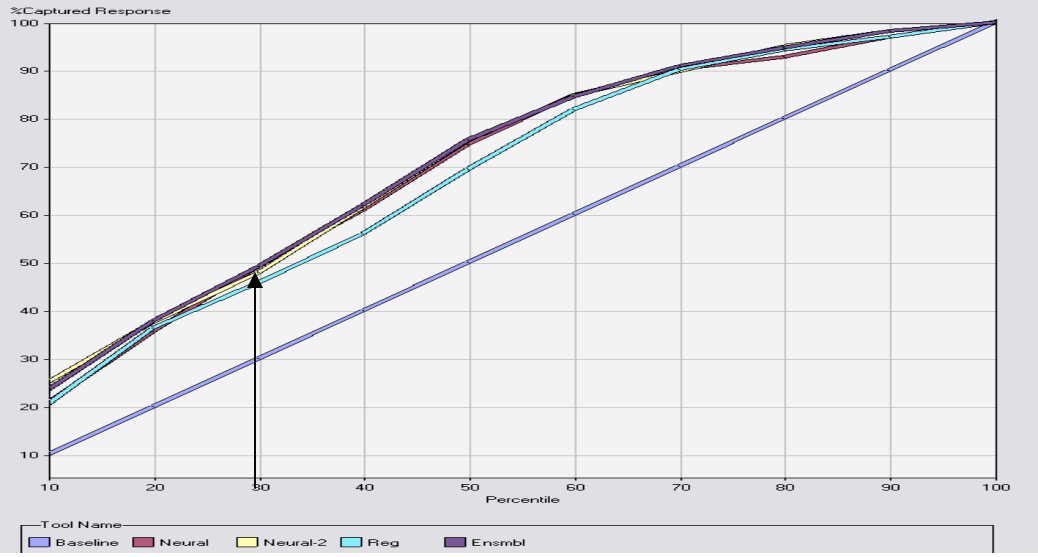

Figure 5-20: Percentage of captured response lift plot for best models belonging to different modeling techniques (input traffic parameters only from Station F)

It may be seen in Figure 5-20 that the performance of the best models from three different modeling techniques is comparable, however, the curve for hybrid model created by averaging the posterior probabilities from these three models is the highest at the 30 percentile and identifies 49.42% of regime 2 rear-end crashes in the validation dataset.

Real-time traffic information from only station of the crash was utilized for the model(s) developed for regime 2 crashes up to now. As the next step in the modeling process potential input variables were increased to include traffic parameters from three stations (Station E, F and G). It was noticed that the average speed downstream of crash site (ASG2) was now the most significant variable. The binary variable "*stationf*" representing location of station F with respect to crash location was significant along with average speed at station of crash (ASF2). The significance of these three parameters (ASG2, *ASF2* and *stationf*) indicates that ASF2 would become significant if "*stationf*" is

downstream of the crash location. Standard deviations of speed at upstream and downstream stations (SSE2 and SSG2) were also found significant. The locations of upstream off ramp and downstream on ramp along with time of the day were the significant static factors.

Table 5-23: Results of variable selection through the classification tree model utilizing Entropy maximization criterion (examined traffic parameters from Stations E, F and G)

Name	Variable Importance Measure (VIM)	Variable Description
ASG2	1.0000	Average Speed at Station G
ASF2	0.6513	Average Speed at Station F
AVF2	0.5723	Average Volume at Station F
DOWNSTREAMON	0.5559	DOWNSTREAMON: Optimal binning for Y =0 if nearest downstream on-ramp is located further than 0.7747 miles =1 if nearest downstream on-ramp is located within 0.7747 miles
CRASHTIME	0.5447	CRASHTIME: Optimal binning for Y =0 if Time of crash between midnight to 12:26 AM =1 if Time of crash between 12:26 AM to 6:46 AM =2 if Time of crash between 6:46 AM to 7:24 PM =3 if Time of crash between 7:24 PM to midnight
UPSTREAMOFF	0.5230	UPSTREAMOFF: Optimal binning for Y =0 if nearest upstream off-ramp is located further than 0.3205 miles =1 if nearest upstream off-ramp is located within 0.3205 miles
SSE2	0.4340	Standard Deviation of Speed at Station E
ASE2	0.4159	Average Speed at Station E
STATIONF	0.3834	Location of Station F relative to crash location =0 if Loop detector station nearest to crash location is located upstream =1 if Loop detector station nearest to crash location is located downstream
SSG2	0.3795	Standard Deviation of Speed at Station G
AVE2	0.3231	Average Volume at Station E
SVG2	0.2456	Standard Deviation of Volume at Station G

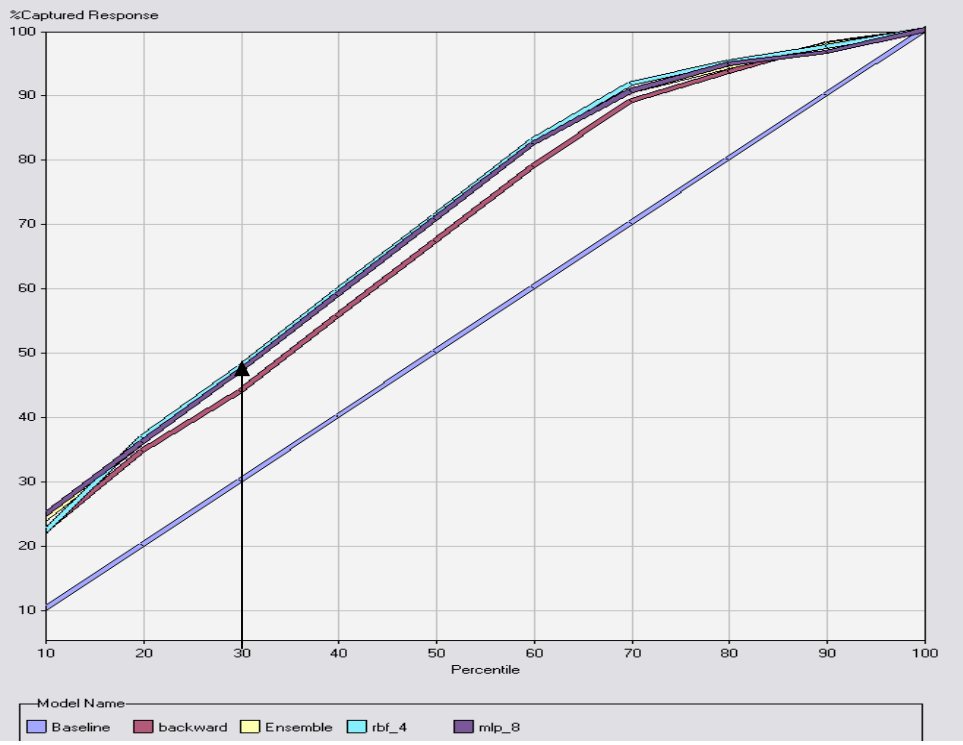

Figure 5-21: Percentage of captured response lift plot for best models belonging to different modeling techniques (input traffic parameters from Station E, F and G)

The percentage captured response lift plots for the best model from each modeling technique (MLP Neural network, NRBF neural network and logistic regression) along with the model created by combining these models (hybrid model) are shown in Figure 5-21. MLP with 8 hidden neurons, NRBF with 4 hidden neurons and backward regression models were found to be the best in their respective categories. It may also be seen that the performance of the NRBF model is the slightly better than the other three (MLP, backward regression and ensemble) models and it identifies almost 48% of the crashes in the first 3 deciles.

In the next step traffic parameters from five loop detectors (Station D to H) were included in the analysis as potential independent variables. Most of the offline factors identified as significant for this set of models were same as the set of models developed with real-time traffic inputs from three stations. Average speed, volume, and occupancy at station H (located at extreme downstream direction) were all found significant indicating that to identify the occurrence of regime 2 rear-end crash at a freeway location the conditions downstream of that site need to be monitored closely. *AVD2* was the only traffic related parameter from extreme upstream station (Station D) that was found significant. The complete list of significant variables is presented in Table 5-24.

Table 5-24: Results of variable selection through the classification tree model utilizing Entropy maximization criterion (examined traffic parameters from Stations D through H)

Name	Importance Measure (VIM)	Variable Description
ASG2	1.0000	Average Speed at Station G
ASF2	0.8970	Average Speed at Station F
ASH2	0.8470	Average Speed at Station H
CRASHTIME	0.6613	CRASHTIME: Optimal binning for Y =0 if Time of crash between midnight to 12:26 AM =1 if Time of crash between 12:26 AM to 6:46 AM =2 if Time of crash between 6:46 AM to 7:24 PM =3 if Time of crash between 7:24 PM to midnight
AVD2	0.6225	Average Volume at Station D
DOWNSTREAMON	0.5984	DOWNSTREAMON: Optimal binning for Y =0 if nearest downstream on-ramp is located further than 0.7747 miles =1 if nearest downstream on-ramp is located within 0.7747 miles
UPSTREAMOFF	0.5361	UPSTREAMOFF: Optimal binning for Y =0 if nearest upstream off-ramp is located further than 0.3205 miles =1 if nearest upstream off-ramp is located within 0.3205 miles
DOWNSTREAMOFF	0.4751	DOWNSTREAMOFF: Optimal binning for Y =0 if nearest downstream off-ramp is located further than 0.0638 miles =1 if nearest downstream off-ramp is located within 0.0638 miles
SSG2	0.4000	Standard Deviation of Speed at Station G
AVH2	0.3048	Average Volume at Station H
BASE_MILPOST	0.2841	BASE_MILPOST: Optimal binning for Y =0 if 0<base_milepost<=11.93 =1 if 11.93<base_milepost<=25.433 =2 if 25.433<base_milepost<=35.18 =3 if 35.181<base_milepost<=36.25
STATIONF	0.2828	Location of Station F relative to crash location =0 if Loop detector station nearest to crash location is located upstream =1 if Loop detector station nearest to crash location is located downstream
AOH2	0.2664	Average Occupancy at Station H

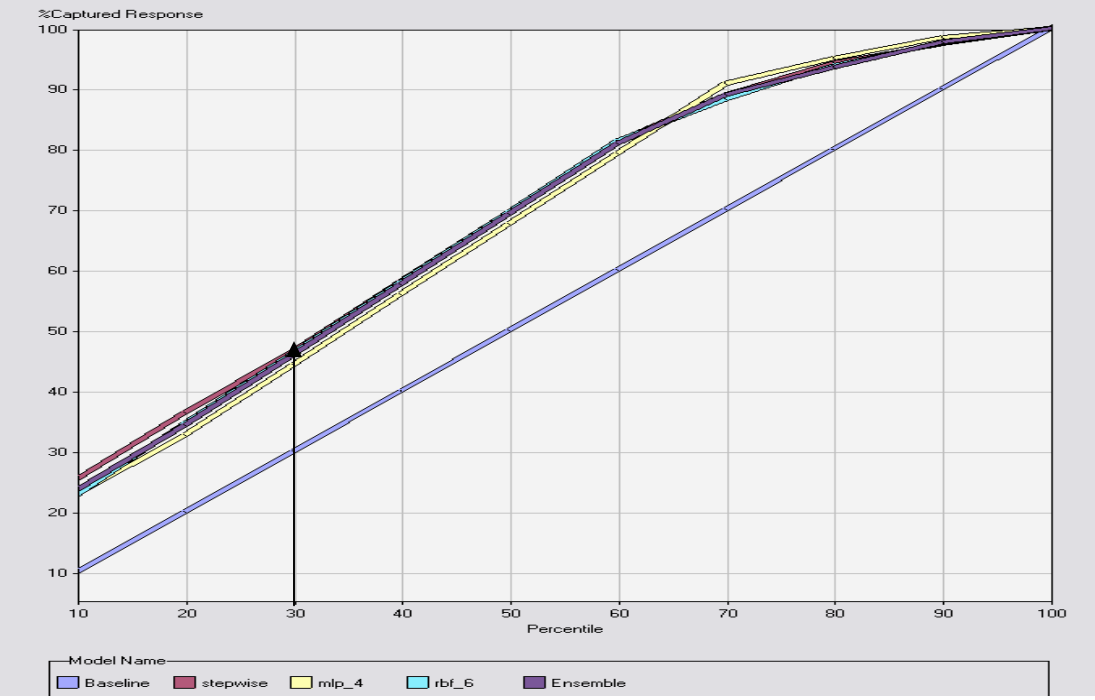

Figure 5-22: Percentage of captured response lift plot for best models belonging to different modeling techniques (input traffic parameters from Station D, E, F, G, and H)

The performance of various models in this set is depicted in Figure 5-22. The optimal performance of MLP and NRBF neural networks over validation dataset was achieved through networks with four and six hidden layer neurons, respectively. The best among the four models, the stepwise logistic regression model captured 46.57% of crashes in the validation dataset in first 30 percentile.

An interesting point to be noted here is that the presence of downstream on-ramp with-in 0.7747 miles remains a significant variable at all three stages of the modeling process; irrespective of the number of stations (1, 3 or 5) from which traffic data is being used. It is in contrast with regime 1 rear-end crashes since there the location of downstream on-ramp was no longer significant if we included data from three or five stations. The reason

for the same is that for regime 1 crashes the traffic conditions had already become congested and then the average occupancies at stations captured those conditions by recording higher occupancy 5-10 minutes before the crash. For regime 2 crashes slow moving traffic from the on-ramp does increase the chances of having a rear-end crash. But since they usually occur when the traffic is still moving at medium to higher speeds, and congested conditions have not expanded on the freeway this effect is not captured by the occupancy from surrounding stations 5-10 minutes before the crash. Therefore, the location of a downstream on-ramp always remains a significant variable.

The summary of the performance of regime 2 rear-end crash models belonging to three sets (utilizing traffic data from 1, 3 or 5 stations) is provided in Table 5-25. The structure (in case of neural networks) or selection procedure (in the case of logistic regression) along with the percentage of crashes captured within first three deciles of posterior probability are shown in table. The "best" model, capturing the highest percentage of crashes within first three deciles is highlighted in each row.

Table 5-25: Structure and percentage of captured response within the first three deciles for best models estimated for different combination of modeling techniques (Regime 2 rear-end crashes)

		Modeling Technique			
		MLP Neural Network	NRBF Neural Network	Logistic Regression	Ensemble Model
Traffic Parameters from	Station F	47.89% (6 hidden nodes)	49.24% (6 hidden nodes)	45.83% (Stepwise selection)	49.24%
	Station E, F, and G	47.20% (8 hidden nodes)	47.68% (4 hidden nodes)	43.87% (Backward selection)	47.20%
	Station D, E, F, G, and H	44.37% (4 hidden nodes)	46.38% (6 hidden nodes)	46.57% (Stepwise selection)	45.97%

The performance of the model utilizing data from only station of the crash is slightly better than the models utilizing data from three or five stations. Note that in case of regime 1 crashes the difference between three sets of model was more significant. However, when the performance of these three sets of models was closely examined it was noticed that the models using data from one, three and five stations respectively were good at identifying certain distinct patterns of crashes and non-crash cases. Therefore, the performance of individual models was indeed expected to improve by combining the best models in each of the three sets. Possible combinations of the highlighted models in each row of Table 5-25 were then estimated by averaging the posterior probabilities of the three individual models. It was found that the hybrid model does improve upon the performance provided by individual models.

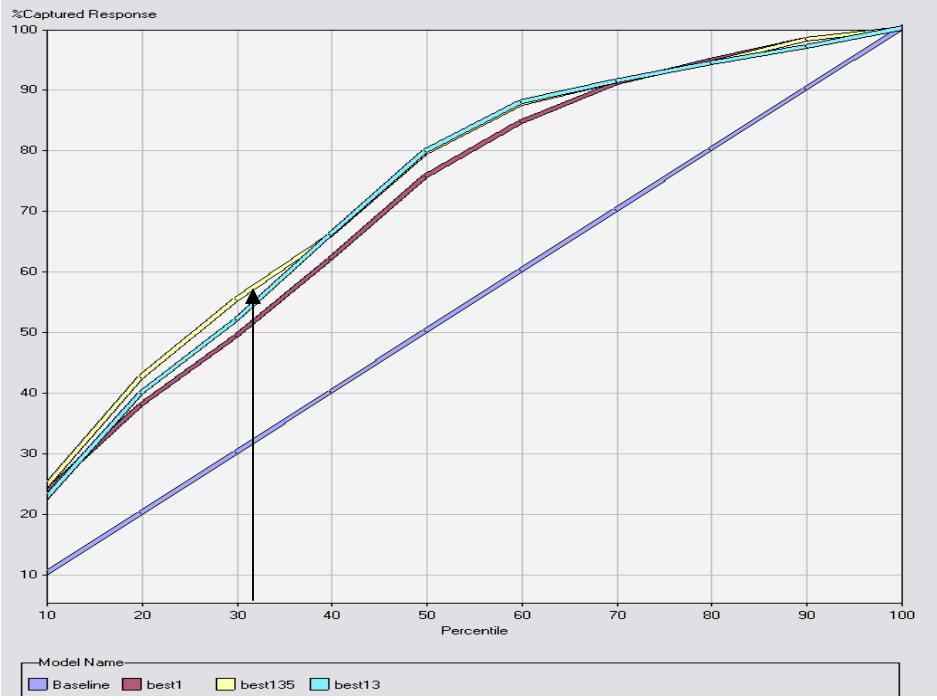

Figure 5-23: Percentage of captured response lift plot for combination of best models for regime 2 rear-end crashes chosen from the three sets

The performance of the two hybrid models titled *best13* and *best135* along with the best individual model (ensemble model using traffic data only from station F) are shown in Figure 5-23. The blue curve represents the model *best13*; created by combining the ensemble model (using traffic data only from station F) with the NRBF having four hidden nodes (using traffic data from stations E-G). The yellow curve represents the ensemble model *best135;* the combination of the all three models highlighted in Table 5-25. It may be seen that the yellow curve representing combination of the three models is significantly higher than the curve for the best individual model (purple color) at 30 percentile. It identifies more than 55% of crashes as compared to 49% of crashes identified by the best individual model. Therefore, the combination of three models is recommended for separating regime 2 rear-end crashes from normal traffic conditions.

5.6.2 Random non-crash data belonging to regime 2

The models developed so far used completely random non-crash data with the aim of identifying traffic and location-specific factors critically associated with two groups of rear-end crashes. As mentioned earlier another way, which is more attractive from a practical application standpoint, to analyze the problem would be to separate the non-crash data as well based on the regime it belongs. Traffic conditions belonging to regime 1 occur very infrequently (only 6% in the randomly selected loop data patterns) on freeways but make up close to 46% of rear-end crashes. Hence, it might be reasonable to classify every pattern that fits into the criterion (specified by the Leaves 1, 2, 4, and 6 in Figure 5-8; also see Table 5-1) of regime 1 rear-end crashes as crash. This way we would identify 46% of rear-end crashes by issuing warning only 6 to 7% of the times. Same procedure however would not work with regime 2 rear-end crashes. Although regime 2 crashes make up bigger portion of rear-end crashes (54%) these conditions are way more frequent (94% in the randomly selected lop data patterns) on the freeway. Hence, sophisticated prediction models might be needed to separate crashes from the non-crash cases within the data identified as regime 2.

In this section a procedure similar to the one adopted above is used to develop another set of prediction model(s) for regime 2 rear-end crashes. The only difference is that before beginning the modeling procedure we scored the non-crash sample using the tree model depicted in Figure 5-8. 259 non-crash data points were identified as belonging to regime 1 out of the 4972 non-crash observations used as non-crash data in the previous section. According to the approach being proposed here these observations would be classified as

a rear-end crash. We understand that in a real-time application such declarations would have led to false alarms but first of all these are reasonable number of false alarms considering that we could identify almost half of rear-end crashes. Also, it should be noted that the real phenomena of interest is crash prone conditions and since these conditions lead to so many crashes even with such little exposure (6% in the random non-crash data) one could classify them as crashes. These 259 observations were removed from the non-crash sample and remaining observations along with regime 2 rear-end crashes were subjected to the modeling procedure used in this chapter. The models resulted in slightly different performance but were comparable to the models developed in the previous section. It was expected since the data used for modeling were only slightly different in the two cases.

The summary of the performance of newly developed regime 2 rear-end crash models belonging to three sets (utilizing traffic data from 1, 3 or 5 stations) is provided in Table 5-26. The number of hidden neurons (in case of neural networks) and selection procedure (in the case of logistic regression) along with the percentage of crashes captured within 30 percentile of posterior probability are also shown in table. The “best” model, capturing the highest percentage of crashes within first three deciles is highlighted in each row.

Table 5-26: Structure and percentage of captured response within the first three deciles for best models estimated for different combination of modeling techniques (Regime 2 rear-end crashes)

		Modeling Technique			
		MLP Neural Network	NRBF Neural Network	Logistic Regression	Ensemble Model
Traffic Parameters from	Station F	48.82% (2 hidden nodes)	50.06% (4 hidden nodes)	49.04% (Stepwise selection)	49.81%
	Station E, F, and G	53.51% (4 hidden nodes)	53.71% (4 hidden nodes)	50.33% (Stepwise selection)	53.60%
	Station D, E, F, G, and H	51.05% (8 hidden nodes)	51.03% (6 hidden nodes)	51.01% (Stepwise selection)	51.01%

The performance of the model utilizing data from three stations (Stations E, F and G) is slightly better than the models utilizing data from one or five stations. Also, the performance of these models was slightly better over the validation dataset compared to the models developed for regime 2 crashes in the previous section. Note that the detailed description of the critical variables identified in the intermediate stages while developing these models are not discussed here because the critical variables associated with regime 2 crashes were thoroughly discussed in the previous section and the focus now is on the crash identification rate of the models.

In the next step the combinations of the highlighted models in each row of Table 5-26 were estimated by averaging the posterior probabilities from the three individual models. It was found that the ensemble models do improve upon the performance provided by individual models.

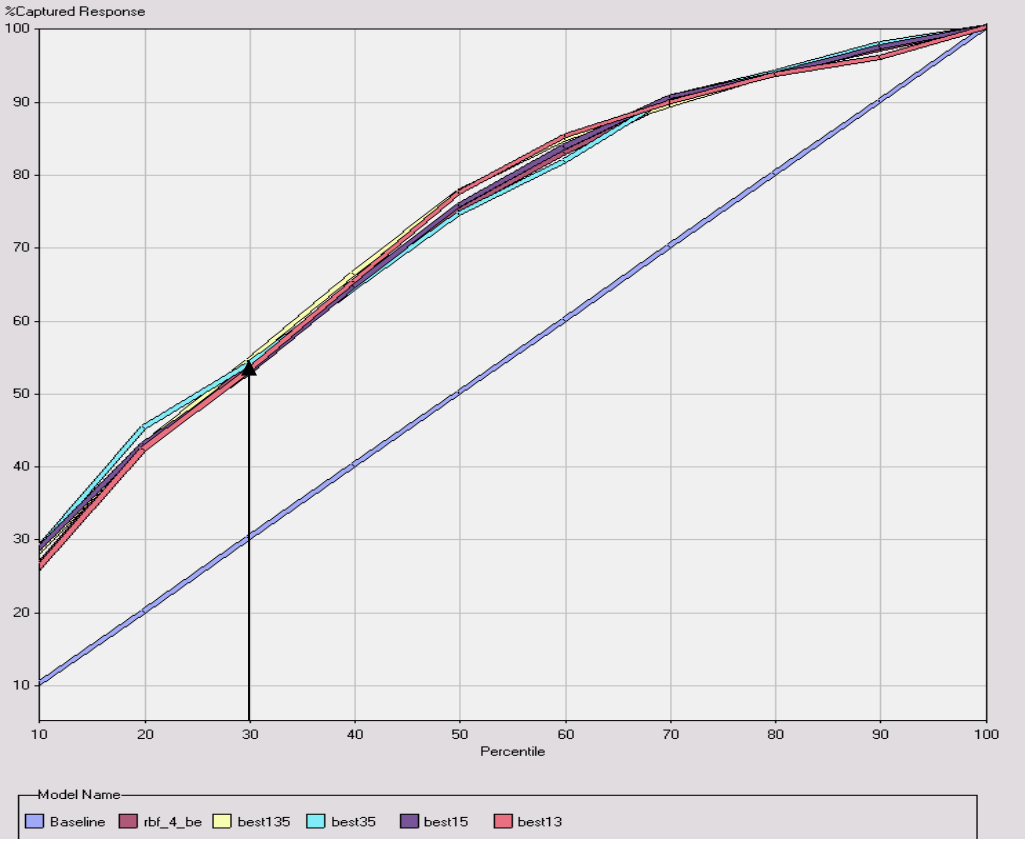

Figure 5-24: Percentage of captured response lift plot for combination of best models for regime 2 rear-end crashes chosen from the three sets

The performances of the all possible hybrid models along with the best individual model (NRBF model with four hidden neurons and traffic parameters from three stations) are depicted in Figure 5-24. The lift plot for model *best35* (representing the combination of best 3-station and best 5-station models) and *best135* (combination of all three models highlighted in Table 5-26) run very close and are slightly higher than the best individual model (representing NRBF network with four hidden nodes using data from station E, F and G). At 30th percentile the combination of the three models has captured maximum percentage of crashes (55.40% of the validation sample) and therefore, the combination

of three models is recommended for identification of regime 2 rear-end crashes. Again it must be noted that due to intermittent failure of the loops, data for the hybrid model may not be always available. These issues related to practical implementation would be discussed in Chapter 8.

5.7 Conclusions

This chapter presents a data mining approach to identify potential freeway crashes of the rear-end type using loop detector data. Random samples of non-crash data have been used alongside pre-crash loop detector data to develop models that can separate crash prone conditions from non-crash cases. These models are intended to be used for real-time detection of crash prone conditions on the 36-mile freeway corridor of Interstate-4 in Orlando. The focus in this chapter is on rear-end crashes which are the single most frequent type of crashes on the study area corridor.

First of all the available sample of 1620 rear-end crashes were divided into two clusters based on prevailing traffic speed configurations within 2-mile stretch of the freeway around crash location. Average speeds right before (0-5 minutes) the crash at stations D, F and H were used to represent the traffic speed configurations. In other words parameters *ASD1*, *ASF1* and *ASH1* were used as input to the clustering algorithm. It was detected through some exploratory analysis that the rear-end crashes belonging to the two clusters/groups differ in their frequency patterns over different times of day, days of week, etc. While cluster 1 crashes were more frequent on weekdays and in downtown Orlando area; cluster 2 crashes had comparatively "*uniform*" distribution over days of

week and along the freeway corridor. It was concluded from the exploratory analysis that crashes belonging to each cluster need to be separately analyzed. Disaggregating crashes by type as well as by prevailing traffic conditions before the crash was a major improvement from our previous work in which generic models were intended to predict all different types of crashes.

The first task in the process was to be able to classify any traffic pattern into cluster 1 or 2 based on prevailing traffic conditions. Two classification tree models were developed in order to separate the two groups of rear-end crashes. The difference between the two models was that one of them used average traffic speed inputs from time-slice 1 (i.e., *ASD1*, *ASF1* and *ASH1*) while the other used the same parameters from time-slice 2 (i.e., *ASD2*, *ASF2* and *ASH2*). It was noticed that although crashes were originally clustered according to the traffic speeds prevailing 0-5 minutes (time-slice 1) before the crashes, the classification tree with inputs from time-slice 2 could identify the cluster to which any crash belongs with sufficient accuracy. The hierarchy of rules to identify the clusters was summarized in Table 5-1. Based on these rules it could be inferred that cluster 1 crashes generally occur when low speed conditions prevail on extended segments of freeway for a relatively long period (at least 10 minutes before the crash) of time; while cluster 2 crashes occur under relatively free-flow traffic operation having high average speeds 5-10 minutes before the crash. It was noticed that if we score the sample of 1620 rear-end crashes with the rules formulated by the tree model 46% of them were identified as cluster 1 and the rest were identified as cluster 2. The traffic conditions belonging to the two clusters are referred to as regime 1 and regime 2. The crashes belonging to the two

regimes are regime 1 and regime 2 rear-end crashes, respectively. A randomly selected non-crash dataset was then also scored using the classification tree model. Only 6% of observations from this dataset fall into the definition of regime 1 rear-end crashes. The rarity of traffic patterns belonging to regime 1 led to a possible implementation approach in which one can use this tree model on real-time data and declare every observation that follows the hierarchy of rules belonging to regime 1 as crash without any further analysis. In which case separate models for regime 2 rear-end crashes should be developed with non-crash data that also identifies itself as regime 2 (i.e., leaves 3, 5 and 7 of the classification tree model).

In this chapter the data mining process has been used for identification of the binary target variable y (equals 1 for crash and 0 for non-crash). The process is repeated three times in this chapter. First, two sets of randomly selected non-crash loop detector data along with two groups of crashes were used to develop separate models that can identify individual groups of rear-end crashes from non-crash data. Model developed with such crash and non-crash sampling would help us identify critical factors associated with each group of rear-end crashes. Although the stated goal of this research is to be able to correctly classify crash prone conditions but since we have this valuable database that includes loop detector data as well as precise geometric features of the freeway, it can be effectively used to identify critical traffic and geometric features of interest that are associated with rear-end crash occurrence. With the real-time practical implementation approach in perspective, however, the mining process for regime 2 rear-end crashes was repeated after removing the non-crash data that belonged to regime 1 from the sample.

