MD-09-SP708B4C

Martin O'Malley, *Governor*
Anthony G. Brown, *Lt. Governor*

State Highway Administration
Maryland Department of Transportation

John D. Porcari, *Secretary*
Neil J. Pedersen, *Administrator*

STATE HIGHWAY ADMINISTRATION

RESEARCH REPORT

AN INTEGRATED KNOWLEDGE-BASED SYSTEM FOR REAL-TIME ESTIMATION OF INCIDENT DURATIONS AND NONRECURRENT CONGESTION DELAY FOR FREEWAY NETWORKS

DEPARTMENT OF CIVIL ENGINEERING

UNIVERSITY OF MARYLAND, COLLEGE PARK, MD 20742

Project number MD-09-SP708B4C
FINAL REPORT

March 5, 2009

Technical Report Documentation Page

1. Report No. MD-09-SP708B4C	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle **An Integrated Knowledge-Based System for Real-Time Estimation of Incident Durations and Nonrecurrent Congestion Delay for Freeway Networks**		5. Report Date March 5, 2009 6. Performing Organization Code
7. Author/s Gang-Len Chang, Woon Kim, and Yi-Shiuan Chang		8. Performing Organization Report No.
9. Performing Organization Name and Address **Department of Civil Engineering, The University of Maryland, College Park, MD 20742**		10. Work Unit No. (TRAIS) 11. Contract or Grant No. SP708B4C
12. Sponsoring Organization Name and Address Maryland State Highway Administration Office of Policy & Research 707 North Calvert Street Baltimore MD 21202		13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code (7120) STMD - MDOT/SHA
15. Supplementary Notes		

16. Abstract

This study presents a set of models for predicting incident duration and identifying associated variables in the state of Maryland. The incident database from Year 2003 to Year 2005 from Maryland State Highway (MDSHA) Administration was used for model development, and Year 2006 data was used for the model validation. This study has employed the Rule-Based Tree method to develop the primary prediction model. To enhance the prediction accuracy for some types of incidents of complex nature or having limited samples, the study has also calibrated several supplemental components based on the Multinomial Logit and Regression methods. Further exploration for fatality incidents has also been conducted with Naïve Bayesian Classifier using the Accident Report database provided by the MD State Police Department. In addition, this study has developed preliminary models for computing incident-induced delay and queue length as a part of applications of incident duration. The developed set of models offers an effective tool for responsible agencies to estimate the approximate predicted duration of a detected incident and its resulting delay as well as queue length.

17. Key Words Incident duration, nonrecurrent congestion, delay, queue	18. Distribution Statement: No restrictions This document is available from the Research Division upon request.		
19. Security Classification (of this report) None	20. Security Classification (of this page) None	21. No. Of Pages	22. Price

Form DOT F 1700.7 (8-72) Reproduction of form and completed page is authorized.

An Integrated Knowledge-Based System for Real-Time Estimation of Incident Durations and Nonrecurrent Congestion Delay for Freeway Networks

To

Office of Policy and Research
Maryland State Highway Administration
707 N. Calvert Street
Baltimore, MD 21202

by

Gang-Len Chang,
Woon Kim,
and Yi-Shiuan Chang
gang@umd.edu

Department of Civil Engineering
The University of Maryland
College Park, MD 20742

March 2009

Table of Contents

Chapter 1 : Introduction………………………………………………………………...…4
1.1 Research Background………….………………………………………..…...4
1.2 Definition of Incident Duration…….………………………………………....4
1.3 Study Purpose and Scope..…..……………………………………….……….5
Chapter 2 : Literature Review…..……………………………………………………9
Chapter 3 : Analysis of Incident Duration Data……………………………………...16
3.1 Introduction………………………………………………………………16
3.2 Data Description…….………………………………………………….......16
3.3 Preliminary Data Analysis……………………………………………...18
3.3.1 Incident Duration……………………………………………………...18
3.3.2 Independent Variables……………………………………………..23
3.4 Average Incident Duration……………………………………………………30
Chapter 4 : Methodology and Analysis………………………………………………37
4.1 Introduction…………………………………………………………………....37
4.2 Preliminary Analysis with CART…………………………………………..…37
4.2.1 Basic Procedures of CART………………………………………………...37
4.2.2 Results and Findings from CART……………………………………....39
4.3 Procedures for a Rule-Based Tree Model (RBTM)……………………………..44
4.4 The Rule-Based Tree Model for Incident Nature of Collision-Fatality (CF)…48
4.4.1 The Tree Structure……………………………………………………....48
4.4.2 Performance and Validation Results…………………………………….....53
4.5 The RBTM for Incident Nature of Collision-Personal Injury (CPI)……………57
4.5.1 The Tree Structure……………………………………………………...….57
4.5.2 Performance and Validation Results………………………………………....62
4.6 The RBTM for Incident Nature of Collision-Property Damage (CPD)……..67
4.6.1 The Tree Structure………………………………………………………67
4.6.2 Performance and Validation Results……………………………………71
4.7 The RBTM for Incident Nature of Disabled Vehicles (DV)……………......77
4.7.1 The Tree Structure………………………………………………………77
4.7.2 Performance and Validataion Results……………………………………..78
4.8 The RBTM for Incident Nature of Others………………………………………79
4.8.1 The Tree Structure………………………………………………………79
4.8.2 Performance and Validation Results………………………………………81
4.9 Overall Findings and Conclusions………………………………………………83
Chapter 5 : Supplemental Models……………………………………………..……86
5.1 Introduction……………………………………………………………………86
5.2 Supplemental Model – I : Multinomial Logit (MNL) Models……………..……89
5.2.1 Multinomial Logit Models………………………………………………...89
5.2.2 Estimation Results with MNL…………………………………………….93
5.3 Supplemental Model – 2 : Multiple Linear Regression Models……………….103
5.3.1 Multiple Linear Regression Models……………………………………....103
5.3.2 Estimation Results with Multiple Linear Regression Models…………....104
5.4 Illustrative Description of the Application of the Developed Model…………..114

Chapter 6 : Advanced Study for Collision-Fatality Incidents......................118
6.1 Introduction..118
6.2 Integrated Data Descriptions...119
6.3 Methodology...122
6.3.1 Naïve Bayesian Classifier (NBC)......................................122
6.4 Model Development and Results..123
6.5 Closure...128
Chapter 7 : Application of Incident Duration Models: Estimations of Incident-Induced Delay and Queue...130
7.1 Introduction...130
7.2 Methodology...130
7.2.1 Experimental Design...130
7.2.2 Description of Variables..131
7.2.3 Model Development and Results.....................................133
7.3 Applications of Developed Models..140
Chapter 8 : Conclusions...142
8.1 Summary of Research Results...142
8.2 Future Research...147
Appendix 1..149
Appendix 2..158
References..179

Chapter 1: Introduction

1.1 Research Background

Traffic incidents have long been recognized as the main contributor of congestion in highway networks. Incidents, as defined in this study, include disabled vehicles, fire, road debris, construction, police activity, and vehicle crashes. On congested highways, any incident, regardless of whether they involve personal fatalities, injuries, or property damages, will cause considerable reduction in roadway capacity due to lane closures or impediments. As reported in the literature, a one-lane blockage on a four-lane road will reduce the roadway capacity by 60 percent (TRB, 2000). The reduction in capacity during the incident duration will inevitably result in heavy congestion and delay, thus leading to enormous socio-economic loss. In day-to-day traffic control and management, with a reliable method for predicting incident duration in real time, responsible agencies can convey information to travelers via the variable message signs (VMS), estimate the resulting queue length and the corresponding total delay, and assess the need to implement detour operations or any other control strategies. Thus, an effective model for predicting the duration of a detected incident is one of the essential tools for traffic agencies in mitigating nonrecurrent congestion in highway networks.

1.2 Definition of Incident Duration

According to the Highway Capacity Manual (TRB, 2000), the entire duration of an incident consists of four phases, as shown in Figure 1.1. The first phase, the detection time, represents the time elapsed from the occurrence of the incident to its detection. The second phase, response time, corresponds to the period of time between the detection of the incident and the arrival of any emergency or incident response unit. The third phase, clearance time, is defined as the time elapsed from the first arrival of response units (e.g., police or emergency vehicles) to the time that the incident is cleared. The last phase, the recovery time, measures the time required for traffic to return to its normal condition.

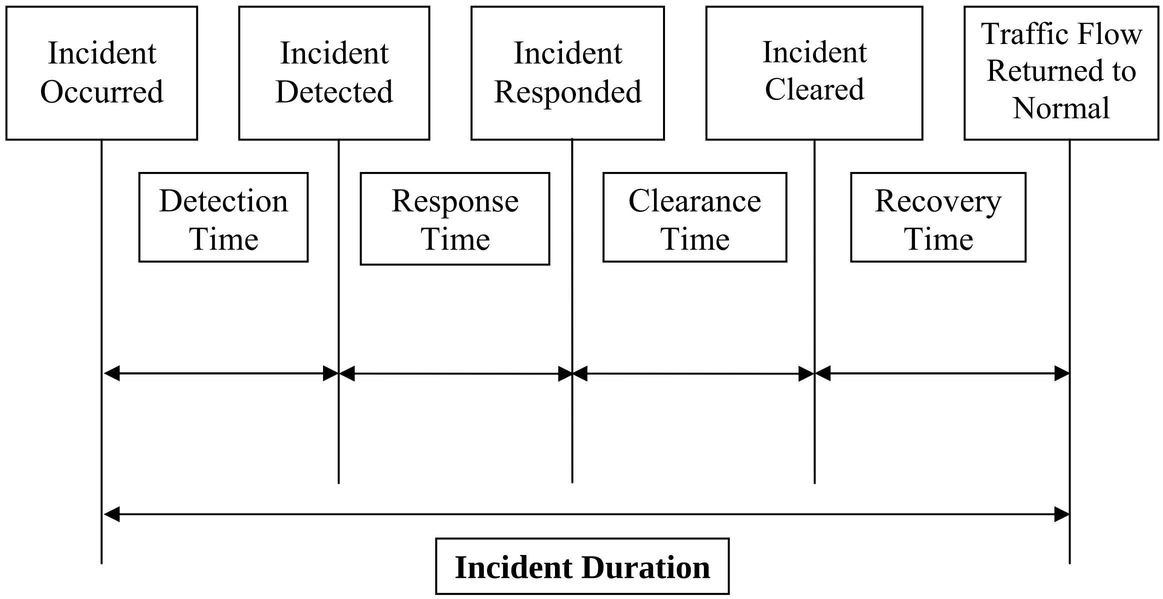


Figure 1.1 Phases of Traffic Incident Duration

In general, it is difficult to know the exact timestamp of incident occurrence; usually, the recovery time is also regarded as being out of scope for incident duration studies. Moreover, the database used for this study includes reliable records only for response and clearance times. Thus, for this study, incident duration is defined as the time elapsed from incident detection to its clearance, which is the sum of the response and clearance times.

1.3 Study Purpose and Scope

Due to the lack of available data, incident duration was usually estimated based on field experience rather than on rigorous statistical models. Improvements in reporting techniques and in incident information databases have facilitated detailed analyses of critical variables that influence incident durations, which has facilitated their prediction. Previous studies in this field have resulted in different prediction methods and models. However, it must be noted that these prediction models were developed based on sets of data derived from different sources. Thus, information available for predicting the duration of an incident may vary among different databases. It has also been observed that incident duration is influenced by various location-specific factors. Hence, to ensure reliable and efficient modeling of incident duration prediction, one needs to calibrate the model from a well-designed database which includes all critical information for that area. Such a model can then be confidently used to implement detour operations or any other control strategies, along with appropriate mitigation measures.

The objective of this study is to develop a set of models for estimating the duration of a detected incident and for identifying variables that may significantly influence the incident duration in the state of Maryland. This study used the CHART (Coordinated Highways Action Response Team) database from the Maryland State Highway Administration (SHA).

This study begins with a review of related literature in Chapter 2, including the most representative approaches for predicting incident durations: (1) probabilistic distributions, (2) conditional probabilities, (3) linear regression models, (4) time-sequential models, (5) decision trees and classification trees, and (6) discrete choice models, and (7) Bayesian classifier.

Chapter 3 focuses on the description of available data and the statistical analyses of interrelations between key variables. This chapter includes a preliminary analysis for the distribution of incident durations, statistical tests for independent variables using the

ANOVA test, Tukey test, and multiple correspondence analyses. The final subsection discusses the average incident duration classified by key variables.

Chapter 4 presents the procedures adopted for model development and evaluation, along with the results of model estimation and validation. This chapter begins with preliminary analyses using the classification and regression tree (CART) model. Based on the findings from the CART, the chapter goes on to explore a new model, named the rule-based tree model (RBTM). Detailed procedures for this model's development and performance, as well as its validation, are also included in the subsections that follow. Chapter 4 concludes with the overall findings from the RBTM, and indicates the necessity of calibrating supplemental models to enhance the performance of the primary model.

Chapter 5 illustrates the two different types of supplemental models for predicting incident duration. It first discusses the calibration of multinomial logit (MNL) models and their performance with a test data set. This is followed by the development of multiple linear regression models, for types of incident with small sample data sets, and their performance. The last section highlights potential applications of supplemental models.

Chapter 6 presents further research on fatality incidents, due to the uniqueness of the data. The study adopts the Naïve Bayesian Classifier (NBC), based on integration of the CHART database with the Maryland State Police Department Accident Report database.

Chapter 7 discusses major applications of the developed incident duration models. It includes models for the total delay estimation and queue length computation based on the estimated incident duration. To better illustrate application of the model, estimated results for delay and queue length are presented using an actual incident from the CHART database.

Chapter 8 summarizes the primary research findings and conclusions of this study. Future research needs are also discussed in this chapter.

Chapter 2: Literature Review

Incident duration has been studied by numerous researchers for several decades using various methodologies. The most representative approaches are (1) probabilistic distributions, (2) conditional probabilities, (3) linear regression models, (4) time-sequential models, (5) decision trees and classification trees, (6) discrete choice models, and (7) Bayesian classifier. Although a variety of techniques exist that provide acceptable results, they cannot be directly applied to incidents occurring at any other locations. Each model was developed with different incident data sources and descriptive variables and thus yields somewhat different results. Therefore, for any target application, it is necessary to develop a new model for different traffic conditions and available data sources.

The first approach to modeling incident durations reviewed in this study is the probabilistic model, which is relatively straightforward to use in forecasting the incident duration. The key aspect of this approach is to view the duration as a random variable and attempt to find a probability density function (PDF) that can fit to the data set. Golob et al. (1987) conducted their research using approximately 530 incidents involving trucks, and found that incident durations could be modeled with a log-normal distribution. Their finding has been supported by other studies by Giuliano (1989), Garib et al. (1997), and Sullivan (1997) for freeway incident durations. In 1999, Ozbay and Kachroo also found that the distribution of incident durations from their data set showed a shape very similar to log-normal distribution, although a few statistical significance tests rejected their hypothesis. However, they realized that when the study data set was subdivided by incident type and severity, these subsets followed a normal distribution. This finding has an important implication, since it supports the theory that the incident duration is a random variable (Smith and Smith, 2002). Similarly, Jones et al. (1991) discovered that a log-logistic distribution could be used to describe their study data set from Seattle. In 2000, Nam and Mannering found that their data set could be illustrated with the Weibull distribution. However, Smith and Smith (2002) could not find an appropriate probability distribution, including log-normal and Weibull distributions, to fit the incident clearance times for their study data.

Probability models for incident duration can be extended to conditional probability models. The key idea of such models is to find the probability distribution of incident durations under certain given conditions; for example, the probability of an incident lasting 30 minutes, given the condition that the incident has already lasted for 10 minutes. Intuitively, it is noticeable that the probability of the end of an incident would vary, depending on how long the incident has lasted (known as duration dependence in Nam and Mannering [2000]) and the incident characteristics. One interesting approach under this concept is the hazard-based duration model. This model allows researchers to formulate incident durations with conditional probability models. Such models have been widely used in the biometrics and industrial engineering fields to determine causality from the duration data. Due to their similarity with the nature of traffic incident duration, their theoretical concepts and models have recently been applied in the transportation field. With such an approach, researchers' interests have expanded from simply estimating and predicting incident durations to computing the likelihood that the incident would finish in the next short time period, given its elapsed duration. One of the most representative studies using this methodology was conducted by Nam and Mannering

(2000), using a two-year data set from Washington State. Their study showed that the duration for each incident phase (i.e., detection/reporting, response, and clearance phases) is significantly affected by numerous factors and that different assumptions of distribution are recommended for different incident phases. They also found that the estimated coefficients were unstable through the two-year data set used in developing their model. As Nam and Mannering concluded, this approach is more useful for determining which variable has greater influence on incident duration than for estimating or predicting the incident duration for a set of given explanatory variables.

Another simple methodology for predicting incident durations is linear regression models. These models usually include a number of binary variables as independent variables to indicate incident characteristics, and a continuous or categorical variable as a dependent variable (i.e., incident duration). One of the best-known linear regression models for incident prediction was developed by Garib et al. (1997) using 277 samples from California. They used various independent variables to represent incident characteristics (e.g., incident type, number of lanes affected by the incident, number of vehicles involved, and truck involvement) and weather conditions (rainy or dry). They also included all possible combinations of the independent variables to develop the best model. The final incident duration model from their research is as follows:

$$Log(Duration) = 0.87 + 0.027X_1X_2 + 0.2X_5 - 0.17X_6 + 0.68X_7 - 0.24X_8$$

where Duration = incident duration (minutes)

X_1 = number of lanes affected by the incident

X_2 = number of vehicles involved in the incident

X_5 = truck involvement (dummy variable)

X_6 = morning or afternoon peak hour indicator (0: morning peak hour; 1: afternoon peak hour)

X_7 = natural logarithm of the police response time (minutes)

X_8 = weather condition indicator (0: no rain; 1: rain)

This model showed 0.81 for adjusted R^2. The logarithm form of incident durations indicated that the incident durations in this data set followed a log-normal distribution, which is supported by the Kolmogorov-Smirnov test. This result is similar to those from Golob et al. (1987) and Giuliano (1988). According to the authors, the police response time was the most significant factor affecting incident durations, followed by weather condition, peak hour, truck involvement, and the combined effect of the number of lanes and vehicles involved in the incident.