In the mining process tree node from the Enterprise Miner (SAS Institute 2001) was used to perform variable selection for the MLP and NRBF neural network based classification models. Standard forward, backward and stepwise procedures were used to select variables for the logistic regression tool. The best individual models belonging to each class of modeling tool were combined by averaging the output posterior probability. It should also be mentioned that even though only the models using data from time slice 2 (5-10 minutes before the crash) are described in this chapter, models using data from time slice 3 and 4 were also attempted but as expected they did not achieve the performance comparable to the models described. If those models would have resulted in better or almost comparable performances they would have been prescribed as potential crash prediction models because they would allow more leverage in terms of time available to process, analyze and disseminate the information that may in turn be used to avoid crashes.

In the modeling process for regime 1 rear-end crashes with completely random non-crash data (irrespective of the regime the non-crash data belong to) it was found that real-time traffic parameters from station F (i.e., the station nearest to crash location) were most critically associated with crash occurrence. The models with traffic parameters exclusively from Station F also resulted in better identification of crashes. The performance of individual MLP, NRBF and logistic regression models suffered when parameters from one or two stations in each direction (upstream and downstream) were involved in the modeling process. Also, it was noticed that the performance of models

belonging to various modeling technique was very comparable if they used traffic data from same stations. Therefore, as expected, combining these models by averaging their output posterior probability did not improve on the performance of individual models. However, there were significant differences in performances of the models for which potential input traffic parameters belonged to different set of stations (i.e., only from Station F, from Stations E-G or Stations D-H). Therefore, these models when combined by averaging their posterior probability slightly improved on the performances of individual models. It was found that the final hybrid model identified more than 88.5% of crashes in the first three deciles of posterior probability.

We now summarize the variables found critically associated with regime 1 rear-end crashes at various stages of the data mining process. At first stage traffic parameters only form station F along with off-line factors were included as potential independent variables. Among the off-line factors mile-post location and presence of a downstream on-ramp within 0.7743 miles were found to be most significant. The other two significant off-line predictors include the presence of off-ramp in upstream and downstream direction. Average occupancy at station of crash indicating congested conditions was among the traffic parameters found most significant. Coefficient of variation in speed along with standard deviation of occupancy at station F indicated frequent formation and dissipation of traffic queues are leading to high variation in speed and occupancy.

In subsequent steps traffic parameters from three (Station E, F and G) or five stations (Station D, E, F, G and H) and not just from station of the crash (Station F) were included

as potential independent variables. It was noticed that mile-post location and presence of an on-ramp downstream were replaced by average occupancy upstream (*AOE2*) and downstream of the station of crash (*AOG2*). The reason for the same is that if we include traffic data from more stations in the analysis the effect of mile-post location and presence of on-ramp is essentially captured by high average occupancies measured at upstream and downstream stations. Coefficient of variation downstream of crash location was also a significant variable. When parameters from five stations (Station D through H) were examined it was noticed that average volume and average occupancy at station D (*AVD2* and *AOD2*) were the significant variable from the extreme upstream station. Average and standard deviation of occupancy (*AOH2* and *SOH2*) were the significant variables from station H. It indicates that during 5-10 minutes period before regime 1 rear-end crashes traffic volume is high approximately 1-mile upstream of the crash location (*AVD2*) with high occupancy conditions at stations E through H . Although it leads to high variation in speed at three stations (*CVSF2*, *CVSG2* and *CVSE2* were all significant and in that order) surrounding the crash location; the most significant of them is measured near the crash location (*CVSF2*).

Note that the factors explicitly accounting for driver population on the freeway corridor under consideration figured in none of the models, however, the optimal binning transformation of "*base_milepost"* (representing the mile-post location of crash and non-crash cases) was significant for models that used traffic inputs only from Station F. It is suspected that driver population related factors might be implicit in the mile post

location. Also, location of station F with respect to crash location (depicted by the binary variable *stationf*) was not found significant at any stage of modeling procedure.

A similar approach of starting with parameters only from the station closest to the crash location and subsequently including parameters from three and five stations was adopted for regime 2 rear-end crashes. The models involving traffic parameters only from station F still performed better than the other models. However, unlike regime 1, the difference in the overall performance of models among three sets was not very significant. The performance of these set of models did differ on individual observations of the validation dataset which indicated that combining the models might improve on the performance of best individual models. Indeed the final hybrid model achieved better performance and identified 55.3% crashes in the validation dataset within the first three deciles of output posterior probability.

For regime 2 rear-end crashes more off-line factors were found significant at different stages of modeling procedure. Average speeds along with average volume (*ASF2* and *AVF2*) were the most significant traffic related variables when parameters from only station of crash were included as potential inputs. Among the off-line factors presence of on-ramp and off-ramp in the downstream direction and presence of off-ramp in the upstream direction were included in the list of critical variables. Location of station F with respect to the crash location was also found significant by the classification tree used for variable selection.

In the subsequent stages when parameters from three stations were considered as potential inputs average speed downstream of station of crash (i.e., *ASG2*) became the most significant traffic parameter. *ASE2*, *ASF2* and *ASG2*, i.e., average speeds at all three stations are significant along with the variable "*stationf*" that indicates the location of station of crash with respect to precise location of crash. These parameters indicate that interplay between the three average speeds might affect the possibility of a crash. For example, higher speeds at station E and F (if it happens to be upstream of crash site) and lower speeds at station G might lead to a rear-end crash. Average volume upstream of crash location (*AVE2*) was also found significant.

When we included traffic data from five stations *AVD2* was the only parameters found significant from the extreme upstream station. It indicated that if there is high demand upstream of a location and high occupancy (indicated by significant *AOH2*) downstream; it could cause a rear-end crash even though the speeds at location around station of crash appear "normal" and no queuing is visible. Average speeds at station of the crash (*ASF2*) and downstream (*ASG2* and *ASH2*) of it were still critical variables. Standard deviation at Station G was also found significant indicative of the unstable traffic that might lead to a rear-end crash upstream.

Comparing the results for two groups of rear-end crashes it was noticed that while most regime 1 crashes may be identified through congested traffic conditions (indicated by high occupancy) in immediate vicinity of the crash location; the geometric and traffic characteristics downstream of crash location play a more significant role for regime 2

crashes. Many location specific variables were in fact identified to be significant by the tree model used for variable selection for these crashes. Also, for regime 2 rear-end crashes the variable indicating presence of on-ramp downstream remained significant even when traffic parameters from three or five stations were included as inputs. From these findings it is inferred that monitoring situation at individual freeway locations through data from each loop detector might be sufficient to anticipate rear-end crashes that belong to regime 1. However, to identify regime 2 rear-end crashes at a certain location the situation at that site must be monitored along with the traffic downstream of it.

Note that these results are obtained by modeling crashes belonging to the two regimes with random non-crash data irrespective of the traffic conditions (regime 1 or regime 2) they belong. It does provide us an insight into factors responsible for rear-end crashes in two traffic regimes. However, from a practical stand point it may be argued that the non-crash data also needs to be classified into two regimes and then the non-crash data only belonging to that particular regime should be used as non-crash sample while modeling individual groups of rear-end crashes. It essentially means that while modeling regime 2 rear-end crashes we should only use random non-crash data that also belong to regime 2. Therefore, it was decided to repeat the mining process for regime 2 rear-end crashes after removing data belonging to regime 1 from the non-crash sample. The mining process showed that the NRBF neural network model with 4 hidden nodes and traffic data from three loop detector stations was the best individual model. It identified close to 54% of the crashes in the validation dataset. In the next step best models in each category (i.e.,

parameters from 1, 3 or 5 stations) were combined to get an improvement in the performance. The models did improve the performance slightly and the combination of best models in each category (titled *best135*) identified almost 55% of the crashes from the validation dataset. The performance of the model was same as that of the ensemble model developed for regime 2 rear-end crashes with completely random non-crash data. It was expected because the only difference between the datasets used to develop the two models was those 6% deleted non-crash observations that belonged to regime 1.

Note that since traffic conditions constituting regime 1 rear-end crashes are a rarity the mining process was not carried out for regime 1 rear-end crashes using non-crash data belonging to regime 1 only. It was decided that any real-time loop data pattern that is classified as regime 1 at the time of application would be declared as a rear-end crash.

In a field implementation plan formulated based on the discussion above we can subject the incoming data to the tree model shown in Figure 5-8 (and Table 5-1). If the tree model results in a regime 1 classification the data pattern may be declared as crash prone and warning for a rear-end crash can be issued. It would be appropriate since conditions associated with regime 1 rear-end crashes (i.e., those identified in Table 5-1) were found to occur in only 6 to 7% of the cases if a random sample of loop detector data is drawn. If data is found belonging to regime 2 traffic speed conditions (identified by leaves 3, 5 and 7 of the classification tree in Figure 5-8) it may be subjected to further models developed in Section 5.6.2. Note that the models presented in that section were developed with

regime 2 non-crash data only and hence are designed to separate crashes within the data satisfying regime 2 traffic conditions.

It is worth mentioning at this point that the performance of the models must be seen in terms of their data requirements as well. Since the data from three or five stations might not be simultaneously available due to failure of the loops, it would be more practical just to use data from one station to identify these crashes. Therefore, even though hybrid model combining best models from the three sets (1-station, 3-station and 5-station models) provide improved performance for regime 2 crashes it doesn't make it an automatic choice for field implementation.

Evaluating crash identification performance of strategies proposed to 'predict' the two groups of rear-end crashes it is apparent that regime 1 rear-end crashes are more readily 'predictable' through real-time traffic conditions. By issuing warnings 6 to 7 % of times we can identify all regime 1 rear-end crashes, however, for remaining 94% cases we would have to issue warnings about 30% of times (i.e., declare data belonging to first three deciles of output probability as regime 2 rear-end crash) to identify 55% of regime 2 crashes using the best hybrid model for regime 2 crashes (Figure 5-24). It indicates that the traffic conditions prevalent 5-10 minutes before regime 1 rear-end crashes are more distinct from normal traffic in general. Note that it is not to argue that methodology to predict regime 2 crashes is less useful than the strategy to predict regime 1 crashes. It is possible that measures such as Variable Speed Limits for reducing the risk of crashes belonging to this regime (regime 2) might be more easily applicable (Dilmore, 2005).

Moreover, since regime 2 crashes would generally occur under higher traffic speeds they may be expected to be more severe. Hence avoiding every single crash in this group might be more beneficial than its counterparts in regime 1. The final classification models recommended for both groups of rear-end crashes (i.e., regime 1 and 2) in this chapter would be recalled in Chapter 8 that discusses the deployment strategy for a reliable crash warning system.

In the next chapter we explore matched case-control logistic regression and PNN models, the techniques evaluated earlier to develop generic models, for classification. The performance of those models would be compared with the data mining based models developed here so that recommendation for optimal real-time identification of rear-end crashes may be made.

CHAPTER 6

PNN AND LOGISTIC REGRESSION MODELS FOR REAR-END CRASHES

6.1 General

In the previous chapter rear-end crashes were classified into two groups. One group of rear-end crashes (i.e., regime 1) was associated with extended congested conditions while the other (regime 2) was associated with the 'disturbances' downstream of the crash site leading to spatial speed differential. While crashes in the former category could be identified because of rarity of the conditions under which they occur on the freeway, complex models were required to identify the crashes belonging to the later. In this regard, multi layer perceptron (MLP) and normalized radial basis function (NRBF) neural network based classification models were developed for regime 2 rear-end crashes. The MLP/NRBF based models were combined with each other and improved crash identification was achieved.

In our previous studies two other modeling techniques, namely, probabilistic neural network (PNN) (Abdel-Aty and Pande, 2005) and matched case-control logistic regression (Abdel-Aty et al., 2004, 2005) were successfully explored to develop generic models for real-time crash identification. Since there remains scope for improvement in identification of rear-end crashes belonging to regime 2, these two modeling techniques are explored in this chapter. The performance of the models developed here would be compared to the models estimated in the previous chapter. The relationship among the outputs from various models would be explored as well.

Note that we have already identified factors associated with regime 1 rear-end crashes and a fairly reliable classification procedure is already available. Therefore, modeling techniques adopted in this chapter are only used for regime 2 rear-end crashes. Also, since the application strategy proposed in the last chapter includes classifying all observations belonging to regime 1 (which makes only 6-7% of the freeway traffic conditions) as a "rear-end crash"; the non-crash data used in this study only belongs to regime 2. In this regard the models developed here are comparable to the models developed in section 5.6.2 of the previous chapter. For developing neural network models, in that section, non-crash observations belonging to regime 1 were removed from the random non-crash database. The datasets used for training and validation in this chapter are identical to the ones used in section 5.6.2. Developing models using the same dataset (training) and then evaluating their performance (also on the same dataset; validation) also allows for making meaningful comparison of the performance of various models.

This chapter is divided into five sections. The next section deals with analysis of regime 2 rear-end crash data in the framework of with-in stratum matched sampling. A logistic regression model is estimated for the binary target. The performance of this model is then examined over the validation dataset used in the previous chapter to assess the performance of the MLP/NRBF models. In the following section, PNN based classification is explored for identification of regime 2 rear-end crashes. Relationships between outputs from PNN models, Hybrid MLP/NRBF models (from the previous chapter) and the matched logistic regression model was explored in the section after that.

The final section summarizes the conclusions from the modeling procedure. These conclusions will be recalled in Chapter 8 while formulating a system for the reliable real-time identification of crashes on the Interstate-4 corridor.

6.2 Matched case-control Logistic Regression

6.2.1 A brief review of methodology

The main advantage of the matched case-control sampling strategy is that it implicitly accounts for various factors such as crash site, time, season, day of the weak, etc. These factors are accounted for by using a within-stratum matched sampling for the binary outcome variable *y* (crash or non-crash) as a function of traffic flow variables $X_1, X_2, \ldots X_k$ from matched crash-non-crash cases where a matched set (referred to as stratum) can be formed using crash site, time, season, day of the weak, etc., as controls. In epidemiological studies, this is known as matched case-control analysis. The sampling technique essentially controls the variability due to matching factors and their effect is implicit in the intercept term. Matched case-control sampling and logistic regression technique was described in detail in Chapter 3.

First of all, simple (involving one covariate) logistic regression models are developed to examine the effect of individual covariates. The analysis from these models is followed by stepwise model selection procedure for estimation of a multivariate model. The sequence adopted for selecting critical variables for the multivariate model was proposed in one of our earlier studies (Abdel-Aty et al., 2005).

6.2.2 Simple models

For each of the seven loop detectors (*C* through *I*) and six time slices (1-6), values of 5-minute averages (*AS, AV, AO*) and standard deviations (SS, SV, SO) of speed, volume and occupancy were available for all crash and the corresponding non-crash cases. The extraction of raw 30-second loop data for crashes and corresponding non-crash cases along with their aggregation to 5-minute level was explained previously. The analysis was carried out with 70% regime 2 rear-end crashes (and their matched non-crash cases) of training dataset used to calibrate the MLP/NRBF models.

Due to data availability, there were different numbers of non-crash cases for each crash. To carry out matched case-control analysis we created a symmetric data sets (i.e., each crash case in the dataset has the same number of non-crash cases as controls) by randomly selecting five non-crash cases for each crash in all four datasets. The choice of selecting five as the number of corresponding non-crash cases was based on one of our earlier findings (Abdel-Aty et al., 2004) which essentially indicated no differences among the results from five different *1: m* datasets (with *1* crash and *m* corresponding non-crash with *m* varying between one to five).

There were 252 potential covariates in all (7 stations * 6 time-slices * 3 parameters (speed, volume, and occupancy) * 2 effects (average and standard deviation)). As part of the preliminary assessment 252 simple models were estimated. The results of simple logistic regression models for the variables pertaining to 5-minute averages of three parameters *(AS, AO, AV)* are shown in Table 6-1. Table 6-2 shows similar results for 5-

minute standard deviations of the same parameters *(SS, SO, SV)* measured at seven loop detectors and six time slices. Based on these results one can identify time duration(s) and location of loop detector(s) whose traffic characteristics are significantly correlated with the binary outcome (crash vs. non-crash in the vicinity of Station F).

The results include the hazard ratio estimated through the proportional hazard regression analysis procedure (*PHREG* of *SAS*) along with the p-value for the chi-square test indicating whether the hazard ratio is significantly different from one. The hazard ratio is an estimate of the expected change in the odds of having a crash. If the output hazard ratio for a variable is significantly different from one (e.g., 2) then increasing the value of this variable by one unit would double the risk of a crash at station *F* (station of the crash). Based on the hazard ratios presented in Tables 6-1 and 6-2 we can conclude that only average speeds downstream (Stations G, H and I) have any significant impact on crash occurrence. P-value for other parameters is greater than 0.05, thereby not rejecting the null hypothesis of no significance of those parameters.

It may be seen from the tables that the average speeds downstream have a significantly negative coefficient (i.e., hazard ratio significantly less than 1) indicating that as the average speed downstream of a freeway location decreases, the odds of crash occurrence increase. Variation (represented by 5-minute standard deviations) of any parameter is not significant at 95% confidence level. Note that none of the parameters (average or standard deviations) have significant hazard ratio beyond time slice 1 at any of upstream station (Station C, D, or E) or station of the crash (Station F). It can be concluded that if

the conditions on a freeway location is relatively free flowing (regime 2 conditions) then the risk of observing a rear-end crash is primarily governed by the average speed measured 1 to 2 miles downstream of that location. Moreover, if one is looking at traffic conditions from a crash identification point of view, i.e., at least 5-10 minutes before crash occurrence, none of the standard deviation parameters are significant. Conclusions from simple models are utilized in the multivariate model building procedure.

Table 6-1: Hazard ratios for AS, AV, and AO measured at 5-minute level during six different time slices and seven stations

		AS			*AV*			*AO*		
Station	**Time slice**	**Hazard Ratio**	**chi-sq.**	**p-value**	**Hazard Ratio**	**chi-sq.**	**p-value**	**Hazard Ratio**	**chi-sq.**	**p-value**
C	1	0.998	0.0895	0.7648	1.002	0.0198	0.888	0.998	0.2904	0.5899
C	2	1.004	0.4311	0.5114	1.026	2.3611	0.1244	1.001	0.7237	0.3949
C	3	1.004	0.6223	0.4302	1.043	2.8371	0.0921	1	0.1236	0.7252
C	4	1.004	0.588	0.4432	1.021	0.9102	0.3401	1.001	0.2095	0.6471
C	5	1.009	2.3806	0.1229	1.034	3.0403	0.0812	1.001	1.0834	0.2979
C	6	1.006	1.237	0.2661	1.016	0.7114	0.399	1	0.0005	0.9818
D	1	1.006	1.2637	0.261	1.01	0.967	0.3254	0.997	0.6128	0.4337
D	2	1.01	2.7889	0.0949	1.006	0.3309	0.5651	0.997	0.4804	0.4882
D	3	1.006	1.1386	0.2859	1.009	0.8413	0.359	0.992	1.6742	0.1957
D	4	1.002	0.0803	0.7769	1.008	0.6932	0.4051	0.995	0.8769	0.349
D	5	1	0.0021	0.9638	1.01	1.0923	0.296	0.997	0.4519	0.5014
D	6	0.999	0.0669	0.7959	1.007	0.5615	0.4537	0.994	1.095	0.2954
E	1	0.989	4.8758	0.0272	0.995	0.0951	0.7578	1.001	0.9571	0.3279
E	2	0.993	1.9633	0.1612	0.989	0.4105	0.5217	1.001	0.3145	0.5749
E	3	0.994	1.3924	0.238	1.019	1.3955	0.2375	1.001	0.5172	0.472
E	4	0.995	0.8257	0.3635	1.015	0.7998	0.3712	1	0.1155	0.7339
E	5	0.992	2.2817	0.1309	1.014	0.6689	0.4134	1.001	0.251	0.6164
E	6	0.995	1.0691	0.3012	1.012	0.5812	0.4458	1	0.0018	0.9666
F	1	0.989	6.0306	0.0141	0.991	0.1627	0.6867	1.003	1.0391	0.308
F	2	1.001	0.0794	0.7781	1.003	0.0235	0.8782	0.993	1.974	0.16
F	3	1	0.0011	0.9738	0.994	0.0628	0.8021	1.001	0.6669	0.4142
F	4	1.001	0.0741	0.7855	1.037	2.1279	0.1446	1.004	1.3271	0.2493
F	5	0.998	0.2083	0.6481	1.033	1.9992	0.1574	0.998	0.3388	0.5605
F	6	0.996	0.7718	0.3797	1.017	0.6527	0.4191	0.998	0.2838	0.5942
G	1	0.977	28.7152	<.0001	0.99	1.1425	0.2851	1.007	2.8297	0.0925
G	2	0.985	12.1697	0.0005	0.995	0.3791	0.5381	1.004	0.4574	0.4988
G	3	0.989	6.1333	0.0133	1	0.0014	0.9704	1	0	0.9958
G	4	0.989	6.5809	0.0103	1	0	1	1.003	0.76	0.3833
G	5	0.989	6.7818	0.0092	0.998	0.0681	0.7942	1.001	0.0897	0.7646
G	6	0.987	8.5604	0.0034	1	0.0009	0.9764	1.004	1.1816	0.277
H	1	0.973	28.429	<.0001	0.988	0.6918	0.4056	1	0.015	0.9025
H	2	0.978	17.8578	<.0001	0.989	0.5788	0.4468	1	0.0103	0.9191
H	3	0.979	17.4421	<.0001	1	0.0001	0.9905	1	0.0552	0.8143
H	4	0.98	16.604	<.0001	0.993	0.3191	0.5722	0.999	0.0501	0.823
H	5	0.979	17.8275	<.0001	0.995	0.1762	0.6746	1.001	0.2211	0.6382
H	6	0.98	15.2274	<.0001	0.992	0.3625	0.5471	0.997	0.5674	0.4513
I	1	0.982	11.4615	0.0007	0.993	0.4659	0.4949	1.003	2.4549	0.1172
I	2	0.987	5.5794	0.0182	0.983	1.5122	0.2188	1.002	0.7184	0.3967
I	3	0.99	3.4816	0.0621	0.992	0.6053	0.4366	1	0	0.9977
I	4	0.994	1.4682	0.2256	0.99	0.9149	0.3388	1.001	0.8078	0.3688
I	5	0.993	1.7042	0.1917	0.991	0.6832	0.4085	0.999	0.1867	0.6657
I	6	0.993	1.6396	0.2004	0.987	1.0904	0.2964	1.003	0.6846	0.408

Table 6-2: Hazard ratios for SS, SV, and SO for 5-minute level during six different time slices and seven stations

		SS			*SV*			*SO*		
Station	**Time slice**	**Hazard Ratio**	**chi-sq.**	**p-value**	**Hazard Ratio**	**chi-sq.**	**p-value**	**Hazard Ratio**	**chi-sq.**	**p-value**
C	1	1.01	0.6801	0.4096	0.999	0.0015	0.9687	1	0.0012	0.9723
C	2	1.001	0.0076	0.9307	1.029	2.708	0.0998	1.001	1.4015	0.2365
C	3	1.008	0.4514	0.5017	1.043	2.6359	0.1045	1.001	1.2228	0.2688
C	4	1.015	1.748	0.1861	1.016	0.5184	0.4715	1	0.155	0.6938
C	5	0.992	0.476	0.4902	1.032	2.7474	0.0974	1.001	0.9846	0.3211
C	6	0.994	0.2866	0.5924	1.049	3.7068	0.0542	1.001	0.952	0.3292
D	1	1.016	1.9909	0.1582	1.007	0.2508	0.6165	1	0.0845	0.7713
D	2	1.014	1.5087	0.2193	1.003	0.0756	0.7834	1	0.0039	0.9503
D	3	1.002	0.0232	0.8789	1.002	0.0147	0.9036	0.996	0.7259	0.3942
D	4	1.013	1.332	0.2484	1.001	0.0106	0.918	0.998	0.4936	0.4823
D	5	1.011	0.8222	0.3645	1.002	0.0387	0.844	0.999	0.3305	0.5654
D	6	1.008	0.5392	0.4628	1.003	0.0402	0.8411	0.997	1.0857	0.2974
E	1	1.007	0.3705	0.5427	0.984	0.3571	0.5501	1.001	1.011	0.3147
E	2	0.998	0.0383	0.8448	0.977	1.0976	0.2948	1	0.4315	0.5113
E	3	1.001	0.0053	0.942	1.013	0.7637	0.3822	1.001	0.7255	0.3943
E	4	0.993	0.3451	0.5569	1.004	0.0256	0.8729	1	0.0428	0.8361
E	5	0.997	0.0705	0.7907	0.997	0.014	0.9057	1	0.1918	0.6614
E	6	0.994	0.273	0.6013	1.018	1.2095	0.2714	1	0.2275	0.6334
F	1	1.018	3.0111	0.0827	1.033	1.513	0.2187	1.002	2.3866	0.1224
F	2	1.001	0.0164	0.8982	1.05	2.3636	0.1242	0.998	0.8487	0.3569
F	3	1.004	0.1273	0.7212	1.011	0.2267	0.634	1.001	1.8707	0.1714
F	4	0.998	0.0261	0.8716	1.03	1.2852	0.2569	1.002	2.9994	0.0833
F	5	1.013	1.5183	0.2179	1.049	4.0484	0.0442	1	0.0303	0.8618
F	6	1.008	0.5139	0.4734	1.026	0.8183	0.3657	0.999	0.201	0.6539
G	1	1.02	3.7177	0.0538	0.989	0.8959	0.3439	1.001	0.5373	0.4636
G	2	1.006	0.3346	0.5629	0.982	1.5955	0.2065	0.993	0.8822	0.3476
G	3	1.002	0.0403	0.841	0.998	0.0444	0.833	1	0.0152	0.9019
G	4	1.004	0.1337	0.7147	0.992	0.5979	0.4394	1	0.0158	0.8999
G	5	1.008	0.5076	0.4762	0.991	0.6128	0.4337	0.999	0.2255	0.6349
G	6	1.022	4.1916	0.0406	0.997	0.0889	0.7656	1.001	0.2165	0.6417
H	1	1.009	0.4987	0.4801	0.988	0.44	0.5071	0.999	0.2503	0.6168
H	2	1.008	0.4268	0.5136	0.976	1.3052	0.2533	1	0.0868	0.7683
H	3	0.996	0.1036	0.7475	0.989	0.4558	0.4996	1	0.0418	0.8379
H	4	1.001	0.0071	0.9329	0.945	2.1338	0.1441	1	0.1386	0.7097
H	5	0.994	0.2624	0.6084	0.988	0.5661	0.4518	1	0.0165	0.8978
H	6	1.006	0.2571	0.6121	0.97	1.2873	0.2565	0.998	1.3366	0.2476
I	1	0.997	0.0637	0.8007	0.988	0.7542	0.3851	1.001	1.4107	0.2349
I	2	0.999	0.005	0.9436	0.954	1.9697	0.1605	1	0.1986	0.6558
I	3	1.01	0.7924	0.3734	0.983	1.2816	0.2576	1	0.0037	0.9514
I	4	1.013	1.3235	0.25	0.973	1.7928	0.1806	1	0.2893	0.5906
I	5	1.017	2.215	0.1367	0.98	1.4206	0.2333	0.999	0.3953	0.5295
I	6	1.014	1.473	0.2249	0.973	1.6658	0.1968	1	0	1

6.2.3 Multivariate model building procedure

First step toward a multivariate logistic regression model was to identify the set of variables most significantly related to the binary outcome variable *y* (*y=0* for non-crash and *y=1* for crash*)* in the dataset. Following the discussion above, the stepwise automatic variable selection option in PHREG procedure of SAS was used in stages to identify significant predictors among the sets of:

(i) 5 AS and (ii) 5 SS variables

(iii) 5 AV and (iv) 5 SV variables

(v) 5 AO and (vi) 5 SO variables

Note that only five average and standard deviation parameters are used for the multivariate model. These parameters are measured from five stations (Stations D through H) during one time slice (time slice2; 5-10 minutes prior to the crash). The choice is based on our findings from the data mining analysis and the simple logistic regression models in the previous section. The most significant predictors found separately in each of these six groups of variables, were then considered together under the stepwise selection procedure and the final set of significant predictors was determined. All parameter estimates and related statistical summary of the coefficients for the model fitted with the final set of significant predictors is provided in Table 6-3.

Table 6-3: Final model developed for regime 2 rear-end crashes using stepwise selection procedures

Analysis of Maximum Likelihood Estimates					
Variable	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**p-value**	**Hazard Ratio**
ASD2	0.03018	0.00787	14.6946	0.0001	1.031
ASG2	-0.02064	0.00657	9.8647	0.0017	0.980
ASH2	-0.02061	0.00680	9.1919	0.0024	0.980

6.2.4 Model interpretation

The final model only includes three average traffic speed parameters while other parameters were found insignificant relative to these variables. The coefficient for two parameters representing speeds downstream of the crash site (ASG2 and ASH2) is negative and almost equal in magnitude while the coefficient for the upstream average speed (ASD2) is positive. Since regime 2 traffic consists of relatively free-flow conditions (identified in Table 5-1 of the previous chapter) it means that under non-congested traffic conditions speed differential between upstream and downstream stations 'causes' a rear-end crash on the freeway section in between. A possible explanation for may be that the drivers under medium to high speed traffic conditions are caught unaware of the congestion that had been building up downstream as suggested by low average speeds at stations D and H; 5-10 minutes prior to the time of crash. The interpretations of the model indicagte that variable speed limit based measures may be used to reduce the potential for rear-end crashes that occur under not-so-congested situation. For example, if the speed downstream of a freeway location appears to be dropping with respect to

speeds one mile upstream then the speed limit at downstream section may be increased to clear up the congestion from that site.

Note that this model could only be estimated by carefully segregating crashes not only by type of collisions (or the first harmful event such as the rear-end collisions) but even further disaggregating the rear-end crashes by prevailing traffic speed regime. Hence, the whole application should be seen in totality with the previous chapter where we identified the two groups of rear-end crashes. This emphasizes the premise of this study about added precision in crash identification that can be achieved by examining crash data as smaller groups at the modeling stage.

6.2.5 Classification performance of the models

As explained in the methodology chapter the log odds ratio of crash occurrence due to traffic flow vector x_{1j} relative to vector x_{2j} are given by the following equation

$$\log\left\{\frac{p(x_{1j})/[1-p(x_{1j})]}{p(x_{2j})/[1-p(x_{2j})]}\right\}=\beta_1(x_{11j}-x_{12j})+\beta_2(x_{21j}-x_{22j})+.....+\beta_k(x_{k1j}-x_{k2j}) \qquad (1)$$

The right hand side of this equation, i.e., the value of log odds ratio, can be estimated using the estimated β coefficients shown in Table 6-3. One may utilize the relative log odds ratio for predicting crashes by replacing x_{2j} by the vector of values of the traffic flow variables in the j^{th} stratum under normal (i.e., non-crash) traffic conditions. One may conveniently use simple average of five (corresponding) non-crash observations within the stratum for each variable. If we let $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j} ..., \bar{x}_{k2j})$ denote the

vector of mean values of non-crash cases of the k variables within the j^{th} stratum, then the log odds of crash relative to non-crash may be approximated by:

$$\log\left\{\frac{p(x_{1j})/[1-p(x_{1j})]}{p(\bar{x}_{2j})/[1-p(\bar{x}_{2j})]}\right\} = \beta_1(x_{11j} - \bar{x}_{12j}) + \beta_2(x_{21j} - \bar{x}_{22j}) + \ldots\ldots\ldots + \beta_p(x_{k1j} - \bar{x}_{k2j}) \qquad (2)$$

The above log odds ratio can then be used to predict crashes by establishing a threshold value that yields that desirable crash classification accuracy (Abdel-Aty et. al, 2004).

To compare the logistic regression model with the neural network models developed in the previous chapter, its classification performance on the validation dataset used there for the regime 2 rear-end crashes must be examined. Hence, we need to obtain an odds ratio for each observation in the dataset, which in turn requires 'matched' cases for every observation in the validation dataset. Note that the validation dataset not only consists of remaining 30% regime 2 rear-end crashes but also has randomly selected non-crash cases. While we have the required matched cases for all the crashes since they were assembled at the data preparation stage (For details on the collection of corresponding matched cases refer Chapter 4) we still need to collect corresponding matched cases in order to calculate the odds ratio for the random non-crash cases as per equation 2 provided above. The correspondence here means that, for example, if a random non-crash belongs to April 19, 1999 (Monday) 9:00 PM, I-4 Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, three loops upstream and three loops downstream of station 30 for half an hour period prior to the estimated time of the crash for all the Mondays of the year at the same time. Hence, this random non-crash

case will have loop data table consisting of the speed, volume, and occupancy values for all three lanes from the loop stations 27-33 (on eastbound direction) from 8:30 PM to 9:00 PM for all the Mondays of the year 1999, with one of them being the day belonging to the original random non-crash case and others being the matched cases for it. The raw 30-second loop data was aggregated as previously and average speed parameters included in the logistic regression were calculated for the matched cases. Out of all available matched cases for each random non-crash, five were randomly selected and three average speed parameters were averaged over those cases to get values of $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j} ..., \bar{x}_{k2j})$ for random non-crash cases. Using the beta coefficients (from Table 6-3) along with $\bar{x}_{2j}$ for crash and non-crash cases on Equation 2, odds ratio was determined for every crash and non-crash case in the validation dataset.

Note that this odds ratio is somewhat analogous to the posterior probability value output by the neural network models developed in the previous chapter. Even though unlike the posterior probability the value of odds ratio is not confined to 1 it is expected to be higher for crashes and lower for non-crash cases. The observations in the validation dataset (only the original crash and random non-crash cases and not the matched cases for either crash or non-crash) were sorted by their odds ratio and lift plot was obtained to examine what percentage of crashes are identified at various within various percentiles of odds ratio. Model performance may be assessed by the percentage of crash identified within first few deciles of odds ratio. Figure 6-1 shows the percentage of crashes from the validation sample identified on the y-axis at various percentiles of odds ratio shown on the x-axis. The lift plot is identical to the ones shown in the previous chapter with the

only difference being that the output used to arrange the observations is from the logistic regression model. Since MLP/NRBF models for regime 2 rear-end crashes in the previous chapter were assessed using the crash identification percentage within first three deciles of posterior probability, same criterion is adopted here. It was noticed that setting the threshold for odds ratio at its 30 percentile value one could identify 41.8% crashes in the validation dataset. This crash identification percentage is considerably lower than the fraction (55.4%) identified by the hybrid model developed in the previous chapter.

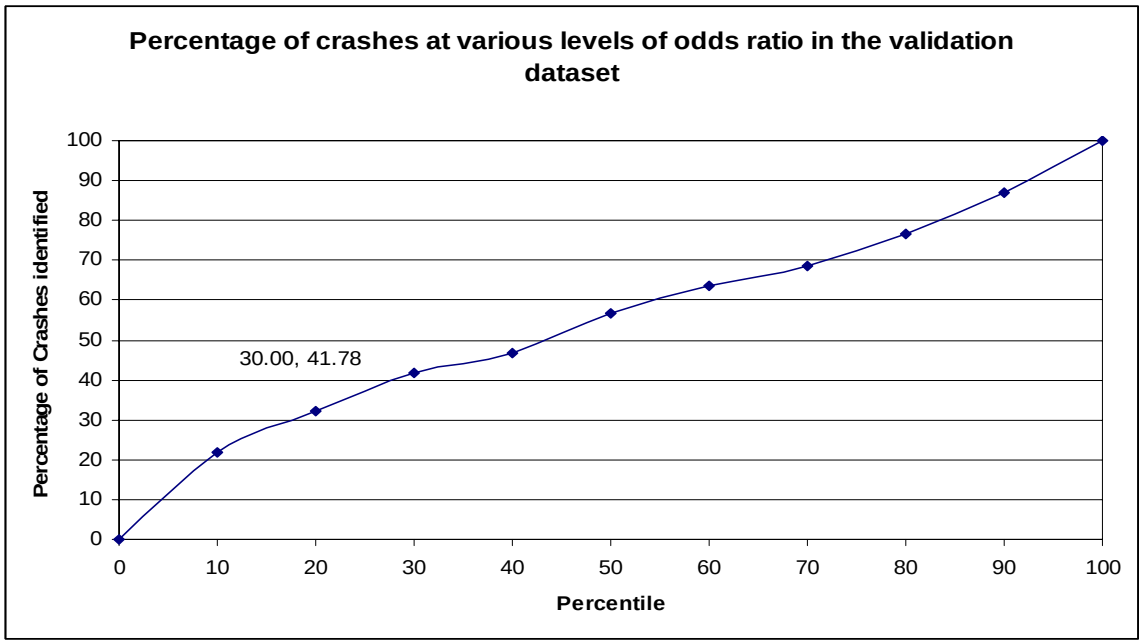

Figure 6-1: Percentage of captured response lift plot for matched case-control sampling based logistic regression model for regime 2 rear-end crashes

It should be noted that the matched case-control model, albeit low on classification accuracy, provides interesting interpretation. The coefficient is positive for the average speed measured at the station location approximately 1-mile upstream of the crash site (i.e., ASD2) while coefficients are negative and almost equal in magnitude for average

speeds measured downstream (ASG2 and ASH2). Note that the model is based on the standard stepwise selection procedure and only identifies the most significant variable having linear relationship with the target. These findings could be useful in developing proactive crash prevention strategies based on variable speed limits.

The main aim of this study is to develop a system for reliable crash identification and the performance of the logistic regression model is not good enough in this regard. In the next section we would explore the PNN based classifiers for identification of regime 2 rear-end crashes. The classification performance of the PNN would be compared to the logistic regression model shown above as well as the MLP/NRBF models developed in the previous chapter.

6.3 Probabilistic Neural Network (PNN) based Classification

6.3.1 A brief review of the methodology

The PNN is a neural network implementation of the well-established multivariate Bayesian classifier, using Parzen estimators to construct the probability density functions for competing classes (Specht, 1996). The principal advantage of using PNN is that it does not require multiple presentations of training data. The training data points once presented to the network get stored in the pattern layer. Hence, the training process for the PNN is much faster than that for MLP or NRBF networks explored earlier in this study.

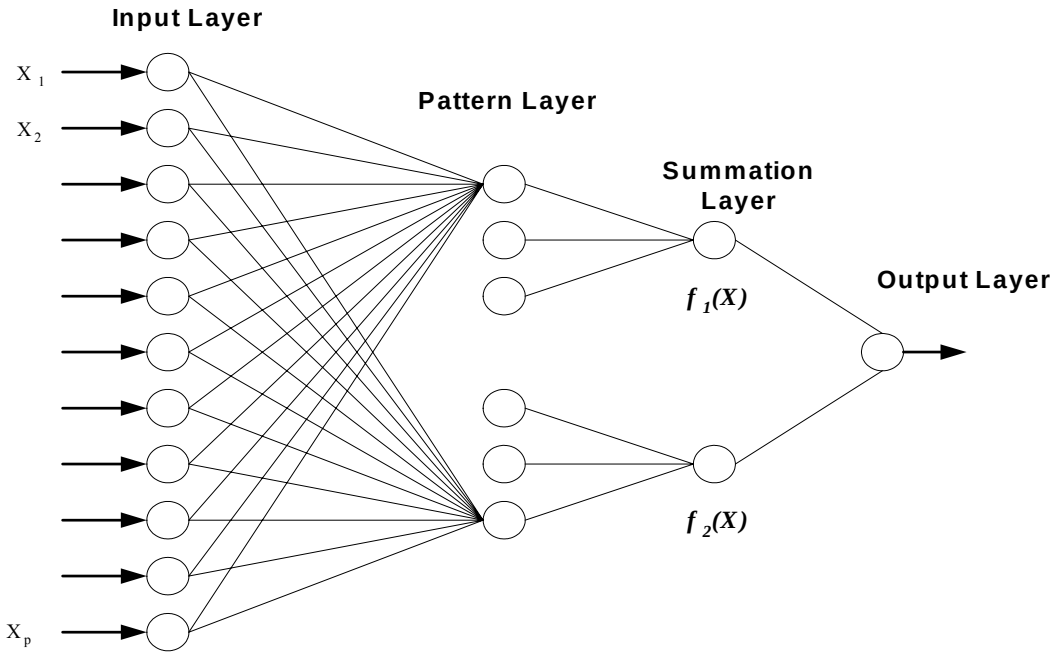

Figure 6-2: The PNN architecture for a two-class classification problem

The network shown in Figure 6-2 depicts *p* dimensional inputs to be classified into two classes. The pattern layer contains one neuron for each training case while the summation layer has one neuron for each class. In the creation (training) phase of the PNN each training case (patterns with known classification) is stored in a neuron of the pattern layer. To classify an unknown input pattern, the execution starts by simultaneously presenting this input vector to all pattern layer neurons. Each pattern neuron then computes a Euclidean distance measure between the input and the training case represented by that neuron. It then subjects the distance measure to neuron's potential function. The following layer contains summation units with a modest task. Each summation layer neuron is dedicated to a single class. It just sums up the pattern layer neurons corresponding to the members of that summation neuron's class. The attained

activation of the summation neuron is the estimated density function value for that population class in vicinity of the unknown input pattern up for classification (Masters, 1995).

To achieve binary classification through the PNN the neuron in the output layer may be used as a threshold discriminator. In fact the PNN architecture used in our previous study (Pande, 2003; Abdel-Aty and Pande, 2005) provided binary output as crash or non-crash through the neuron in the output layer. Note that using output layer neuron as threshold discriminator is equivalent of using posterior probability of 0.5 as the threshold. It essentially means classifying all observations above 0.5 as crash and below 0.5 as non-crash. It, however, reflects the performance of the model at a predetermined threshold on output posterior probability. As argued in the previous chapter, due to inherent imbalance (between crash and non-crash cases) in the training and validation sample, classification based on a predetermined threshold is inappropriate. A continuous measure of performance evaluation is needed, instead. Therefore, it was decided to compare the models using the cumulative percentage of captured response lift plot over validation dataset. To accomplish this, the function of output layer neuron must be modified.

The PNN architecture used in this study is modified to provide a measure of attained activation by the summation neuron for crashes. To estimate it the transfer function (from summation layer to output layer) was changed from the original comparison function to the modified "softmax" function. The "softmax" function essentially normalizes the attained activation of the summation layer neuron for the "crash" class between 0 through

1. The closer this modified output for an observation is to unity the more likely it is for that observation to be a crash. This output is analogous to the posterior probability output obtained through the MLP and NRBF neural network models in the previous chapter. With this output we can compare the performance of the PNN models with the models developed in the previous chapter.

6.3.2 Inputs to classification models

The step by step modeling procedure followed to estimate optimal PNN models was identical to the one used in the previous chapter for MLP/NRBF models. The first set of models were developed using real-time traffic parameters only from station of the crash along with the offline factors (such as the presence of the ramps, mile-post location etc.). In the subsequent steps traffic parameters from three (i.e., station E, F, and G) and five (i.e., Stations D through H) were included as potential independent variables. The first task was to decide on which of the parameters (traffic and location characteristics) should be used as inputs to the PNN models. Since the PNN is essentially a neural network based classifier, it is appropriate to use the variables identified through classification tree based algorithm in the previous chapter.

For the three sets of models (with traffic parameters from one, three, or five stations) variables used as input to the PNN models are shown in Table 6-4. Note that the parameters shown for the three sets of models are not the same as the parameters shown in Tables 5-22 through 5-24 in Chapter 5. The list of variables in those tables were the results of variable selection algorithm on a dataset with regime 2 crashes and completely

random non-crash data which did include small percentage of observations belonging to regime 1. As mentioned earlier, according to the proposed application strategy all regime 1 observations would be classified as crash and therefore we are working with crashes as well as non-crash data only belonging to regime 2. The variables (shown in the Table 6-4) were used as inputs to models developed in section 5.6.2 of the previous chapter. Models in that section were developed after excluding observations belonging to regime 1 from the training and validation dataset, which also happens to be the case here.

Table 6-4: List of variables used as inputs to the 1-station, 3-station and 5-station PNN models for identification of regime 2 rear-end crashes

List of traffic factors selected through tree model with		
Traffic parameters only from station F	**Traffic parameters from stations E, F, and G**	**Traffic parameters from stations D, E, F, G, and H**
ASF2, AVF2, SOF2, AOF2	*ASG2, ASF2, AOF2, AVF2, SSG2, SOG2*	*ASG2 ,ASF2, ASH2, AOF2, SOG2*
List of off-line factors selected through tree model with		
Traffic parameters only from station F	**Traffic parameters from stations E, F, and G**	**Traffic parameters from stations D, E, F, G, and H**
DOWNSTREAMOFF =0 if nearest downstream off-ramp is located further than 0.0638 miles =1 if nearest downstream off-ramp is located within 0.0638 miles *DOWNSTREAMON* =0 if nearest downstream on-ramp is located further than 0.7747 miles =1 if nearest downstream on-ramp is located within 0.7747 miles *BASE_MILPOST* =0 if 0<base_milepost<=11.93 =1 if 11.93<base_milepost<=25.43 =2 if 25.43<base_milepost<=35.18 =3 if 35.18<base_milepost<=36.25 *STATIONF* =0 if Loop detector station nearest to crash location is located upstream =1 if Loop detector station nearest to crash location is located downstream *CRASHTIME* =0 if Time of crash between midnight to 12:26 AM =1 if Time of crash between 12:26 AM to 6:46 AM =2 if Time of crash between 6:46 AM to 7:24 PM =3 if Time of crash between 7:24 PM to midnight	*DOWNSTREAMON* =0 if nearest downstream on-ramp is located further than 0.7747 miles =1 if nearest downstream on-ramp is located within 0.7747 miles *DOWNSTREAMOFF* =0 if nearest downstream off-ramp is located further than 0.0638 miles =1 if nearest downstream off-ramp is located within 0.0638 miles *CRASHTIME* =0 if Time of crash between midnight to 12:26 AM =1 if Time of crash between 12:26 AM to 6:46 AM =2 if Time of crash between 6:46 AM to 7:24 PM =3 if Time of crash between 7:24 PM to midnight *UPSTREAMOFF* =0 if nearest upstream off-ramp is located further than 0.3205 miles =1 if nearest upstream off-ramp is located within 0.3205 miles *BASE_MILPOST* =0 if 0<base_milepost<=11.93 =1 if 11.93<base_milepost<=25.43 =2 if 25.43<base_milepost<=35.18 =3 if 35.18<base_milepost<=36.25	*CRASHTIME* =0 if Time of crash between midnight to 12:26 AM =1 if Time of crash between 12:26 AM to 6:46 AM =2 if Time of crash between 6:46 AM to 7:24 PM =3 if Time of crash between 7:24 PM to midnight *DOWNSTREAMON* =0 if nearest downstream on-ramp is located further than 0.7747 miles =1 if nearest downstream on-ramp is located within 0.7747 miles *UPSTREAMOFF* =0 if nearest upstream off-ramp is located further than 0.3205 miles =1 if nearest upstream off-ramp is located within 0.3205 miles *BASE_MILPOST* =0 if 0<base_milepost<=11.93 =1 if 11.93<base_milepost<=25.43 =2 if 25.43<base_milepost<=35.18 =3 if 35.11<base_milepost<=36.25

6.3.3 Calibration of PNN models

A critical problem while 'training' the PNN was that observations belonging to crash category were only 15% of the sample used for training the models. To appropriately design the PNN a balanced training dataset is required. One idea was to randomly select non-crash data points equal to the 15% crashes out of the complete random non-crash sample and use them for training along with the crash data points. The problem with this approach was that we would loose key contribution from a lot of available non-crash data points. Hence it was decided to reduce the observations belonging to non-crash cases by means of a clustering procedure. Subtractive clustering procedure was used in order to reduce 2568 non-crash observations into 426 cluster centers which was the number of regime 2 rear-end crashes in the training dataset. The clustering procedure essentially involves identifying an appropriate cluster radius such that 426 cluster centers are selected to represent all observations belonging to the region within that particular radius. It should be noted, however, that the non-crash data points in the validation dataset were not clustered and were used as is. The application of subtractive clustering procedure at the training stage of PNN was proposed earlier by Pande (2003) and was subsequently used by Abdel-Aty and Pande (2005).

The most critical issue while creating a MLP/NRBF network based classifiers was to estimate the optimal number of neurons in the hidden layer. This is not the case with the PNN. In a PNN, the network decides the number of hidden nodes automatically and it's the same as the number of training patterns. As explained in Chapter 3 the critical parameter for the PNN is σ representing the spread value. For small values of the spread

parameter the PNN reduces to nearest neighbor classifier with each individual case exerting too much influence on the performance of the network. Higher values of σ cause the PNN to loose the details of density functions being estimated. The range examined to search for the optimal spread parameter was from 0.001 through 0.1 with an increment of 0.001. It essentially means that 100 PNN models, all with different value of the spread parameters, were estimated and the validation dataset was scored using these models. Hence, within each of the three sets (including traffic parameters from 1, 3 or 5 stations) 100 PNN models were estimated with varying values of the spread parameter.

The performance of 100 models within each set was evaluated based on criterion similar to the one used for neural networks presented in the previous chapter. As mentioned earlier, the output for the PNN models was also in the form of a posterior probability (and not in the binary form). One could examine the percentage of crash cases in the validation sample within first three deciles of this output. The spread parameter value yielding the highest percentage of crashes in the top 30% observations was selected as the optimal value. Using this criterion the models with optimal value of the spread parameter σ were identified in each of the three sets (i.e., the PNNs with input traffic parameters from 1, 3 or 5 stations). The percentage of crashes identified within first three deciles of the output posterior probability along with the optimal spread parameter is provided in Table 6-5. Note that the Table also shows the performance of the optimal individual model for each of the three sets between the other two neural network architectures (MLP and NRBF). The performance of optimal PNN models is comparable to the best of the models from other two neural network architectures.

Table 6-5: Percentage of regime 2 rear-end crashes captured within first three deciles of output posterior probability through the best models within different neural network architectures

		Percentage of crashes identified and spread parameter to obtain maximum crash identification within 30 percentile of "posterior probability" PNN Model	The optimal model between the other two neural network architectures along with the number of hidden neurons (From Chapter 5 Table 5-26)
Traffic Parameters from	Station F	49.21% ($\sigma = 0.041$)	50.06% (NRBF: 4 hidden nodes)
	Station E, F, and G	52.90% ($\sigma = 0.060$)	53.71% (NRBF: 4 hidden nodes)
	Station D, E, F, G, and H	53.20% ($\sigma = 0.083$)	51.05% (MLP: 8 hidden nodes)

In the next step all combinations (i.e., the hybrid models) of the best PNN models (shown in Table 6-5) were created by averaging the posterior probabilities estimates by the individual models for each observation in the validation dataset. It was found that the hybrid models do improve upon the performance provided by individual models.

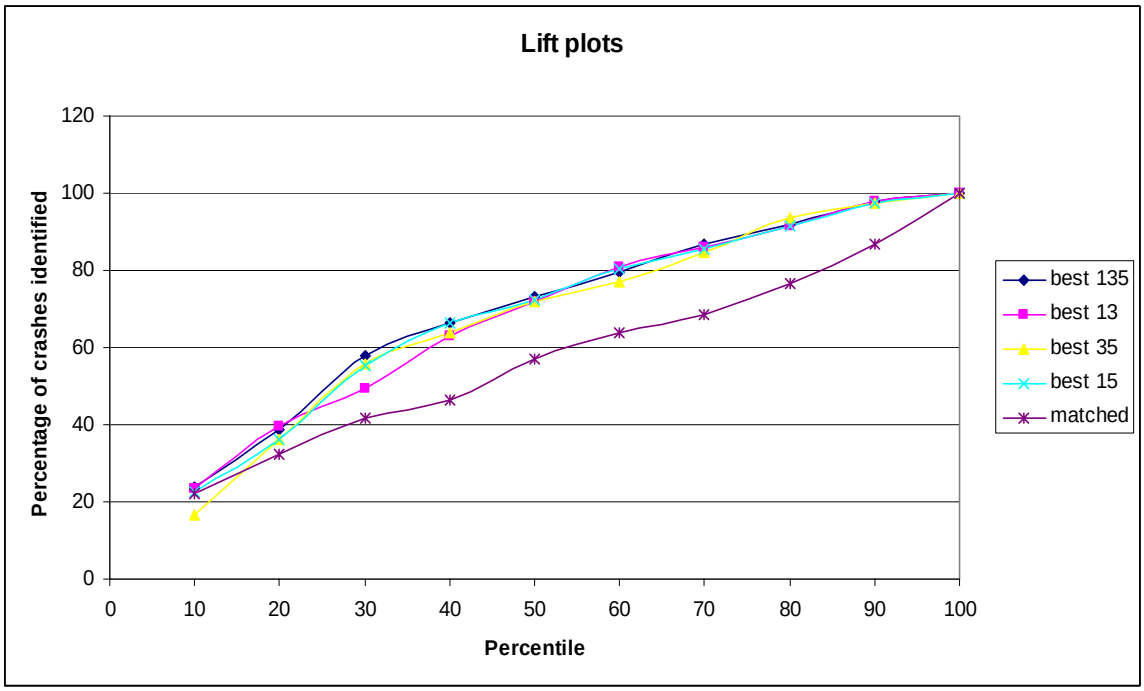

Figure 6-3: Percentage of captured response lift plot for combination of best models for regime 2 rear-end crashes chosen from the three sets

The lift plots depicting the performances of all possible hybrid models are shown in Figure 6-3. The curve shown as *best135* (combination of three, i.e., best 1-station, 3-station and 5-station PNN models from Table 6-5) runs higher than other lift curves in the vicinity of 30 percentile region. It is slightly above the curve belonging to the hybrid model titled *best35* (representing the combination of best 3-station and 5-station PNN models), *best15* (representing the combination of best 1-station and 5-station PNN models) and *best13* (representing the combination of best 1-station and 3-station PNN models). At 30th percentile the combination of the three models (i.e., best135) captures maximum percentage of crashes (57.89% of the crashes from the validation sample) and is, therefore, recommended for identification of regime 2 rear-end crashes in the PNN category. The percentage of crashes identified by the hybrid PNN model is higher, albeit

comparable, than the best hybrid neural network model proposed in the previous chapter that identified 55.4% crashes in the same dataset. Note that performance of the matched logistic regression model at various deciles of odds ratio is also shown in Figure 6-3. Except for the first 10 percentile its performances is significantly lower than the PNN model combinations.

6.4 Relationship between Outputs from Best Models in each Category

In this study three distinct modeling approaches are explored for identification of regime 2 rear-end crashes. The three modeling approaches include data mining based neural network architectures (i.e., the MLP and NRBF), matched case control logistic regression, and probabilistic neural network (PNN). Performance of the models in each category is evaluated based on its output for the observations from the same validation dataset. A proposed contribution of this study includes combining results from different models to achieve reliable real-time identification of crashes. In this regard it would be interesting to examine how the outputs of the best models from each modeling technique correlate with each other.

A perfect (or very high) positive correlation would indicate that the output from the models convey the same information and combining these models would not yield any additional benefit in terms of crash identification. However, we do expect the output from these models to have significant positive correlation with each other since outputs form the models (be it odds ratio or posterior probability) are supposed to be higher for crashes and lower for non-crash cases. Table 6-6 shows the estimates of the correlation

coefficients between outputs of different models for the observations in the validation dataset. Note that we could only examine this correlation matrix since we used the same training and validation dataset throughout the modeling procedure irrespective of the modeling technique.

Table 6-6: Correlation between the odds ratio (output from logistic regression model) and posterior probability (output from best PNN and NRBF/MLP hybrid model) for the validation dataset observations

	Posterior probability PNN (Hybrid of best 1-station, 3-station and 5-station models)	Posterior probability MLP/NRBF(Hybrid of best 1-station, 3-station and 5-station models)	Odds ratio Matched case-control logistic regression
Posterior probability PNN (Hybrid of best 1-station, 3-station and 5-station models)	1	0.63355 <.0001	0.15704 <.0001
Posterior probability MLP/NRBF (Hybrid of best 1-station, 3-station and 5-station models)	0.63355 <.0001	1	0.31202 <.0001
Odds Ratio Matched case-control logistic regression	0.15704 <.0001	0.31202 <.0001	1

It may be observed from Table 6-6 that the correlation between output of the hybrid PNN and hybrid MLP/NRBF models was higher than their correlations with odds ratio form the matched logistic regression model. Hybrid PNN model output has the least correlation (coefficient estimate) with the odds ratio. It indicates that it might be worth while to combine the two models in a real-time application as they might capture different crashes and their combination would yield higher crash identification.

The idea, however, had to be dropped as it was noticed that 41.8% crashes identified within 30 percentile of odds ratio were a subset of the 57.89% identified through the hybrid PNN models. Based on these observations it was decided the only hybrid PNN and hybrid MLP/NRBF models would be used later in the real-time system to classify the observations into crash vs. non-crash.

6.5 Conclusions

The procedure to efficiently identify rear-end crashes belonging to regime 1 was proposed in Chapter 5. While models were developed for identification of regime 2 crashes as well, there was sufficient scope of improvement. In this regard rear-end crashes and non-crash data belonging to regime 2 traffic conditions were analyzed in this chapter.

First a logistic regression model was estimated for these crashes based on the within stratum matched study design. The model clearly showed the speed difference between upstream and downstream of crash location to be "responsible" for regime 2 rear-end crashes. However, based on the classification performance of the model it was argued that it is not ideal for accurate real-time identification of crashes. It was further verified by the fact that the crashes identified through this model were a subset of the crashes identified by the other modeling approach (i.e., the PNN) explored in this chapter. Due to its poor classification performance over first 30 percentile of odds ratio it had limited application for crash identification which happens to the main aim of this research. The findings from the model may be utilized, however, for devising variable speed limit strategies to reduce the differential in average speeds and calm the conditions prone to

regime 2 rear-end crashes. Individual PNN models along with the hybrid models developed by combining them identified slightly higher percentage of crashes in the validation set than the hybrid models developed by combing MLP/NRBF models from the previous chapter. Hence the hybrid models from this chapter as well as from the last chapter will be recalled while demonstrating the application of multiple models in the form of a system identifying crash prone conditions on the freeway.

As we have seen in the preliminary analysis that although frequency of other types of crashes is significantly less than the rear-end collisions, their number is by no means insignificant. In the next chapter we would shift our focus from rear-end crashes and develop models for identification of lane-change related crashes that happen to be the second most frequent type of crash on the study area corridor.

CHAPTER 7

ANALYSIS OF LANE CHANGE RELATED CRASHES

7.1 General

As mentioned earlier, most of the existing work on proactive real-time traffic management has been generic in nature. Majority (up to or more than 50%) of crashes on freeways are rear-end collisions. Thus, the real-time traffic parameters identified as indicative of a potential crash through these generic models can by in large be associated with rear-end crashes. It is consistent with the fact that some of the factors identified in this research for rear-end crashes (Chapter 5) were actually also found significant in the generic models developed in our previous studies (Abdel-Aty et al., 2004).