Khattak et al. (1995) realized that the full set of variables for incident forecasts would be available at the moment the incident is cleared. Although prediction models based on this complete set of variables would be more accurate and reliable, they are less practical for the real-time prediction of incident durations because this full set of variables can only be available after the incident is cleared. Thus, Khattak et al. introduced a time-sequential model, based on the idea that the prediction of incident duration made earlier in the incident life would be more informative to incident management, even with lower accuracy and reliability. The model that they developed has ten distinct stages of incident duration, based on the availability of information. Each stage indicates different ranges of incident duration and has a separate truncated regression model. At each stage, progressively more variables are included to explain the stage duration. Despite its originality and reasonability, this model was not tested or validated, due to the lack of field data. The authors also mentioned that the intention of their study was to introduce and demonstrate the time-sequential model rather than to prove the performance of their model in traffic operations.

Another approach available in the literature is the decision tree model. The purpose of applying this methodology is to discover patterns in a given data set without considering the fundamental probabilistic distribution (Smith and Smith, 2001). Smith and Smith (2001) pointed out that the pattern-recognition model has been used recently to develop incident duration models. One representative model was developed by Ozbay and Kachroo (1999) for the Northern Virginia region. They began by developing a model to predict clearance times using linear regression, based on a large sample size. Unfortunately, they completed the analysis with a poor result ($R^2 \approx 0.35$) and learned that incident durations follow neither a log-normal nor a log-logistic distribution. As an alternative method, they explored a decision tree model and finally generated the relation patterns shown in Figure 2.1 for predicting clearance times.

It can be noted that the incident tree consists of a series of decision variables. For instance, the tree uses incident type as the first variable to decide if the detected incident type is known or not. Once it is classified as an unknown type, the tree immediately provides 45 minutes for the clearance time. Otherwise, it goes to the next level to decide which type of incident it falls into. After that, it will face the next decision variable (e.g., "Is wrecker used?") and so on. Also, the outcome from this tree is an average clearance time under current conditions, as estimated from past records.

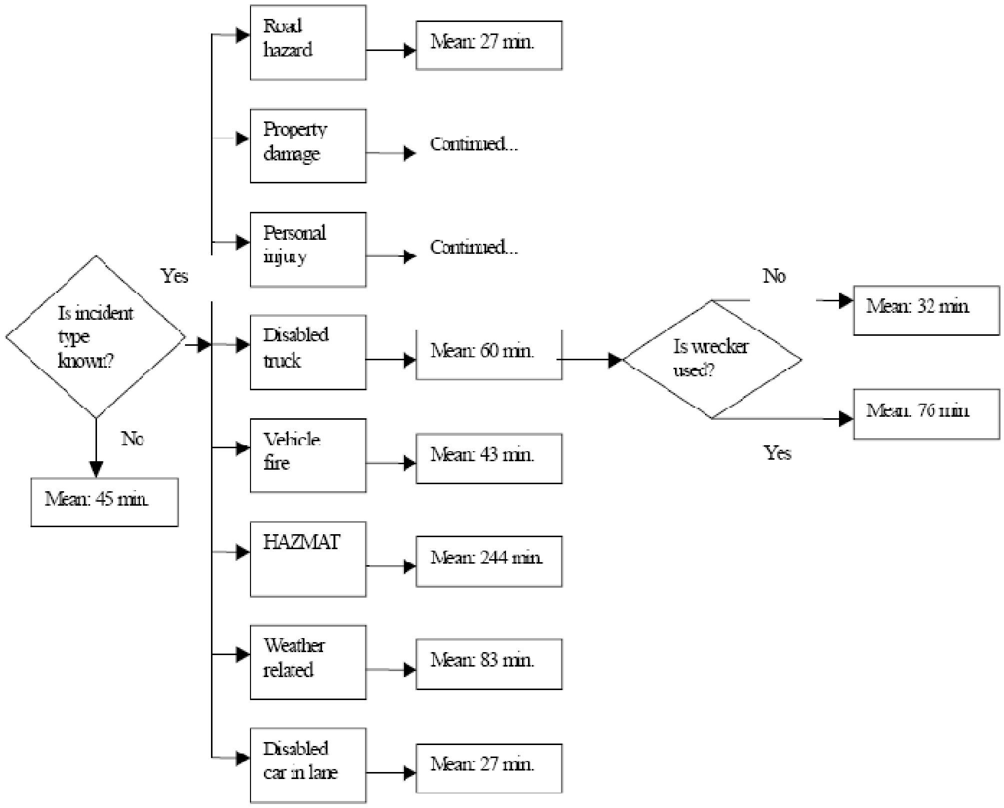


Figure 2.1 Part of the Complete Decision Tree to Predict Clearance Time, by Ozbay and Kachroo (1999)

Ozbay and Kachroo were satisfied with the new tree, based on the test results, since about 57.14 percent (44 out of 77) of tested incidents were predicted within ten minutes of prediction error. They also found that the large differences between predicted and actual clearance times were caused by numerous outliers.

Smith and Smith (2001), who were inspired by the Ozbay and Kachroo study, tried to develop a similar classification tree. They concluded that a classification tree developed on the basis of a reliable and sufficient database performs well, even though the results of their classification tree were not satisfactory due to poor data quality. A detailed discussion regarding classification trees will be presented in Chapter 4.

Another approach reviewed for this study is the discrete choice model. Most studies in the literature have treated incident duration as a continuous variable. Lin et al. (2004) developed a system that integrates the discrete choice model and the rule-based model to predict incident duration. They first adopted ordered probit models to classify sample data for incident duration into several time intervals, and then developed a rule-based supplemental model to enhance the accuracy of prediction results.

One of the most recent studies published develops a model using the Naïve Bayesian Classifier. Boyles et al. (2007) found that the model results from the NBC were satisfactory compared with the results from the linear regression model.

Building on the work by Lin et al., and using an enriched data set, this study explored the integrated application of a set of new models, including a rule-based tree model, a discrete choice model, and a multiple regression model. For further research dedicated to incidents associated with fatalities, the NBC was applied to enhance model performance. The proposed methodology will be discussed in more detail in Chapters 4, 5 and 6.

Chapter 3: Analysis of Incident Duration Data

3.1 Introduction

This chapter presents the description of data used for this study and the statistical analyses of interrelations between key variables. It includes the distribution of incident duration, statistical tests for independent variables using the ANOVA test, Tukey test, and multiple correspondence analyses (MCA). The final section discusses the average incident duration, classified by key variables.

3.2 Data Description

To evaluate the performance of its incident response operations, the SHA has developed an incident management database called CHART (Coordinated Highways Action Response Team). Since 1996, CHART has collected major and minor incidents that have occurred in Maryland, and the highway system of CHART-II is its most recently upgraded database. This study is based on highway incident data extracted from CHART-II from the years 2003 to 2005 for model development, and from year 2006 for model validation. The data set from CHART-II for this research includes:

- Incident duration: detected, responded, and cleared timestamps;
- Incident characteristics: number of shoulder lanes blocked, total number of lanes at the incident location, and number of lanes blocked (for the same direction, the opposite direction, and for both directions);
- Ratio of lane blockage: number of lanes blocked (for the same direction, the opposite direction, or for both directions)/total number of lanes at the incident location;

- Type of incident: property or personal damage by collision and fatality by collision, as well as debris, disabled vehicle, vehicle fire, police activities, off-road activities, and emergency roadwork;
- Response team information: participation of SHA patrol;
- Information about involved vehicles: number of vehicles involved, type of vehicles involved (truck-trailer, single-unit truck, or pickup van);
- Time: peak time (AM peak and PM peak) indicators, weekend indicator, night indicator, and time in hours when an incident was detected;
- Location information: county, road name, and exit number for I-495, I-95, I-695, and I-270 only; and
- Pavement condition: dry, wet, snow/ice, chemical wet, and unspecified.

In this study, any record that included a missing value for any information was excluded from consideration for statistical analysis, model development, and validation. Since CHART-II records the exit number of the incident location only for four major interstate roads, I-495, I-95, I-695, and I-270, the specified location information is available only for part of the entire sample. As mentioned earlier, the incident duration represents the sum of the response time and clearance time, since the incident occurance time is not available. In addition, records with durations below five minutes were excluded, since such short durations seemed unreasonable. After cleaning up the raw database, 6765 records were left for statistical analysis and model development, and 6501 for the model validation.

3.3 Preliminary Data Analysis

3.3.1 Incident Duration

As mentioned in the literature review, incident durations were found to follow several different, but similar, shapes of distribution. Golob et al. (1987) discovered that, by using data for vehicle crashes with trucks involved, the total of incident durations fit in the log-normal distribution, while, according to Jones et al. (1991), incident durations could be illustrated by the log-logistic distribution. The findings of Golob et al. have been supported by several researchers in the subsequent years (Giuliano, 1989; Garib et al., 1997; and Sullivan, 1997). Ozbay and Kachroo (1999) found that the durations of incidents of similar type and severity showed a normal distribution, while Nam and Mannering (2000) suggested a Weibull distribution for incident durations. Except for the normal distribution, the common feature of those distributions is a shift to the left so that a large portion of the duration data is concentrated on shorter durations, as shown in Figure 3.1, below (Smith and Smith, 2002).

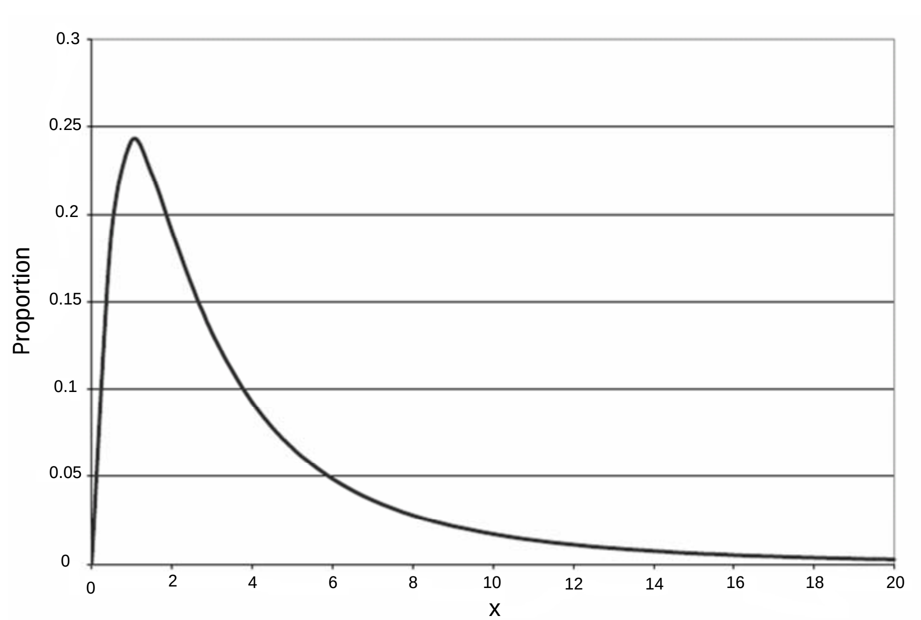


Figure 3.1 General Shape of Log-Normal Distribution (Smith and Smith, 2002)

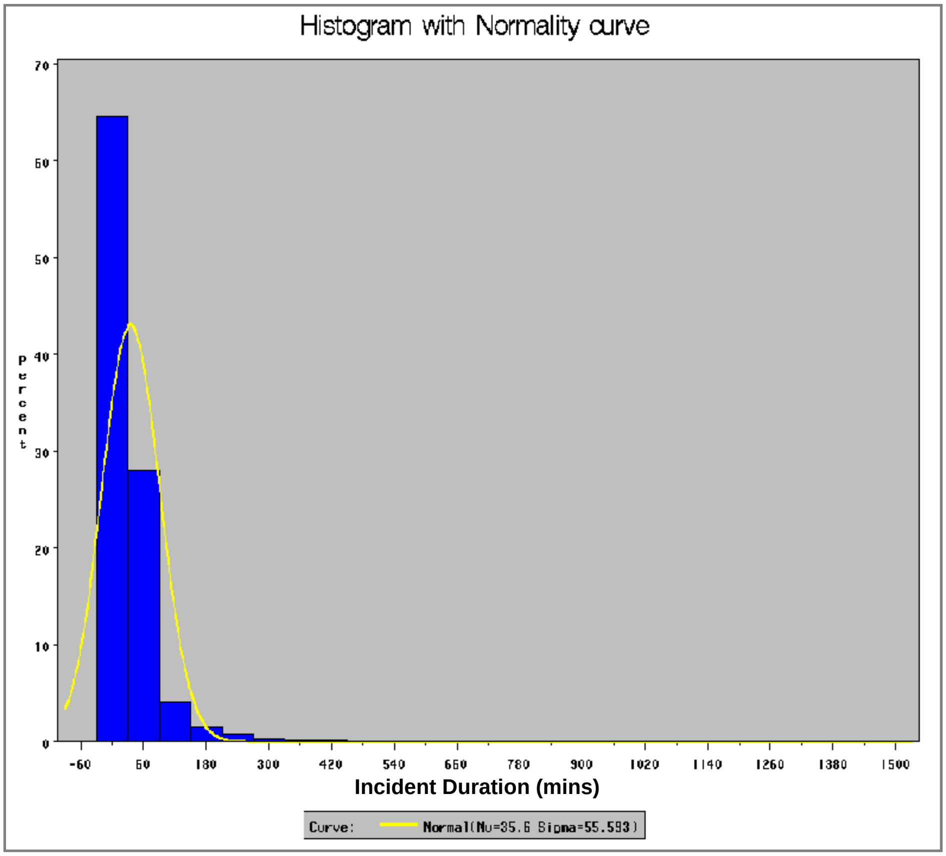


Figure 3.2 Histogram with a Normality Curve of the Incident Duration Used in This Study

To understand the distribution of incident durations, the entire available data set (including data with incident durations of less than five minutes) is plotted in the histogram shown in Figure 3.2. The available incident durations clearly form a shape of distribution similar to that shown in Figure 3.1. Considering the quantile-quantile plot (Q-Q plot) and probability plot (P-P plot) for log-normal distribution (Figure A1.1 in Appendix 1) and Weibull distribution (Figure A1.2 in Appendix 1), the data suggest a distribution that is closer to a log-normal distribution than to the Weibull distribution. However, the hypothesis tests, such as the Kolmogorov-Smirnov test, Anderson-Darling test, and chi-square test for distributions of log-normal, log-logistic, Weibull, and so on, all reject the log-normal distribution hypothesis at 0.01 and 0.05 significance levels.

Since the following statistical tests are performed under the assumption of normality of the data set, it is essential to transform the original data to fit a normal distribution. Although various transformation techniques exist, Johnson and Wichern (1993) and Dimakos suggested that power transformations would be appropriate when the selection of transformation is not really obvious. Box and Cox (1964) stated that power transformations shrink large values of a variable X and, at the same time, they enlarge small values. The family of power transformations, which is defined with λ, has the following general form (Dimakos):

$$x^{\lambda} = \frac{x^{\lambda} - 1}{\lambda}, \text{ where } \lambda \neq 0 \text{ and } x > 0 \qquad \text{(Eq. 3.1)}$$

$$x^{\lambda} = \ln x, \text{ where } \lambda = 0 \text{ and } x > 0 \qquad \text{(Eq. 3.2)}$$

The value of λ is selected in order to maximize the following function:

$$l(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{j=1}^{n}(x_j^{\lambda} - \overline{x^{\lambda}})^2\right] + (\lambda - 1)\sum_{j=1}^{n}\ln x_j \qquad \text{(Eq. 3.3)}$$

where n is the number of observations, x_j is the original value of the jth observation, and $\overline{x^{\lambda}}$ is an arithmetic average of the transformed observation and is defined as:

$$\overline{x^{\lambda}} = \frac{1}{n}\sum_{j=1}^{n} x_j^{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_j^{\lambda} - 1}{\lambda}\right) \qquad \text{(Eq. 3.4)}$$

By using the Box-Cox macro introduced by Dimakos, the optimal value of λ found for the data set of this study is 0.1. The transformed data set is much closer to fitting a normal distribution, as shown in the descriptive statistics (e.g., histograms (Figure 3.3), Q-Q plots (Figure A1.3), or P-P plots (Figure A1.4)). In a histogram, the overall shape of the distribution of the transformed data set becomes nearly symmetrical (see Figure 3.3). The Q-Q plot and the P-P plot also show that the Box-Cox power transformation helps the original data set convert to a normal distribution, because the transformed observations are placed near the diagonal dashed line (see Figure A1.3 and A1.4, Appendix 1). Though the descriptive statistics demonstrate that the Box-Cox power transformation works quite well to alter the original distribution to a normal distribution, the hypothesis tests still reject the null hypothesis (H_0 : the data follow a normal distribution) at 0.01 and 0.05 significance levels. Appendix 1 presents the results of basic statistical measures and hypothesis tests by SAS.