A possible reason for the generic nature of the models could be that often times the database assembled for modeling purposes is not large enough for a disaggregate analysis by type of crash (e.g., Oh et al., 2001 etc.). A sufficiently large database comprising crashes over 5-year period from the 36.25-mile Interstate-4 corridor has been assembled for this study. The rear-end crashes make up about 51% of the crashes in this database. Other crashes such as sideswipe, angle or collision with a guard rail, etc. make up between 0 to 11.0 % of the crash sample. Crashes most commonly observed after rear-ends may be categorized as sideswipe, angle and single vehicle crashes (which include a variety of collisions involving guard rails, parked vehicles and road-side barriers), respectively.

In this research we have so far focused on rear-end crashes which are the single most frequent type of collisions on freeways. However, since there are substantial numbers of other crashes (such as collisions related to lane-changing maneuvers) as well, there is a need to incorporate models identifying conditions prone to such crashes in the intended real-time crash prediction system. In particular, with 11% share sideswipes are the second most common type of collisions on Interstate-4 corridor under consideration. In this chapter, loop data corresponding to historical sideswipe crashes are analyzed in order to develop the sideswipe component of the real-time crash 'prediction' system. The models from this component can be applied in parallel to the models for the rear-end crashes and remedial measures depending upon the output from the two components may be applied to calm the conditions and avoid crashes. For example, a temporary "no lane-changing" sign can be used to reduce the probability of an impending sideswipe crash.

7.2 Crash Data Description

According to the database maintained by the FDOT there were 4189 crashes reported on the Interstate-4 corridor under consideration over the five year period (1999 through 2003). However, out of these, only 3124 had any loop data available. Among these, about 11% were identified as sideswipe while 10% of them were classified as angle crashes. Based on a study by Wang and Knipling (1994) it could be safely assumed that the crashes classified as sideswipe crashes occurred when one vehicle intentionally changes lane and sideswipes or is sideswiped by a vehicle in the adjacent lane. This postulation was verified by examining the actual reports filed by enforcement officers on the scene of

these crashes. Among the angle crashes, those on the inner through lanes (the center and left lane) of the freeway were hypothesized to be lane changing related because of the rare interaction of the vehicles on these lanes with the vehicles approaching from other directions. A closer examination of reports for angle crashes led to the conclusion that these crashes on the center and left through lanes, although reported as angle crashes, in fact show more resemblance to sideswipe crashes in their mechanism and can be associated with lane changing (Lee et al., 2006). Hence, the crashes that are intended to be identified by the models developed in this chapter include crashes that can be attributed to lane changing, i.e., all sideswipe crashes and the angle crashes on center and left lane. These crashes make up about 16% of the 3124 crashes with some corresponding loop data available and are referred to as lane-change crashes in this study.

Again, the loop detector data surrounding crash location would be used as a (surrogate) measure of traffic conditions along with the geometric design of the freeway to identify the conditions prone to lane-change related crashes. Data mining based modeling techniques along with matched case-control logistic regression are applied in this chapter towards that objective. Variables included for the rear-end crashes would be explored as potential inputs besides the measures of traffic flow variation across the three through lanes. The reason for including measures of across lane variation is that the interaction between flows in individual lanes might affect the lane changing behavior of drivers as well as the risk involved in lane changing maneuvers. Both these factors can potentially affect the occurrence of lane-change related crashes.

7.3 Sampling Issues

In the analysis presented in this chapter traffic flow parameters from all three lanes would be required to deduce the input variables. Hence, crash and non-crash cases with data from only one/two of the three lanes can not be used for the lane-change related crashes. Therefore, the data requirements for this analysis are less liberal than those for the rear-end crashes since for rear-end crashes we could use observations (crash and non-crash) with loops on at least one lane functioning.

The non-crash data used for the analysis is selected from a sample of 150000 random non-crash cases. The process of generating these 150000 cases was described in Chapter 4. Out of these 150000 random non-crash cases, a non-crash sample may be drawn depending on the data requirements of the methodology used for analysis. As expected, the histogram distribution of variables time, station and direction over these 150000 cases appeared to be uniform. Out of these 150000 cases, 95922 cases had partial loop data available. It was noticed that the distribution of these cases over all stations and time was also uniform. Samples drawn from these 95922 cases were used in the analysis of the rear-end crashes. Since all freeway locations were uniformly represented we could analyze the variable such as the milepost location, time of day and presence of ramp as independent variables for the rear-end crashes. However, the same sample can not be used for the lane change related crashes since the partial loop data would not be sufficient to calculate the flow ratios and reliable estimates of variation of speed/flow/occupancy across the three lanes.

One possible way to deal with the missing values at three lanes would be to make use of the information of the two lanes if available to impute the information on the third unavailable lane. The relationships used to get the information (speed/volume/occupancy) for the third lane would depend on the presence of ramps and geometric characteristics of the freeway etc. Therefore relationship at any particular location (i.e., station) can not be used at any other location (station). It means the stations at which we have complete data available for a very few cases would still not have sufficient information to develop imputation models. Besides, it would not be possible to impute the information for two lanes where data from only one lane is available. Moreover, the time and effort needed to implement the imputation procedure would be beyond the scope of the current project. A weighted sampling procedure among the available 'complete' cases is adopted instead.

Out of the 95922 cases we extracted 47693 cases which had loop data available from all three lanes. However, it was noticed that unlike the cases with partial data, the cases with 'complete' data were not uniformly distributed at all stations. In fact some stations had no cases with data from three lanes. The histogram distribution of various loop detector stations over these non-crash cases is provided in Figure 7-1.

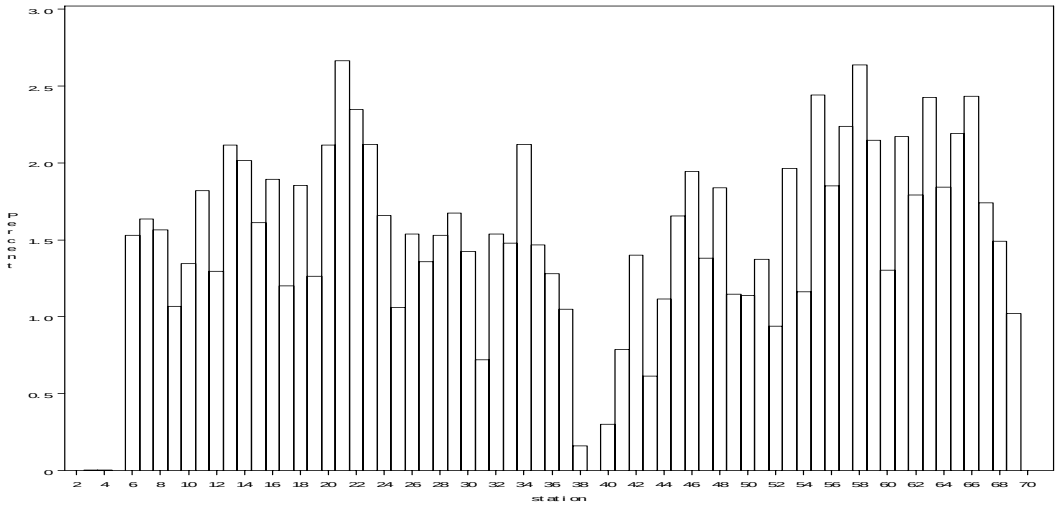

Figure 7-1: Distribution of stations over all cases with complete lane by lane data available

It is clear from Figure 7-1 that the non-crash distribution is not uniform and some stations such as station 38 and 40 have almost no observations with complete lane by lane data. Moreover, stations on the periphery of the study area (stations 2 through 6 and station 69 through 71) also have no lane by lane data available. Hence it was decided to limit the scope of the system used to predict lane-change related crashes between stations 6 and 69. Moreover, stations 38 to 41 were also excluded because of their failures in reporting data from all three lanes. Hence, the analysis was limited to freeway sections in the vicinity of 59 stations rather than all 69 stations that were part of the original study area.

Table 7-1: Frequency of non-crashes at various stations and its comparison with the frequency as per uniform random distribution

Station	Original frequency of non-crashes with complete data	Proportion as per uniform distribution	Frequency as per uniform distribution	Frequency in the boosted sample	Ratio to get the final random sample from the boosted sample
6	180	0.0163	376	360	1
7	342	0.0168	389	342	1
8	399	0.0165	381	399	1
9	344	0.0175	405	344	1
10	270	0.0179	413	540	0.765
11	386	0.016	370	386	1
12	358	0.0173	401	358	1
13	378	0.0175	404	378	1
14	532	0.0183	424	532	0.796
15	431	0.017	393	431	0.912
16	422	0.0174	402	422	1
17	356	0.0175	405	356	1
18	321	0.0177	408	642	0.636
19	401	0.0162	374	401	1
20	420	0.0164	379	420	1
21	560	0.0178	411	560	0.734
22	582	0.0171	395	582	0.679
23	443	0.0162	375	443	0.847
24	389	0.0154	356	389	1
25	359	0.0176	406	359	1
26	305	0.0167	386	610	0.634
27	310	0.0148	341	620	0.55
28	324	0.0162	374	648	0.576
29	381	0.0174	401	381	1
30	394	0.0165	382	394	1
31	290	0.0168	388	580	0.669
32	193	0.0172	396	386	1
33	351	0.0152	352	351	1
34	501	0.0168	388	501	0.775
35	500	0.0182	421	500	0.841
36	434	0.0164	378	434	0.871
37	409	0.016	371	409	1
42	397	0.0174	402	397	1
43	258	0.0162	375	516	0.727
44	305	0.017	394	610	0.645
45	424	0.0183	423	424	1
46	371	0.0173	399	371	1
47	386	0.0173	401	386	1

Station	Original frequency of non-crashes with complete data	Proportion as per uniform distribution	Frequency as per uniform distribution	Frequency in the boosted sample	Ratio to get the final random sample from the boosted sample
48	329	0.0171	395	329	1
49	357	0.0183	423	357	1
50	298	0.0177	409	596	0.686
51	218	0.0168	388	436	1
52	190	0.0171	395	380	1
53	386	0.0163	377	386	1
54	349	0.0171	395	349	1
55	429	0.0182	421	429	1
56	526	0.0174	402	526	0.764
57	517	0.017	392	517	0.758
58	569	0.0171	394	569	0.693
59	531	0.0165	381	531	0.718
60	327	0.017	394	654	0.602
61	289	0.0164	379	578	0.655
62	460	0.0162	375	460	0.816
63	500	0.0174	401	500	0.802
64	547	0.0174	403	547	0.737
65	506	0.0166	383	506	0.757
66	504	0.0173	399	504	0.791
67	502	0.017	392	502	0.782
68	372	0.0162	375	372	1

To clarify the differences between the data requirements/availability for the rear-end and lane-change crashes it can be said that if one considers the failure of loop on any one lane among the three then the pattern of failure is not random and some stations would have more representation in such data than the others. It is the case here since we need data from all three lanes to do a meaningful analysis for lane-changing related crashes. However, if we consider a loop station as failed only when none of the three lanes are reporting data then the failure patterns may be considered random. It was the case when we sampled random non-crash cases (with partial loop data) to develop models for the rear-end crashes.

To analyze the effect of location characteristics along with the traffic data measured at loop detectors on crashes the non-crash sample should be distributed similar to the original distribution of the random non-crash cases (i.e., it should be uniform as was for 150000 cases). To get a uniform distribution of all stations weighted sampling of 'complete' cases is required. It means over sampling of observations from stations which have less data available. However, some of the stations had none of the cases with data from three lanes available and hence these stations (Stations 2 through 6, 38 through 41 and 69 through 71) had to be dropped from the analysis.

The frequency of the eastbound stations among the 'complete' cases is provided in Table 7-1. The table also shows the proportion of the cases belonging to these stations in the uniform random sample of 95922 cases which was used to sample non-crash cases to model rear-end crashes. The third column in the table shows the expected frequency according to these proportions. Note that the numbers of cases available for the stations belonging to the rows highlighted in red are less than what it should be according to a uniform distribution, while the rows highlighted in green have the frequency higher than what is should be according to the random distribution. The cells highlighted in yellow have frequency among the 'complete' cases very close to the frequency as per a uniform distribution. A "boosted" sample was created in which we duplicated the records for the stations belonging to red highlighted cell, which made the frequency of cases at these stations either more or comparable to the frequency as per a uniform distribution. From this "boosted" sample we selected all cases for stations which had frequency comparable to the uniform distribution. For the stations which had more cases we randomly selected

the required sample. The ratio of frequency in the "boosted sample" vs. the frequency in the final random sample is provided in the last column. Note that the ratio is less than or equal to 1.

The distribution now was uniform over all stations and a sufficiently large sample randomly drawn from this sample would appropriately represent all freeway locations. Since we introduced a systematic bias in the random non-crash data we needed to introduce the same bias in the crash sample as well. The premise of this sampling exercise is that since this weighted sampling changed the distribution of 'complete' non-crash sample into uniform random (which was the distribution of original non-crash sample) one could use the same weights on crash data to change the distribution of the crashes to whatever their original distribution was without taking loop data availability into consideration.

For the crashes the stations corresponding to the cells highlighted in red were first duplicated to boost their sample to account for more loop failures at these stations. Note that these are the same stations for which we duplicated the non-crash cases in the previous step. Then the ratio which was used for the non-crash cases for each station to get final sample from the boosted sample was used again in order to get the final crash sample. Note that the forth column of Table 7-2 is identical to the last column of the Table 7-1. Using this ratio final crash sample was determined and was used along with the non-crash sample for analysis. We have a sample which is balanced in terms of

freeway location and it is now possible to analyze geometric characteristics of the freeway along with the loop data on crash occurrence.

Table 7-2: Frequency of crashes at various stations and its comparison with the frequency as per uniform random distribution

Station	Original frequency of crashes with complete data	Frequency in the boosted sample	Ratio to get the final sample from the boosted sample	Final crash sample
6	1	2	1	2
7	0	0	1	0
8	2	2	1	2
9	1	1	1	1
10	2	4	0.765	3
11	3	3	1	3
12	0	0	1	0
13	1	1	1	1
14	0	0	0.796	0
15	2	2	0.912	2
16	1	1	1	1
17	0	0	1	0
18	1	2	0.636	1
19	0	0	1	0
20	0	0	1	0
21	0	0	0.734	0
22	8	8	0.679	5
23	5	5	0.847	4
24	3	3	1	3
25	0	0	1	0
26	0	0	0.634	0
27	0	0	0.55	0
28	4	8	0.576	5
29	2	2	1	2
30	4	4	1	4
31	2	4	0.669	3
32	3	6	1	6
33	9	9	1	9
34	2	2	0.775	2
35	4	4	0.841	3
36	4	4	0.871	3
37	3	3	1	3
42	1	1	1	1
43	2	4	0.727	3
44	2	4	0.645	3
45	1	1	1	1
46	3	3	1	3

Station	Original frequency of crashes with complete data	Frequency in the boosted sample	Ratio to get the final sample from the boosted sample	Final crash sample
47	1	1	1	1
48	0	0	1	0
49	2	2	1	2
50	2	4	0.686	3
51	1	2	1	2
52	3	6	1	6
53	1	1	1	1
54	1	1	1	1
55	0	0	1	0
56	3	3	0.764	2
57	2	2	0.758	2
58	1	1	0.693	1
59	2	2	0.718	1
60	2	4	0.602	2
61	0	0	0.655	0
62	1	1	0.816	1
63	0	0	0.802	0
64	2	2	0.737	1
65	0	0	0.757	0
66	0	0	0.791	0
67	2	2	0.782	2
68	1	1	1	1

There was another way of sampling the data without duplicating observations; i.e., to include all observations from the stations with least observations and randomly pick these many observations from other stations. For example, in the eastbound direction pick all 180 observations from station 6 which has the least observations among the 'complete' case data and then randomly pick 180 observations from other stations to make a uniform non-crash sample. In which case we would be selecting approximately 1/3 observations from station 22 which happens to be the most heavily represented station in the eastbound direction. Note that the proportion we use to sample the non-crash cases from a station we have to apply the same proportions for the crash cases as well. In which case station 22 E that has eight crashes with complete loop data would have only three crashes remaining for analysis. It would result in removing a lot of crash cases with all data

available. Hence we adopted this alternative sampling strategy so that not a lot of crash data are thrown out at random. Note that this weighted sampling procedure was applied separately for the eastbound and westbound directions since loop detector failure in one direction has no bearing on the other direction. Only Eastbound cases are shown here for illustration.

The sampling procedure adopted here indeed makes sense only if the more failure of certain loops is the only reason of the under-representation of some stations and the distribution of the data from each station over time of day is uniform. In fact we did examine the distributions of all available crashes by time of day and they were found to be uniform leading to inference that only failure of certain detector stations is the contributing cause of non-uniformity of distribution.

However, some potential problems associated with the sampling procedure should still be acknowledged. One of them is that for duplicated crashes the loop data is assumed to be identical to the available crashes which creates (perfect) correlation between some observations. Although this concern is somewhat alleviated by the fact that 189 of the 219 crashes in the final sample were unique.

Another concern is that the stations which have zero crashes with complete data available would still have zero crashes after boosting the sample since cases are duplicated using a multiplication factor. For example in the eastbound direction there are three such stations (Station 26, 27, and 61) out of which Station 61 has no lane change related crashes at all

over the five year period. For Eastbound and Westbound directions we calculated the proportions of crashes from each station in the final sample which would be used for variable selection analysis in the following section. These proportions were compared with the proportional representation of these stations in the original lane-change related crash sample (the later proportions were based on actual frequency of crashes without taking loop data availability into consideration). It was found that at 95% confidence level the matched proportion test indicated no difference between the two samples which addresses some of the concerns raised about the sampling procedure.

In the next section preliminary analysis of crash and non-crash cases sampled in this section is presented. Given the composition of the sample created for this analysis it may be used to make reliable inferences about contribution of location characteristics (off-line factors) on lane-change related crashes.

7.4 Preliminary Analysis

Besides data requirements there are some more aspects in which the analysis of lane-change crashes differs from that of the rear-end crashes. For example, the rear-end crashes being related to queue, required analysis of data from a series of stations upstream and downstream of crash location. For lane-change crashes we are more interested in traffic conditions at or very near to the crash location. Therefore, using the variable "*stationf*" generated based on the location of station of the crash with respect to the crash (or assigned non-crash) location we determined the stations located immediate

upstream and downstream of crash location and only data from those two stations is used as inputs to the models for lane-change related crashes.

In the sample prepared in the previous section there were 219 crashes and 44000 non-crash cases. Note that this sample was obtained through weighted sampling procedure with the goal of retaining the population structure for crashes as well as non-crashes as it was without taking complete loop data availability into consideration. Crashes on this dataset were over sampled to include all cases available and not loose out on any information available prior to historical crashes. As for the non-crash cases an appropriate size sample could be drawn from 44000 cases. Note that a very small sample might result in non-uniform distribution of time of day, location etc. over the non-crash cases. It was decided to use 5% of crash and 95% non-crash cases in the sample used for analysis. To get 95% portion of this dataset we sampled 4380 non-crash cases from the 44000 available cases. The modeling procedure adopted was similar to the one used for either of the two groups (regime 1 and regime 2) of rear-end crashes.

First transformations were applied to some critical off-line factors; such as "*base_milepost*" (representing mile post location of the crash and non-crash cases), distances of the nearest on and off ramp in the upstream and downstream directions from crash location, namely, "*upstreamon*", "*upstreamoff*", "*downstreamon*", and "*downstreamoff*" along with "*timeofcrash*". In their original continuous form these variables were not suitable for real-time crash prediction system aimed in this research because their value would change continuously through the freeway corridor. These

variables were transformed into ordinal variables having maximum association with the binary target variable. The data was then portioned into training and validation dataset before being subjected to the tree model to perform variable selection for subsequent neural network models. A standard 70:30 split was used to obtain training and validation datasets, respectively. Note that a stratified random sampling with binary target variable *y* as the stratification variable was used to partition the data, so that 5:95 crash vs. non-crash ratio is maintained in both training and validation datasets.

The frequency distributions of the six transformed ordinal variables with respect to the binary target variable (*y=0* for non-crash and *y=1* for crash) are provided in Tables 7-3 through 7-8. Note that along the rows on the first column these tables depict the range of continuous variables that constitute the optimal bins. The two subsequent columns show the frequency (and row percentage) of crash and random non-crash cases, respectively, in the bin represented by corresponding row. The sample used for this analysis consists of 219 (≈ 5%) crashes and 4380 (≈ 95%) non-crash cases. Therefore, the bins with greater than 5% crash cases may be considered more crash prone while the bins with less than 5% may be considered relatively safer.

It may be seen from Table 7-3 that based on occurrence of lane change crashes the corridor is divided into four segments with cutoff points located at milepost 12.013, 27.05 and 31.43 miles. Note that the region with mileposts between 12.013 and 27.05 miles is has the maximum row percentage for the crash cases at 6.73% which is only marginally higher than the crash percentage (5%) used for the sample. Note that for a group of rear-

end crashes row frequency was as high as 31% in the 10-mile stretch located in the downtown Orlando area. It indicates that the risk of having a lane change crash is somewhat less associated with the corridor location as compared to the rear-end crashes. The number of vehicles, i.e., the exposure, which is supposed to higher in the downtown Orlando area has little or no impact on lane change crashes.

Table 7-3: Frequency table of the variable created through optimal binning transformation of "base_milpost" for crash (lane-change crashes) and non-crash cases

Optimal binning of "*base_milepost*" with respect to target variable	y		Total
	0 (non-crash Cases)	1(crash cases)	
Minimum - 12.013	1204 96.63	42 3.37	1246 (100)
12.013-27.059	1996 93.27	144 6.73	2140 (100)
27.059-31.431	597 96.14	24 3.86	621 (100)
31.431-high	583 98.48	9 1.52	592 (100)
Total	4380 (95)	219 (5)	4599 (100)

Table 7-4 provides similar information for "*timeofcrash*"; which was divided into two optimal bins (categories) that were created with cut-off point at 34019 (9:27 AM). The table indicates that the period between midnight to 9:27 AM there are fewer lane-change related crashes as compared to the later period. However, the highest row percentage is only 5.65% indicating that time of the day also might not be significantly associated with lane-change crashes.

Table 7-4: Frequency table of the variable created through optimal binning transformation of "time of crash" for crash (lane-change crashes) and non-crash cases

Optimal binning of "*time of crash*" (expressed in terms of seconds past midnight) with respect to target variable	Y		Total
	0 (non-crash case)	1(crash case)	
Minimum – 34019 (midnight to 9:27 AM)	1557 96.89	50 3.11	1607 (100)
34019 – Maximum (9:27 AM to midnight)	2823 94.35	169 5.65	2992 (100)
Total	4380 (95)	219 (5)	4599 (100)

Next off-line factors to be transformed were the location of on and off ramps. The location of these ramps with respect to the location at which crash risk is being assessed could potentially be critical. For example, an off-ramp located upstream of a freeway location would effect the odds of crash occurrence in a different way than an on-ramp located downstream. For every crash and non-crash case the distances of nearest on and off ramp in upstream and downstream direction are available from the geometric design database created for this study (See Chapter 4 for details). These continuous variables were named "*downstreamon*", "*downstreamoff*" "*upstreamon*" and "*upstreamoff*".

Four ordinal variables with two levels each were created by transforming these variables. These newly created variables are shown in Tables 7-5 through 7-8 along with crash and non-crash frequencies in the resulting categories. The distances of nearest ramps (of both types in both directions) are essentially divided based on a threshold value. This threshold value is obtained with the objective of maximizing the association of the resulting

categories of transformed variable with the target variable and the transformation procedure is identical to the one used in Chapter 5 for rear-end crashes.

Cut-off for the distances for categorization of nearest downstream on-ramp and upstream off-ramp, respectively, are 0.5192 and 0.8987 (Column 1 of Tables 7-7 and 7-8). Thresholds for the downstream off-ramp and upstream on-ramp are 2.497 miles (Table 7-5) and 1.91 miles (Table 7-6), respectively. Threshold as high as 1.91 (for upstream on-ramp) or 2.497 (for downstream off-ramp) would mean that one category of the transformed variable would encompass most of the observations. Hence, the presence of an on-ramp up to 1.89 miles upstream has the same impact on occurrence of a lane-change crash as an on-ramp located 0.1, 0.2, or 1.2 mile upstream. Similarly it does not matter whether an off-ramp is located near of far (from 0 up to 2.4973 miles downstream). It essentially means that the presence of ramps belonging to these two categories has no impact on lane-change crash occurrence. It is interesting to note that the presence of an off-ramp downstream does not affect the probability of lane-change crash occurrence significantly.

For the other two categories of the ramps (i.e., the off-ramp located upstream of crash location and the on-ramp located downstream of crash location) the optimal thresholds were found to be much lower and both categories of the resulting ordinal variables encompassed significant number of observations.

To transform continuous variable "*downstreamon*" (distance of nearest on-ramp in the downstream direction) the 'optimal' threshold was estimated to be 0.5192. In the category with observations having nearest downstream on-ramp within 0 through 0.5192 miles there are almost 7% crashes; while in observations with nearest downstream on-ramp greater than 0.5192 miles there are only 3.4% crashes (Table 7-7). It indicates that sites within 0.5192 miles upstream of an on-ramp are at higher risk of a lane change crash than other freeway locations. This result could be explained by the fact that as the vehicles approach an on-ramp the congestion caused by the incoming vehicles force the vehicles to make lane-changing maneuvers in order to avoid it. It could potentially result in higher chances of a related crash.

The distance between crash (and non-crash) location and the nearest upstream off-ramp (i.e. the variable "upstreamoff") has the threshold of 0.8987 miles (Table 7-8). It indicates that for distance of approximately 0.9 miles, according the categorization obtained here, downstream of an off-ramp there is higher probability of a lane change related crash. A possible explanation for the same might be that as the drivers, who want to exit from the freeway with in next few miles, drive besides an off-ramp they might feel the need to get to the right lane so that when the appropriate ramp arrives they don't get stuck in the inner lanes. Experience of passing an off-ramp might act as an indication, a signal in a sense for the drivers to change lanes. This lane changing behavior could potentially lead to related crashes.

Table 7-5: Frequency table of the variable created through optimal binning transformation of "downstreamoff" for crash (lane-change crashes) and non-crash cases

Optimal binning of "*downstreamoff*" (distance of nearest downstream off ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0001:low-2.4973	4276 95.36	208 4.64	4484 (100)
0002:2.4973-high	104 90.43	11 9.57	115 (100)
Total	4380 (95)	219 (5)	4599 (100)

Table 7-6: Frequency table of the variable created through optimal binning transformation of "upstreamon" for crash (lane-change crashes) and non-crash cases

Optimal binning of "*upstreamon*" (distance of nearest upstream off ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0001:low-1.9087	3896 94.89	210 5.11	4106
0002:1.9087-high	484 98.17	9 1.83	493
Total	4380 (95)	219 (5)	4599 (100)

Table 7-7: Frequency table of the variable created through optimal binning transformation of "downstreamon" for crash (lane-change crashes) and non-crash cases

Optimal binning of "*downstreamon*" (distance of nearest downstream on ramp from crash location) with respect to target variable	**y**		**Total**
	0 (non-crash case)	**1 (crash case)**	
0001:low-0.5192	1653 93.07	123 6.93	1776
0002:0.5192-high	2727 96.6	96 3.4	2823
Total	4380 (95)	219 (5)	4599 (100)

Table 7-8: Frequency table of the variable created through optimal binning transformation of "upstreamoff" for crash (lane-change crashes) and non-crash cases

Optimal binning of "*upstreamoff*" (distance of nearest upstream off ramp from crash location) with respect to target variable	y		Total
	0 (non-crash case)	1 (crash case)	
0001:low-0.8987	2534 93.96	163 6.04	2697
0002:0.8987-high	1846 97.06	56 2.94	1902
Total	4380 (95)	219 (5)	4599 (100)

Following appropriate transformations of off-line factors variable selection procedure was initiated. First the data was partitioned into 70:30 training and validation datasets. The independent variables used in the study included the average and standard deviation of the speed, volume, and occupancy, during 5-10 minutes prior to the crash occurrence from two stations, immediately upstream and downstream of the crash location. The variable *"SSU2"* shown for example represents the standard deviation of 30-seconds speed observations during the 5-minute period of 5-10 minutes prior to a crash at station located upstream of the crash location. According to the nomenclature shown in Figure 7-2 the same parameter measured at downstream of crash site would be named "SSW2".

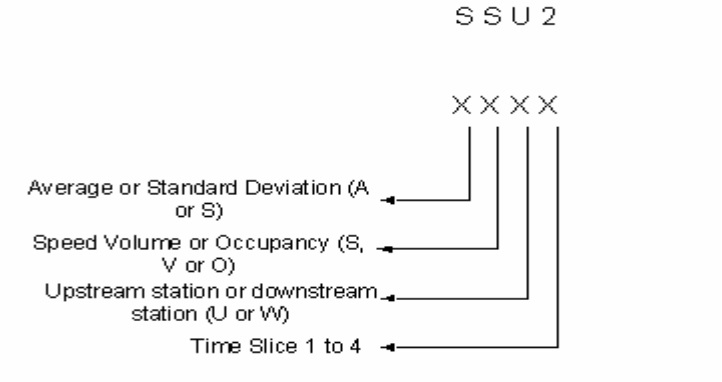

Figure 7-2: Nomenclature for the factors used for lane-change related analysis

Beside these factors flow ratios, representing a measure of intensity of lane-changing, identified by Chang and Kao (1991) and Lee et al. (2006) were also used. The reason for examining these parameters was that, these parameters with a measure for actual number of lane-changing might have significant association with the occurrence of lane-change related crashes.