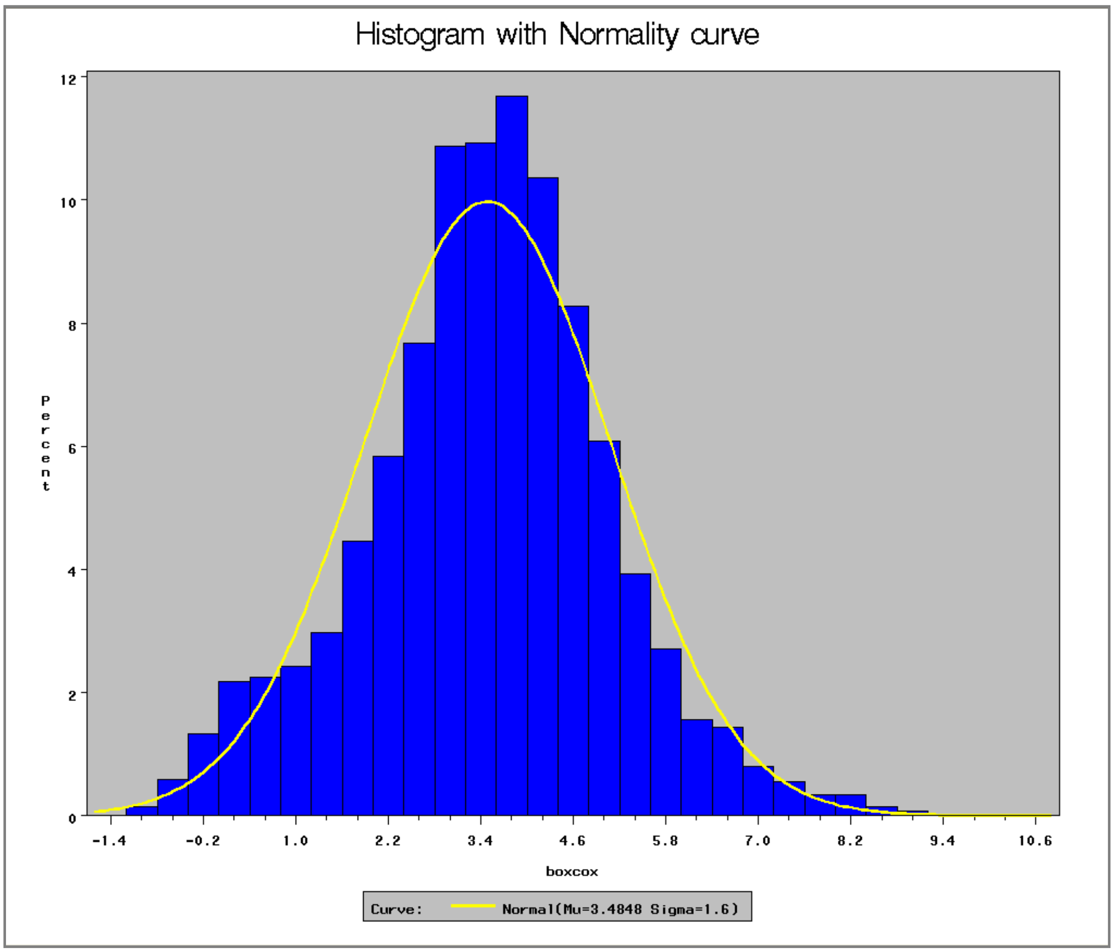


Figure 3.3 Histogram of the Box-Cox Power-Transformed Data Set

The same procedure is performed with the data set which excludes incident durations of less than five minutes. The optimal value of λ for the truncated data set is found to be -0.2. Even though the descriptive statistics for this case also show that the distribution is quite close to a normal distribution, all of the hypothesis tests reject its normality at 0.01 and 0.05 significance levels. However, the statistics of tests become much smaller than those for the original data set which includes incident durations of less than five minutes (see Table 3.1). This means that the truncated data set fits better to a normal distribution when compared to the original data set.

Table 3.1 Summary of Hypothesis Tests Statistics

Using the Original Data Set								
Model	Parameter	N	Chi-Sq	P-value	A-D	P-value	K-S	P-value
$x^{\lambda} = \frac{x^{\lambda} - 1}{\lambda}$	λ = 0.1	7798	393.6	0.00	19.24	< 0.005	0.03745	< 0.01
Using the Truncated Data Set (Incident Duration >= 5 min)								
Model	Parameter	N	Chi-Sq	P-value	A-D	P-value	K-S	P-value
$x^{\lambda} = \frac{x^{\lambda} - 1}{\lambda}$	λ = -0.2	6765	250.9	0.00	3.607	< 0.005	0.01616	< 0.01

Incident Duration as Categorical Variables

Although the incident duration variable is continuous in nature, it is more useful and practical to predict the duration by interval, such as between 20 and 30 minutes, rather than with a precise prediction of, for example, 26.5 minutes.

This study employs the following procedures to categorize the continuous variable. First, incident durations are categorized based on the cumulated percentage of the available samples. A category is defined by the range that covers approximately 15 percent of total samples, while the records with durations longer than 120 minutes form the last category. Smith and Smith (2002) classified their dependent variable (clearance time) into three categories – short, middle and long – for applying the CART. Since the CART was used to build a preliminary model in this research, a three-category variable similar to the one used by Smith and Smith (2002) was also considered to be one option for classifying the dependent variable set. A four-category variable (short, middle, long, and very long) was explored as well. For the more detailed analysis using the primary model, the RBTM, incident durations were also categorized for every five minute interval up to 120 minutes. As in the first categorization, records with durations longer than 120 minutes constitute the last category. Categories of the dependent variable used for this study are summarized in Table 4.1 in Chapter 4.

3.3.2 Independent Variables

Specifications of Independent Variables

Unlike the previous study by Lin et al. (2004), this study specifies independent variables as discrete variables, such as 0, 1, 2 and 3 if they have an inherent order, based on the actually recorded values, rather than represented as dummy variables. This specification can help reflect the possibility of different impacts when the condition becomes more severe. Independent variables without inherent order such as counties and hours incident occurred remains as dummy variables. These independent variables are summarized, along with other variables, in Table 3.2.

Statistical Tests for Independent Variables

In this study, a one-way analysis of variance (ANOVA) test is first carried out to see the effect of each independent variable on the incident duration. For the multicategorical variables showing significantly different impacts on the incident duration, a further analysis (Tukey Test) is carried out to regroup the categories of the variables. Furthermore, an MCA is implemented to determine a set of most significant variables which can explain most parts of the entire data set.

1. ANOVA Test

ANOVA tests were performed to test whether any of the descriptive variables had significant effects on the incident duration. Each of the descriptive variables was tested with transformed incident durations, and all of them showed significant effects, except the indicator of *Pick Up Van Involvement* at the 0.01 and 0.05 significance levels. The p-value of the ANOVA test for this variable is 0.094, so that the null hypothesis, the mean of incident durations with pick up vans involved is equal to the mean of those without pick up vans involved, cannot be rejected. However, at the 0.1 significance level, this variable can still be included in the model development.

Table 3.2 Independent Variables Used for the Model Development

Variables	Original Range (Value or Category)	Regrouped Range (Value or Category)
Incident Nature	Collision-Fatality Collision-Personal Injury Collision-Property Damage Disabled Vehicle Debris Fire Others (Police Activity, Emergency Road Work, Off-Road Work)	Collision-Fatality Collision-Personal Injury Collision-Property Damage Disabled Vehicle Others (Debris, Fire, Police Activity, Emergency Road Work, Off-Road Work)
Pavement Condition	Dry Wet Snow/Ice Chemical wet Unspecified	Dry Not Dry
Road Name	I-495 IL, OL I-95 N, S I-695 IL, OL I-270 N, S I-370 E, W I-68 E, W I-795 N, S I-83 N, S I-895 E, W I-97 N, S MD-295 N, S 70 E, W US 1 N, S US 50 E, W Other	G1 : I-495 IL, OL G2 : I-895 E, W MD-295 N, S I-270 N, S G3 : I-695 IL, OL I-95 N, S I-97 N, S US 50 E, W G4 : I-795 N, S I-370 E, W I-83 N, S 70 E, W US 1 N, S Other G5 : I-68 E, W
CHART Involved	0, 1	N/A
Single-Unit Truck Involved	0, 1	N/A
Pick-Up Van Involved	0, 1	N/A
Tractor-Trailer Involved	0, 1	N/A
No. of Single-Unit Trucks Involved	0, 1, 2, 3, 4	0, 1, >=2
No. of Pick-Up Vans Involved	0, 1, 2, 3, 4, 5, 6, 8	(0 or 1), >=2
No. of Tractor-Trailers Involved	0, 1, 2, 3, 4, 5, 6	0, 1, >=2
Weekend	0, 1	N/A
Peak Hour	0, 1	N/A

Variables (cont')	Original Range (Value or Category)	Regrouped Range (Value or Category)
No. of Vehicles Involved	> 0	1, (2 or 3), >=4
No. of Same-Direction Lane Blockages	0, 1, 2, 3, 4, 5, 6, 7	0, 1, 2, >=3
No. of Opposite-Direction Lane Blockages	0, 1, 2, 3, 4, 5	0, 1, >=2
No. of Shoulder Blockages	0, 1, 2, 3, 4	0, 1, >=2
Shoulder Blockage Indicator	0, 1	N/A
Total Lane Blockages	0, 1, 2, 3, 4, 5, 6, 7, 8, 12	0, 1, 2, >=3
Ratio of Same-Direction Lane Blockages	0.00 ~ 1.00	N/A
Ratio of Opposite-Direction Lane Blockages	0.00 ~ 1.00	N/A
Ratio of Total Direction Lane Blockages	0.00 ~ 1.00	N/A
No. of Lanes (One Direction)	2, 4, 8	N/A
Hour Incident Occurred	1, 2, 3, …., 23, 24	Day : 6 ~ 20 Night : everything else
Response Time (minutes)	> 0.00	N/A
County	32 different counties	N/A

2. Regrouped Independent Variables Using Tukey Test

To figure out which groups have similar properties so that they can be combined into one group, this study applied the Tukey HSD (honestly significant difference) test, which is designed for pairwise comparisons based on the studentized range proposed by Tukey in 1952. The test starts by sorting the means of groups in ascending order to calculate the difference in means for each pair of groups. Then, it computes the minimum pairwise difference required using the following formula (Tukey, 1952, 1953).

$$HSD_{\min} = Q_a \sqrt{\frac{MS_{wg}}{S}} \qquad \text{(Eq. 3.5)}$$

where Q_a is a critical value from a studentized range statistic table at a level, MS_{wg} is the mean square error within group from ANOVA, and S is the number of samples per group.

In the above formula, HSD_{min} represents the minimum pairwise difference between the means of any two particular groups considered to be significant. Q_α depends upon parameters k (the number of groups in the original analysis) and df_{wg} (the number of degree of freedom associated with MS_{wg} in the original analysis) at α level. When the number of samples is not equal for each group, S is replaced with the harmonic mean of the grouped samples. Lastly, HSD_{min} is compared to the actual difference in means (M_L-M_S, where M_L is the larger mean value, while M_S is the smaller mean value in two groups) for each pair of groups. If the actual difference is greater than HSD_{min}, the two groups are significantly different with respect to their means.

When the Tukey test is implemented, one should be aware of the increment of the error rate, α, due to the repeating of procedures. To adjust this error rate, the Bonferroni inequality (Rencher, 2002) has been widely applied. The adjusted error rate by Bonferroni inequality is α/c, where c is the number of comparisons. The regrouped independent variables, using the Tukey test with the Bonferroni inequality adjustment on α, are summarized in Table 3.2, along with the original categories.

Initially, the incident nature was categorized into seven classes. The Tukey test showed that two incident types, *Debris* and *Fire*, are not significantly different from the incident type *Others*. Hence, those three incident types (i.e., *Debris*, *Fire*, and *Others*) could be grouped as one large incident type. The number of data having single-unit trucks and tractor-trailers was recategorized into three groups (*0*, *1*, and *>=2*), whereas the number of pick up vans was recategorized into two groups (*0* or *1*, and *>=2*).

3. Variable Selection Using MCA

The correspondence analysis (CA) was originally developed by Jean-Paul Benzécri in France in the early 1970's (Benzécri, 1973). It has the same function as the factor analysis but is used mainly for categorical variables. Since this technique was first introduced in French, it took some time to reach popularity in English-speaking countries (Carrol et al., 1986; Hoffman and Franke, 1986). Similar techniques were also developed independently in other countries under different names, such as optimal scaling, quantification method, or homogeneity analysis (Hill and Lewicki, 2005). As the first step in performing CA, one must compute the relative frequencies for the frequency table of two variables, such that the sum of all entries of the table equals 1.0. The row or column totals in the relative frequency table are referred to as the row *mass* or column *mass*, respectively (Greenacre, 1984). In the table, rows and columns are completely independent, and the entries of the rows and columns can be recreated by the totals of rows and columns, which are referred as row and column *profiles* in CA (Hill and Lewicki, 2005).

Under the condition that the rows and columns of the frequency table are completely independent of each other, the expected frequencies in the table can be derived from the respective column total times the row total, divided by the grand total, based on the well-known formula of the chi-square statistic for two-way tables. The differences (or deviations) from the expected values contribute to the overall chi-square. From this perspective, CA can be viewed as a technique to decompose the total chi-square statistics, or an *inertia*, which is defined as chi-square divided by the grand total of the frequency in CA (Greenacre, 1984), by expressing a small number of dimensions that represent the deviations from the expected values.

A statistical software package (in this study, SAS) can produce the results of CA, including dimensions, corresponding values, eigenvalues, percent of inertia, and chi-square. The dimensions are extracted to maximize the distances between row and column points.

While CA is based on the two-way table, MCA is designed for more than two variables. Since MCA can be regarded as an extension of simple CA, the characteristics and interpretations of results are the same as those in CA.

Since this study included more than two categorical predictors MCA was performed to find the most significant independent variables that could explain deviations from the expected values. Thus, regrouped variables were input to MCA, and 32 dimensions, which contain all information in the input table, were extracted. Each dimension forms by linear relationship between coefficients and corresponding variables, e.g., $Dim_i = \sum_j \beta_j X_j$, where β_j is a coefficient, and X_j is a corresponding variable.

In a dimension, the variable with the largest absolute value of coefficient represents the most significant variable and dominates that dimension (Jolliffe, 1972 and 1973). Table A1.1 in Appendix 1 summarizes the largest coefficient values and the corresponding variables for these 32 dimensions. As shown in the table, the most significant factor in the first and second dimensions, which was also the most significant factor for the entire study, was the number of blocked lanes for the opposite direction greater than or equal to two. This result reflects that the incidents involving more than one lane blockage in the opposite direction were more likely to be severe and have a longer duration. Although the total number of dimensions is 32, the variables representing all dimensions can be summarized as the following 11 variables, since some variables repeatedly appear in different dimensions. The categories which make the variable significant in MCA are indicated in parentheses.

- No. of Lane Blockage for Opposite Direction (>=2)
- No. of Single-Unit Trucks Involved (1 and >=2)
- No. of Lane Blockage for Same Direction (2 and >=3)
- Incident Nature (Others: Debris, Fire, Police Activity, Emergency Road Work, Off-Road Work)
- Regrouped Road: Group 5 (I-68)
- Incident Nature (collision fatality)
- No. of Shoulder Blockage (>=2)
- No. of Pick-Up Van Involved (>=2)
- No. of Vehicles Involved (=1)
- Shoulder Blockage Indicator (=0)
- No. of Total Lane Blockage (>=3)

3.4 Average Incident Duration

Before starting the model development, the average incident duration was computed to investigate its relationships with explanatory variables. Tables 3.3(a)-3.3(c) summarize the statistical results of incident durations under different classifications. As shown in Table 3.3(a), the incident durations increased with the number of heavy vehicles (e.g., tractor-trailers, single-unit trucks, or pickup vans) involved. The same relation is also shown in Table 3.3(b), where the incident duration increases with the number of blocked lanes. Incident durations on weekends and at night were generally longer than the durations on weekdays and in the daytime, due to the longer response and clearance times.

It is noticeable that incidents occurring on the four major freeways, I-495, I-95, I-695, and I-270, had relatively shorter durations than others. This can be explained by the locations of operations centers, which determine the accessibility of the response units. Maryland has six operations centers — one statewide operations center, and five traffic operations centers. Among them, five operations centers are located near those four major roads, because they are primary roads around the two metropolitan areas — Washington D.C. and Baltimore.

Also, the incident durations were found to exhibit remarkable differences between different incident types. As shown in Table 3.3(c), the incidents caused by disabled vehicles had the shortest durations on average (22.47 minutes), followed by incidents involving property damage, others (fire, debris, emergency road work, police activities and off-road activities), and personal injuries. As expected, incidents causing fatalities usually resulted in the longest durations (208.66 minutes). Figure 3.4 illustrates the distribution of frequency across incident duration intervals for each type of incident. In the category of incidents with disabled vehicles, 96.3 percent of their durations were distributed between 5 and 70 minutes, and 63.3 percent were between 5 and 20 minutes. This reflects that incidents involving disabled vehicles are likely to have shorter durations.

Incidents with property damage also showed a similar shape of distribution, and 90.2 percent of such incidents took between 5 and 70 minutes. However, unlike the incidents with disabled vehicles, they were quite evenly distributed up to 30 minutes. Incidents causing personal injuries and fatalities were more likely to have longer durations. For example, 94.2 percent of incidents resulting in fatalities lasted over 70 minutes, and 78.6 percent of them lasted over 120 minutes. Note that 80.8 percent of incidents causing personal injuries resulted in durations longer than 20 minutes, while 60.9 percent of all personal injury incidents took between 20 and 70 minutes. In the category of incidents classified as *Others*, incident durations distributed quite evenly across all intervals. These results are consistent with the observation that the distribution of incident durations varies with incident nature. Therefore, incident nature emerges as one of the most significant factors for classifying incidents of different durations.