The flow ratio devised by Chang and Kao (1991) was based on their field studies to identify "macroscopic" traffic factors related to lane changing behavior. Lee et al. (2006) proposed some modifications to the above flow ratios to overcome the limitations of applying this factor to investigate its effects on lane change related crashes. It was noted that work by Chang and Kao (1991) only relates the number of lane changes in specific lane to *AFR* (average flow ratios) in the corresponding lane but does not consider the total number of lane changes in all lanes in the form of overall *AFR (OAFR)*. However,

OAFR might be important in representing general traffic stability on freeways and its consequent impact on crash risk. Therefore, *AFR* calculated for each subject lane should be combined to reflect the total number of lane changes (Lee et al., 2006).

The objective of the study by Lee et al. (2006) was to be able to differentiate between rear-end and lane-change related crashes. Two measures of overall flow ratio (OAFR) based on the 5-minute average vehicle counts on three lanes of the Interstate-4 corridor were used. First, the average flow ratios for the individual lanes were defined as follows:

$$AFR_1(t) = \frac{v_2(t)}{v_1(t)} \times \left(\frac{NL_{2,1}(t)}{NL_{2,1}(t) + NL_{2,3}(t)} \right)$$

$$AFR_2(t) = \frac{v_1(t)}{v_2(t)} + \frac{v_3(t)}{v_2(t)}$$

$$AFR_3(t) = \frac{v_2(t)}{v_3(t)} \times \left(\frac{NL_{2,3}(t)}{NL_{2,1}(t) + NL_{2,3}(t)} \right) \quad (1)$$

where,

$AFR_1(t)$ = average flow ratio in lane 1 (left lane) during time interval t;

$AFR_2(t)$ = average flow ratio in lane 2 (center lane) during time interval t;

$AFR_3(t)$ = average flow ratio in lane 3 (right lane) during time interval t;

$v_1(t)$, $v_2(t)$, $v_3(t)$ = average flow in lane 1, 2 and 3, respectively, during time interval t;

$NL_{2,1}(t)$, $NL_{2,3}(t)$ = the number of lane changes from lane 2 to 1 and from lane 2 to 3, respectively, during time interval t.

In above equations, since the fractions of lane changes from lane 2 to lanes 1 and 3 are unknown in this study, they were assumed to be equal (i.e. $NL_{2,1}/(NL_{2,1}+ NL_{2,3}) = NL_{2,3}/(NL_{2,1}+ NL_{2,3}) = 0.5$). In case of *AFR* in lane 2, since there is only one way of lane change from lanes 1 and 3, there is no need to estimate the fractions of lane changes and *OAFR* (overall average flow ratio) can be calculated using the following expression:

$$OAFR(t) = \sqrt[3]{0.5\left(\frac{v_2(t)}{v_1(t)}\right)\times\left(\frac{v_1(t)+v_3(t)}{v_2(t)}\right)\times 0.5\left(\frac{v_2(t)}{v_3(t)}\right)} \tag{2}$$

Equation 2 in a more general form for an n-lane freeway may be represented as follows:

$$OAFR(t) = \sqrt[n]{AFR_1(t)\times AFR_2(t)\times\cdots\times AFR_n(t)} = \left(\prod_{i=1}^{n} AFR_i(t)\right)^{1/n} \tag{3}$$

Another way of combining the three flow ratios to obtain a measure of overall average flow ratio was proposed and the resulting OAFR was represented in the following form:

$$OAFR(t) = \frac{AFR_1(t) + AFR_2(t) + \cdots + AFR_n(t)}{n} = \frac{\left(\sum_n AFR_i(t)\right)}{n} \tag{4}$$

Note that Equations 3 and 4 represent geometric and arithmetic means, respectively, of the individual average flow ratios shown in Equation 1 (Lee et al., 2006). In this study we would be comparing both these measures for crash and random non-crash cases to examine if any of them have a significant association with the binary output.

After including these flow ratios training and validation dataset were subjected to the classification tree based variable selection process. Variables included as potential inputs were the average and standard deviation of the speed, volume, and occupancy (SSU2, SSW2 etc.). In addition the flow ratios (represented by Equations 2 through 4) from the station located upstream of the crash location one at a time were also subjected to the selection process. None of the two overall flow ratios turned out to be significantly associated with the binary target, however. The list of significant variables identified by classification tree model employing entropy maximization splitting criterion is provided in Table 7-9. All parameters subjected to the variable selection process belong to time slice 2. Note that we did try to use parameters from other time slices (i.e., time slice3 and 4, 10-15 and 15-20 minutes prior to the crash, respectively), too. Their association with the binary target variable y, however, was not as significant as the parameters from time slice 2. Also note that none of the off-line factors, including the factors explicitly related to driver population composition (from Chapter 4) had significant VIM.

Table 7-9: Results of variable selection through the classification tree model utilizing entropy maximization criterion

Name	Variable Importance Measure (VIM)	Variable Description
AVW2	1	Average volume at station downstream of crash location
ASW2	0.9052	Average speed at station downstream of crash location
SVW2	0.9049	Standard deviation of volume at station downstream of crash location
AVU2	0.5713	Average volume at station upstream of crash location
SVU2	0.457	Standard deviation of volume at station upstream of crash location
AOU2	0.406	Average occupancy upstream of crash location
SSU2	0.3622	Standard deviation of speed at station upstream of crash location
SOU2	0.1745	Standard deviation of occupancy at station upstream of crash location

It essentially means that the factors used to create the strata (i.e., the control parameters) for matched case-control sampling (See Chapter 3 for details) are not very critical for crashes related to lane-changing. As the next step we estimated the stepwise logistic regression model for binary target 'y' based on the matched sample to verify this postulation. Note that for matched analysis we could use all crashes which had 'complete' data available and there was no need to modify the sample because in the matched sampling design non-crash data were in fact required only from locations which had crash data available. The parameters of the logistic regression model estimated using stepwise variable selection procedure are shown in Table 7-10.

It may be seen that not all the variables included in the classification tree models enter the logistic regression model. The reason why the tree model identifies more variables is its flexible selection criterion. Parameters identified as significant through the stepwise procedure (Table 7-10) are the subset of variables selected through classification tree (Table 7-9). It shows that the control variables (location, time of day, day of week etc.) in fact have little significant influence on crash occurrence.

Table 7-10: Logistic regression model resulting for backward variable selection procedure on the matched data for the lane change crashes

Variable	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
SSU2	-0.03627	0.02003	3.2785	0.0702	0.964
AVU2	0.09914	0.05048	3.8570	0.0495	1.104
ASW2	-0.02937	0.00852	11.8812	0.0006	0.971
AVW2	-0.06197	0.04952	1.5659	0.2108	0.940

From the lists of variables found significant through the classification tree (Table 7-9) and stepwise selection procedure (Table 7-10) it may be inferred that parameters corresponding to intensity of lane changing (such as the flow ratios and presence of ramps inducing the drivers to make more lane changing maneuvers) are not associated with crash occurrence. Note that the overall flow ratio represented by Equation 2 was successfully employed by Lee et al. (2006) for separating rear-end crashes from those related to lane-changing. OAFR, however, was not indicative of the risk of observing a lane-change related crash. Parameters depicting stable flow conditions across the three lanes (low variation of volume, speed, and occupancy) upstream of the crash site were

found critical in this regard. It indicates that if low temporal as well as across lane variation of traffic flow parameters is observed at a certain location it might be risky to change lanes.

In short, two critical conclusions may be drawn from this preliminary analysis; one, geometric characteristics of the freeway segments are not as significantly associated with lane-change crashes as they are with the rear-end crashes. Second, the ratio of flows measured at 5-minute level are not sufficient to separate crashes from random non-crash cases and therefore the across-lane variation of traffic parameters must be examined in more detail.

The former conclusion is very interesting since one can take the argument further to infer that a model developed using data from a segment of the freeway may be applied to freeway segments loop data belonging to which were not used at the modeling stage. It can act as the basis for excluding crashes from certain locations (for which we had to duplicate the data in the previous section e.g., Stations 26, 27 61 etc. in the Eastbound direction) at the modeling stage and still be able to assess crash risk at those locations in real-time provided loop detectors from three lanes are reporting data in the future. However, note that this is still not enough to include locations of station 2-6, 37-41 and 69-71 in the system assessing risk of lane-changing crashes since these locations were not even the part of the preliminary analysis. Therefore, it can not be said definitively that the characteristics of these locations also would not be associated with lane-change related crashes. Hence in the next step of the analysis we excluded crash and non-crash data from

stations belonging to the rows highlighted red in Table 7-1 and Table 7-2 (e.g., 26, 27, 28, 31, and 32 etc. on eastbound). Such stations on westbound direction were excluded as well from the modeling procedure. After this exclusion there were 219 crashes in the sample were reduced to 162 unique crashes.

7.4.1 Variables representing across lane variation

In the analysis presented in the previous section 5-minute standard deviation of speed, volume and occupancy (SS/SV/SOU2) were representing the across lane variation of traffic parameters at the upstream station. Note that, for example, SVU2 is the standard deviation of thirty 30-second volume observations observed from the three lanes at the station located upstream of the crash location during 5 minute period 5-10 minutes prior to a lane-change crash. Note that this parameter (SVU2) has two sources of variation, one, the difference between observations across lane and the other, the temporal variation. If the value of this parameter is low then one can conclude that across lane as well as temporal variation is down, however if this parameter is high then the across lane variation may not necessarily be high. It could be high purely because of high temporal variation in speed. Therefore, in the next section variables more precisely representing across lane variation in traffic flow parameter would be examined for their effect on lane-change related crashes. Two sets of such parameters were calculated. The first set of parameters measuring 5-minute average of between-lane variations of speed/volume/occupancy are defined in the following equation:

$$ABLVSU2 = \frac{1}{10}\sum_{i=1}^{10}\left|LS - (LS+CS+RS)/3\right| + \left|CS - (LS+CS+RS)/3\right| + \left|RS - (LS+CS+RS)/3\right|$$
$$ABLVVU2 = \frac{1}{10}\sum_{i=1}^{10}\left|LV - (LV+CV+RV)/3\right| + \left|CV - (LV+CV+RV)/3\right| + \left|RV - (LV+CV+RV)/3\right| \quad (5)$$
$$ABLVOU2 = \frac{1}{10}\sum_{i=1}^{10}\left|LO - (LO+CO+RO)/3\right| + \left|CO - (LO+CO+RO)/3\right| + \left|RO - (LO+CO+RO)/3\right|$$

LS, CS, and RS represent left, center, and right lane speed values observed every thirty seconds. First, the average of 30-second speeds over the three lanes is calculated as (LS+CS+RS)/3. The absolute value of the difference between individual lane speeds and this average is then added together which happens to be the term inside the summation in Equation 5. The parameter is then averaged over ten 30-second observations that are recorded during the five minute slice of 5-10 minutes period before the crash. These parameters shown are calculated for station located upstream of the crash location for time slice 2 as indicated by the term "U2" at the end of each parameter. The term "ABLV" represents "average between lane variations". ABLV for volume and occupancy are calculated in an identical manner. Note that this is just one way to examine the across lane variation of traffic parameters and the second set of parameters calculated to represent them is provided below:

$$ADALSU2 = \frac{1}{10}\sum_{i=1}^{10}\left|LS - CS\right| + \left|CS - RS\right|$$
$$ADALVU2 = \frac{1}{10}\sum_{i=1}^{10}\left|LV - CV\right| + \left|CV - RV\right| \quad (6)$$
$$ADALOU2 = \frac{1}{10}\sum_{i=1}^{10}\left|LO - CO\right| + \left|CO - RO\right|$$

In Equation 6 the absolute difference between speeds from adjacent lanes is added together and averaged over the five-minute slice. The term "ADAL" represents "average difference between adjacent lanes". Of course the two sets of parameters are related with each other and it was noticed that correlation coefficients between corresponding parameters from Equation 5 and 6 were in the vicinity of 0.95. Therefore, these parameters were not attempted together in the variable selection/modeling procedure and were tried one at a time.

7.5 Preliminary analysis with unique crashes

The dataset with 162 crashes and 3650 non-crash (all unique) cases was then partitioned into training (70%) and validation (30%) datasets. The datasets were subjected to classification tree based variable selection process. The variable included as potential inputs at this stage were the average and standard deviation of the speed, volume, and occupancy at the downstream station (AS/SS/AV/SV/AO/SOW2). In addition, we subjected the three sets of across-lane variation (in speed/volume/occupancy) measures at the upstream station one set at a time to the selection process. The three sets include, one, the same set of measures used in the previous section i.e., SSU2, SVU2, and SOU2 and the other two represented by Equations 5 and 6, respectively. The list of significant variables identified by classification tree models employing entropy maximization criterion for optimal split is provided in Table 7-11.

Table 7-11: Results of variable selection procedure with only unique crashes for lane-change crashes

Name	Variable Importance Measure (VIM)	Variable Description
ASW2	1.0000	Average speed at station downstream of crash location
ASU2	0.6179	Average speed at station upstream of crash location
AOW2	0.5142	Average occupancy at station downstream of crash location
ADALOU2	0.2692	Average of absolute difference between 30-second occupancy observations on adjacent lanes
SVW2	0.2591	Standard deviation of volume at station downstream of crash location
SSW2	0.2006	Standard deviation of speed at station downstream of crash location

By examining the classification tree model closely it was noticed that with high average speed downstream of crash site (ASW2) along with low average speeds upstream (ASU2) the likelihood of lane-change related crashes increases. It indicates that as drivers perceive a chance to increase speed as they travel from low average speed regime (measured at station upstream) to high average speeds (measured at station located downstream of the crash site) they might make lane-changing maneuvers increasing chances of conflictions. It was also noticed that if both upstream and downstream are operating at high speeds (around or greater than 50 mph) small average differences between adjacent lane occupancies upstream of the crash site involve more risk than the sites with this parameter (ADALOU2) being high. Hence if the difference in occupancy across adjacent lanes is lower then caution should be exercised while changing lanes.

Standard deviation of volume and speed (SVW2 and SSW2) downstream of crash site were found to be positively associated with lane-change related crashes.

7.6 Modeling and results

Following variable selection neural network based modeling procedure was initiated with variables shown in Table 7-11 as inputs. As described in the one of the previous chapters the best models were identified through the lift plot having cumulative percentage of captured response for the validation dataset on the vertical axis. The output of the neural network based classification models for any observation is termed as the posterior probability of the event (i.e., a lane-change crash in this case). Posterior probability is a number between 0 and 1. The closer it is to unity the more likely, according to the model, it is for that observation to be a crash. In a lift chart, the observations in the validation dataset are sorted from left to right by the output posterior probability from each model. The sorted group is lumped into ten deciles[3] (one decile represents 10 percentile) along the horizontal axis. The left-most decile is the *10%* of observations with highest posterior probability i.e., most likely to be a lane-change related crash. The performance of each model may be measured by determining how well the models capture the target event across various deciles. From a practical application point of view it must be understood that crashes are rare events and one would need to be parsimonious in issuing warnings for crashes. Therefore, it might be not be reasonable to assign more than 20-30% of observations as crashes and it was decided to evaluate the model performances based on

[3] Decile is defined as any of nine points that divided a distribution of ranked scores into equal intervals where each interval contains one-tenth of the scores

percentage of crashes identified within first three deciles (deciles = 10 percentiles) of posterior probability. It should be noted it (the posterior probability) is not the probability of crash occurrence at a given point in time but is a measure providing the relative likelihood of crash occurrence given the composition of the sample. That is the reason in this research we have examined the performance of the models on validation dataset based on percentiles rather than setting a specific threshold on posterior probability.

The first neural network architecture explored for classification is the multi-layer perceptron (MLP) with Levenberg-Marquardt training algorithm. The training procedure starts with an arbitrary randomly chosen set of interconnection weights and then it tries to minimize the difference between network output and the desired outputs for the training dataset. All runs have been carried out with a maximum number of epochs (a complete list presentation) of 1500, and error goal of 0.01. It has been proven in the literature that an MLP structure with one hidden layer and nonlinear activation functions for the hidden nodes can learn to approximate virtually any function to any degree of accuracy (Cybenko, 1989). The most critical issue then, was to estimate the number of neurons in the hidden layer. The underestimation of hidden neurons leads to a network having an incomplete representation of inputs and by contrast, the over representation reduces the network to a simple look-up table. The methodology adopted for selecting appropriate number of nodes in the hidden layer was to evaluate the performance of the models having hidden nodes varying from 2 through 8. To achieve this seven separate *"neural network"* nodes were used in the Enterprise Miner process flow diagram (See details in the Chapter 5).

Similar to the rear-end crashes unconstrained normalized radial basis function neural network (NRBF) were used for classification of lane change related crashes as well. To select appropriate number of nodes in the hidden layer performance of seven different NRBF networks, with hidden nodes varying from 2 through 8, was examined.

Table 7-12 depicts the performance of various NRBF and MLP neural networks having varied number of hidden neurons. It may be seen that NRBF network with three hidden neurons and MLP network with four hidden neurons provide the best crash identification in the first three deciles of posterior probability. The row corresponding to the two models are highlighted in the table.

Table 7-12: Structure and percentage of captured response within the first three deciles for best models estimated for different combination of modeling techniques (Crashes attributed to lane-changing)

Neural network architecture	Number of hidden neurons	Crash identification in first three deciles (Percentage)
NRBF	2	31.42
NRBF	3	48.00
NRBF	4	32.87
NRBF	5	44.00
NRBF	6	44.29
NRBF	7	32.00
NRBF	8	37.26
MLP	2	38.73
MLP	3	44.44
MLP	4	50.00
MLP	5	40.44
MLP	6	33.26
MLP	7	34.26
MLP	8	45.90

In the next step the two models were hybridized by averaging posterior probabilities from these two models, it was noticed that a significant improvement in crash identification was achieved through the hybrid model.

Figure 7-3 shows the lift plot for the two individual models (NRBF-3 and MLP-4) highlighted in Table 7-12 along with the model created by averaging the putput from the two models. The curve shows the percentage of the lane-change crashes in the validation dataset captured within various deciles of posterior probability by each model on y-axis. On the x-axis the percentiles are shown at equal interval of 10. Figure 7-3 also demonstrates "*performance*" of a random baseline model which represents the percentage of crashes identified in the validation sample if one randomly assigns observations as crash and non-crash. A model can be assessed by examining the separation of corresponding lift curve from the random baseline curve.

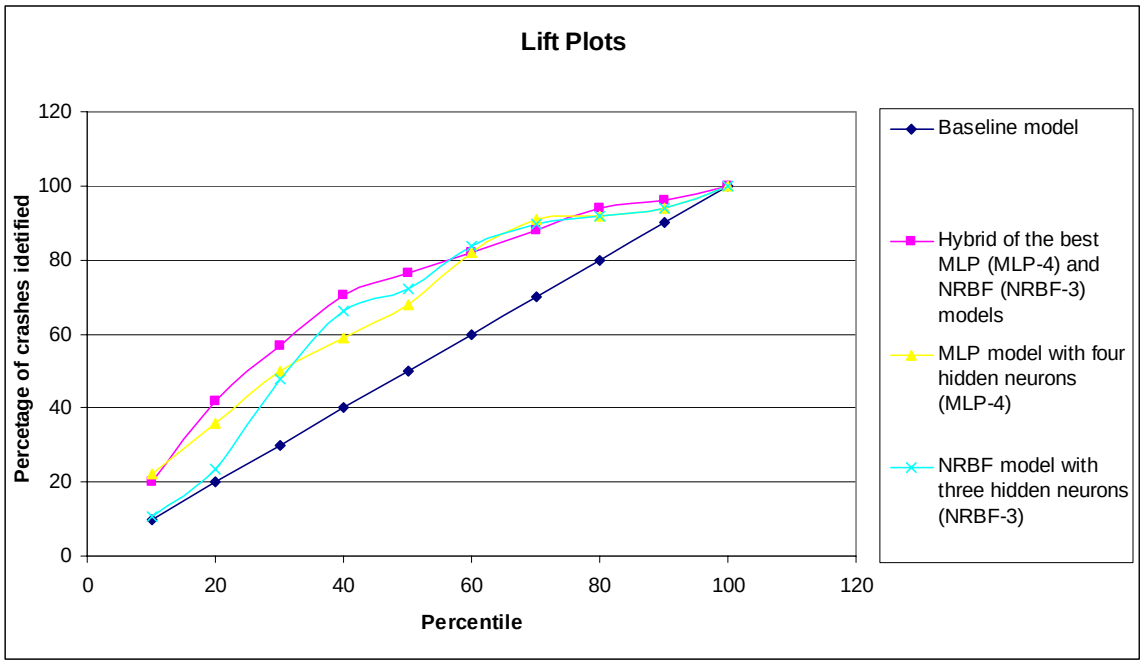

Figure 7-3: Percentage of captured response lift plot for best models belonging to different modeling techniques along with hybrid (ensemble) model

It may be seen in Figure 7-3 that the hybrid model idetifies 57% crashes from the validation dataset within first three deciles. It roughly translates into 57% classification accuracy over crash cases with 70% accuracy over non-crash cases. This model can be used to identify conditions prone to lane change crashes between any of the two loop detector stations included in the modeling procedure.

Note that the logistic regression model based on the matched case control sampling (a model analogous to the one shown in Table 7-10) is not estimated to calculate the odds ratio for classification of observations in the validation dataset. The reason for the same is that while within stratum sampling is very attractive in terms of controlling for variables

not measured in real-time, the logistic regression modeling procedure is not good at identifying variables having non-linear association with the target. In any case, since none of the off-line factors (used to form the stratum in the matched sampling) were significantly associated with lane change related crashes the advantages of matched sampling were limited. Moreover, the matched procedure did not provide classification accuracy comparable to the neural network models in the case of regime 2 rear-end crashes in the previous chapter. Hence, only neural network based models are recommended for real-time identification of lane-change related crashes.

7.7 Summary and Conclusions

This chapter presents a data mining based approach to identify potential lane-change related freeway crashes through loop detector data. Lane-change related crashes include all sideswipe crashes and angle crashes on inner lanes of the freeway that may be attributed to lane changing maneuvers.

Due to the nature of the crashes to be identified loop data from all three lanes of the freeway were required for the analysis. It was noticed if one considers a loop as failed if any one of three lanes is not reporting data; the failure patterns are not random and some locations (i.e., stations) tend to fail more often than others. To analyze the effect of freeway location characteristics (off-line factors) on lane-change related crashes the sample of non-crash cases should uniformly represent all freeway segments. A simple random sampling from the non-crash cases that have all three lanes data (from the upstream station) available would not yield such a sample. Therefore, we over-sampled

from some stations while under-sampling from others to get a uniform distribution over all stations for non-crash cases. We then used the same sampling proportions for the stations to sample the crash data as well with the assumption that this sampling would recover the underlying distribution for the lane change related crashes. Note that some stations for which three-lane loop data from upstream stations were never available had to be excluded from the analysis. Using the sampling procedures (described in detail earlier in this Chapter) we created a sample of 219 (5%) crashes and 4380 (95%) non-crash cases for analyses.

After some preliminary exploration and transformation of critical off-line factors tree node from the Enterprise Miner (SAS Institute, 2001) was used to perform variable selection. It was noticed that intensity of lane changes, measured in terms of overall flow ratios, was not significant to separate crashes from non-crashes. The factors such as presence of ramps, which might induce drivers to change lane, were also found insignificant. On the other hand, it was noticed that low variation in speed/volume across lanes under 'at capacity' flow, involves considerable risk. To verify the findings from the classification tree models we also estimated matched logistic regression model based on standard stepwise variable selection procedure. The variables included in the logistic regression model were a subset of the variable identified by the tree model with 'random' non-crash data. Since these off-line factors work as the external controls in the matched analysis, it indicates that no additional advantages would be achieved through the matched sampling based on these factors. Furthermore, it could also be argued that location specific characteristics need not be used as inputs to classification models.

Note that some of the stations with almost no complete data were removed earlier from the analyses. Based on the findings from exploratory analysis we removed some more locations that were under-represented in the random non-crash sample. Due to this modification we did not have to duplicate any crash or non-crash case in the sample. The final sample now consisted of 162 crashes. Since across lane variations (or lack there of) of traffic parameters found significant in the preliminary analysis more parameters representing differences in speed/volume/occupancy across lanes were included as inputs.

The new sample along with additional variables representing across lane variations was subjected to classification tree based variable selection. The variables found significant were average speeds upstream and downstream of the crash site. Average differences between adjacent lane occupancies upstream of the crash site (ADALOU2) along with standard deviation of volume and speed (SVW2 and SSW2) downstream were found to be positively associated with lane-change related crashes.

These variables (shown in Table 7-12) were used as inputs to classification models based on two neural network architectures (MLP and NRBF). Seven models belonging to each of the two architectures were developed by varying the number of hidden neurons from 2 through 8. It was found that the MLP model with four and NRBF model with three hidden neurons were the best individual models. The hybrid model created by combining these two models bettered the performance of individual models and identified 57% of crashes within first three deciles of posterior probability output. This model is

recommended to assess the risk of a lane-change crash between two loop detector stations on the freeway.

It should be mentioned that even though only the models using data from time slice 2 (5-10 minutes before the crash) are described in this chapter, models using data from time slice 3 and 4 were also attempted but as expected they did not achieve the performance comparable to the models described. If those models would have resulted in better or almost comparable performances they would have been prescribed as potential crash prediction models because they would allow more leverage in terms of time available to process, analyze and disseminate the information that may in turn be used to avoid crashes. Also, we suspect that the models with parameters from time slice 1 would have resulted in better crash identification. However, time slice 1 being too close to actual time of crash they can not be used in a real-time application due to practical considerations.

Multiple models for identification of rear-end and lane-change related crashes are now available to us. As the final step simultaneous application of these models will be demonstrated over loop data from a whole day of the year 2004. Since we have used loop data from 1999 through 2003 for developing the models it is appropriate to use data from the year 2004 to demonstrate the application. The final classification models recommended in this chapter as well as Chapter 6 would be recalled in the next chapter to coalesce them in the form of a reliable crash warning system.

CHAPTER 8

STRATEGY FOR REAL-TIME IMPLEMENTATION

8.1 General

In this research binary classification models have been developed using traffic and geometric data for crash and non-crash cases. For more precise identification of crash prone conditions crashes were segregated into groups based on the harmful event associated with the crash. The two major sets of models developed in this research include rear-end and lane-change related crashes. The output from the models was obtained in the form of posterior probability of observing a crash. The models were evaluated based on their classification performance over validation datasets that were set aside at the modeling stage and were not used in the training of the individual neural network models. Final models recommended for real-time classification were the hybrid models created by averaging the posterior probability output from multiple individual models (Chapter 5, 6, and 7).

Crashes are rare events, hence, instead of using the actual proportion of crash vs. non-crash cases; crashes were over-sampled in all the datasets used to develop the models. The composition (crash and non-crash ratio) of the validation dataset was not similar to the actual data on which the models would be applied in real-time. Hence, it would not have been appropriate to examine the performance of the models on validation dataset through contingency classification tables (actual vs. predicted for the two classes) based

on a specific threshold on posterior probability. Lift plots depicting percentage of crashes (in the validation dataset) identified within various percentiles of posterior probability were used instead. The advantage of this approach is that no specific threshold value has to be specified at the modeling (i.e., validation) stage. A percentile threshold was recommended instead which remains largely independent of the sample composition. For example, assume that a model identifies 60% crashes in the validation dataset by assigning 20% observations (with maximum posterior probability) as crashes. If in the future we apply this model in real-time and classify observations with posterior probability value more than the 20^{th} percentile (from the top) as crash it would translate into approximately 80% accuracy over non-crash data and identify 60% of crashes over a sufficiently long period of time. However, we still need to get the estimates for percentile(s) used as threshold for separating crashes from 'normal' conditions in real-time. These threshold estimates may be established using a 'sufficiently' large randomly selected loop detector data.

In this chapter these thresholds are estimated and then used to classify observations from one whole day of loop detector data in a virtual real-time application framework. To establish the thresholds, a random sample of loop detector data from the five year period has been used while data from February 6, 2004 have been used for a virtual real-time application. Best models for individual groups of crashes developed earlier in this study would be summarized in the next section. Then these models will be applied on a large dataset consisting of randomly selected representative loop data to establish percentile distributions for output of various models. Before applying these models along with

estimated thresholds to demonstrate their application on 'virtual' real-time loop data, we would briefly discuss the issue of how single vehicle crashes fit into this crash 'prediction' framework.

8.2 Summary of Classification Models

In this research, models for two major groups of crashes are developed; namely, rear-end and sideswipe crashes. These two groups constitute 51% and 16% of all reported crashes on the I-4 corridor under consideration. The details of the modeling procedure for rear-end crashes was described in Chapters 5 and 6 while classification models for lane-change related crashes were developed in Chapter 7. The models for these two groups of crashes are summarized in this section.

8.2.1 Rear-end crashes

Rear-end crashes may be grouped into two equally frequent groups (cluster) based on prevailing traffic speed configurations within the 2-mile stretch of the freeway around the crash location (Chapter 5). These groups may be identified through traffic speeds observed 5-10 minutes prior to crashes. The classification tree based rules for separating one cluster from the other were summarized in Table 5-1. If we apply these rules to randomly selected non-crash data it was noticed that only 6% of observations from a random dataset fall into the definition of regime 1 rear-end crashes. It may be recalled from Chapter 5 that regime 1 traffic conditions are associated with congestion and intermittently forming ephemeral queues over 1-2 mile section of the freeway (Table 5-

1). Due to the rarity of traffic patterns belonging to regime 1 one can classify every observation belonging to regime 1 as a rear-end crash without any further analysis. Remaining 94% of the loop detector data patterns on the freeway belong to regime 2 traffic conditions. Neural network based classification models were developed for separating regime 2 rear-end crashes from 'normal' (i.e., not crash prone) traffic belonging to regime 2.