Table 3.3(a) Summary of Average Incident Duration Classified by Key Variables

Variables	Avg. Duration (minutes)	Frequency
No. of Tractor-Trailers		
0	34.89	5809
1	51.95	780
2	164.18	152
>= 3	257.36	24
No.of Single-Unit Trucks		
0	38.97	6101
1	49.95	574
2	81.66	77
>=3	124.72	13
No. of Pickup Vans		
0	41.5	5006
1	35.6	1365
2	43.57	333
>=3	56.52	61
No. of Vehicles Involved		
1	34.2	3090
2	43.42	2393
3	47.19	823
4	51.61	278
>=5	63.83	181
Day/Night		
Day	36.06	5917
Night	71.87	848
Day of Week		
Weekday	39.34	6103
Weekend	51.7	662
Hour of Day		
Off-Peak Hour	45.3	4058
Peak Hour	33.44	2707

Table 3.3(b) Summary of Average Incident Duration Classified by Key Variables (cont'd)

Variables	Avg. Duration (minutes)	Frequency
Number of Lanes (One Direction)		
2	61.79	802
4	37.85	5727
8	34.02	236
No. of Lanes blocked (In Same Direction)		
0	35.21	2623
1	32.04	2656
2	60.84	976
3	71.58	342
>=4	77.46	168
No. of Lanes blocked (In Opposite Direction)		
0	39.41	6430
1	50.5	221
2	87.18	88
3	91.66	19
>=4	50.2	7
Total number of Lanes Blocked (Same+Opposite direction)		
0	34.1	2511
1	32.11	2632
2	59.37	1034
3	66.46	340
>=4	81.45	248
Shoulder Blockage		
No Blockage	38.84	2837
Is Blocked	41.79	3928

Table 3.3(c) Summary of Average Incident Duration Classified by Key Variables (cont'd)

Variables	Avg. Duration (minutes)	Frequency
Incident Nature		
Disabled Vehicle	22.47	1713
Collision-Property Damage (CPD)	35.73	2662
Collision-Personal Injury (CPI)	53.96	1971
Collision-Fatality (CF)	208.66	84
Others	50.25	335
CHART		
Not Involved	34.77	898
Involved	41.43	5867
Pavement Condition		
Unspecified	56.61	469
Dry	37.73	4864
Wet	44.95	977
Snow/Ice	44.61	447
Chemical Wet	50.68	8
Road Name		
I-895	28.93	137
I-495	30.75	2051
I-695	34.98	1252
I-95	36.67	946
US 50	36.89	510
MD 295	38.43	239
I-270	39.15	319
I-97	44.18	118
I-795	44.55	85
I-370	54.21	2
I-83	56.61	248
I-70	69.88	191
Others	72.41	597
US 1	89.71	45
I-68	182.88	25

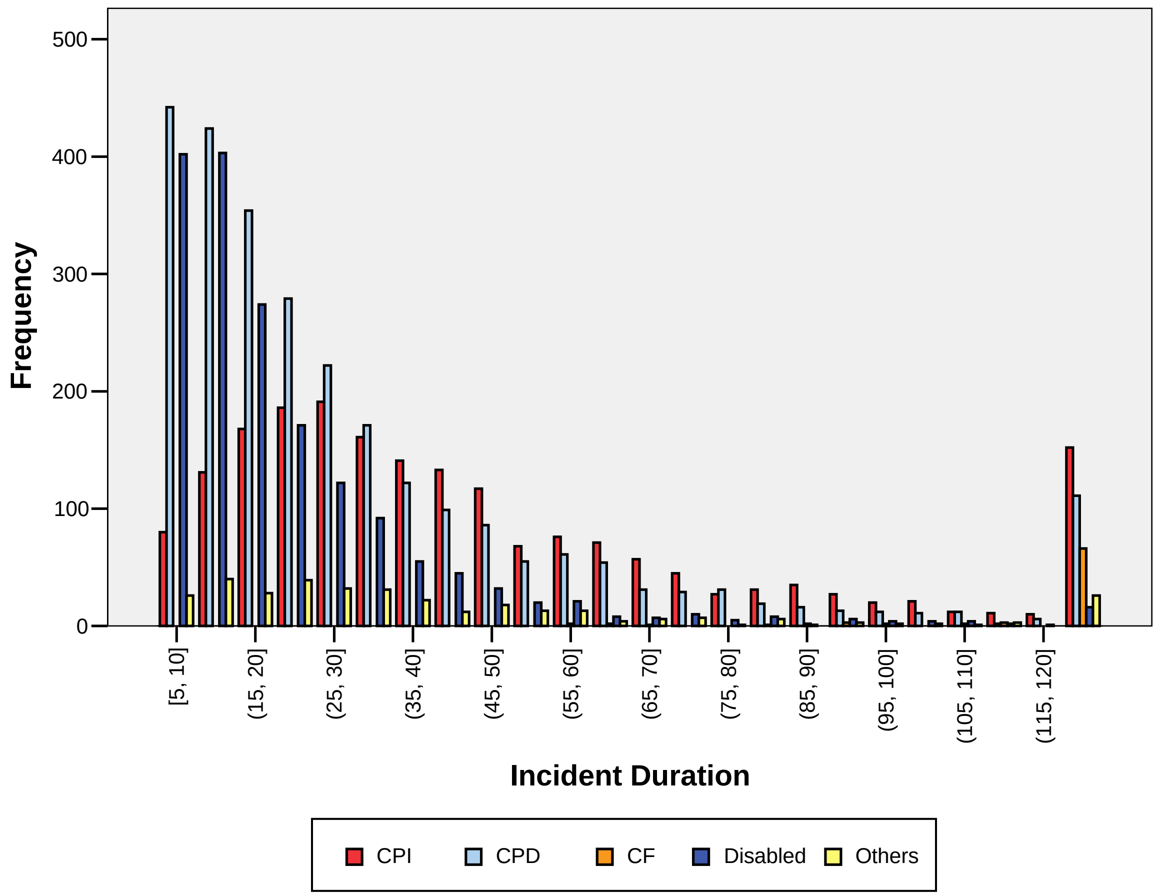


Figure 3.4 Distribution of Incident Duration Frequency by Each Incident Nature

Chapter 4: Methodology and Analysis

4.1 Introduction

This chapter explores several potential methods for developing an effective prediction model for the duration of incidents in Maryland. It begins with a discussion of the preliminary analyses with CART. Based on the findings from CART, this study has further developed a rule-based tree model (RBTM) in Section 4.3, along with its calibration procedures. All model structures, along with their performances and validations, are presented in Sections 4.4 to 4.8. Overall findings and conclusions are discussed in the last section.

4.2 Preliminary Analysis with CART

4.2.1 Basic Procedures of CART

CART, also known as C&RT, is a type of decision tree technique which was introduced and popularized by Breiman et al. (1984). This nonparametric statistical method first determines a sequence of if-then logic conditions developed based on an analysis of the relationships between the dependent and independent variables. Based on the set of logic conditions, it builds a classification tree for categorical dependent variables, and a regression tree for continuous dependent variables.

CART consists of four steps – tree building, stopping the tree building, pruning, and optimal tree selection. Using a learning data set, the optimal tree is built for the outcome and predictor variables. The test data set is required to validate the classification and decision rules.

In the tree-building step, the root node, including all data sets, is first split into two child nodes according to the best possible variable to split, called a splitter. The best splitter is used to maximize the average "purity" of the two child nodes. Among various available measures of purity, the most commonly used measure is the "Gini," followed by "Twoing" (Lewis, 2000). After splitting, each node, including the root node, is assigned a predicted outcome category, based on a function shown below.

Node is category i, if $\dfrac{C(j\,|\,i)\pi(i)N_i(t)}{C(i\,|\,j)\pi(j)N_j(t)} > \dfrac{N_i}{N_j}$ for all values of j,

where, $C(j|i)$ is the cost of classifying i as j,

$\pi(i)$ is the prior probability of i,

N_i is the number of category i in the data set,

and $N_i(t)$ is the number of category i in the node.

Procedures of node splitting and assigning for a predicted category are repeated for each node until it is impossible to continue.

To stop building a tree, at least one of the following criteria should be satisfied:

(1) Only one observation is left in each child node.

(2) The distributions of predictor variables for all observations within each child node are identical, which makes further splitting impossible.

(3) The maximum tree level, externally set by users, has been reached.

Usually, a tree created by the aforementioned procedures is likely to be overfit. That may result in making it difficult for users to read and interpret, and so the process of tree pruning is recommended. To prune the overfit tree, generally the method of "cost complexity" is used. In this method, the complexity parameter, α, is gradually increased during the pruning process; α is the measure of how much additional accuracy is needed to demand the additional complexity for the additional split (Lewis, 2000). As α increases, the tree gets simpler, with more nodes pruned. While pruning, the optimal tree is selected with the optimal value of α, so that the information in the training data set is well fit but not overfit (Lewis, 2000). Detailed discussions of CART are available in the literature (Breiman et al., 1984; Lewis, 2000; Yohannes and Hoddinott, 1999; Lemon et al., 2003).

4.2.2 Results and Findings from CART

Table 4.1 presents three different ways for proceeding with the design of the classification tree. The results and findings, based on the optimal trees developed for each type of dependent variable are summarized below.

1. Among 25 independent variables, the nature of the incident was selected as the first splitter to build a tree. The selected optimal trees show that incident durations for *Collision-Property Damage* and *Disabled Vehicles* were relatively short, since about 53 percent of these incidents had duration of between 5 and 20 minutes. On the other hand, incident durations for *Collision-Personal Injury*, *Fatality*, and *Others* were likely to be longer, because about 59 percent of these incidents distribute between 20 and 70 minutes. These relations are consistent with the frequency distribution of incident durations (see Figure 3.4 in Chapter 3).

Table 4.1 Summary of Dependent Variables Used for Design of the Classification Tree

Type of Dependent Variable	Number of Categories	Definition (Ranges of duration for each category)	Percentage (%)
Basic	9	1: [5, 10] mins 2: (10, 15] mins 3: (15, 20] mins 4: (20, 30] mins 5: (30, 45] mins 6: (45, 70] mins 7: (70, 90] mins 8: (90, 120] mins 9: > 120 mins	14 15 12 18 16 12 4 3 5
Recategorized DV[1] 1 (RCDV1)	3	Short: [5, 20] mins Middle: (20, 70] mins Long: > 70 mins	41 47 12
Recategorized DV[1] 2 (RCDV2)	4	Short: [5, 20] mins Middle: (20, 70] mins Long: (70, 120] mins Very Long: > 120 mins	41 47 7 5

[1] Dependent Variable

2. Without the information for classification costs and prior probabilities, each node is assigned to a predicted outcome category which has the highest frequency (i.e., the highest probability).

3. Based on the experimental results, the difference in tree performance between using the original independent variables and regrouped independent variables was trivial. Also, the CART algorithm itself has the ability to choose the most significant variable as the best splitter; it can also find the best regrouped categories within the selected variable.

4. Tables 4.2 to 4.4 summarize the prediction result for each tree. Table 4.2 shows that the tree (Tree 1) developed for nine categorized (*Basic*) dependent variables had an overall percentage of correct predictions of 30.2 percent. About 71 percent of the incidents with durations of between five and ten minutes were predicted

correctly. But the tree at this level could not predict correctly for incidents having durations of 70 to 90 minutes and 90 to 120 minutes. Trees developed for three categorized (*RCDV1*) dependent variable and for four categorized (*RCDV2*) dependent variable (Tree 3) reflect the similar trend, but achieve better levels of performance, with overall percentages of correct prediction of 63.5 and 63.1 percent for Tree 2 and Tree 3, respectively. Neither tree, however, would suffice for use in predicting incident durations exceeding 70 minutes. For example, Tree 2 predicted 22.8 percent correctly for incidents lasting longer than 70 minutes. In Tree 3, incident durations of 70 to 120 minutes were not predicted correctly at all, and this tree correctly predicted only 31.1 percent of incidents lasting for more than two hours.

Overall, CART performs quite well for short or middle ranges of incident duration, especially for those lasting between five and ten minutes. However, it does not provide satisfactory results for incidents of long duration (e.g., longer than one hour). Smith and Smith (2001) obtained similar results in their research, although their tree was developed to forecast the clearance time. The overall prediction accuracy of their classification tree was 58.47 percent, and they concluded that this accuracy level was not good enough for use in traffic incident management.

Table 4.2 Prediction Result of the Tree Developed for the 9-Categorized (*Basic*) Dependent Variable (Tree 1)

Observed	**Predicted**									
Incident Duration (mins)	[5,10]	(10,15]	(15, 20]	(20, 30]	(30, 45]	(45, 70]	(70, 90]	(90, 120]	>120	Percent Correct
[5, 10]	673	137	0	93	30	13	0	0	4	70.8%
(10, 15]	446	331	6	140	48	22	0	0	5	33.2%
(15, 20]	297	192	85	157	71	20	0	0	2	10.3%
(20, 30]	352	165	60	449	161	50	0	0	5	36.2%
(30, 45]	281	96	36	349	249	64	0	0	9	23.0%
(45, 70]	171	51	18	297	153	107	0	0	27	13.0%
(70, 90]	55	21	11	89	42	41	0	0	14	0.0%
(90, 120]	35	13	2	50	40	33	0	0	27	0.0%
>120	22	19	8	72	50	53	0	0	146	39.5%
Overall Correct Percentage	34.5%	15.2%	3.3%	25.1%	12.5%	6.0%	0.0%	0.0%	3.5%	30.2%

Table 4.3 Prediction Result of the Tree Developed for the 3-Categorized (*RCDV1*) Dependent Variable (Tree 2)

Observed	**Predicted**			
Incident Duration (mins)	short: [5, 20]	middle: (20, 70]	long: > 70	Percent Correct
short: [5, 20]	1998	761	13	72.1%
middle: (20, 70]	1000	2108	42	66.9%
long: > 70	138	513	192	22.8%
Overall Correct Percentage	46.4%	50.0%	3.7%	63.5%

Table 4.4 Prediction Result of the Tree Developed for the 4-Categorized (*RCDV2*) Dependent Variable (Tree 3)

Observed	**Predicted**				
Incident Duration (mins)	short: [5, 20]	middle: (20, 70]	long: (70, 120]	very long: >120	Percent Correct
short: [5, 20]	1985	777	0	10	71.6%
middle: (20, 70]	961	2168	0	21	68.8%
long: (70, 120]	92	354	0	27	0.0%
very long: >120	31	224	0	115	31.1%
Overall Correct Percentage	45.4%	52.1%	0.0%	2.6%	63.1%

4.3 Procedures for a Rule-Based Tree Model (RBTM)

From the outcome of CART, it is clear that the incident nature is the most significant variable for classification of incident durations. Based on this finding, along with the other analysis results from CART discussed previously, this study has redesigned a classification tree, called a rule-based tree model (RBTM), using the following procedures. Note that incident durations, which were grouped into five-minute intervals, are used in this approach.

Step 1: Set the regrouped incident nature as the first splitter.

As discussed in Chapter 3 (see Table 3.2), incidents with *Debris, Vehicle Fire, Police Activity, Emergency Road Work,* and *Off-Road Work* do not show statistically significant differences in their durations. In addition, the number of records available for incidents with *Police Activity, Emergency Road Work,* and *Off-Road Work* is somewhat small for developing a separate model. Thus, the regrouped incident nature was considered as a more appropriate splitter than the original one.

Step 2: Determine the next splitter for each node.

This step generates a cross-tabulation table (Hill and Lewicki, 2005) to determine the next splitter for each node. That table can display the number of cases in each category defined by two or more specified variables. For each independent and dependent variable (e.g., incident durations), this step creates a cross-tabulation table, along with a bar chart to show the distribution of frequency for different categories of the independent variable potentially associated with the incident durations. Then, the independent variable that exhibits the most different kind of distribution in different categories is selected as the next splitter.

Step 3: <u>Split the node based on the determined splitter in each category.</u>

The focus of this step is to convert each splitting node into *If-then; Else-then* statements; these statements will constitute the set of rules for determining the incident duration for the node.

Step 4: <u>Assign the predicted incident duration range for each split node.</u>

This step determines the best representative range of incident durations for each node. To achieve this requires first searching for an interval that is less than or equal to 30 minutes and which covers at least 70 percent of all cases within a node. If no such interval exists within the node, then the shortest interval covering at least 60 percent of all cases within the node is assigned as the predicted incident duration for that node.

Step 5: <u>Repeat Steps 2 to 4 for all nodes until the predetermined criteria for stopping the tree growth are satisfied.</u>

When a node satisfies one of the following criteria, one can stop the tree at that node:

1. No independent variable is available as a splitter.
2. Only one observation is left in a node.

To evaluate the performance of rules for each node, this study adopted the concepts of ***support*** and ***confidence*** developed for Associate Rules (Hill and Lewicki, 2005). The ***support*** for the rule refers to the number of cases satisfying the *If-Then* rule. The ***confidence*** of the rule is defined as the ratio of the number of cases satisfying the *If-*

Then rule (i.e., the ***support***) to the number of cases satisfying the *If* statement only. The indicator of ***confidence*** is conceptually the same as the conditional probability of the *Then* statement, given the *If* statement of the rule.

Based on the findings through the aforementioned model development procedure, it is clear that the second splitter is *County*, which is a spatial factor. After splitting the data set by *County*, one can repeat the same procedures to complete the RBTM for each *County* of each incident nature. Due to the constraints of samples, this study analyzed only the data from Montgomery County. Figure 4.1 shows the structure of the RBTM.