First, two neural network architectures, namely, MLP and NRBF were calibrated with repeated presentation of the training dataset. Three sets of neural network models were developed. In the first set traffic parameters only from the station closest to the crash location were used and in the two subsequent sets parameters from three and five stations around crash location were used as potential inputs.

Based on their performance over validation dataset NRBF neural networks with four hidden neurons were found be the best models among the first two sets. MLP neural network with eight hidden neurons was found to be the optimal when traffic parameters from five stations were used as inputs.

These best models (MLP/NRBF) from the three sets were then hybridized by averaging their output posterior probability for individual observations. It was noticed that the hybrid model created by combining three models (i.e., with the best models in each of the three sets as its constituents) provided the best performance over the validation dataset.

As mentioned earlier, the best performance of this model for regime 2 crashes doesn't make it an automatic choice for field implementation because this hybrid model would require data from five stations to be available simultaneously for assessing the risk of regime 2 rear-end crashes in real-time. Due to intermittent loop failures these data might not be available and therefore one might have to rely on the hybrid model that uses output from the best models in 1-station and 3-station category. Similarly if data from three stations are not available one might need to switch to the NRBF with four hidden neurons (the best model in the 1-station category) that uses traffic data from only the station closest to the crash location.

Note that hybrid models for regime 2 rear-end crashes were also developed using PNN architecture. The reason for separately exploring this architecture is that the calibration of the PNN is not similar to MLP/NRBF and repeated presentations of training data are not required for its calibration. Moreover, the parameter influencing the performance of the PNN is the spread parameter and not the number of neurons in the hidden layer. Three sets of models along with the hybrids of the best among them (a procedure similar to MLP/NRBF architecture) were developed using the PNN as well. The performance of the best individual PNN models was summarized in Table 6-5.

The performance of the hybrid PNN models must also be seen in the context of data requirements. As discussed in Section 6.3.3 the model created by hybridizing best 1-station, 3-station and 5-station PNNs provides optimal crash identification. However, if data from five stations are not available then the hybrid of best 1-station and 3-staiton

models or just the best 1-station model may be applied depending upon the data availability.

Before formulating a final strategy for the identification of regime 2 rear-end crashes one must also consider how the posterior probability output is obtained for a new pattern by MLP/NRBF and PNN models. MLP/NRBF neural network architectures have their connection weights calibrated during the iterative training process and therefore the estimation of posterior probability is not computationally extensive. For the PNN models the training phase, being non-iterative, is very fast. However, computing every single output posterior probability would require accessing the entire training data stored in the pattern layer. Hence estimating the crash risk through PNNs (i.e., hybrid models with multiple PNNs as its constituents) could be very time and resource consuming. Hence in a real-time application we would prefer to employ hybrids of MLP/NRBF models first, and if the patterns are declared crash prone then the location could be flagged immediately. If the models return a "non-crash" decision only then the data may be subjected to the corresponding PNN models. The complete real-time application strategy is discussed later in this chapter.

8.2.2 Classifying lane-change related crashes

Lane-change related crashes are significantly less frequent than the rear-end crashes. These crashes may be attributed to conditions in which lane changing maneuver involves more risk. Measures representing variation in traffic parameters (speed, volume, and occupancy) across lanes were required as inputs to account for such traffic conditions. Therefore, for these models, loop data from all three lanes from the station located

upstream of the crash location were required. Also, based on the preliminary analyses it was concluded that the freeway location characteristics are not critically associated with crashes attributed to lane-changing. Therefore, only traffic parameters from stations located upstream and downstream were used as potential inputs to the classification models.

Among the neural network models it was noticed that the MLP model with four hidden neurons and NRBF model with three hidden neurons were the best individual models. The two models were hybridized to improve their performance and indeed the combined model resulted in much better performance than either of the individual models. It identified 57% crashes from the validation dataset within 30 percentile of posterior probability (refer Chapter 7 for details).

In the case of rear-end crashes differences between the percentage of crashes identified through MLP/NRBF models and the PNN models was marginal (1-2%). For more frequent rear-end crashes even such a small improvement would result in a significant number of additional crashes identified. In the case of lane-change crashes a comparable improvement in terms of percentage would result in identification of very few additional crashes. Therefore, the marginal benefit of PNN models would be very limited, more so, due to their computationally exhaustive scoring algorithm.

It is worth repeating here that crash and non-crash cases only with data from all three lanes available at upstream stations were used in the analysis. Hence, in a real-time application the risk of lane-change crashes may only be assessed for locations where loop detectors are reporting data from all three lanes. Also, note that some locations, stations in the vicinity of which very rarely had historical data available from all three lanes, had to be excluded from the analysis (Chapter 7; Section 7.3).

8.3 Threshold Estimates for the Models

The performance of the models over validation dataset was evaluated based on the percentage of crashes having posterior probability more than its 30^{th} percentile values. For classification of traffic patterns in real-time an estimate of these 30^{th} percentile values is required so that it may be used as threshold to separate crash prone and 'normal' conditions. These thresholds may be estimated based on historical traffic data. To estimate these thresholds we used the sample of 150000 observations extracted randomly from the 36.25-mile corridor of Interstate-4 over the 5-year period.

8.3.1 Threshold for lane-change crashes

In the sample with 150000 random patterns loop data were not available for all cases. Only Slightly more than 96000 cases had partial loop data available. The number of cases with loop data from all three lanes available was 47963. To establish the threshold for the lane-change related crashes only 47963 cases could be used. Out of these cases data belonging to stations 2 through 6, 38 through 41, and 69 through 71 were also excluded

because these locations were not part of the input data used to estimate classification models for lane-change related crashes (Refer Chapter 7; Section 7-3). The hybrid model having MLP with four hidden neurons and NRBF with three hidden neurons as its constituents was used to score the random loop data.

Table 8-1: Distributions of the percentiles of output posterior probabilities obtained by the hybrid model for lane-change crashes over random loop data and all lane-change crashes

Percentile	Value of Posterior Probability over random data	Value of Posterior Probability over all lane-change related crashes
100 (Minimum)	0.02114	0.025574
90	0.028311	0.0355
80	0.031657	0.039745
70	0.034494	0.043697
60	0.037574	0.046417
50 (Median)	0.04106	0.049883
40	0.045027	0.052767
30	0.048779	0.057179
20	0.052327	0.070378
10	0.057307	0.10301
0 (Maximum)	0.30155	0.24214

It may be noticed from Table 8-1 that the estimated posterior probability by the model over randomly selected cases varies from 0 through 0.30155. The 30^{th} percentile value (from the top) is 0.048779; hence if posterior probability of lane-change crash is assessed to be greater than 0.048779 for any real-time traffic pattern one could flag that particular location. As per the performance of the model on the validation dataset this threshold is expected to identify approximately 57% of lane-change related crashes. The last column in the Table above shows the percentile of posterior probability as measured over all historical lane-change related crashes from 1999 through 2003. As expected, values for all deciles are higher for the crash cases with the exception of the maximum value.

8.3.2 Threshold for rear-end crashes

A similar estimation of threshold can be done for rear-end crashes as well. However, to establish the threshold for rear-end crashes we can not subject all available random cases to the hybrid models. Note that these models are designed to separate crashes from 'normal' conditions under regime 2 traffic regime. Therefore, we must fist apply the rules shown in Table 5-1 to the data and filter cases belonging to regime 1 out of the sample. It was found that as expected a little over 6% such (regime 1) observations were filtered out of the random dataset and rest of the observations (all belonging to regime 2 as per Table 5-1) may be used for estimation of thresholds on posterior probability that may be used to classify regime 2 rear-end crashes in real-time. The thresholds were estimated for the output posterior probabilities from six different models that may potentially be employed for real-time identification of rear-end crashes (Section 8.5). Out of the six, three models were either individual MLP/NRBF neural network (if data from only station of crash were available) or their hybrids (if data from three or five stations were available). The other three were either best 1-station PNN model or the hybrids of best 1-station, 3-station and 5-station PNN models. The rationale for estimating threshold for all these models would be explained later in the chapter where the real-time application framework for the hybrid models is discussed (Section 8.5).

Table 8-2: Distributions of the percentiles of output posterior probabilities from MLP/NRBF based regime 2 rear-end crash classification models over random loop data

Percentile	Posterior Probability (Best 1-station MLP/NRBF model, i.e., NRBF with 4 hidden neurons)	Posterior Probability (Hybrid of best 1-station and best 3-station MLP/NRBF model)	Posterior Probability (Hybrid of best 1-station, best 3-station and best 5-station MLP/NRBF model)
100 (Minimum)	0.020274	0.029889	0.023445
90	0.046069	0.044248	0.038826
80	0.058259	0.051523	0.047067
70	0.07024	0.059882	0.056064
60	0.0839	0.071638	0.066359
50 (Median)	0.09936	0.084971	0.078738
40	0.12052	0.10289	0.094604
30	0.14205	0.13142	0.11658
20	0.17811	0.16588	0.15642
10	0.24781	0.25035	0.23412
0 (Maximum)	0.90054	0.74091	0.74046

Table 8-3: Distributions of the percentiles of output posterior probabilities from PNN based regime 2 rear-end crash classification models over random loop data

Percentile	Posterior Probability (Best 1-station PNN model)	Posterior Probability (Hybrid of best 1-station and best 3-station PNN model)	Posterior Probability (Hybrid of best 1-station, best 3-station and best 5-station PNN model)
100 (Minimum)	0.000016127	0.000056878	4.91E-13≈0
90	0.0038456	0.018936	1.05E-08≈0
80	0.041863	0.060698	0.00000038
70	0.10996	0.13381	0.000006363
60	0.19961	0.22769	0.000084439
50 (Median)	0.29581	0.32255	0.00108185
40	0.39099	0.41458	0.013705
30	0.47584	0.48466	0.12324
20	0.50858	0.5	0.43978
10	0.58655	0.51469	0.52792
0 (Maximum)	0.98542	0.88166	0.95964

Tables 8-2 and 8-3 show the percentile distributions of the model outputs (i.e., posterior probabilities of observing a regime 2 rear-end crash) over the random loop data for MLP/NRBF and PNN based models, respectively. Since these models were evaluated based on their 30th percentile value as threshold; the same percentile may be used as the

threshold for classifying real-time patterns if any of these models in applied to the real-time data. In the tables the row corresponding to 30th percentile posterior probability values has been highlighted. If any of the models during a real-time application estimated the posterior probability values higher than the highlighted thresholds then the concerned location may be flagged for a (regime 2) rear-end crash.

8.3.2.1 Distributions of observations with high risk of rear-end crash

Having established the thresholds for outputs of various models that would be part of the real-time implementation strategy, in this section we analyze the distribution of the observations in the random data that either belong to regime 1 or have their posterior probability for regime 2 rear-end crash greater than or equal to 0.11652 (30th percentile value for the best hybrid MLP/NRBF model; Column 4 Table 8-2).

Note that in the random database observations that are identified as belonging to regime 1 would have been classified as a rear-end crash if these data were collected in real-time. Similarly regime 2 observations, scored with the best MLP/NRBF based hybrid model, having posterior probability more than 0.11652 would have been classified as a rear-end crash. It will be interesting to compare the distribution of actual regime 1 and regime 2 rear-end crashes with the distributions of the observations identified as such in the random database. As noticed in the modeling procedure time of day and mile post location were critical for both groups of rear-end crashes. The distribution of these two parameters have been examined for actual rear-end crashes from both clusters (groups). These distributions have been compared with the distributions of these parameters over

the observations from the random dataset that would have been identified as rear-end crashes.

Table 8-4: Frequency table of regime identified by the tree model (shown in Table 5-1) for a large sample of random loop data

Traffic Regime	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6155	6.63	6155	6.63
2	86643	93.37	92798	100.00

The frequencies of the two traffic regimes in the random dataset are shown in Table 8-4. As expected the proportion of the regime 1 in the data is close to 6.5%. Since only a little more than 6.5 % traffic patterns need to be identified as crashes to correctly classify all rear-end crashes in this group (which are more than 45% of all rear-end crashes); this group of rear-end crashes is (the most) readily identifiable.

Figure 8-1 shows the distributions of mile-post location for actual regime 1 rear-end crashes (at the bottom) and observations belonging to regime 1 in the random dataset (at the top). Figure 8-2 shows the distributions of same observations over time of day (measured in terms of seconds past midnight). Note that, not surprisingly, the distributions appear almost identical.

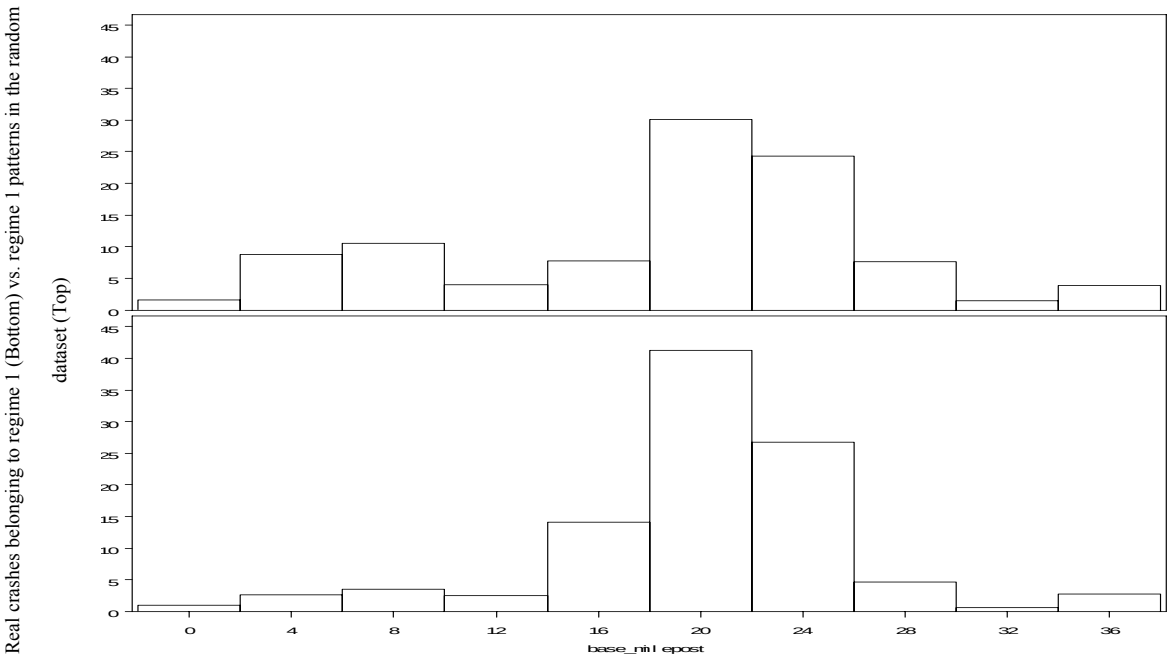

Figure 8-1: Distributions of mile-post location for real regime 1 rear-end crashes (at the bottom) and observations from random dataset belonging to regime 1 traffic conditions (on the top)

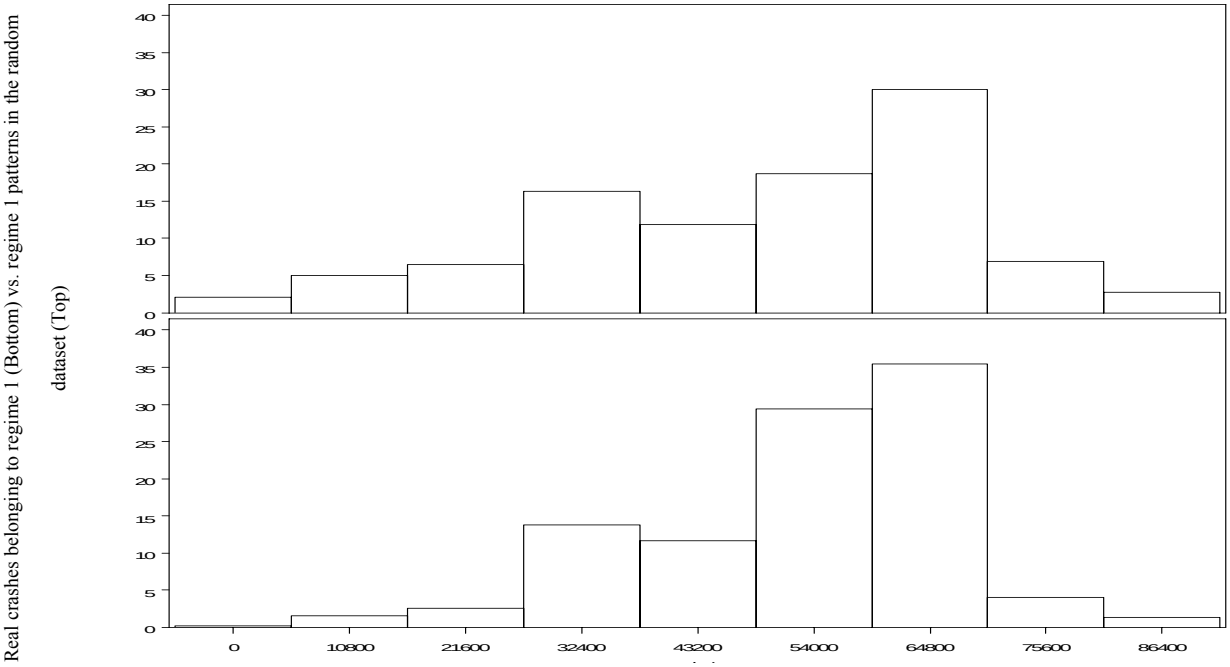

Figure 8-2: Distributions of time of crash for real regime 1 rear-end crashes (at the bottom) and observations from random dataset belonging to regime 1 traffic conditions (on the top)

While all regime 1 rear-end crashes may be identified using the classification tree rules shown in Table 5-1; for regime 2 crashes we have to rely on the output from the models created by hybridizing individual MLP/NRBF neural network models. We now examine the distributions of 30% observations in the random dataset that have maximum posterior probability of being regime 2 rear-end crashes.

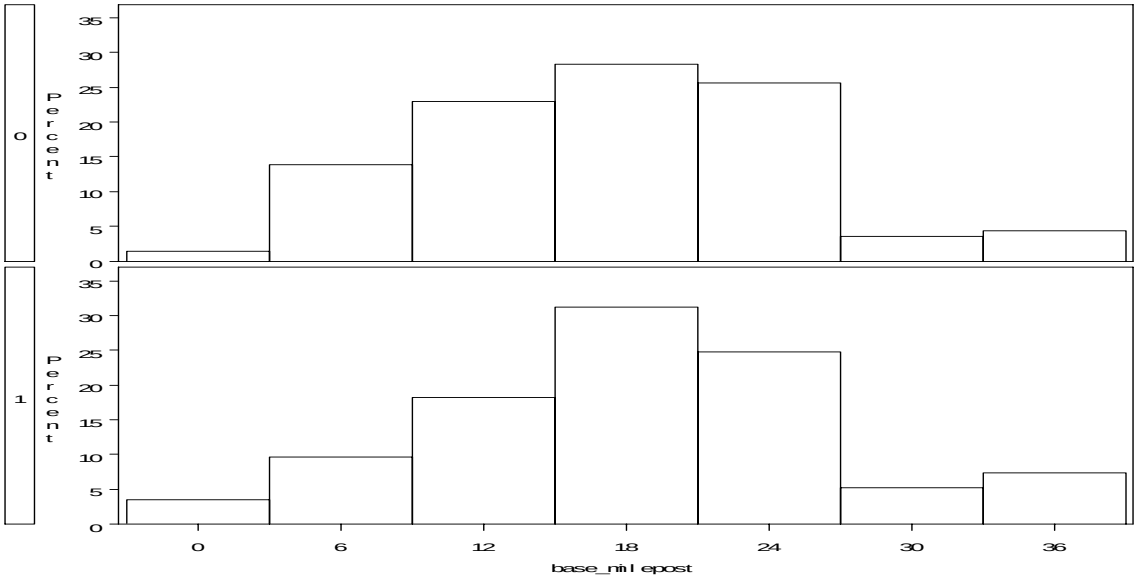

Figure 8-3: Distributions of mile-post location for real regime 2 rear-end crashes (at the bottom) and 30% observations from the random dataset with maximum risk of observing a regime 2 rear-end crash (on the top)

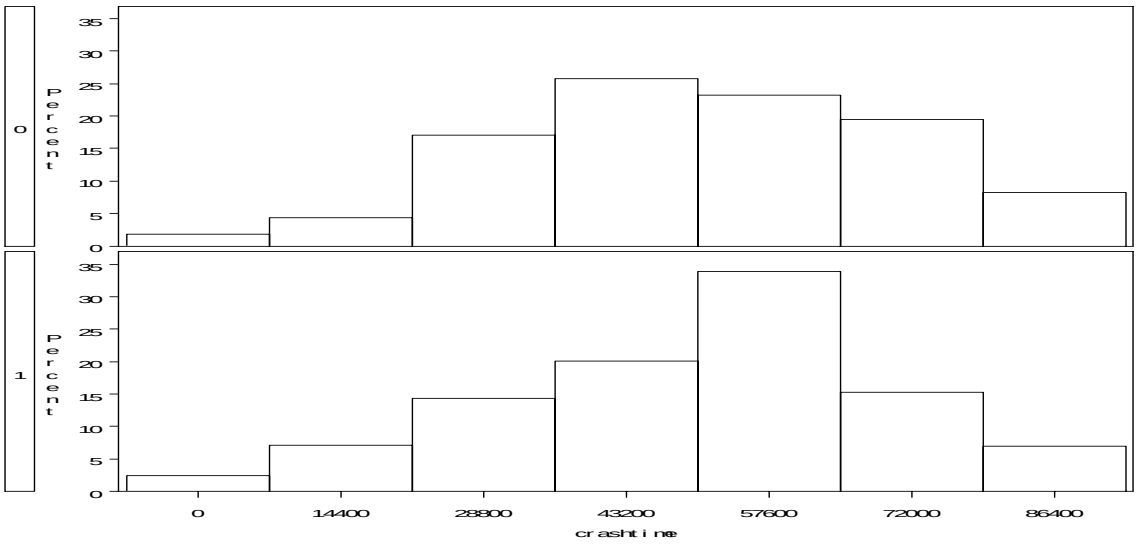

Figure 8-4: Distributions of time of the crash for real regime 2 rear-end crashes (at the bottom) and 30% observations from the random dataset with maximum risk of observing a regime 2 rear-end crash (on the top)

Figure 8-3 shows the distributions of mile-post location for actual regime 2 rear-end crashes (at the bottom) and 30% observations from the random dataset with maximum posterior probability (at the top). It may be seen that the two distributions appear to be very similar indicating that model is in fact able to distinguish locations with high risk of regime 2 rear-end crash. Figure 8-4 shows a similar distribution over time of day (measured in terms of seconds past midnight). Unlike the regime 1 crashes (Figure 8-2) the distribution over time of the day is not identical for the two histograms for regime 2 rear-end crashes (Figure 8-4); partially because the identification accuracy of regime 2 crashes is not as good as their regime 1 counterparts. Note that a similar visual comparison for lane-change crashes and randomly selected observations with high risk of such crashes would not have been meaningful since off-line factors, e.g., the time of day and mile-post location were not found to be significantly associated with lane-change crashes.

8.4 What about Single Vehicle Crashes?

The models developed in this study are for rear-end and lane-change related crashes. Single vehicle crashes, which are the next most frequent group (≈16%), are conspicuously missing from the analysis presented so far in this research. Single vehicle crashes are characterized by the involvement of only one moving vehicle and are initiated by the events such as vehicle hitting the guard rail, overturning or running off the road. Very rare types of crashes on Interstate-4 such as collision with a parked car may also be grouped into this category. Table 8-5 shows the frequency of some of the more frequent

crashes involving only one moving vehicle. Single vehicle crashes that make up at least 1% of the overall crash data are shown in the table.

Table 8-5: Frequency of single vehicle crashes by first harmful event on the I-4 corridor over the 5-year period from 1999 through 2003

Description (MV=Motor Vehicle)	Frequency	Percentage in complete crash data
18 MV Hit Guardrail	165	3.79
20 MV Hit Concrete Barrier Wall	142	3.26
31 Overturned MV	123	2.83
29 MV Ran Into Ditch/Culvert	61	1.40
17 MV Hit Utility Pole/Light Pole	44	1.01
Total	**635**	**12.93**

Among the five distinct types of crashes shown in Table 8-5; crashes other than overturned vehicles may be considered as "ran-off-the-road" type. Note that in this research we are only interested in the crashes that occur on the mainline of the freeway and not in the crashes on off and on-ramps. Ran-off-the-road crashes on the mainline of the freeway are expected to result from last minute driver action to avoid a collision with the vehicle(s) in front or on adjacent lane. This scenario is unlike the ran-off-road crashes that occur on the off-ramps due to excessive speeds.

A preliminary analysis of loop data belonging to single vehicle crashes and some randomly selected non-crash data was carried out to spot traffic patterns for single vehicle crashes. Although almost all single vehicle crashes did occur under off-peak hours of the day and outside of the downtown Orlando region of the corridor; they were not concentrated in high speed traffic regime. It was observed that average traffic speeds measured at the nearest station and (sharp) radius at the crash location affected the

occurrence of crashes involving overturned vehicles. Moreover, the fraction of overturned vehicles was significantly higher during late-night and early morning hours. Other major categories of single vehicle crashes had no real association with these average speeds and curvature. It is therefore reasonable to assume that "ran-off-the-road" type crashes are in fact resulting from evasive action on the part of the driver. If that is the case then one might expect to identify a good portion of such crashes through the models we have developed here for rear-end and lane-change related crashes.

8.4.1 Identification of single vehicle crashes through the models developed for rear-end crashes

Out of 635 crashes shown in Table 8-5; 392 had corresponding loop data partially available. If we apply the rules to identify traffic regime (regime 1 or 2) then 7% of crashes were found to be regime 1 and 93% of them were found to be regime 2. According to the strategy adopted here 7% of crashes that belong to regime 1 traffic conditions would have been identified as rear-end crashes. We next applied the hybrid of best 1-station, best 3-station and best 5-station MLP/NRBF models to single vehicle crashes (from Table 8-5) belonging to regime 2 traffic conditions. The output of this hybrid model was the posterior probability; single vehicle crashes with high posterior probability would be identified as rear-end crash by the model.

Table 8-6 shows the percentile distributions of the model outputs (i.e., posterior probabilities of observing a regime 2 rear-end crash) over random loop data in the second column. In subsequent columns percentile distributions of the posterior probability

estimates are shown for all single vehicle crashes, single vehicle rollovers and single vehicle crashes belonging to the "ran-off-the-road" category, respectively. Note that the second column in Table 8-6 is identical to the last column in Table 8-2 because they represent the output of the same model over same dataset. Note that according to the model output on this random dataset 30 percentile threshold for separating 'normal' conditions from rear-end crashes was set at 0.11658. It is clear from the last column of the table that this value is less than the 50 percentile value of the posterior probability for "ran-off-the-road" single vehicle crashes (i.e., 0.1184 from the last column of Table 8-6). Hence, we expect that more than 50 percent of such crashes to be identified as regime 2 rear-end crashes. To be precise, 53% of such crashes had their posterior probability greater than 0.1184 and they would have been identified by the hybrid model.

On the other hand, if we carefully examine the output for rollover crashes (Column 4 Table 8-6) its percentile distribution appears to be very comparable to the percentile distribution of posterior probability over the random data. In fact only 31% percent of the rollover crashes have their posterior probability greater than the threshold value of 0.11658. Hence, while the hybrid model can correctly identify potential "ran-off-the-road" crashes; it has no discriminatory power to 'predict' the rollovers.

Table 8-6: Percentiles of regime 2 rear-end crash (posterior) probability estimates (based on best MLP/NRBF based hybrid model) for random data and different categories of single vehicle crashes

Percentile	All random data from regime 2	All single vehicle crashes from to regime 2	Rollover single vehicle crashes from regime 2	Single vehicle crashes other than rollover from regime 2
100 (Minimum)	0.023445	0.028973	0.0341	0.029
90	0.038826	0.050293	0.0362	0.0517
80	0.047067	0.060232	0.0524	0.0605
70	0.056064	0.074756	0.067	0.0748
60	0.066359	0.085702	0.0764	0.0872
50 (Median)	0.078738	0.11629	0.0824	0.1184
40	0.094604	0.1628	0.1064	0.1669
30	0.11658	0.22625	0.1157	0.2298
20	0.15642	0.28443	0.1479	0.2844
10	0.23412	0.35327	0.2139	0.3463
0 (Maximum)	0.74046	0.4773	0.3477	0.4773

8.4.2 Identification of single vehicle crashes through model developed for lane-change related crashes

An evasive driver action might also result from an effort to avoid vehicle(s) in adjacent lanes. Therefore, it would be interesting to conduct an analysis similar to the one in previous section with the hybrid model for lane-change related crashes replacing the regime 2 rear-end crash model. Note that the lane-change crashes were developed without dividing the data into exclusive traffic regimes (regime 1 and regime 2). Data from all available single vehicle crashes (irrespective of their regime affiliation) were subjected to the hybrid model for lane-change crashes instead of only regime 2 single vehicle crashes that were subjected to the hybrid model for rear-end crashes.