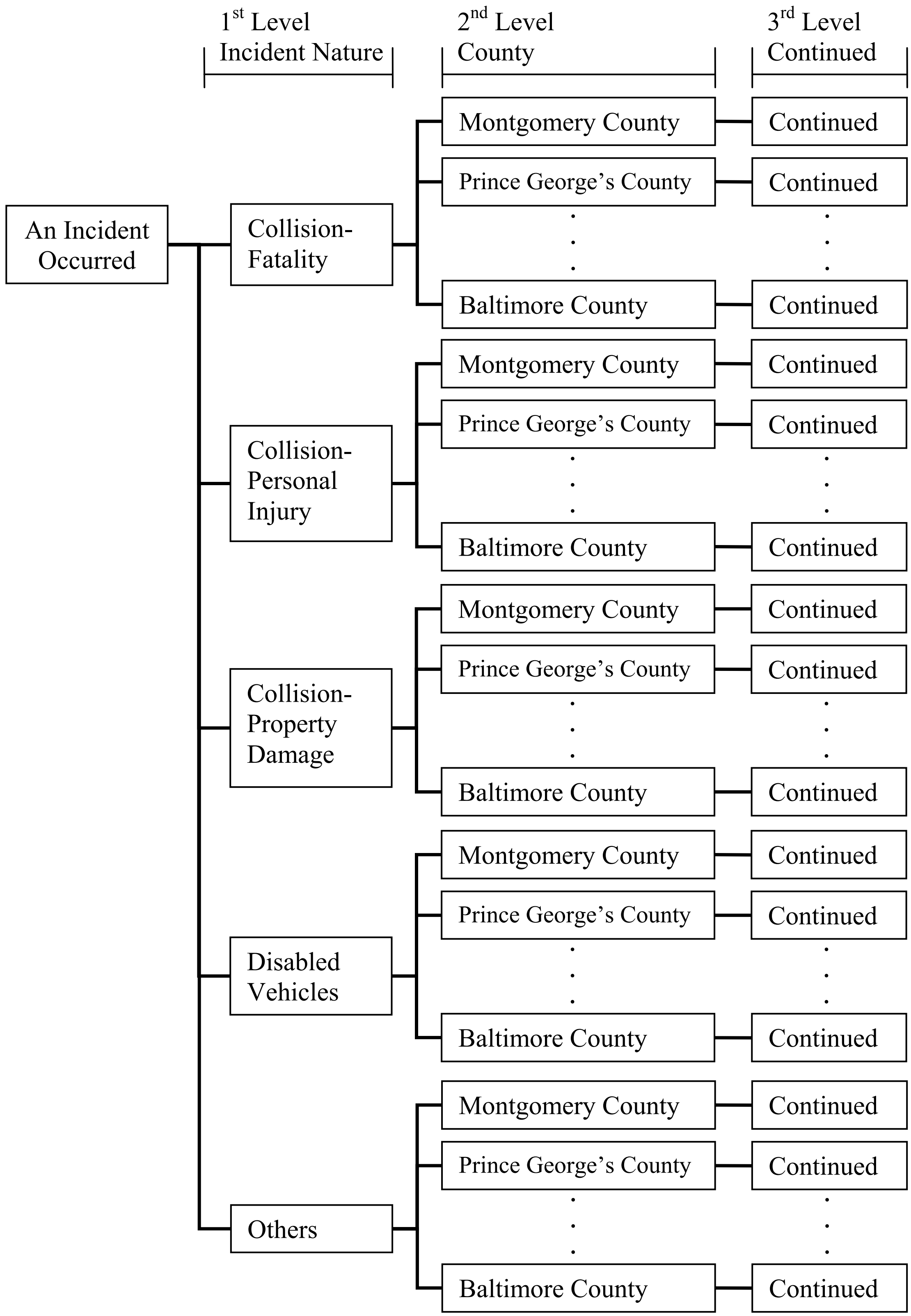


Figure 4.1 The Structure of Rule-Based Tree Models

4.4 The Rule-Based Tree Model for Incident Nature of Collision-Fatality (CF)

4.4.1 The Tree Structure

For those incidents resulting in *Collision-Fatality (CF)*, their distributions over 300 minutes are scattered over a wide range (300 to 1500 minutes) of durations, while the distribution in the range of 60 to 300 minutes is condensed and nearly symmetric (see Figure 4.2). Most cases lasting over 300 minutes occurred on roads which are out of scope for this study, and about 78 percent of those cases show a ratio of blocked lanes in the same direction greater than or equal to 0.5. This means that those incidents resulted in an extreme level of severity. In addition, about 68 percent of these occurred between midnight and 6 AM. One extreme case involved 73 vehicles, including five tractor-trailers, and it resulted in the longest duration, 1501 minutes. Since these cases require special response and operational efforts, this study has excluded them from the model development.

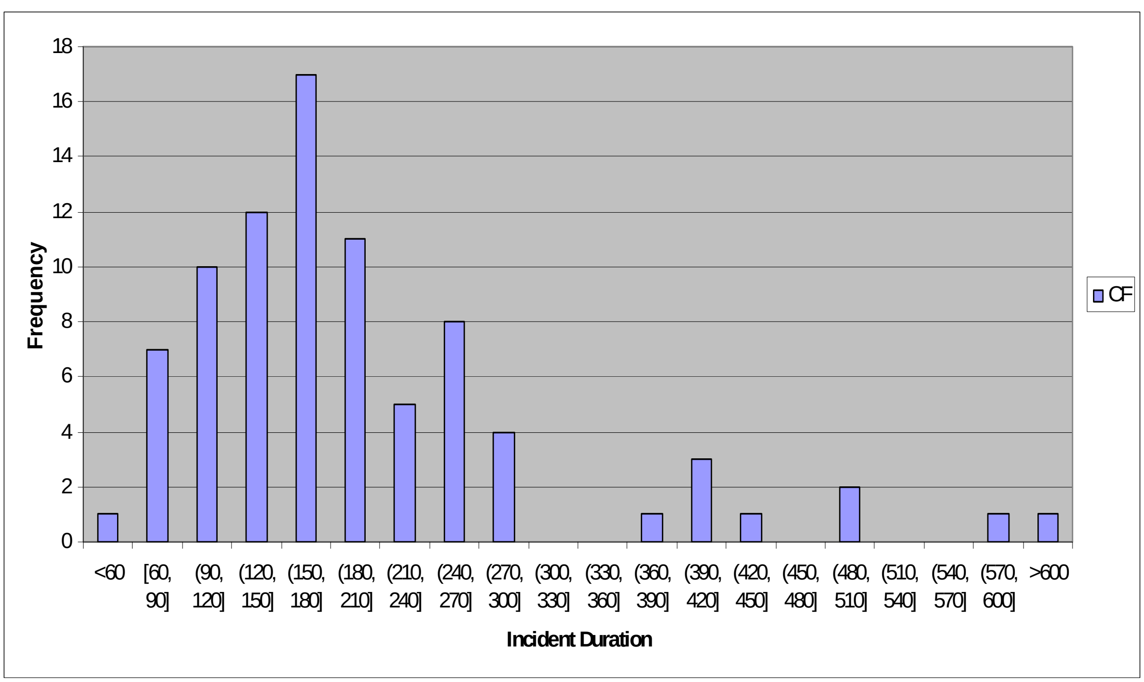


Figure 4.2 Distribution of Frequencies for Incidents resulting in *CF*

The RBTM for fatality incidents consists of the following rules. Unlike the other incident natures, fatality incidents do not include *County* as the first splitter, due to the deficiency of sample size. Hereafter, “*IncD*” stands for incident duration in minutes.

1st Level

Rule 1: **If** *Weekend*, **then** Rule 2-a; **Else** Rule 2-b

2nd Level

Rule 2-a: **If** *Pickup Van* is not involved, **then** Rule 3-a; **Else** Rule 3-b

Rule 2-b: **If** *Tractor-Trailer* is not involved, **then** Rule 3-c; **Else** Rule 3-d

At this level, heavy vehicles — such as pickup vans, single-unit trucks and tractor-trailers — had a noticeably strong effect on the resulting durations of incidents involving fatalities.

3rd Level

Rule 3-a: **If** *Shoulder* is not blocked, **then** 180<*IncD* <=200; **Else** Rule 4-a

Rule 3-b: **If** *Shoulder* is not blocked, **then** 180<*IncD* <=200; **Else** 160<*IncD* <=180

Rule 3-c: **If** occurs during *Off-Peak Hours*, **then** Rule 4-b; **Else** Rule 4-c

Rule 3-d: **If** *No. of vehicles involved* < 4, **then** Rule 4-d; **Else** 260<*IncD* <=300

This level, as well as the following levels, captures the effect of shoulder blockage on the durations of incidents that involve fatalities. When a shoulder lane was blocked, the incident duration was likely to be shorter than without such a blockage, and this is not consistent with the average incident duration classified by shoulder blockage presented in Chapter 3 (see Table 3.3(b)). This may be attributed to the fact that a shoulder lane blockage generally provides a wider working space for the incident response units to better perform the necessary tasks.

<u>4th Level</u>

Rule 4-a: **If** occurs in the *Daytime*, **then** Rule 5-a; **Else** 160<*IncD* <=180

Rule 4-b: **If** *Pickup Van* is not involved, **then** Rule 5-b; **Else** Rule 5-c

Rule 4-c: **If** *No. of vehicles involved* = 1, **then** Rule 5-d; **Else** Rule 5-e

Rule 4-d: **If** *No. of blocked lanes in the same direction* <= 1, **then** Rule 5-f

; **Else** Rule 5-g

<u>5th Level</u>

Rule 5-a: **If** *Ratio of blocked lanes in the same direction*<=0.5, **then** 260<*IncD* <=280

; **Else** 80<*IncD* <=100

Rule 5-b: **If** *Shoulder* is not blocked, **then** Rule 6-a; **Else** Rule 6-b

Rule 5-c: **If** *Shoulder* is not blocked, **then** Rule 6-c; **Else** Rule 6-d

Rule 5-d: **If** *Road* is I-695, I-95, MD 50 or I-97, **then** 80<*IncD* <=140

; **Else if** *Road* is I-795, I-83, I-70, I-370, US 1 or others, **then** 140<*IncD* <=160

Rule 5-e: **If** *No. of lanes in the same direction* = 2, **then** 60<*IncD* <=80

; **Else** Rule 6-e

Rule 5-f: **If** occurs in the *Daytime*, **then** 180<*IncD* <=240; **Else** 240<*IncD* <=300

Rule 5-g: **If** *Ratio of blocked lanes in the same direction* <=0.5, **then** Rule 6-f

; **Else** Rule 6-g

At this level, a noticeable relation was found to exist between *Road* (i.e., the highway segment) and incident durations.

6th Level

Rule 6-a: **If** *No. of blocked lanes in the same direction* <=2, **then** Rule 7-a

; **Else** 55<*IncD* <=80

Rule 6-b: **If** *Pavement* is wet, **then** Rule 7-b; **Else** Rule 7-c

Rule 6-c: **If** occurs in the *Daytime*, **then** 220<*IncD* <=240; **Else** 280<*IncD* <=300

Rule 6-d: **If** occurs in the *Daytime*, **then** 120<*IncD* <=180; **Else** 160<*IncD* <=200

Rule 6-e: **If** *Pickup Van* is not involved, **then** Rule 7-d; **Else** 180<*IncD* <=200

Rule 6-f: **If** *Shoulder* is not blocked, **then** 240<*IncD* <=260; **Else** Rule 7-e

Rule 6-g: **If** occurs in the *Daytime*, **then** Rule 7-f; **Else** 140<*IncD* <=160

At this level, clearly, the duration of fatality-related incidents occurring in the daytime was likely to be shorter than those at night. One may attribute this outcome to the fact that the number of response units available at night is less than during the daytime.

7th Level

Rule 7-a: **If** occurs in the *Daytime*, **then** Rule 8-a; **Else** Rule 8-b

Rule 7-b: **If** occurs in the *Daytime*, **then** 60<*IncD* <=120; **Else** 140<*IncD* <=160

Rule 7-c: **If** *Ratio of blocked lanes in the same direction* <= 0.5,

then 160<*IncD* <=180; **Else** 100<*IncD* <=160

Rule 7-d: **If** *Ratio of blocked lanes in the same direction* <= 0.5,

then 220<*IncD* <=260; **Else** 180<*IncD* <=200

Rule 7-e: **If** *Single-Unit Truck* is not involved, **then** 60<*IncD* <=180

; **Else** 200<*IncD* <=220

Rule 7-f: **If** *Ratio of blocked lanes in the same direction* <=0.75,

then 180<*IncD* <=200; **Else** 80<*IncD* <=140

At this level, one can observe that, as the number of lanes blocked in the same direction increased, the incident duration generally decreased. This can be explained by the fact that more blocked lanes during operations may provide a wider working space for incident response units to efficiently clear an incident.

<u>8th Level</u>

Rule 8-a: **If** *No. of vehicles involved*=1, **then** 120<*IncD* <=140

; **Else** 180<*IncD* <=200

Rule 8-b: **If** *Pavement* is wet, **then** 140<*IncD* <=160; **Else** 180<*IncD* <=260

One interesting result seen at this level is about the pavement conditions. In general, the wet pavement condition reflects inclement weather, which tends to increase the number of incidents and incident durations. However, in the study data set for *CF*, the relationship between wet pavement and incident duration was opposite to what was expected. This data set shows that incident durations in the wet pavement condition were likely to be shorter than those in the non-wet pavement condition, and this observation is consistent with the results in Table 4.5, Average Fatality Incident Durations for Different Pavement Conditions. This can be explained by the fact that, in inclement weather, incident response units are on alert and more staff is available for emergency medical services (EMS).

Table 4.5 Average Fatality Incident Durations for Different Pavement Conditions

Pavement Condition	Avg. Incident Duration (mins)	Standard Deviation of Incident Duration	Frequency
Unspecific	184.70	82.43	8
Dry	173.00	57.65	52
Wet	127.66	40.67	14
Snow/Ice	173.03	N/A	1
Chemical Wet	N/A	N/A	N/A

4.4.2 Performance and Validation Results

Tables 4.6 and 4.6(a) summarize the estimation results of RBTMs using the data set collected from years 2003 to 2005.

While most samples for other incident natures are distributed within two hours (i.e., 5 to 120 minutes), samples for *CF* are scattered between 60 and 300 minutes. In addition, the sample size was very small (i.e., 84), although such incidents had been collected for three years. Thus, the ranges of incident durations assigned at many of the terminal nodes (highlighted cells) in RBTMs (*Then* statement in rules) are likely to be wider (e.g., about 60 minutes) than those for other incident natures (e.g., about 25 minutes in *Collision-Personal Injury*). Although the predicted incident durations fell in a relatively wide range, the ***confidences*** for most of the rules are acceptable.

Table 4.6 Summary of Estimation Results for the RBTM for *CF* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases	Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases
1	Rule 1	(80, 200]	93.75	15	16	(60, 200]	71.19	42	59
2	Rule 2-a	(160, 200]	60.00	6	10	(80, 180]	100.00	6	6
3	Rule 2-b	(100, 200]	75.00	30	40	(180, 300]	63.16	12	19
4	Rule 3-a	(180, 200]	100.00	2	2	(80, 180]	87.50	7	8
5	Rule 3-b	(180, 200]	100.00	1	1	(160, 180]	80.00	4	5
6	Rule 3-c	(100, 240]	78.57	22	28	(120, 260]	75.00	9	12
7	Rule 3-d	(80, 260]	88.24	15	17	(260, 300]	100.00	2	2
8	Rule 4-a	(80, 100]	66.67	2	3	(160, 180]	80.00	4	5
9	Rule 4-b	(100. 200]	78.68	14	19	(120, 240]	88.89	8	9
10	Rule 4-c	(80, 160]	100.00	5	5	(180, 260]	85.71	6	7
11	Rule 4-d	(220, 300]	66.67	4	6	(80, 200]	72.73	8	11
12	Rule 5-a	(260, 280]	100.00	1	1	(80, 100]	100.00	2	2
13	Rule 5-b	(120, 200]	62.50	5	8	(100, 180]	81.82	9	11
14	Rule 5-c	(220, 240]	66.67	2	3	(120, 180]	83.33	5	6
15	Rule 5-d	(80, 140]	100.00	3	3	(140, 160]	100.00	2	2
16	Rule 5-e	(60, 80]	100.00	1	1	(180, 160]	100.00	6	6
17	Rule 5-f	(180, 240]	100.00	3	3	(240, 300]	100.00	3	3
18	Rule 5-g	(160, 260]	75.00	3	4	(80, 200]	100.00	7	7
19	Rule 6-a	(120, 160]	100.00	6	6	(55, 80]	100.00	2	2
20	Rule 6-b	(60, 160]	100.00	4	4	(100, 180]	100.00	7	7
21	Rule 6-c	(220, 240]	100.00	2	2	(280, 300]	100.00	1	1
22	Rule 6-d	(120, 180]	100.00	4	4	(160, 200]	100.00	2	2
23	Rule 6-e	(220, 260]	80.00	4	5	(180, 200]	100.00	1	1
24	Rule 6-f	(240, 260]	100.00	1	1	(60, 220]	100.00	3	3
25	Rule 6-g	(80, 140]	66.67	4	6	(140, 160]	100.00	1	1
26	Rule 7-a	(120, 240]	88.89	8	9	(140, 200]	66.67	2	3
27	Rule 7-b	(60, 120]	100.00	3	3	(140, 160]	100.00	1	1
28	Rule 7-c	(160, 180]	100.00	2	2	(100, 160]	100.00	5	5
29	Rule 7-d	(220, 260]	100.00	4	4	(180, 200]	100.00	1	1
30	Rule 7-e	(60, 180]	100.00	2	2	(200, 220]	100.00	1	1
31	Rule 7-f	(180, 200]	100.00	1	1	(80, 140]	80.00	4	5

Table 4.6(a) Summary of Estimation Results for the RBTM for *CF* Incidents Occurring in Montgomery County (cont'd)

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf. [1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf. [1] (%)**	**Support**	**Total Cases**
32	Rule 8-a	(120, 140]	100.00	2	2	(180, 200]	100.00	1	1
33	Rule 8-b	(140, 160]	100.00	1	1	(180, 260]	100.00	2	2

Note: 1. Sample size is 75.
2. Highlighted cells are terminal nodes in the RBTM.
[1] Conf. stands for ***confidence***.

However, the overall validation results shown in Tables 4.7 and 4.7(a), using a data set collected in year 2006 (sample size is 70), indicate that only two nodes show a ***confidence*** over 70 percent. Many validation results of the nodes that appear close to terminal nodes show a low ***confidence***. Some of the terminal nodes (highlighted cells) cannot be validated, since no records in the validation data set satisfy *If* conditions given in those nodes. Models for *CF* show unsatisfactory performance, even with the larger data set for model development. Hence, exploring some supplemental models and additional explanatory variables (e.g., the number of fatalities, severity of injuries, or driver condition) seem essential for further capturing the relations between incident duration and incidents involving fatalities.

The supplemental models for incidents resulting in fatality are discussed in detail in Chapter 5.

Table 4.7 Summary of Validation Results for the RBTM for *CF* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases	Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases
1	Rule 1	(80, 200]	60.00	12	20	(60, 200]	56.00	28	50
2	Rule 2-a	(160, 200]	29.41	5	17	(80, 180]	0.00	0	4
3	Rule 2-b	(100, 200]	42.50	17	40	(180, 300]	80.00	8	10
4	Rule 3-a	(180, 200]	33.33	1	3	(80, 180]	64.29	9	14
5	Rule 3-b	(180, 200]	0.00	0	2	(160, 180]	0.00	0	2
6	Rule 3-c	(100, 240]	54.84	17	31	(120, 260]	66.67	6	9
7	Rule 3-d	(80, 260]	66.67	6	9	(260, 300]	50.00	1	2
8	Rule 4-a	(80, 100]	0.00	0	6	(160, 180]	10.00	1	10
9	Rule 4-b	(100. 200]	50.00	13	26	(120, 240]	60.00	3	5
10	Rule 4-c	(80, 160]	25.00	1	4	(180, 260]	60.00	3	5
11	Rule 4-d	(220, 300]	66.67	4	6	(80, 200]	42.86	3	7
12	Rule 5-a	(260, 280]	0.00	0	2	(80, 100]	0.00	0	4
13	Rule 5-b	(120, 200]	40.00	4	10	(100, 180]	50.00	8	16
14	Rule 5-c	(220, 240]	0.00	0	2	(120, 180]	66.67	2	3
15	Rule 5-d	(80, 140]	50.00	1	2	(140, 160]	0.00	0	2
16	Rule 5-e	(60, 80]	0.00	0	3	(180, 160]	0.00	0	2
17	Rule 5-f	(180, 240]	100.00	1	1	(240, 300]	0.00	1	0
18	Rule 5-g	(160, 260]	0.00	0	1	(80, 200]	50.00	3	6
19	Rule 6-a	(120, 160]	40.00	4	10	(55, 80]	N/A	N/A	0
20	Rule 6-b	(60, 160]	0.00	0	1	(100, 180]	53.33	8	15
21	Rule 6-c	(220, 240]	0.00	0	1	(280, 300]	0.00	0	1
22	Rule 6-d	(120, 180]	N/A	N/A	0	(160, 200]	0.00	0	3
23	Rule 6-e	(220, 260]	N/A	N/A	0	(180, 200]	0.00	0	2
24	Rule 6-f	(240, 260]	N/A	N/A	0	(60, 220]	0.00	0	1
25	Rule 6-g	(80, 140]	50.00	1	2	(140, 160]	25.00	1	4
26	Rule 7-a	(120, 240]	75.00	3	4	(140, 200]	16.67	1	6
27	Rule 7-b	(60, 120]	0.00	0	1	(140, 160]	N/A	N/A	0
28	Rule 7-c	(160, 180]	0.00	0	4	(100, 160]	0.00	0	11
29	Rule 7-d	(220, 260]	N/A	N/A	0	(180, 200]	N/A	N/A	0
30	Rule 7-e	(60, 180]	0.00	0	1	(200, 220]	N/A	N/A	0
31	Rule 7-f	(180, 200]	N/A	N/A	0	(80, 140]	0.00	0	2

Table 4.7(a) Summary of Validation Results for the RBTM for *CF* Incidents Occurring in Montgomery County (cont'd)

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases	Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases
32	Rule 8-a	(120, 140]	50.00	1	2	(180, 200]	0.00	0	2
33	Rule 8-b	(140, 160]	100.00	1	1	(180, 260]	0.00	0	5

Note: 1. Sample size is 64.
2. Highlighted cells are terminal nodes in the RBTM.
[1] Conf. stands for ***confidence***.

4.5 The RBTM for Incident Nature of Collision - Personal Injury (CPI)

4.5.1 The Tree Structure

The following rules construct the RBTM for incidents causing personal injuries, based on the data from Montgomery County.

1st Level

Rule 1: **If** *Total no. of blocked lanes* <=2, **then** Rule 2-a; **Else** Rule 2-b

2nd Level

Rule 2-a: **If** *No. of blocked lanes in the opposite direction*=0, **then** Rule 3-a

; **Else** Rule 3-b

Rule 2-b: **If** *Total no. of blocked lanes* = 3, **then** Rule 3-c; **Else** Rule 3-d

At the first and second levels, the number of blocked lanes is selected as a significant factor.

3rd Level

Rule 3-a: **If** *Single-Unit Truck* is not involved, **then** Rule 4-a; **Else** Rule 4-b

Rule 3-b: **If** *No. of blocked lanes in the opposite direction*=1, **then** Rule 4-c

; **Else** 10< *IncD* <=35

Rule 3-c: **If** *No. of Tractor-Trailer*=0, **then** Rule 4-d

; **Else if** *No. of Tractor-Trailer*=1, **then** Rule 4-e

; **Else if** *No. of Tractor-Trailer*>=2, **then** 75< *IncD* <=100

Rule 3-d: **If** *Pickup Van* is not involved, **then** 5<=*IncD* <=45; **Else** 30< *IncD* <=70

At the third level, involvement of heavy vehicles (i.e., single-unit trucks, pickup vans, and tractor-trailers) showed a strong relation to incident duration.

<u>4th Level</u>

Rule 4-a: **If** *Pickup Van* is not involved, **then** Rule 5-a; **Else** Rule 5-b

Rule 4-b: **If** *No. of Single-Unit Trucks*=1, **then** Rule 5-c; **Else** Rule 5-d

Rule 4-c: **If** *Road* is I-495, **then** 10< *IncD* <=30; **Else** 20 < *IncD* <=40

Rule 4-d: **If** *Road* is I-495, **then** Rule 5-e; **Else** 5<= *IncD* <=45

Rule 4-e: **If** *Shoulder* is blocked, **then** 15< *IncD* <=35

; **Else if** occurs in the *Daytime* 15< *IncD* <=40

At this level, the durations of incidents occurring on I-495 fell into a different range than those on other roads.

<u>5th Level</u>

Rule 5-a: **If** *Tractor-Trailer* is not involved, **then** Rule 6-a; **Else** Rule 6-b

Rule 5-b: **If** *No. of Pickup Van* =1, **then** Rule 6-c

; **Else if** *No. of Pickup Van* =2, **then** Rule 6-d

; **Else** 15 <*IncD*<=35

Rule 5-c: **If** *Pickup Van* is not involved, **then** Rule 6-e; **Else** 25 <*IncD*<=50

Ruel 5-d: **If** *Pickup Van* is not involved, **then** 35 <*IncD*<=40

;**Else** 185 <*IncD*<=190

Rule 5-e: **If** *Pavement* is not wet, **then** 15< *IncD* <=45

; **Else** Rule 6-f

Overall, this level selects the information regarding pickup van involvement as a key splitter.

6th Level

Rule 6-a: **If** *No. of Vehicles* involved=1, **then** Rule 7-a; **Else** Rule 7-b

Rule 6-b: **If** *Pavement* is dry, **then** Rule 7-c; **Else** 15<*IncD*<=25

Rule 6-c: **If** *Total no. of lanes blocked*=0, **then** Rule 7-d

; **Else if** *Total no. of lanes blocked*=1, **then** Rule 7-e

; **Else** Rule 7-f

Rule 6-d: **If** *Road*= I-270 N, **then** 40 <*IncD*<=65

; **Else if** *Road*= I-270 S, **then** 25 <*IncD*<=40

; **Else if** *Road*= I-495, **then** Rule 7-g

Rule 6-e: **If** occurs during *Off-Peak Hours,* **then** 25 <*IncD*<=45

; **Else** 30 <*IncD*<=50

Rule 6-f: **If** *Pickup van* is not involved, **then** 5<= *IncD* <=35; **Else** 20<*IncD*<= 50

7th Level

Rule 7-a: **If** occurs during *Off-Peak Hours*, **then** Rule 8-a; **Else** Rule 8-b

Rule 7-b: **If** *No. of Vehicles involved*=2, **then** Rule 8-c

; **Else if** *No. of Vehicles involved* =3, **then** Rule 8-d

; **Else** 20<*IncD*<=40

Rule 7-c: **If** *No. of Vehicles involved* <=2, **then** 5<=*IncD*<=25; **Else** 15<*IncD*<=40

Rule 7-d: **If** *Shoulder* is not blocked, **then** Rule 8-e; **Else** Rule 8-f

Rule 7-e: **If** *Shoulder* is not blocked, **then** 15 <*IncD*<=40; **Else** 10 <*IncD*<=45

Rule 7-f: **If** occurs during *Off-Peak Hours*, **then** Rule 8-g; **Else** 15 <*IncD*<=35

Rule 7-g: **If** *Tractor-Trailer* is not involved, **then** 10 <*IncD*<=30

; **Else** 2 hours <*IncD*<=3.5 hours

At this level, as the number of vehicles involved with an incident increased, the incident duration was likely to increase.

<u>8th Level</u>

Rule 8-a: **If** *Pavement* is dry, **then** Rule 9-a

; **Else if** *Pavement* is wet, **then** 10<*IncD*<=30

; **Else if** *Pavement* is snow/ice, **then** 40<*IncD*<=55

Rule 8-b: **If** *Pavement* is dry, **then** 10<*IncD*<=25; **Else** 15<*IncD*<=35

Rule 8-c: **If** *Weekend*, **then** 15<*IncD*<=30

; **Else** Rule 9-b

Rule 8-d: **If** *Road* is I-495, **then** Rule 9-c; **Else** Rule 9-d

Rule 8-e: **If** occurs during *Off-Peak Hours,* **then** 40 <*IncD*<=65

; **Else** 5 <=*IncD*<=25

Rule 8-f: **If** occurs during *Off-Peak Hours,* **then** 5 <=*IncD*<=25

; **Else** 25 <*IncD*<=45

Rule 8-g: **If** *Pavement* is dry, **then** 15 <*IncD*<=45 ; **Else** 25 <*IncD*<=45

At this level, one can observe that incidents occurring in dry pavement conditions were likely to be shorter than those in other conditions, as expected. Also noticeable is the effect of *Off-Peak Hours* on incident duration differs within the subsets. For example, in Rule 8-e, incidents occurring during off-peak hours resulted in shorter durations, while in Rule 8-f, they resulted in longer durations.

<u>9th Level</u>

Rule 9-a: **If** *Shoulder* is not blocked, **then** Rule 10-a; **Else** Rule 10-b

Rule 9-b: **If** *Pavement* is dry, **then** Rule 10-c; **Else** Rule 10-d

Rule 9-c: **If** *Shoulder* is not blocked, **then** 45<*IncD*<=60; **Else** 35<*IncD*<=55

Rule 9-d: **If** *Ratio of blocked lanes in the same direction* < 0.5, **then** 15<*IncD*<=40

; **Else** 5<=*IncD*<=15

Note that, at this level, information regarding a lane blockage, including shoulder lanes, became a significant factor in determining incident durations.

<u>10th Level</u>

Rule 10-a: **If** *Number of Lanes*=4, **then** 5 <*IncD* <= 20; **Else** 35<*IncD*<=50

Rule 10-b: **If** *Road* is I-270, **then** 20<*IncD*<=30,

; **Else if** *Road* is I-495, **then** 10<*IncD*<=35

Rule 10-c: **If** occurs during *Off-Peak Hours*, **then** 5<=*IncD*<=30; **Else** 10<*IncD*<=35

Rule 10-d: **If** occurs during *Off-Peak Hours*, **then** 15<*IncD*<=40; **Else** 10<*IncD*<=35

At this level, the *Peak Hour* factor shows a different degree of influence in different subsets. With Rule 10-c, the duration of incidents occurring during peak hours was likely to be longer than that during off-peak hours, and vice versa with Rule 10-d.

To complete the RBTM for incidents caused by collisions with personal injury, this study has built the tree up to the tenth level. This reflects the complexity of predicting the durations of various types of incidents.

4.5.2 Performance and Validation Results

As shown in Tables 4.8 and 4.8(a), the overall performance results for this model are quite satisfactory, even with the validation data set.

However, with Rules 3-d and 4-d, the predicted range of incident durations was over 30 minutes, with unsatisfactory ***confidences***, which are lower than 70 percent. Therefore, a supplemental model was needed. Due to the limited sample size, the supplemental model has been developed with the sub-data set that was used for developing Rule 2-b. Similarly, rules for 6-b, 8-c, 8-d, 10-a, and 10-b demonstrate a low ***confidence,*** i.e., a wide range of predicted incident duration. Thus, the sub-data set including all these cases (i.e., a subset for Rule 5-a) was used to develop a separate supplemental model.

Lastly, another supplemental model has been developed using a subset satisfying Rule 5-b, since this subset includes rules with unsatisfactory results, such as Rules 7-e, 7-g, and 8-g.

Table 4.8 Summary of Estimation Results for the RBTM for *CPI* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf. [1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf. [1] (%)**	**Support**	**Total Cases**
1	Rule 1	(10, 50]	83.17	257	309	[5, 45]	65.30	64	98
2	Rule 2-a	[5, 50]	88.00	257	292	(10, 40]	94.10	16	17
3	Rule 2-b	[5, 45]	66.67	40	60	[5, 45]	63.16	24	38
4	Rule 3-a	(10, 50]	81.92	222	271	(25, 50]	80.95	17	21
5	Rule 3-b	(10, 30]	84.61	11	13	(10,35]	75.00	3	4
6	Rule 3-c	(15, 45]	60.00	30	50	(15 ,40]	71.40	5	7
						(75,100]	100.00	3	3
7	Rule 3-d	[5, 45]	72.22	13	18	(30,70]	60.00	12	20
8	Rule 4-a	[5, 50]	89.00	168	189	(10 ,45]	76.83	63	82
9	Rule 4-b	(25, 50]	84.21	16	19	N/A	N/A	N/A	2
10	Rule 4-c	(10, 30]	100.00	9	9	(20,40]	100.00	4	4
11	Rule 4-d	(15, 45]	67.87	19	28	[5,45]	63.64	14	22
12	Rule 4-e	(15, 35]	80.00	4	5	(15,40]	83.30	5	6
13	Rule 5-a	[5, 40]	77.53	138	178	[5, 25]	63.64	7	11
14	Rule 5-b	[5, 40]	74.07	40	54	(15, 50]	68.18	15	22
						(15 ,35]	66.70	4	6
15	Rule 5-c	(25, 45]	69.23	9	13	(25 ,50]	100.00	6	6
16	Rule 5-d	(35 ,40]	100.00	1	1	(185 ,190]	100.00	1	1
17	Rule 5-e	(15,45]	75.00	6	8	(20, 40]	70.00	14	20
18	Rule 6-a	[5, 35]	70.91	39	55	(10,40]	77.20	44	57
19	Rule 6-b	[5, 25]	62.50	5	8	(15, 25]	66.67	2	3
20	Rule 6-c	[5, 45]	88.24	15	17	(10, 40]	81.25	13	16
						(15, 45]	80.95	17	21
21	Rule 6-d	(40 ,65]	80.00	4	5	(25 ,40]	80.00	4	5
						(10, 50]	83.33	10	12
22	Rule 6-e	(25 ,45]	87.50	7	8	(30 ,50]	60.00	3	5
23	Rule 6-f	[5, 35]	91.67	11	12	(20, 50]	75.00	6	8
24	Rule 7-a	[5, 40]	70.45	31	44	(10, 25]	72.73	8	11
25	Rule 7-b	[5, 40]	83.75	67	80	(10, 50]	76.00	19	25
						(20, 40]	83.30	10	12
26	Rule 7-c	[5, 25]	80.00	4	5	(15, 40]	100.00	3	3
27	Rule 7-d	[5, 45]	81.82	9	11	(20, 45]	66.67	4	6
28	Rule 7-e	(15 ,40]	100.00	8	8	(10, 45]	87.50	7	8

Table 4.8(a) Summary of Estimation Results for the RBTM for *CPI* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases	Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases
29	Rule 7-f	(15, 45]	73.33	11	15	(15, 35]	83.33	5	6
30	Rule 7-g	(10 ,30]	80.00	8	10	2 ~ 3.5 hrs	100.00	2	2
31	Rule 8-a	(20, 50]	73.33	22	30	(10, 30]	80.00	8	10
						(40, 55]	100.00	3	3
32	Rule 8-b	(10,25]	83.30	5	6	(15, 35]	80.00	4	5
33	Rule 8-c	(15, 30]	100.00	1	1	[5 , 40]	83.54	66	79
34	Rule 8-d	(35, 60]	63.64	7	11	(10, 40]	71.43	10	14
35	Rule 8-e	(40 ,65]	75.00	3	4	[5 , 25]	85.71	6	7
36	Rule 8-f	[5 ,25]	100.00	3	3	(25 ,45]	75.00	3	4
37	Rule 8-g	(15 ,45]	70.00	7	10	(25 ,45]	80.00	4	5
38	Rule 9-a	[5, 50]	100.00	13	13	(10, 35]	70.59	12	17
39	Rule 9-b	(10, 35]	71.67	43	60	[5, 30]	68.42	13	19
40	Rule 9-c	(45, 60]	75.00	3	4	(35, 55]	57.14	4	7
41	Rule 9-d	(15, 40]	72.73	8	11	[5, 15]	66.67	2	3
42	Rule 10-a	[5, 20]	63.60	7	11	(35, 50]	100.00	2	2
43	Rule 10-b	(20,30]	100.00	4	4	(10, 35]	66.67	8	12
						(10, 35]	70.59	12	17
44	Rule 10-c	(15, 40]	73.68	14	19	(10, 35]	73.17	30	41
45	Rule 10-d	[5, 30]	70.00	7	10	(10, 35]	77.78	7	9

Note: 1. Sample size is 407.
2. Highlighted cells are terminal nodes in the RBTM.