Table 8-7: Percentile of lane-change crash posterior probability estimates for random data and different categories of single vehicle crashes

Percentile	All random data	All single vehicle crashes	Rollover single vehicle crashes	Single vehicle crashes other than rollover
100 (Minimum)	0.02114	0.0247	0.0265	0.0247
90	0.028311	0.0304	0.0269	0.0316
80	0.031657	0.0338	0.0315	0.0346
70	0.034494	0.0378	0.0352	0.0391
60	0.037574	0.0422	0.0387	0.0438
50 (Median)	0.04106	0.0463	0.0422	0.0471
40	0.045027	0.0517	0.0498	0.0517
30	0.048779	0.054	0.0534	0.0542
20	0.052327	0.0559	0.0561	0.0558
10	0.057307	0.0636	0.0598	0.0638
0 (Maximum)	0.30155	0.0984	0.0894	0.0984

Table 8-7 shows the percentile distributions of the model outputs (i.e., posterior probabilities of observing a lane-change related crash) over the random loop data in second column. In subsequent columns percentile distributions of the posterior probability estimates are shown for all single vehicle crashes, rollover related single vehicle crashes and "ran-off-the-road" single vehicle crashes, respectively. Note that the second column is the identical to the second column in Table 8-1 because they represent the output of the same model over same dataset. According to the model output on this random dataset the 30 percentile threshold for separating 'normal' conditions from lane-change related crashes was set at 0.048779. It is clear from the last column of the table that this value is only slightly more than the median posterior probability for "ran-off-the-road" single vehicle crashes (i.e., 0.0471 from the last column of Table 8-7). Hence, we

expect that slightly less than 50 percent of such crashes to be identified as lane-change related crashes. To be precise, 48% of such crashes had their posterior probability greater than 0.048779 and they would have been identified by the hybrid model.

8.4.3 Conclusions from identification of single vehicle crashes through the models developed for other types of collisions

Based on the discussion provided here it may be argued that even though single vehicle crashes were not included in the sample at modeling stage due to their diverse characteristics; a sizeable portion of them (belonging to "ran-off-the-road" category in particular) could be identified through the models developed in this research. In fact 58.82% crashes (with available data) belonging to "ran-off-the–road" category could be 'predicted' using the procedure described above. It should, however, be acknowledged that single vehicle rollovers would not be predictable by examining the loop data. Curved freeway sections during free-flow traffic regime (leading to speeding) usually experience high frequency of such crashes and these conditions may be identified as 'crash prone'. However, note that we are working with 30-second averages (which will have to be further aggregated in order to use it in a modeling framework) of the speeds at specific locations (i.e., loop detect stations) and not with speed data for individual vehicles. Therefore, it is difficult to ascertain if a fraction of vehicles passing through a curved section during the 30-second period were speeding, more so if the data are aggregated to 5-minute level. It is difficult to 'predict' rollovers crashes through the models similar to the ones developed in this research because loop detector data measured every 30-seconds from stations ½ mile apart do not provide spatial resolution equivalent to detailed

vehicle by vehicle movement data. More detailed data are required to capture human behavior and vehicle characteristics that may largely be held responsible for roll-over crashes.

8.5 Real-time Application Framework

Based on the modeling procedure and the results from classification models a framework for real-time implementation is proposed here. In the proposed framework models developed for rear-end and sideswipe crashes are applied in parallel and locations would be flagged for any type of crash independent of the flag for the other type of crash. It is therefore possible for locations to be flagged for rear-end crash or lane-change related crash or both. The framework in the form of a flow chart is shown in Figure 8-5.

For rear-end crashes the application first starts by applying classification tree model based rules shown in Table 5-1 (Chapter 5). Those rules may be used to identify whether traffic data belong to regime 1 or regime 2. If the patterns belong to regime 1 a rear-end crash warning is issued for the location without any further application. If the patterns are identified to be regime 2 then we need to apply the neural network based hybrid models. As mentioned earlier, the hybrid models that combines best 1-station, 3-staton and 5-station MLP/NRBF models provided optimal crash identification over the validation dataset and hence is preferred over other models. This model, of course, would need data from five stations around the section where we are trying to assess the crash risk. Therefore, in the next step check for data availability over five stations is applied. If data

from five stations are available then the data are subjected to the hybrid model. The posterior probability output obtained from the model is then compared with the 30 percentile threshold estimates highlighted in last column of Table 8-2. If the output is greater than the threshold values of 0.11658 then the location may be flagged for a rear-end crash. If the output is less than the threshold then one may subject the data to the hybrid model having best 1-station, 3-station and 5-station PNN models as its constituents. If the output from the PNN hybrid model is also less than the corresponding threshold (which is 0.12324 and is highlighted in the last column of Table 8-3) then the location need not be flagged for a rear-end crash. This approach is based on the need to be conservative and issue a crash warning, even if one of the two (hybrid of MLP/NRBF models and hybrid of PNN models) models finds the conditions to be crash prone. The reason why the hybrid MLP/NRBF model is preferred in the hierarchy is that the individual constituents of this hybrid model use iterative training procedures and the process of estimating an output for new pattern is fairly quick. On the other hand for PNN application major computations are carried out in the application phase. Therefore, the PNN based hybrid model is only applied when MLP/NRBF does not result in a crash warning.

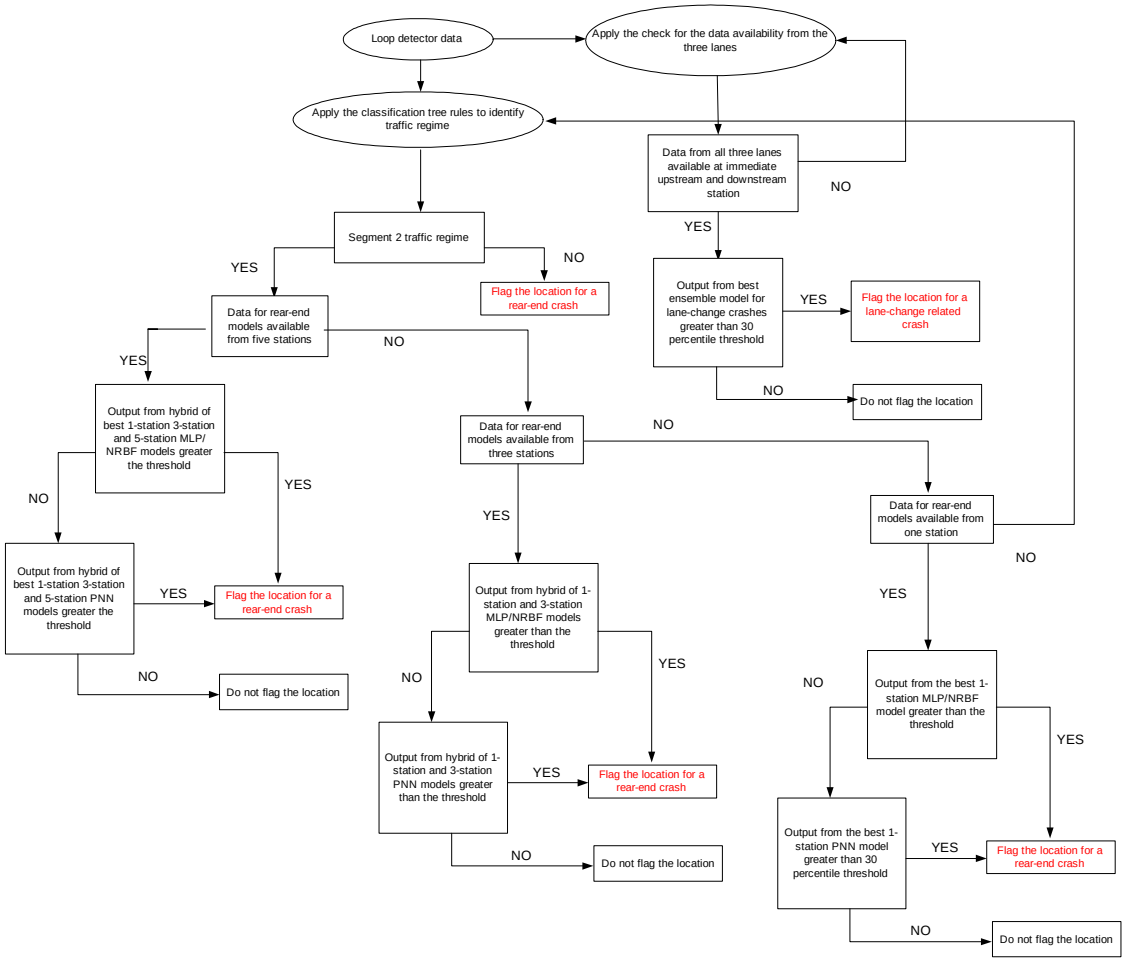

Figure 8-5: Proposed framework for real-time identification of crash prone conditions

If data from five stations are not available due to intermittent loop failures a check for data availability is applied for three stations. If data from three stations are available then respective MLP/NRBF and/or PNN hybrids (created using best individual 1-station and 3-station models) models would be applied to the real-time data. If data are not available from three stations then best individual 1-station PNN and/or MLP/NRBF models may be applied for assessing the risk of a regime 2 rear-end crash. The decision process to flag (or not to flag) the location would be identical to the one used when data from five stations were available. If data from even the nearest station are not available then it

would not possible to assess the risk of rear-end crash at that location. The thresholds for these different hybrid models, which may potentially be applied in real-time in case of missing loop data, were shown in Tables 8-2 and 8-3.

Note that the process described in the previous section is just to assess the risk of a rear-end crash. To assess the risk of a lane-change related crash first the check on the data is applied. If all three lanes at the upstream stations are functioning then the hybrid of the two of the best neural network models (NRBF with 7 hidden neurons and MLP with 3 hidden neurons) is applied to the input parameters. If the output posterior probability is greater than the threshold established by the random sample of loop detector data (0.048779; refer 30 percentile value from Table 8-1) then warning for a lane-change related crash may be issued.

Note that even though there are no specific models in the system for single vehicle crashes; a portion of such crashes, especially which are caused by evasive action of the drivers to avoid other vehicles, would be identified by the system through the flags for rear-end and/or lane-change crashes. In the next section issues pertaining to application of this framework in real-time are discussed.

8.6 Issues Relevant for Real-time Implementation

To apply the models as part of the framework discussed in the previous section, the whole corridor may be divided into sections as per Figure 8-6. The figure shows the freeway segments of approximately ¼ mile in length. The freeway is divided into

sections such that within each section the definitions of Station F (station nearest to the crash location) and parameter "*stationf*" (binary variable depicting if station F is upstream or downstream of crash location) are identical. The inputs to the neural network models (constituting the hybrid models) are function of these two parameters.

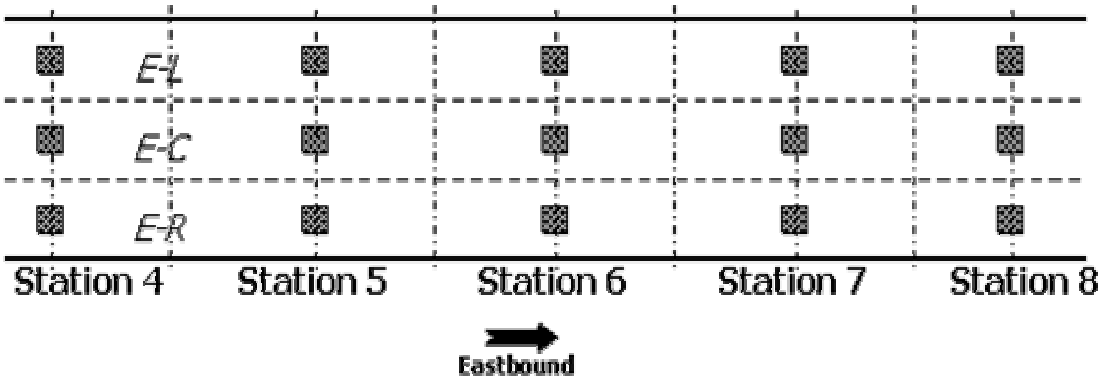

Figure 8-6: Arrangement of freeway sections with respect to real-time application of the framework proposed

The figure shows the arrangement of sections for a small portion of the whole corridor in the eastbound direction for demonstration. To assess the risk in real-time at any instant, one can use the proposed implementation framework for each of the sections shown in Figure 8-6. Note that as one moves from one segment to the other the definition of "Station F" would change which in turn would change the traffic parameters to be used as inputs to the classification models.

The second issue associated with the real-time application was that of the duration of update. Since traffic parameters from time slice 2 were used as inputs, the models assess the crash risk within next 5-10 minutes. However, it does not necessarily mean that the

update must be done every five minutes. The update may be done on a continuous basis as soon as new observations come in. For example, traffic parameters (average and standard deviation) may be calculated based on ten most recent observations available and then after 30-seconds as the latest observations (since loop data is collected every 30 seconds) come in they may be included in the calculation of traffic parameters replacing the foremost observations. This update strategy was proposed in one of our earlier study (Pande et al., 2005). In that study application of simple (one-covariate logistic regression) models for preliminary assessment of crash risk was discussed. Since multiple neural network models would be needed to implement the framework proposed here; 30-second update might not be practical in terms of the resource and processing time requirements. Therefore a 5-minute update is recommended which would give ample processing time for application of multiple models.

8.7 Demonstration of Virtual Real-time Implementation

The application of all models is demonstrated over data from Friday, February 6, 2004. According to Figure 8-6 every station would be "station of crash" for two ¼-mile freeway sections (upstream and downstream) associated with it. In all there would be 276 (69 stations* 2 directions* 2 sections) freeway sections on which the crash risk should be assessed according to the implementation plan is shown in Figure 8-5. This section provides a discussion on the continuously updated assessment of crash risk, wherever possible, for rear-end and lane-change related crashes. Four crashes were reported on February 6, 2004 on study area corridor. The rarity of crashes is signified by the fact that for four crashes there would be 79484 (276 sections * 24 hours * 12 5-minute

periods=79488 – 4) 'non-crash' patterns in the day if we update the crash risk every five minutes. The number of warning issued would of course be higher than the actual number of crashes because according to the application strategy proposed here we expect to identify more than 30% of patterns as 'crash prone'.

The logic behind issuing more warnings is that every 'non-crash' pattern does not necessarily represent 'normal' conditions. Although crashes, being rare event, might not occur after each warning; based on our analysis of historical data a sizeable proportion of crashes occurred under such traffic conditions. It is logical to assume that crash prone conditions worth issuing warning(s) to the drivers are more common than the crashes themselves. The primary purpose of demonstrating the application is to show that the thresholds estimated for the hybrid models based on random data in fact make sense, i.e., the number of warnings issued based on 30 percentile thresholds over the random data is in the vicinity of 30% over the course of a complete day.

It is worth mentioning that some patterns from February 6, 2004 could not be scored due to missing data. Moreover, the hybrid mode for lane-change related crashes can not be applied at some locations that were excluded at the time of analysis and therefore the hybrid models can not be used to assess the crash risk at those locations. In this section we summarize the performance of the system under the constraints of data availability for the day we have chosen to demonstrate the real-time application.

8.7.1 Application of the models for crashes reported on February 6, 2004

There were four crashes reported on February 6, 2004; the day chosen to demonstrate the real-time implementation plan. One out of the four crashes had absolutely no corresponding loop data available. The crash with missing data involved collision with a parked vehicle. The other three had at least partial loop data available for analysis. The details of these crashes are provided in Table 8-8.

Table 8-8: Details of the crashes reported on February 6, 2004 on the study area corridor

Crash report number	Time of crash	Direction	Station of crash	First harmful event	Mile post location	Location of Station of the crash
728266770	18:15	W	41	6 (Sideswipe)	21.194	1 (upstream of crash location)
753088170	2:05	W	30	6 (Sideswipe)	16.77	0 (downstream of crash location)
758869780	18:25	E	70	1 (rear-end)	36.063	1 (upstream of crash location)

Of course partial data are not sufficient to obtain an assessment for real-time identification. Station 41 was one of the stations excluded for the component of the system that assesses the risk of lane-change related crashes. Therefore, crash report number "728266770" would not be identified by the system.

As we can see one of the remaining crashes was sideswipe while the other was rear-end. Therefore, we would be interested in the 'prediction' obtained by the system at the

sections located downstream of station 70 E and upstream of the section located 30 W based on loop data observed 5-10 minutes prior to the reported time of these crashes.

For the crash with report number "758869780" the value of parameter ASD2 (average speed at station D 5-10 minutes before the crash, i.e., average speed at Station 68 E during the period 6:15 to 6:20 PM) was 16.61 MPH. The parameter ASF2 (average speed at station F 5-10 minutes before the crash, i.e., speed at Station 70 E during the period 6:15 to 6:20 PM) was reported to be 37.944 MPH. Based on the first row of Table 5-1 it may be inferred that conditions in the vicinity of station 70 during the slice between 18:15 to 18:20 belonged to regime 1. It essentially means that warning for the impending rear-end would have been issued according to the implementation plan proposed here. In fact it was noticed that at the stations 69 and 70 had been experiencing regime 1 traffic conditions for at least ½ hour before this crash. It again emphasizes the fact that it is not necessary that as soon as the system issues the warning a crash would occur; crash prone conditions might prevail for some time before the crash actually occurs. One can even view this as an advantage since the crash was actually identified much 'earlier'. Also, note that until about 15-20 minutes later the conditions remained regime 1 in the vicinity. It might be due to congestion caused by the crash and according to the system crash warning would remain in place due until even after the crash. The reason for the same is that the incident related congestion would keep the traffic conditions in regime 1. It again is a good symptom because even secondary crashes which were not included in this research as part of the crash sample could potentially be identified based on the traffic speed configurations prevailing after the crash.

The posterior probability estimate of observing a lane-change crash 5-10 minutes before the crash with report number "753088170" could not be estimated because at the time loop detectors at only two lanes were functioning at the Station 31 which happens to be the station upstream of crash location.

The analysis of the loop data prior to the crash gives an idea that how the data availability might limit application of the system. On the other hand it also shows that the one crash for which data were available was correctly identified much ahead of its time. It should be re-emphasized that manifestation of crash identification capabilities of various models is not the main purpose of demonstrating application of the models in a 'virtual' real-time scenario. Crashes being rare events one can not possibly gauge the identification performance of models with data from just one day (i.e., only four crashes), even if data were to be available for all crashes. The crash identification performance of the models was satisfactorily demonstrated through the validation datasets which consisted of sufficiently large crash and non-crash sample. The main aim of this virtual real-time application was to observe the distribution of the posterior probably estimates obtained by various models and compare that to the distribution we obtained by applying the same models on random data. We examine the two distributions for similarity; with particular interest at 30 percentile value. If 30 percentile values for the model outputs over Friday (February 6, 2004) data are close to the thresholds established using random data then it can be claimed with certain confidence that the models are performing as per expectation.

8.7.2 Application of the models over the complete data

It was important to note how many of the crashes with available data are identified correctly by the system. However, as explained earlier it is even more critical to examine output from various models for the complete data in comparison to their outputs over the randomly selected data. It is expected that the behavior of the model outputs would be somewhat similar to their output over a complete day. It will ensure that the numbers of warnings are in the vicinity of their expected value. The rationale for examining performance of the models over random data was to represent all traffic conditions in the data that one might experience on a typical day over the whole corridor. For example, we expect slightly more than 6.5% patterns to belong to regime 1 based on the output from the tree model over random data (Table 8-4). If the proportion of regime 1 traffic conditions over data from a complete day is significantly more than 6.5% then the models are not performing as per expectations. The performance of the models is evaluated based on the percentage of crashes identified within a certain number of warnings. If the number of warnings exceeds its expected value then the performance of models would not be reliable.

Some reasonable sources of differences might be acceptable, however; for example since the application is being demonstrated over loop data from a Friday. Fridays have been known to experience more crashes than any other weekday; especially regime 1 rear-end and lane change related crashes (See Figure 5-11 and Figure 8-8).

8.7.2.1 Distribution of traffic regimes in loop data

As per the implementation strategy first step in identification of rear-end crashes would be to score the data with the classification tree based rules to identify traffic regimes. It was observed that distribution of the traffic regimes over the course of the day was consistent with the frequency patterns observed in the random data. Table 8-9 shows the distribution of the two regimes over eastbound and westbound directions.

Table 8-9: Distribution of the two traffic regimes over the Friday data

Direction	Regime 1	Regime 2
WB	8.3453	91.6547
EB	4.8467	95.1533
Total	**6.6087**	**93.3913**

It may be seen that while the overall distribution of the two regimes is consistent with the proportion of the two traffic regimes in the random data; there are variation across the two directions (eastbound and westbound). Moreover, it is worth mentioning that the distribution was not uniform across locations. For example in the eastbound direction station 22 was the first station to observe any patterns belonging to regime 1; while station 46 observed as much as 12% patterns belonging to regime 1. These 12% patterns would have induced a warning of a rear-end crash; even though in hindsight we know that no such crash was reported. It should be understood that the warning would have been issued because the patterns are worth warning the drivers since almost half the historical rear-end crashes occurred under such traffic conditions. Also, note that the regime 1 conditions are a measure of congested traffic conditions and therefore locations near downtown Orlando during peak hours and around attractions during the evenings

might experience such conditions. It is argued here that supervising such conditions based on traffic speed configurations identified through the tree model would help in 'predicting' a significant number of rear-end crashes ahead of their occurrence.

8.7.2.2 Posterior probability distribution for regime 2 rear-end crashes

According to the implementation plan depicted in Figure 8-5 if the data are available from five stations then the model used to estimate the crash risk would be the hybrid of the best 1-station, 3-station and 5-station MLP/NRBF models. The rationales for putting this model at the top of stepwise implementation hierarchy were discussed earlier. In case of unavailability of the data we can switch to other models that use data from only one or three stations. This is the reason why the threshold for separating crashes from non-crash cases was estimated for all possible models. In this section we exhibit the performance of our most preferred model over complete loop data.

To show that the model is performing as expected we would compare the distributions of posterior probability estimates (of observing a regime 2 rear-end crash) over the whole day with those over the randomly selected observations from the five year period. Table 8-10 shows the percentile distribution of the most preferred hybrid MLP/NRBF model (i.e., hybrid of best 1-station, 3-station and 5-station individual models) over all regime 2 observations in the Friday data. The table also repeats, in last column, the percentiles shown in Table 8-3 for the same model; which was the result of scoring randomly selected regime 2 observations over the course of five year.

Table 8-10: Distribution of the percentiles of output posterior probability of regime 2 rear-end crashes (based on the best hybrid MLP/NRBF based model) over random sample of loop data and a complete day loop data from February 6, 2004

Percentile	Posterior Probability (Hybrid of best 1-station, best 3-station and best 5-station MLP/NRBF model) from the random data	Posterior Probability (Hybrid of best 1-station, best 3-station and best 5-station MLP/NRBF model) over Friday data
100 (Minimum)	0.023445	0.0279
90	0.038826	0.0397
80	0.047067	0.0472
70	0.056064	0.0546
60	0.066359	0.0628
50 (Median)	0.078738	0.075
40	0.094604	0.0914
30	0.11658	0.1102
20	0.15642	0.15
10	0.23412	0.2497
0 (Maximum)	0.74046	0.6142

It may be seen that the value for various percentiles matched closely except for the 100 (minimum) or 0 (maximum) percentiles. The advantage of using percentile threshold is that such individual outliers do not exert a lot of influence. The 30 percentile threshold based on the random data is very close to the 30 percentile value based on the Friday data. In fact the fraction of observations that had the posterior probability over the threshold of 0.11658 (highlighted in Table 8-10) was found to be 27.90% which is very close its expected value of 30%. Therefore, it may be inferred that the threshold estimated earlier (Table 8-3) is suitable for classification. Similar proximity between the estimated thresholds and 30 percentile values over Friday data was observed for the five other models that may potentially be used for regime 2 rear-end crash identification in the event of missing data. The five other models constitute; i) hybrid of best 1-station and 3-station MLP/NRBF models, ii) best 1-station individual NRBF model, iii) hybrid of best

1-station, 3-station and 5-station PNN models iv) hybrid of best 1-station and 3-station PNN model, and v) best 1-station individual PNN model.

We now demonstrate time series of the posterior probability estimated by the model through out the day for three of the possible 238 sections characterized in Figure 8-6. These three locations from the eastbound corridor of the freeway are chosen from three different regions; section upstream of station 62 is near the attractions, station 26 is in downtown Orlando area and Station 16 is further east of downtown Orlando. Note that the posterior probability estimates change every five minutes because input traffic parameters are updated based on the most recent ten observations at the corresponding loop detectors. Since the estimates are obtained from the same model and geometric design parameters (i.e., the location characteristics) are used as inputs to the model (and not as any sampling control factors) one can even compare various locations. It allows us to draw conclusions such as one location is experiencing more risk of observing a crash as compared to other locations based on the high value of estimated posterior probability. The figure also shows the 30 percentile threshold established to separate crashes from normal conditions as a horizontal straight line.

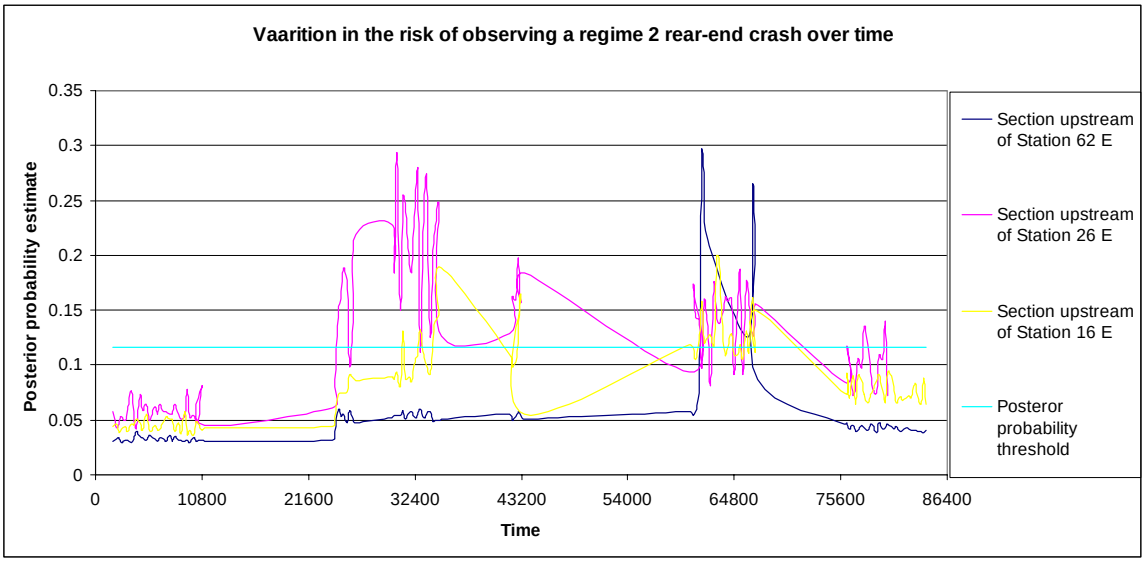

Figure 8-7: Variation of posterior probability of observing a regime 2 rear-end crash over time for three sections on February 6, 2004

Time on the x-axis is represented in terms of seconds past midnight. It may be seen in the figure that section located in downtown (upstream of station 26) is over the 30 percentile threshold for more duration than the other two locations. At the section upstream of station 16 the posterior probability estimated through the hybrid model only exceeds the threshold for small periods just after the morning and afternoon peak hours. For section nearest to attractions (section upstream of Station 62 E) crash risk only exceeds the threshold in the vicinity of 6:00 PM. The fraction of the patterns having posterior probability higher than the threshold over the course of the day was found to be somewhat correlated with the frequency of rear-end crashes over the three regions of the freeway to which the chosen sections belong. For the section in the region east of downtown Orlando (Section upstream of Station 16 E) 19.49% of the patterns were crash prone; while the same percentage was 47.76 and 5.71 respectively for sections located in downtown Orlando (Section upstream of Station 26 E) and near attractions (Section

upstream of Station 62 E), respectively. It may be argued that some such patterns on freeway sections might be recurring. However, there is sufficient amount of temporal variation in the measure of risk estimated based on updated traffic data to suggest that a crash frequency based disaggregate analysis would not have been sufficient to flag certain freeway sections during certain times of the day.

8.7.2.3 Posterior probability distribution for lane-change related crashes

To examine the performance of the hybrid model for lane-change related crashes percentile distributions of posterior probability estimates (of observing a lane-change related crash) over the whole day were compared with those over the randomly selected observations from the five year period. The comparison is shown in Table 8-11. It may be observed from the table that for the same deciles values of posterior probability are higher for the Friday (February, 6 2004) data. It indicates that in a real-time application the number of warnings for lane-change related crashes would have been higher than its expected value of 30%.

Table 8-11: Distributions of the percentiles of output posterior probabilities obtained by the hybrid model for lane-change crashes over random loop data and all lane-change crashes

Percentile	Value of Posterior Probability over random data	Value of Posterior Probability over Friday data
100 (Minimum)	0.02114	0.0219
90	0.028311	0.029
80	0.031657	0.0337
70	0.034494	0.0376
60	0.037574	0.0413
50 (Median)	0.04106	0.0447
40	0.045027	0.048
30	0.048779	0.0507
20	0.052327	0.0532
10	0.057307	0.0568
0 (Maximum)	0.30155	0.2702

The 30 percentile threshold based on the random data was estimated to be 0.048779. However, 30 percentile value for the Friday data is 0.0507. The fraction of observations with posterior probability higher than the estimated threshold (i.e., 0. 048779; highlighted in Table 8-11) was found to be 37.13% which is more than 30%. It was, however, noticed that the frequency of historical lane-change related crashes was maximum over Fridays during the five year period from 1999 through 2003 (See Figure 8-8). This observation makes the more than expected number of warnings acceptable.