[1] Conf. stands for ***confidence***.

Table 4.9 Summary of Validation Results for the RBTM for *CPI* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases	Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases
1	Rule 1	(10, 50]	71.97	113	157	[5, 45]	57.14	20	35
2	Rule 2-a	[5, 50]	75.82	116	153	(10, 40]	75.00	3	4
3	Rule 2-b	[5, 45]	51.85	14	27	[5, 45]	75.00	6	8
4	Rule 3-a	(10, 50]	72.54	103	142	(25, 50]	36.36	4	11
5	Rule 3-b	(10, 30]	33.33	1	3	(10,35]	100.00	1	1
6	Rule 3-c	(15, 45]	52.00	13	25	(15 ,40]	N/A	0	0
						(75,100]	0.00	0	2
7	Rule 3-d	[5, 45]	50.00	2	4	(30,70]	25.00	1	4
8	Rule 4-a	[5, 50]	65.09	69	106	(10 ,45]	75.00	27	36
9	Rule 4-b	(25, 50]	57.14	4	7	N/A	N/A	N/A	4
10	Rule 4-c	(10, 30]	33.33	1	3	(20,40]	N/A	0	0
11	Rule 4-d	(15, 45]	60.00	9	15	[5,45]	40.00	4	10
12	Rule 4-e	(15, 35]	N/A	0	0	(15,40]	N/A	0	0
13	Rule 5-a	[5, 40]	71.11	64	90	[5, 25]	25.00	4	16
14	Rule 5-b	[5, 40]	76.00	19	25	(15, 50]	50.00	4	8
						(15 ,35]	0.00	0	3
15	Rule 5-c	(25, 45]	42.86	3	7	(25 ,50]	N/A	0	0
16	Rule 5-d	(35 ,40]	0.00	0	0	(185 ,190]	N/A	0	0
17	Rule 5-e	(15,45]	100.00	3	3	(20, 40]	41.67	5	12
18	Rule 6-a	[5, 35]	68.00	17	25	(10,40]	64.62	42	65
19	Rule 6-b	[5, 25]	30.77	4	13	(15, 25]	N/A	0	0
20	Rule 6-c	[5, 45]	33.33	1	3	(10, 40]	100.00	3	3
						(15, 45]	50.00	1	2
21	Rule 6-d	(40 ,65]	N/A	0	0	(25 ,40]	50.00	1	2
						(10, 50]	66.67	4	6
22	Rule 6-e	(25 ,45]	25.00	1	4	(30 ,50]	100.00	3	3
23	Rule 6-f	[5, 35]	33.33	2	6	(20, 50]	50.00	3	6
24	Rule 7-a	[5, 40]	64.71	11	17	(10, 25]	12.50	1	8
25	Rule 7-b	[5, 40]	68.09	32	47	(10, 50]	80.00	8	10
						(20, 40]	37.50	3	8
26	Rule 7-c	[5, 25]	80.00	4	5	(15, 40]	0.00	0	4
27	Rule 7-d	[5, 45]	0.00	0	2	(20, 45]	100.00	1	1
28	Rule 7-e	(15 ,40]	33.33	1	3	(10, 45]	60.00	3	5

Table 4.9(a) Summary of Validation Results for the RBTM for *CPI* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf.[1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf.[1] (%)**	**Support**	**Total Cases**
29	Rule 7-f	(15, 45]	50.00	1	2	(15, 35]	100.00	4	4
30	Rule 7-g	(10 ,30]	80.00	4	5	2 ~ 3.5 hrs	0.00	0	5
31	Rule 8-a	(20, 50]	53.33	8	15	(10, 30]	0.00	0	2
						(40, 55]	N/A	0	0
32	Rule 8-b	(10,25]	0.00	0	2	(15, 35]	N/A	0	0
33	Rule 8-c	(15, 30]	0.00	0	3	[5 , 40]	72.73	32	44
34	Rule 8-d	(35, 60]	18.52	5	27	(10, 40]	71.43	10	14
35	Rule 8-e	(40 ,65]	100.00	2	2	[5 , 25]	N/A	0	0
36	Rule 8-f	[5 ,25]	0.00	0	1	(25 ,45]	100.00	1	1
37	Rule 8-g	(15 ,45]	71.43	5	7	(25 ,45]	0.00	0	2
38	Rule 9-a	[5, 50]	75.00	3	4	(10, 35]	45.45	5	11
39	Rule 9-b	(10, 35]	62.16	23	37	[5, 30]	71.43	5	7
40	Rule 9-c	(45, 60]	0.00	0	2	(35, 55]	66.67	2	3
41	Rule 9-d	(15, 40]	100.00	3	3	[5, 15]	50.00	1	2
42	Rule 10-a	[5, 20]	50.00	2	4	(35, 50]	N/A	0	0
43	Rule 10-b	(20,30]	14.29	1	7	(10, 35]	100	4	4
						(10,35]	72.73	8	11
44	Rule 10-c	(15, 40]	58.82	10	17	(10, 35]	60.00	12	20
45	Rule 10-d	[5, 30]	50.00	1	2	(10, 35]	60.00	3	5

Note: 1. Sample size is 192.
2. Highlighted cells are terminal nodes in the RBTM.
[1] Conf. stands for ***confidence***.

4.6 The RBTM for Incident Nature of Collision - Property Damage (CPD)

4.6.1 The Tree Structure

The rules constituting the RBTM for incidents with property damage in Montgomery County are summarized below.

1st Level

Rule 1: **If** *Tractor-Trailer* is not involved, **then** Rule 2-a; **Else** Rule 2-b

Tractor-trailer involvement is selected as the first splitter for incidents causing property damage, since it emerges as a factor that can clearly divide the available samples into distinctly different distributions.

2nd Level

Rule 2-a: **If** *Pickup Van* is not involved, **then** Rule 3-a; **Else** Rule 3-b

Rule 2-b: **If** *No. of Tractor-Trailer* =1, **then** 5<= *IncD* <=30

; **Else if** *No. of Tractor-Trailer* =2, **then** Rule 3-c

; **Else** *No. of Tractor-Trailer* >=3, **then** 90< *IncD* <=200

At this level, additional information regarding heavy vehicle involvement plays a key role in determining the resulting incident durations.

3rd Level

Rule 3-a: **If** *Shoulder* is not involved, **then** Rule 4-a; **Else** Rule 4-b

Rule 3-b: **If** *No. of Pickup Van* =1, **then** Rule 4-c; **Else** Rule 4-d

Rule 3-c: **If** *Road* is I-495, **then** 5<= *IncD* <=110

; **Else if** *Road* is I-270, **then** 60< *IncD* <=240

; **Else** *Road* is Others, **then** 40< *IncD* <=60

4th Level

Rule 4-a: **If** *Road* is I-495, **then** Rule 5-a; **Else if** *Road* is I-270, **then** Rule 5-b

; **Else** *Road* is Others, **then** Rule 5-c

Rule 4-b: **If** *Road* is I-495, **then** 5<= *IncD* <=30

; **Else if** *Road* is I-270, **then** Rule 5-d

; **Else** *Road* is Others, **then** Rule 5-e

Rule 4-c: **If** *Shoulder* is not blocked, **then** Rule 5-f; **Else** Rule 5-g

Rule 4-d: **If** *Shoulder* is not blocked, **then** Rule 5-h; **Else** Rule 5-i

The variable of *Road* or *Shoulder Blockage* is used as the next splitter.

5th Level

Rule 5-a: **If** occurs during *Off-Peak Hours*, **then** Rule 6-a; **Else** 6-b

Rule 5-b: **If** occurs during *Off-Peak Hours*, **then** Rule 6-c; **Else** 6-d

Rule 5-c: **If** *Pavement* is dry, **then** 5<=*IncD* <=20; **Else** 60 <*IncD* <=85

Rule 5-d: **If** *Pavement* is dry, **then** Rule 6-e; **Else if** *Pavement* is wet, **then** Rule 6-f

; **Else if** *Pavement* is snow/ice, **then** Rule 6-g

; **Else** 120 <*IncD* <=180

Rule 5-e: **If** *Ratio of total lanes blocked* <0.5, **then** 5 <=*IncD* <=20

; **Else** 120 <*IncD* <=180

Rule 5-f: **If** *No. of vehicles involved* =1, **then** 5 <=*IncD* <=15

; **Else if** *No. of vehicles involved* =2, **then** 5 <=*IncD* <=20

; **Else** Rule 6-h

Rule 5-g: **If** *No. of total lanes blocked* =0, **then** Rule 6-i

; **Else if** *No. of total lanes blocked* =1, **then** Rule 6-j

; **Else** Rule 6-k

Rule 5-h: **If** *Ratio of blocked lanes in the same direction* <0.5, **then** Rule 6-l

; **Else** Rule 6-m

Rule 5-i: **If** *Road* is I-495 IL, **then** Rule 6-n

; **Else if** *Road* is I-495 OL, **then** 5 <=*IncD* <=20

; **Else** Rule 6-o

The 5th level shows that the duration of incidents was likely to be shorter in dry pavement conditions than in other pavement conditions.

<u>6th Level</u>

Rule 6-a: **If** *Pavement* is dry, **then** Rule 7-a; **Else** Rule 7-b

Rule 6-b: **If** *Single-Unit Truck* is not involved, **then** Rule 7-c; **Else** 25 <*IncD* <=40

Rule 6-c: **If** *Pavement* is dry, **then** Rule 7-d; **Else** Rule 7-e

Rule 6-d: **If** *Pavement* is dry, **then** 5 <=*IncD* <=30

; **Else if** *Pavement* is wet, **then** 5 <=*IncD* <=20

; **Else if** *Pavement* is snow/ice, **then** 90 <*IncD* <=150

; **Else** 5 <=*IncD* <=15

Rule 6-e: **If** *No. of vehicles involved* <=1, **then** 5 <=*IncD* <=30

; **Else if** *No. of vehicles involved* is 2 or 3, **then** 5 <=*IncD* <=30

; **Else** *No. of vehicles involved* >=4, **then** 25 < *IncD* <=45

Rule 6-f: **If** 12 <=*Incident Hour* <=23, **then** 5 <=*IncD* <=25; **Else** 65 <*IncD* <=85

Rule 6-g: **If** *Ratio of total lanes blocked* <=0.25, **then** 30 <*IncD* <=55

; **Else** 90 <*IncD* <=150

Rule 6-h: **If** *Pavement* is snow/ice, **then** Rule 7-f; **Else** 5 <=*IncD* <=30

Rule 6-i: **If** occurs during *Off-Peak Hours*, **then** 5 <=*IncD* <=25

; **Else** 5 <=*IncD* <=25

Rule 6-j: **If** occurs during *Off-Peak Hours*, **then** Rule 7-g; **Else** 5 <=*IncD* <=25

Rule 6-k: **If** *Ratio of total lanes blocked* <0.5, **then** 5 <=*IncD* <=25

; **Else** 20 <*IncD* <=45

Rule 6-l: **If** *Exit no.* is 27 or 28, **then** 15 <*IncD* <=25

; **Else if** *Exit no.* is 31, 34 or 39, **then** 25 <*IncD* <=35

Rule 6-m: **If** *Ratio of blocked lanes in the opposite direction*=0, **then** 30 <*IncD* <=45

; **Else** 45 <*IncD* <=60

Rule 6-n: **If** *Ratio of blocked lanes in the same direction* <0.25, **then** 5 <=*IncD* <=25

; **Else** 10 <*IncD* <=30

Rule 6-o: **If** *Ratio of total lanes blocked* =0, **then** 5 <=*IncD* <=15

; **Else** Rule 7-h

At this level, information about pavement conditions and blocked lanes play significant roles in determining the durations of incidents resulting in property damage. Incident durations clearly increased as the lane-blockage ratio increased. In addition, the time at which an incident occurred had a significant relation with its resulting incident duration.

7th Level

Rule 7-a: **If** *Ratio of total lanes blocked* <=0.25, **then** 5 <=*IncD* <=25

; **Else** 35 <*IncD* <=50

Rule 7-b: **If** occurs in the *Daytime*, **then** Rule 8-a; **Else** 35 <*IncD* <=65

Rule 7-c: **If** *Pavement* is wet, **then** 35 <*IncD* <=55; **Else** 10 <*IncD* <=35

Rule 7-d: **If** *No. of total lanes blocked* =0, **then** 10 <*IncD* <=35; **Else** 30 <*IncD* <=45

Rule 7-e: **If** *Pavement* is wet, **then** 10 <*IncD* <=15; **Else** 40 <*IncD* <=60

Rule 7-f: **If** *Road* is I-495, **then** 10 <*IncD* <=20

; **Else if** *Road* is I-270, **then** 50 <*IncD* <=110

Rule 7-g: **If** *Road* is I-495 IL, **then** 5 <=*IncD* <=20

; **Else if** *Road* is I-495 OL, **then** 10 <*IncD* <=30

; **Else if** *Road* is I-270, **then** 30 <*IncD* <=45

Rule 7-h: **If** *Ratio of blocked lanes in the same direction*<0.5, **then** 30 <*IncD* <=45

; **Else** 45 <*IncD* <=70

One noticeable impact on the incident duration at this level is due to the factor of *Road*. According to Rules 7-f and 7-g, incidents occurring on I-495 are likely to be shorter than those same types of incidents occurring on I-270. The same relations have also been observed in developing Rule 3-c at the 3rd level.

<u>8th Level</u>

Rule 8-a: **If** *Response Time* < 30 mins, **then** 5 <=*IncD* <=30; **Else** 40 <*IncD* <=60

4.6.2 Performance and Validation Results

Tables 4.10 and 4.10(a) show the summary of model performance for incidents with property damage. Most of terminal nodes demonstrate quite satisfactory results for both of the range of incident durations and the ***confidence***. The performance of Rule 1, with just the *If* condition, demonstrates satisfactory results without any additional splitter. The predicted range of incident durations is less than 30 minutes, and the probability (***confidence***) is greater than 0.7 (70 percent). Since one of the main research purposes is to discover relations between incident duration and associate factors, this study continues to build the tree to its next level.

However, the *Else* condition in the Rule 1 showed unsatisfactory performance results. Even with additional splitters, the performance for this sub-data set was not improved, as shown in Table 4.10 (see Rules for 2-b and 3-c). Since the durations of incidents within this subset of small size (i.e., 46) are distributed over a wide range, the RBTM could not yield definitive results. This suggests the need to calibrate a supplemental model.

In addition, since Rules 5-c, 7-b, and 7-d could not perform up to the expected level, they also need supplemental models. However, due to the limited sample data for these subsets, this study has developed a supplemental model for these cases with the higher-level subset used for Rule 4-a.

Tables 4.11 and 4.11(a) summarize the model validation results. Note that a large number of rules at levels 1, 2, 3, and 4 demonstrated satisfactory results in the validation data set, while many rules at lower levels did not perform as expected, due either to the need for additional factors or the lack of sufficient sample data.