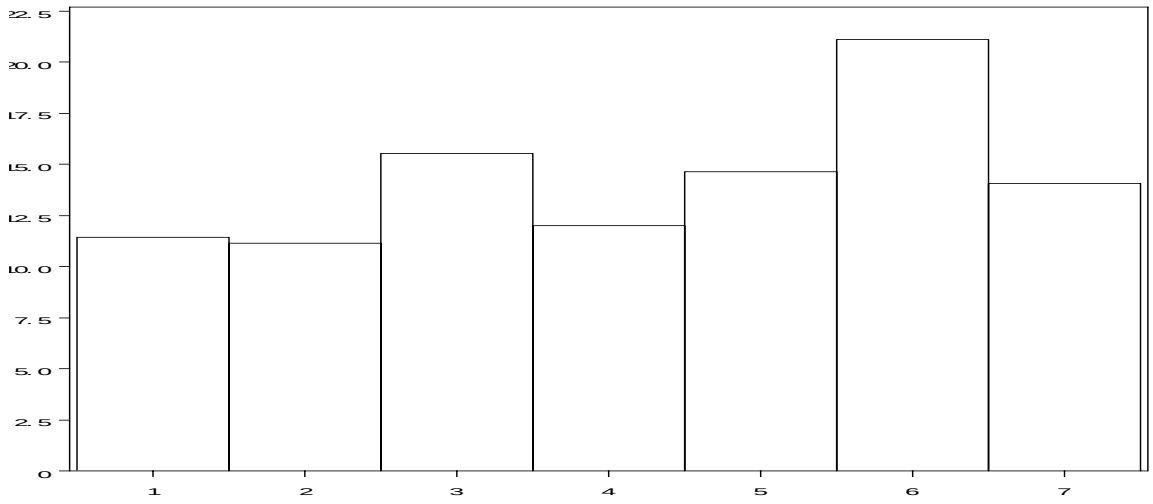

Figure 8-8: Histogram distribution for frequency of lane-change related crashes over day of the week (1: Sunday to 7: Saturday)

8.8 Conclusions

Models belonging to different categories of crashes have been assembled in this Chapter in the form of a system to reliably identify crash prone conditions in real-time. It was also shown that the system can identify a portion of single vehicle crashes that could possibly occur under conditions prone to rear-end and/or lane change related crashes. It was noticed that some crashes in which vehicles hit the object on the side of the road or ran into median were result of the drivers taking evasive action to avoid hitting vehicle(s) in front of them. While the choice of separating such crashes at the modeling stage from rear-end and lane-change crashes was logical it is fascinating to observe that models designed for these crashes could identify significant proportion of single vehicle crashes (other than rollovers).

The system of course has its limitations due to missing data and often times the decisions about flagging a location can not be made. It is more of a problem with the lane-changing component of the system where data from all three lanes of upstream stations are required. In the future we expect advancements in technology leading to improved functionality of loop detectors, which would in turn enable the system to work more regularly.

It was demonstrated that output from various models based on data from a typical Friday (February 6, 2004) had a distribution comparable to the distributions of outputs estimated over random historical loop data. It signifies that the thresholds based on historical random data may be used for future real-time application. Note that this research focuses on real-time identification of crash prone traffic patterns. The system developed here may be used to flag freeway locations; what to do next with this information still remains a matter of research. In the final chapter future scope of this system are discussed along with the summary and conclusions from this research.

CHAPTER 9

CONCLUSIONS AND FUTURE SCOPE

9.1 General

In this research, classification models for rear-end and lane-change related crashes are developed. The models use traffic surveillance data (obtained from dual loop detectors) and geometric design parameters of the freeway as inputs and provide a measure of risk for a specific type of crash in terms of posterior probability. Binary classification may be obtained by applying thresholds on output posterior probability to separate crash prone conditions from 'normal' freeway traffic. These models are then integrated in the form of a system for real-time crash risk assessment. In this chapter we summarize conclusions from this study. The contributions of this research are also discussed along with the future scope.

9.2 Summary and Conclusions

The main contribution of this research is the systematic identification of relationships between traffic/geometric characteristics of the freeway and historical crash occurrences of specific types. Unlike the traditional traffic safety studies that use aggregate measure of traffic flow (e.g., AADT / Peak hour volume); traffic characteristics for this research were measured through under ground loop detectors right before the crash occurrences. The advantage of using traffic surveillance data as input is that the variation in risk of observing a crash may be measured and updated in real-time. Such models can

potentially revamp the incident detection based reactive approach to traffic management into more proactive strategies aimed at crash prevention.

Most of the past research in the area has relied upon generic models developed by analyzing all types of crashes simultaneously. To reliably estimate crash risk in real-time, however, we split the crash data into smaller groups such that the crashes within each group are similar to each other while crashes across groups possess distinct characteristics. To end up with a sufficient sample size in these groups a large sample of crashes was required. Therefore, in this research we collected the crash data for a five year period (1999 through 2003) from the 36.25-mile instrumented corridor of Interstate-4 in Orlando metropolitan area. The crash data were systematically collated with traffic and geometric characteristics of the freeway. We also estimated driver population composition on various sections of the freeway corridor (by time of day and day of week) to examine its effect on real-time crash potential. The database with crashes and corresponding traffic/geometric/driver population characteristics for Interstate-4 is one of the valuable by-products of this research

Following the data collection crashes were segregated based on the harmful event responsible for crash occurrence. The subsequent analysis in this research may be divided into three parts; analysis of rear-end crashes, analysis of lane-change related crashes and integration of the models developed for these two groups of crashes in a real-time application framework. In following sections we summarize the findings from these three components of this research.

9.2.1 Analysis of rear-end crashes

Based on preliminary analysis it was concluded that rear-end crashes on the freeway may be grouped into two distinct clusters based on average speeds in approximately 2-mile region around the crash location 5-10 minutes before a crash. One group of crashes belongs to extended congestion on the freeway (regime 1) while the average speeds are relatively higher during the 5-10 minute period before regime 2 crashes (refer Table 5-1 for specific conditions stipulating the two traffic regimes).

Essentially, regime 1 crashes are the ones which occur when the congested conditions have already set in and could be observed at loop detectors at least 5-10 minutes before the crash. Five to ten minutes before regime 2 crashes conditions at the crash location are not very congested but due to the presence of a downstream on-ramp or otherwise a differential between traffic speeds upstream and downstream starts to build up. This is indicated by the significance of both upstream and downstream traffic speeds and high occupancy at station approximately 1 mile downstream of crash location for this group of rear-end crashes.

It was noticed that overall proportion of regime 1 traffic conditions on the freeway is only 6 to 7% but they make up 46% of rear-end crashes. The rarity of patterns belonging to regime 1 in the random sample of the freeway traffic data led to the conclusion that one can apply the classification tree model used for regime identification on real-time data. Every observation that follows the hierarchy of classification tree rules belonging to

regime 1 may be declared as crash without any further analysis. Hence, all regime 1 crashes may be identified with only 6 to 7% warnings. Although rear-end crashes belonging to regime 2 traffic conditions make up bigger portion of rear-end crashes (54%); these conditions were more frequent (94% in the randomly selected loop data). Hence, separate classification models were needed to separate crashes from the non-crash cases within the traffic data belonging to regime 2.

Three sets of MLP/NRBF neural network models were developed; traffic parameters only from the station located nearest to the crash location (Station F) were included as potential input variables in the first set. In the two subsequent sets traffic parameters from three (Station E, F, and G) and five stations (Station D, E, F, G, and H) were included as potential inputs. For the first two sets, NRBF neural networks with four hidden neurons was found to be the optimal architecture; while MLP with eight hidden neurons provided optimal performance among the third set of models. The output posterior probability from best individual models was averaged to estimate the hybrid models which provided slight improvement over the crash identification performance of individual models. The hybrid of the three aforementioned models identified 55.40% crashes in the validation dataset within 30% observations with maximum posterior probability.

Similarly three sets of PNN classification models were developed, i.e., models using traffic parameters from one, three or five stations as inputs. The parameter varied to search for the optimal PNN model was the spread value. The optimal values of spread parameter for the three sets of models were found to be 0.041, 0.060, and 0.083,

respectively (Table 6-5). In the next step combinations (i.e., the hybrid models) of the best PNN models were created by averaging the posterior probabilities estimates from individual models. It was found that the best hybrid model, which is the combination of the three aforementioned models, captured 57.89% crashes from the validation dataset within 30% observations with maximum posterior probability.

If we examine the crash identification performance for the two groups of rear-end crashes (regime 1 and regime 2) it is apparent that by issuing warnings only 6 to 7 % of times we can identify all regime 1 rear-end crashes. For the remaining 94% cases we would have to issue warnings for about 30% cases (i.e., declare 30% observations with maximum output posterior probability as rear-end crash) to identify less than 60% of regime 2 crashes. It indicates that the traffic conditions prevalent 5-10 minutes before regime 1 rear-end crashes are more distinct from 'normal' traffic. Note that this fact cannot be used to argue that the strategy to 'predict' regime 1 crashes is more useful than the strategy to identify conditions prone to regime 2 rear-end crashes. It is possible that measures, such as variable speed limits (VSL), for reducing the risk of crashes belonging to regime 2 are more easily applicable (Dilmore, 2005). Moreover, since regime 2 crashes would generally occur under higher traffic speeds they may be expected to be more severe. Hence, avoiding each additional crash in this group might be more beneficial than avoiding a crash from regime 1.

Also, note that the best hybrid models for regime 2 rear-end crashes would need data from five surrounding stations to be available. In case the data from five stations are not

available, hybrid or individual models that use data from only three or one station may be used. It is worth mentioning that matched case-control logistic regression model was also estimated based on within stratum matched sampling and stepwise variable selection procedure. It indicated that speed differential between upstream and downstream of crash site is significantly associated with regime 2 rear-end crashes. However, its classification accuracy over the validation dataset was inferior to the neural network based hybrid models. Therefore, the logistic regression model was not used as part of the system developed for real-time identification of conditions prone to rear-end crashes.

9.2.2 Analysis of lane-change related crashes

Most common lane-change related crashes on the freeways are classified as sideswipe crashes. However, they are not the exclusive constituents of lane-change related crashes. Crashes on the inner lanes of the freeway that are recorded as angle crashes can also be attributed to lane-changing (Lee et al., 2006). This was verified by examining the crash reports for these angle crashes.

Lane-change related crashes are expected to be influenced by interaction between traffic parameters measured on different lanes of the same station. Therefore, crash and non-crash cases with loop data available from all three lanes at the upstream station were used for analysis. It was observed that none of the off-line factors (geometric characteristics) were significant according to the classification tree based variable selection procedure. Average speeds upstream and downstream of the crash site were significant variables. Average differences between adjacent lane occupancies upstream of the crash site

(ADALOU2) along with standard deviation of volume and speed (SVW2 and SSW2) downstream were found to be associated with lane-change related crashes. After identifying the critical variables we subjected the data to multiple neural network models. It was noticed that the MLP model with four hidden neurons and NRBF model with three hidden neurons were the best individual models. The two models were hybridized to improve their performance and the combined model resulted in identification of 57% lane-change related crashes from the validation dataset. It was significantly higher than either of the individual models. Note that PNN classifiers were not created for lane-change related crashes.

9.2.3 Assembling multiple models: Real-time application framework

The results from the models summarized in the previous sections may be used to classify real-time traffic patterns. The models for rear-end and lane-change related crashes are suggested to be applied in parallel so that a warning for rear-end crash is independent of the warning for a lane-change related crash.

The issues relevant to application of these models on real-time data were discussed in the previous chapter. For example, six different models are available to assess the risk for regime 2 rear-end crashes. The model providing the best crash identification at the evaluation stage required data from a series of five loop detector stations to be simultaneously available. The most preferred model may be replaced by models with more tolerant data requirements, even though it would mean sacrificing on the classification accuracy, in case any of the five stations is not reporting data. It was also

shown that the distribution of the output from various models over data from one complete day was akin to their outputs over a large sample of randomly selected loop data. Therefore, thresholds established through application of model(s) on the random data may be used over the course of the day to separate crash prone conditions from normal conditions. Note that the performance of the models was assessed based on crash identification with 30 percentile posterior probability as the threshold. Therefore, the real-time application is also demonstrated based on the 30 percentile threshold which essentially means 30 percent positive decisions (i.e., warnings). The threshold value of the posterior probability may be increased to a lower percentile if desired.

The day used for 'virtual' real-time application was February 6, 2004. On the Interstate-4 corridor under consideration four crashes were reported on that day. The crash, which had the required data available, could be identified ahead of its time through the real-time application plan proposed here. Promising crash identification capabilities of the system were also indicated by the fact that the models part of the proposed system were evaluated based on their performance on validation datasets that consisted of the observations not part of the training data.

For the other three crashes data pre-requisites were not met and the application of the system was restricted due to unavailability of data. While at the modeling stage problems due to missing data were overcome through extensive data collection efforts; some improvement in the hardware technology would be desirable at the application stage.

9.3 Additional Comments and Future Scope

In this research we have analyzed the crash data by type of crash and developed a system of classification models that can identify conditions prone to certain types of crashes from 'normal' conditions. Crashes are rare events and involve significant human factor. Driving behavior of individuals is expected to play some role, even a significant one, in a crash. However, it should be noted that there is no way to factor performance of individual drivers in real-time. Detailed vehicle by vehicle movement data is necessary to achieve that degree of surveillance, which being impossible to obtain we are essentially dealing with a measure of traffic flow available to us at certain pre-specified locations (i.e., the loop detector stations). Data from these locations are correlated with crash occurrence in the vicinity. The objective is to try and identify patterns observed at the loop detector stations prior to historical crashes. These patterns may then be described as "turbulent" conditions on freeway sections in which a crash is more likely to occur and the drivers need to be more attentive in order to avoid crashes. Geometric design parameters are also included in this study; therefore, if a certain combination of "turbulent" conditions and freeway geometric design is observed in the future a crash occurrence may be expected.

It is also worth mentioning that the role of human factors varies by type of crash under consideration. For example, under regime 1 traffic conditions (extended congestion over a long stretch of the freeway) only a slight error on part of the driver would lead to a rear-end crash while a single vehicle rollover on the mainline of the freeway would almost never occur unless reckless driving or vehicle malfunction is involved. This essentially

explains why the performance of the model used to identify regime 1 rear-end crashes is so efficient (all crashes identified with only 6% positive decisions) or why the statistical link between the average speeds measured at loop detectors and rollover crashes can not be used to 'predict' them. While these two groups of crashes are extremes in either direction; regime 2 rear-end and lane-change related crashes are somewhere in the middle. These groups of crashes albeit identifiable, do not enjoy the accuracy with which regime 1 rear-end crashes are identified. These observations again emphasize the precision in crash identification that we are able to achieve by separating crashes into smaller groups at the modeling stage.

Of course there are other advantages of segregating the crash data into smaller groups. The patterns of loop data used to identify rear-end crashes were not similar to those used to identify lane-change related crashes. Moreover, due to the segregation the more frequent rear-end crashes could be analyzed in more detail than the other groups since the incentive of improving rear-end crash identification is significantly more than other types of crashes. These advantages justify the extensive crash data collection effort put in for this study.

The application of models is demonstrated such that the fraction of positive decisions (i.e., warnings) is about 30%. In the previous chapter, while only 30 percentiles values were used as threshold for various models; all deciles (0 through 100 percentiles with an increment of 10) were established (Tables 8-1, 8-2, and 8-3). In case at application stage it is felt that 'too many' warnings are being issued then the threshold may be increased to

20 or 10 percentile (10 or 20 percentile values would be higher). How many warnings are 'too many' depends on the application. For example, if the model output is supplied, in some form, to the drivers then authorities must ensure that the high number of warnings does not lead the drivers to disregard the information.

It opens questions on how the information may be best utilized and in what form, if at all, should it be transferred to the drivers using the facility. It has been shown that the models in this research, with a certain degree of accuracy, can identify crash prone conditions on the freeway. It essentially is the primary component of proactive traffic management. The next logical step towards that aim would be to devise measures to use this research for crash prevention. Warnings could be issued to the drivers on flagged sections of the freeway through variable message signs (VMS). Separate models developed here would help in devising specific countermeasures for different groups of crashes. For example, warnings for rear-end crashes could take the form "exercise caution while following" or warning for a lane-change related crash could be "no lane-changing next x miles or y minutes". These proactive measures need to be applied carefully so that drivers pay attention to the warnings but do not overreact in panic. Therefore, impact of such warnings on driver behavior needs to be thoroughly studied, possibly through a driver survey.

The concept of variable speed limits could also be used to intervene and reduce the risk of rear-end crashes. Higher speed limits on downstream with lower speed limits on the upstream of potential 'black spots' identified by the model(s) would be the fundamental

approach towards variable speed limits. To precisely understand when and how far on sections upstream/downstream to decrease/increase the speed limits, detailed further analysis is required. Microscopic traffic simulation may be employed to assess the benefits (i.e., the achieved reduction in crash risk) of variable speed limits (Dilmore, 2005).

The study demonstrates the applicability of loop detector data for identifying crash prone conditions on the freeway in real-time. Traffic data used in this study are collected using dual magnetic induction loop detectors, which are one of the most common traffic surveillance apparatus. Therefore findings from this research are transferable to other freeways as well. On freeways with better hardware capabilities (hence, less missing data) the system might perform even more efficiently. While it would be advisable to recalibrate some of the neural networks; inputs to those models could be adopted from the list of significant variables identified for various groups of crashes in this research.

REFERENCES

Abdel-Aty, M., and Abdalla, F., Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes using generalized estimating equations for correlated data. Presented at the *83rd Annual Meeting of the Transportation Research Board* (TRB), Washington D.C., 2004.

Abdel-Aty, M., Chen, C., and Schott, J., An Assessment of the effect of driver age on traffic accident involvement using log-linear models. *Accident Analysis & Prevention Journal, Volume 30, No. 6*, 1998, pp. 851-861.

Abdel-Aty, M., and Pande, A., Identifying crash propensity using specific traffic speed conditions. *Journal of Safety Research, Vol. 36*, 2005, pp. 97-108.

Abdel-Aty, M., Uddin, N., and Pande, A., Split models for predicting multi-vehicle crashes under high speed and low speed operation conditions on freeways. Forthcoming in the *Transportation Research Record,* 2005.

Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., and Hsia, L., Predicting freeway crashes based on loop detector data using matched case-control logistic regression. *Transportation Research Record, No. 1897, National Research Council,* Washington, D.C., 2004, pp. 88-95.

Abdelwahab, H., and Abdel-Aty, M., Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record, No. 1746, Transportation Research Board, National Research Council,* Washington, D.C., 2001, pp. 6-13.

Abdelwahab, H., and Abdel-Aty, M., Traffic safety analysis for toll plazas using artificial neural networks and logit models. *Transportation Research Record, No. 1784, Transportation Research Board, National Research Council,* Washington, D.C., 2002, pp. 115-125.

Abdulhai, B., and Ritchie, S., G., Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. *Transportation Research Part C: Emerging Technologies, Volume 7, Issue 5*, 1999, pp. 261-280.

Adeli, H., and Karim, A., Fuzzy-wavelet RBFNN model for freeway incident detection. *Journal of Transportation Engineering*, *Volume 126, No. 6*, 2000, pp. 464-471.

Agresti, A., Categorical data analysis, 2nd Ed. *John Wiley and Sons, Inc.,* 2002.

Al-Deek, H., Garib, A., and Radwan, A., E., A New Method for Estimating Freeway Incident Congestion. *Transportation Research Record, No. 1494, Transportation Research Board, National Research Council,* Washington, D.C., 1995, pp. 30-39.

Al-Deek, H., Ishak, S., and Khan, A., Impact of freeway geometric and incident characteristics on incident detection. *The ASCE Journal of Transportation Engineering, Vol. 122 No. 6,* 1996, pp. 440-446.

Awad, W., and Janson, B., Prediction models for truck accidents at freeway ramps in Washington state using regression and artificial intelligence techniques. *Transportation Research Record, No. 1635, Transportation Research Board, National Research Council,* Washington, D.C., 1998, pp. 30-36.

Breiman, L., Freidman, J., H., Olshen, R., A., and Stone, C., J., Classification and Regression Trees *Wadsworth & Brooks/Cole Advanced Books & Software,* Monterey, California, 1984.

Chandra, C., and Al-Deek, H., New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Transportation Research Record, No. 1867, National Research Council,* Washington, D.C., 2004, pp. 116-126.

Chang, G., and Kao, Y., An empirical investigation of macroscopic lane-changing characteristics on uncongested multilane freeways. *Transportation Research Part A, Volume 25, No. 6,* 1991, pp. 375-389.

Cheu, R., L., and Ritchie, S., G., Automated detection of lane-blocking freeway incident using artificial neural networks. *Transportation Research Part C: Emerging Technologies, Volume3, Issue 6*, 1995, pp. 371-388.

Christodoulou, C., and Georgiopoulos, M., Applications of Neural Networks in Electromagnetics, *Artech House*, Boston, 2001.

Collett, D., Modeling binary data. *Chapman and Hall*, 1991.

Cybenko, C., Approximations by superposition of sigmoid functions. *Mathematics of Control Signals and Systems, Vol. 2,* 1989, pp. 303-314.

Dilmore, J., Implementation strategies for real-time traffic safety improvements on urban freeways. *University of Central Florida*, 2005.

Golob, T., F., and Recker, W., W., Relationships among urban freeway accidents, traffic flow, weather and lighting Conditions. *California PATH Working Paper UCB-ITS-PWP-2001-19, Institute of Transportation Studies.* University of California, Berkeley, 2001.

Golob, T., F., and Recker, W., W., A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A, Volume 38, No. 1,* 2004, pp. 53-80.

Golob, T., F., Recker, W., W., and, Alvarez, V., M., Freeway safety as a function of traffic flow. *Accident Analysis & Prevention*. Vol. 36, no. 6, 2004 (a), pp. 933-946.

Golob, T., F., Recker, W., W., and Alvarez, V., M., Tool to evaluate the safety effects of changes in freeway traffic flow. J*ournal of Transportation Engineering. Volume130, Issue 2,* 2004 (b), pp. 222-230.

Hagan, M., T., and Menhaj, M., Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks, Volume 5, No. 6,* 1994, pp. 989-993.

Hand, D., Mannila, H., and Smyth, P., Principles of Data Mining. *M.I.T Press*, Massachusetts, 2001.

Haykin, S., Neural networks: A comprehensive foundation. *Macmillan Publishing company*, New York, 1999.

Henderson, M., Human Factors In Traffic Safety: A Reappraisal. *Traffic Accident Research Unit, Department of Motor Transport*, New South Wales, Australia, 1971.

Hosner, D., W., and Lemeshow, S., Applied Logistic Regression, *Wiley & Sons*, 1989.

Hughes, R., and Council, F., On establishing relationship(s) between freeway safety and peak period operations: Performance measurement and methodological considerations. Presented at *the 78th annual meeting of Transportation Research Board*, Washington, D.C., 1999.

Ishak, S., and Al-Deek, H., Performance of automatic ANN-based incident detection on freeways. *Transportation Engineering Journal of ASCE, Vol. 125, No. 4*, 1999, pp. 281-290.

Ishak, S., and Alecsandru, C., Analysis of freeway pre-incident, post-incident, and non-incident conditions using second-order spatio-temporal traffic performance measures. Presented at *the 84th annual meeting of Transportation Research Board*, Washington, D.C., 2005

Kockelman, K., M., and Ma, J., Freeway speeds and speed variations preceding crashes, within and across lanes. Presented at *the 83rd annual meeting of Transportation Research Board,* Washington, D.C., 2004.

Kohonen, T., Learning vector quantization. *Neural Networks, 1 (suppl 1),* 1988, pp. 303.

Lee, C., Abdel-Aty, M., and Hsia, L., Potential real-time indicators of sideswipe crashes on freeways. *Manuscript prepared for presentation at 85th annual meeting of Transportation Research Board,* Washington, D.C., 2006.

Lee, C., Hellinga, B., and Saccomanno, F., Assessing safety benefits of variable speed limits. *Transportation Research Record, No. 1897*, *National Research Council,* Washington, D.C., 2004, pp. 183-190.

Lee, C., Saccomanno, F., and Hellinga, B., Analysis of crash precursors on instrumented freeways. *Transportation Research Record, No. 1784, Transportation Research Board, National Research Council,* Washington, D.C., 2002, pp. 1-8.

Lee, C., Saccomanno, F., and Hellinga, B., Real-time crash prediction model for the application to crash prevention in freeway traffic. *Transportation Research Record, No. 1840*, *National Research Council,* Washington, D.C., 2003, pp. 68-77.

Madanat, S., and Liu, P., A prototype system for real-time incident likelihood prediction. *IDEA Project Final Report (ITS-2), Transportation Research Board, National Research Council,* Washington, D.C., 1995.

Masters, T., Advanced algorithms for neural networks: A C++ sourcebook. *John Wiley and Sons, Inc.,* 1995.

Moody, J., and Darken, C., J., Fast learning in networks of locally-tuned processing units. *Neural Computation, 1*, 1989, pp. 281-294.

Mussone, L., Ferrari, A., and Oneta, M., An analysis of urban collisions using an artificial intelligence model. *Accident Analysis and Prevention, Vol. 31,* 1999, pp. 705-718.

Oh, C., Oh, J., Ritchie, S., and Chang, M., Real time estimation of freeway accident likelihood. Presented at *the 80th annual meeting of Transportation Research Board,* Washington, D.C., 2001.

Pande, A., Classification of real-time traffic speed patterns to predict crashes on freeways. MS Thesis, *University of Central Florida*, 2003.

Pande, A., Abdel-Aty, M., and Hsia, L. Spatio-temporal variation of risk preceding crash occurrence on freeways. Forthcoming in the *Transportation Research Record*, 2005.

Park, S., and Ritchie, S., G., Exploring the relationship between freeway speed variance, lane changing, and vehicle hetrogenity. Presented at *the 83rd annual meeting of Transportation Research Board,* Washington, D.C., 2004.

Powell, M., J., D., Radial basis function approximations to polynomials. *Proceedings of 12th Biennial Numerical Analysis Conference*, Dundee, 1987, pp. 223-241.

Quinlan, J., R., C4.5: Programs for Machine Learning. *Morgan Kaufmann,* San Mateo, California, 1993.

Rumelhart, D., Hinton, G., and Williams, R., Learning internal representation by error propagation, parallel distributed processing, *Explorations in the Microstructure of Cognition, Vol. 1: Foundations,* D. E. Rumelhart and J. L. McClelland, editors, *MIT Press*, Cambridge, MA, 1986, pp. 318-362.

SAS Institute, Getting Started with Enterprise Miner Software. *Release 4.1, SAS Institute*, Cary, NC, 2001.

Sayed, T., and Abdelwahab, W., Comparison of fuzzy and neural classifiers for road accident analysis. *Journal of Computing in Civil Engineering*, *ASCE, Vol. 12, No. 1,* 1998, pp. 42-47.

Sohn, S., and Shin, H., Pattern recognition for road traffic accident severity in Korea. *Ergonomics, Vol. 44, No. 1*, 2001, pp. 107-117.

Specht, D., F., Probabilistic neural networks and general regression neural networks. In: Chen, C.H. (Ed.), *Fuzzy Logic and Neural Network Handbook. McGraw-Hill,* Berlin, 1996, pp. 3.1–3.37.

Stamatiadis, N., and Deacon, J., A., Quasi-induced exposure: methodology and insight. *Accident Analysis and Prevention, Volume 31,* 1999, pp. 705-718.

Tao, K., M., A closer look at the radial basis function (RBF) networks. *Conference Record of the 27th Asilomar Conference on Signals, Systems and Computers (Singh, A., ed.), Volume 1, IEEE Comput. Soc. Press,* Los Alamitos, CA, 1993, pp. 401-405.

Tarassenko, L., and Roberts, S., Supervised and unsupervised learning in radial basis function classifiers. *IEE Proceedings-- Vis. Image Signal Processing, 141,* 1994, pp. 210-216.

Traffic Safety Facts: National Highway Traffic Safety Administration, *National Center for Statistics and Analysis.* US Department of Transportation, 2002.

Vorko, A., and Jovic, F., Multiple attribute entropy classification of school-age injuries. *Accident Analysis and Prevention, Vol. 32,* 2000, pp. 445-454.

Wang, J., and Knipling, R., Lane change/merge crashes problem size assessment and statistical description. *Department of Transportation, National Highway Traffic Safety Administration, Report No. DOT HS 808 075*, 1994.

Wilamowski, B., Iplikci, S., Kayank, O., and Efe, O., M., An algorithm for fast convergence in training neural networks. Presented at *the International Joint Conference on Neural Networks (IJCNN'01)*, Washington DC, July 15-19, 2001.

Zhang, C., Ivan, J., N., El-Dessouki, W., M., and Anagnostou, E., N., Relative risk analysis for studying the impact of adverse weather conditions and congestion on traffic accidents. Presented at *the 84th annual meeting of Transportation Research Board,* Washington, D.C., 2005.

Zhou, M., and Sisiopiku, V., P., Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record, No. 1581, National Research Council,* Washington, D.C., 1997, pp. 47-52.