Table 4.10 Summary of Estimation Results for the RBTM for *CPD* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf.[1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf.[1] (%)**	**Support**	**Total Cases**
1	Rule 1	[5, 30]	75.00	249	392	[5, 30]	60.87	28	46
2	Rule 2-a	[5, 45]	86.03	234	272	N/A	N/A	N/A	120
3	Rule 2-b	[5, 30]	78.79	26	33	[5, 75]	72.73	8	11
						(90, 200]	100.00	2	2
4	Rule 3-a	[5, 45]	83.49	91	109	[5, 30]	78.53	128	163
5	Rule 3-b	[5, 30]	82.22	74	90	[5, 35]	86.67	26	30
6	Rule 3-c	[5, 110]	66.67	4	6	(60, 240]	100.00	2	2
						(40, 60]	66.67	2	3
7	Rule 4-a	[5, 45]	82.81	53	64	[5, 45]	85.29	29	34
						[5, 45]	81.82	9	11
8	Rule 4-b	[5, 30]	82.05	96	117	[5, 30]	72.09	31	46
						(120, 180]	66.67	2	3
9	Rule 4-c	[5, 30]	88.00	22	25	[5, 30]	80.00	52	65
10	Rule 4-d	(15, 35]	77.78	7	9	[5, 35]	90.48	19	21
11	Rule 5-a	[5, 45]	78.95	30	38	[5, 40]	88.46	23	26
12	Rule 5-b	(10, 45]	93.75	15	16	[5, 30]	77.78	14	18
13	Rule 5-c	[5, 20]	75.00	6	8	(60, 85]	66.67	2	3
14	Rule 5-d	[5, 30]	83.87	26	31	[5, 25]	66.67	4	6
						(30, 55]	60.00	3	5
						(120, 180]	100.00	1	1
15	Rule 5-e	[5, 20]	100.00	1	1	(120, 180]	100.00	2	2
16	Rule 5-f	[5, 15]	100.00	3	3	[5, 20]	85.71	12	14
						[5, 30]	75.00	6	8
17	Rule 5-g	[5, 40]	96.67	29	30	[5, 30]	79.17	19	24
						[5, 30]	81.82	9	11
18	Rule 5-h	(15, 35]	100.00	6	6	(30, 60]	100.00	3	3
19	Rule 5-i	[5, 30]	91.67	11	12	[5, 20]	100.00	5	5
						(10, 35]	75.00	3	4
20	Rule 6-a	[5, 40]	80.00	20	25	[5, 35]	61.54	8	13
21	Rule 6-b	[5, 25]	76.19	16	21	(25, 40]	80.00	4	5
22	Rule 6-c	(25, 45]	62.50	5	8	(10, 25]	62.50	5	8

Table 4.10(a) Summary of Estimation Results for the RBTM for *CPD* Incidents Occurring in Montgomery County (cont'd)

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf.[1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf.[1] (%)**	**Support**	**Total Cases**
23	Rule 6-d	[5, 30]	78.57	11	14	[5, 20]	100.00	2	2
						(90, 150]	100.00	1	1
						[5, 15]	100.00	1	1
24	Rule 6-e	[5, 30]	100.00	7	7	[5, 30]	85.71	18	21
						(25, 45]	100.00	3	3
25	Rule 6-f	[5, 25]	100.00	4	4	(65, 85]	100.00	2	2
26	Rule 6-g	(30, 55]	75.00	3	4	(90, 150]	100.00	1	1
27	Rule 6-h	(55, 105]	66.67	2	3	[5, 30]	100.00	5	5
28	Rule 6-i	[5, 25]	82.35	14	17	[5, 25]	69.23	9	13
29	Rule 6-j	[5, 30]	75.00	9	12	[5, 25]	83.33	10	12
30	Rule 6-k	[5, 25]	77.78	7	9	(20, 45]	100.00	2	2
31	Rule 6-l	(15, 25]	100.00	2	2	(25, 35]	75.00	3	4
32	Rule 6-m	(30, 45]	100.00	2	2	(45, 60]	100.00	1	1
33	Rule 6-n	[5, 25]	83.33	5	6	(10, 30]	83.33	5	6
34	Rule 6-o	[5, 15]	100.00	2	2	(30, 70]	100.00	2	2
35	Rule 7-a	[5, 25]	80.00	8	10	(35, 50]	100.00	2	2
36	Rule 7-b	[5, 30]	66.67	6	9	(35, 65]	75.00	3	4
37	Rule 7-c	(35, 55]	100.00	2	2	(10, 35]	84.21	16	19
38	Rule 7-d	(10, 35]	80.00	4	5	(30, 45]	66.67	2	3
39	Rule 7-e	(10, 15]	100.00	4	4	(40, 60]	75.00	3	4
40	Rule 7-f	(10, 20]	100.00	1	1	(50, 110]	100.00	2	2
41	Rule 7-g	[5, 20]	71.43	5	7	(10, 30]	100.00	4	4
						(30, 45]	100.00	1	1
42	Rule 7-h	(30, 45]	100.00	1	1	(45, 70]	100.00	1	1
43	Rule 8-a	[5, 30]	85.71	6	7	(40, 60]	100.00	2	2

Note: 1. Sample size is 438.
2. Highlighted cells are terminal nodes in the RBTM.

[1] Conf. stands for ***confidence***.

Table 4.11 Summary of Validation Results for the RBTM for *CPD* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf. [1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf. [1] (%)**	**Support**	**Total Cases**
1	Rule 1	[5, 30]	69.41	177	255	[5, 30]	48.72	19	39
2	Rule 2-a	[5, 45]	88.24	165	187	N/A	N/A	N/A	68
3	Rule 2-b	[5, 30]	60.00	18	30	[5, 75]	50.00	4	8
						(90, 200]	0.00	0	1
4	Rule 3-a	[5, 45]	86.89	53	61	[5, 30]	73.81	93	126
5	Rule 3-b	[5, 30]	68.00	34	50	[5, 35]	66.67	12	18
6	Rule 3-c	[5, 110]	60.00	3	5	(60, 240]	0.00	0	2
						(40, 60]	0.00	0	1
7	Rule 4-a	[5, 45]	91.89	34	37	[5, 45]	81.82	18	22
						[5, 45]	50.00	1	2
8	Rule 4-b	[5, 30]	75.79	72	95	[5, 30]	69.23	18	26
						(120, 180]	0.00	0	3
9	Rule 4-c	[5, 30]	61.54	8	13	[5, 30]	70.27	26	37
10	Rule 4-d	(15, 35]	100.00	2	2	[5, 35]	68.75	11	16
11	Rule 5-a	[5, 45]	86.96	20	23	[5, 40]	85.71	12	14
12	Rule 5-b	(10, 45]	87.50	7	8	[5, 30]	50.00	7	14
13	Rule 5-c	[5, 20]	50.00	1	2	(60, 85]	N/A	N/A	0
14	Rule 5-d	[5, 30]	68.18	15	22	[5, 25]	50.00	2	4
						(30, 55]	N/A	N/A	0
						(120, 180]	N/A	N/A	0
15	Rule 5-e	[5, 20]	N/A	N/A	0	(120, 180]	0.00	0	3
16	Rule 5-f	[5, 15]	0.00	0	5	[5, 20]	57.14	4	7
						[5, 30]	100.00	1	1
17	Rule 5-g	[5, 40]	87.50	14	16	[5, 30]	70.59	12	17
						[5, 30]	50.00	2	4
18	Rule 5-h	(15, 35]	100.00	2	2	(30, 60]	N/A	N/A	0
19	Rule 5-i	[5, 30]	66.67	4	6	[5, 20]	55.56	5	9
						(10, 35]	100.00	1	1
20	Rule 6-a	[5, 40]	86.67	13	15	[5, 35]	75.00	6	8
21	Rule 6-b	[5, 25]	69.23	9	13	(25, 40]	0.00	0	1
22	Rule 6-c	(25, 45]	40.00	2	5	(10, 25]	33.33	1	3

Table 4.11(a) Summary of Validation Results for the RBTM for *CPD* Incidents Occurring in Montgomery County (cont'd)

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf.[1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf.[1] (%)**	**Support**	**Total Cases**
23	Rule 6-d	[5, 30]	50.00	5	10	[5, 20]	33.33	1	3
						(90, 150]	N/A	N/A	0
						[5, 15]	N/A	N/A	0
24	Rule 6-e	[5, 30]	50.00	1	2	[5, 30]	66.67	12	18
						(25, 45]	0.00	0	2
25	Rule 6-f	[5, 25]	100.00	2	2	(65, 85]	0.00	0	2
26	Rule 6-g	(30, 55]	N/A	N/A	0	(90, 150]	N/A	N/A	0
27	Rule 6-h	(55, 105]	N/A	N/A	0	[5, 30]	100.00	1	1
28	Rule 6-i	[5, 25]	75.00	6	8	[5, 25]	50.00	4	8
29	Rule 6-j	[5, 30]	55.55	5	9	[5, 25]	75.00	6	8
30	Rule 6-k	[5, 25]	0.00	0	1	(20, 45]	100.00	3	3
31	Rule 6-l	(15, 25]	N/A	N/A	0	(25, 35]	N/A	N/A	0
32	Rule 6-m	(30, 45]	N/A	N/A	0	(45, 60]	N/A	N/A	0
33	Rule 6-n	[5, 25]	80.00	4	5	(10, 30]	0.00	0	1
34	Rule 6-o	[5, 15]	0.00	0	1	(30, 70]	N/A	N/A	0
35	Rule 7-a	[5, 25]	50.00	6	12	(35, 50]	0.00	0	3
36	Rule 7-b	[5, 30]	57.14	4	7	(35, 65]	N/A	N/A	0
37	Rule 7-c	(35, 55]	0.00	0	2	(10, 35]	45.45	5	11
38	Rule 7-d	(10, 35]	100.00	3	3	(30, 45]	50.00	1	2
39	Rule 7-e	(10, 15]	0.00	0	3	(40, 60]	0.00	0	2
40	Rule 7-f	(10, 20]	N/A	N/A	0	(50, 110]	N/A	N/A	0
41	Rule 7-g	[5, 20]	100.00	3	3	(10, 30]	33.33	2	6
						(30, 45]	N/A	N/A	0
42	Rule 7-h	(30, 45]	N/A	N/A	0	(45, 70]	0.00	0	1
43	Rule 8-a	[5, 30]	57.14	4	7	(40, 60]	N/A	N/A	0

Note: 1. Sample size is 294.
2. Highlighted cells are terminal nodes in the RBTM.

[1] Conf. stands for ***confidence***.

4.7 The RBTM for Incident Nature of Disabled Vehicles (DV)

4.7.1 The Tree Structure

The following rules construct the RBTM for incidents with disabled vehicles occurring in Montgomery County.

1st Level

Rule 1: **If** *Weekend*, **then** 5<=*IncD* <=25; **Else** Rule 2-a

2nd Level

Rule 2-a: **If** occurs during *Off-Peak Hours*, **then** 5<=*IncD* <=35; **Else** Rule 3-a

3rd Level

Rule 3-a: **If** *Shoulder* is not blocked, **then** 5<=*IncD* <=30; **Else** Rule 4-a

4th Level

Rule 4-a: **If** *No. of shoulders blocked*=1, **then** Rule 5-a; **Else** 5<=*IncD* <=20

5th Level

Rule 5-a: **If** *Pickup Van* is not involved, **then** 5<=*IncD* <=25; **Else** 5<=*IncD* <=20

Note that incidents occurring during peak hours were more likely to be cleared in a shorter duration than those during off-peak hours. Also, when any shoulder lane was blocked at peak hours due to a disabled vehicle, the average incident duration was slightly shorter than for incidents without a shoulder blockage.

When a disabled vehicle is a pickup van, the estimated range for incident duration is 5 to 20 minutes. But for other types of disabled vehicles, the incident can be cleared within 25 minutes from the time of detection.

4.7.2 Performance and Validation Results

Since most incidents due to *Disabled Vehicle* (83.3 percent for Montgomery County alone) fell in a relatively short range of 5 to 30 minutes, a simple rule could predict their resulting durations. Also, even after applying a series of additional splitters to subdivide the data set, the ***confidence*** for each subset does not show any noticeable change. This is due mainly to the fact that the incidents caused by disabled vehicles involved only a single vehicle.

As shown in Table 4.12, most of these developed rules showed satisfactory results for their ***confidence*** and the estimated range of incident durations. Their validation results, reported in Table 4.13, were also at an acceptable level, except for those having only very small samples.

Table 4.12 Summary of Estimation Results for the RBTM for *Disabled Vehicles* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases	Incident Duration (mins)	Conf. [1] (%)	Support	Total Cases
1	Rule 1	[5, 25]	81.82	9	11	[5, 35]	89.51	274	306
2	Rule 2-a	[5, 35]	88.76	158	178	[5, 30]	85.16	109	128
3	Rule 3-a	[5, 30]	83.95	68	81	[5, 25]	85.11	40	47
4	Rule 4-a	[5, 25]	83.72	36	43	[5, 20]	100.00	4	4
5	Rule 5-a	[5, 25]	82.35	28	34	[5, 20]	88.89	8	9

Note: 1. Sample size is 317.
2. Highlighted cells are terminal nodes in the RBTM.

[1] Conf. stands for ***confidence***.

Table 4.13 Summary of Validation Results for the RBTM for *Disabled Vehicles* Incidents Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	Conf. (%)	Support	Total Cases	Incident Duration (mins)	Conf. (%)	Support	Total Cases
1	Rule 1	[5, 25]	0.00	0	1	[5, 35]	88.61	140	158
2	Rule 2-a	[5, 35]	93.51	72	77	[5, 30]	76.54	62	81
3	Rule 3-a	[5, 30]	68.09	32	47	[5, 25]	85.29	29	34
4	Rule 4-a	[5, 25]	85.29	29	34	[5, 20]	N/A	N/A	0
5	Rule 5-a	[5, 25]	85.19	23	27	[5, 20]	85.71	6	7

Note: 1. Sample size is 159.
2. Highlighted cells are terminal nodes in the RBTM.
[1] Conf. stands for ***confidence***.

4.8 The RBTM for Incident Nature of Others

4.8.1 The Tree Structure

The rules used to construct the RBTM for *Incident Nature* of *Others* are presented below.

1st Level

Rule 1: **If** *Shoulder* is not blocked, **then** Rule 2-a; **Else** Rule 2-b

2nd Level

Rule 2-a: **If** *Tractor-Trailer* is not involved, **then** Rule 3-a; **Else** *IncD*=493

Rule 2-b: **If** occurs during *Off-Peak Hours*, **then** Rule 3-b; **Else** Rule 3-c

The rules at this level reflect clearly that incidents involving tractor-trailers generally resulted in longer incident durations than those with any other types of vehicles.

3rd Level

Rule 3-a: **If** *Single-Unit Truck* is not involved, **then** Rule 4-a; **Else** *IncD*=105

Rule 3-b: **If** *Pickup Van* is not involved, **then** Rule 4-b; **Else** Rule 4-c

Rule 3-c: **If** *Ratio of total lanes blocked* <0.25, **then** 5<= *IncD* <=20

; **Else** 30< *IncD* <=50

All of the rules at this level collectively capture the fact that the number of lanes being blocked during the response operation positively correlated with the resulting incident durations. Incident durations also positively correlated with the involvement of heavy vehicles or trucks.

4th Level

Rule 4-a: **If** *Total no. of lanes blocked* <=1, **then** Rule 5-a; Else 25< *IncD* <=40

Rule 4-b: **If** *Road* is I-495, **then** Rule 5-b; **Else if** *Road* is I-270, **then** Rule 5-c

; **Else** *IncD*=607

Rule 4-c: **If** *Road* is I-495, **then** 30< *IncD* <=40

; **Else if** *Road* is I-270, **then** 10< *IncD* <=25

The rules constructed at this level reflect the fact that the response efficiency for the same incident type may vary significantly among all highways under the coverage of emergency incident response operations.

5th Level

Rule 5-a: **If** *Road* is I-495, **then** 5<= *IncD* <=25

; **Else if** *Road* is I-270, **then** 20< *IncD* <=35

Rule 5-b: **If** occurs in the *Daytime*, **then** Rule 6-a; **Else** Rule 6-b

Rule 5-c: **If** *Tractor-Trailer* is not involved, **then** Rule 6-c; **Else** Rule 6-d

As expected, the time of detection was one of the critical factors contributing to the resulting incident durations. In general, the duration of incidents occurring in the daytime was likely to be shorter than that at night.

6th Level

Rule 6-a: **If** *Pavement* is dry, **then** Rule 7-a; **Else** 60< *IncD* <=75

Rule 6-b: **If** *Tractor-Trailer* is not involved, **then** 45< *IncD* <=60

; **Else** *IncD* > 120

Rule 6-c: **If** *Exit no.* is 1, **then** 15< *IncD* <=40

; **Else** 10< *IncD* <=20

Rule 6-d: **If** *Ratio of blocked lanes in the same direction* < 1, **then** 80< *IncD* <=100

; **Else** 240< *IncD* <=300

Information at this level reveals that incidents occurring at some locations may have longer durations than those of the same type occurring at other locations. For instance, Rule 6-c indicates that the incidents occurring at Exit 1 on I-270 were likely to last longer than those at other locations.

7th Level

Rule 7-a: **If** *Tractor-Trailer* is not involved, **then** 5<= *IncD* <=25

; **Else** *IncD* > 120

4.8.2 Performance and Validation Results

Since the sample size for these cases was relatively small, it was difficult to develop a reliable RBTM. It was even more challenging to validate this model, because the validation data set had only 18 records of such incidents. As a result, more than 50 percent of rules could not be validated (see Table 4.15). Nevertheless, the overall performance is promising, except with some rules shown in Table 4.14. A supplemental model for enhancing the performance level was thus developed and is presented in the next chapter.

Table 4.14 Summary of Estimation Results for the RBTM for *Incident Nature – Others* Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf.[1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf.[1] (%)**	**Support**	**Total Cases**
1	Rule 1	[5, 40]	78.57	11	14	[5, 40]	63.64	21	33
2	Rule 2-a	[5, 30]	69.23	9	13	493	100.00	1	1
3	Rule 2-b	[5, 40]	63.64	14	22	[5, 35]	63.64	7	11
4	Rule 3-a	[5, 30]	75.00	9	12	105	100.00	1	1
5	Rule 3-b	[5, 40]	63.64	14	22	(15, 40]	100.00	4	4
6	Rule 3-c	[5, 20]	85.71	6	7	(30, 50]	75.00	3	4
7	Rule 4-a	[5, 25]	70.00	7	10	(25, 40]	100.00	1	1
8	Rule 4-b	[5, 50]	63.64	7	11	(10, 40]	66.67	4	6
						607	100.00	1	1
9	Rule 4-c	(30, 40]	100.00	3	3	(10, 25]	100.00	1	1
10	Rule 5-a	[5, 25]	75.00	6	8	(20, 35]	100.00	2	2
11	Rule 5-b	[5, 40]	75.00	6	8	(45, 60]	66.67	2	3
12	Rule 5-c	(10, 40]	100.00	4	4	(90, 300]	100.00	2	2
13	Rule 6-a	[5, 25]	71.43	5	7	(60, 75]	100.00	1	1
14	Rule 6-b	(45, 60]	100.00	2	2	> 120	100.00	1	1
15	Rule 6-c	(15, 40]	100.00	2	2	(10, 20]	100.00	2	2
16	Rule 6-d	(80, 100]	100.00	1	1	(240, 300]	100.00	1	1
17	Rule 7-a	[5, 25]	80.00	4	5	> 120	50.00	1	2

Note: 1. Sample size is 47.
2. Highlighted cells are terminal nodes in the RBTM.

[1] Conf. stands for ***confidence***.

Table 4.15 Summary of Validation Results for the RBTM for *Incident Nature – Others* Occurring in Montgomery County

No	Rule	IF				ELSE			
		Incident Duration (mins)	**Conf.[1] (%)**	**Support**	**Total Cases**	**Incident Duration (mins)**	**Conf.[1] (%)**	**Support**	**Total Cases**
1	Rule 1	[5, 40]	66.67	4	6	[5, 40]	75.00	9	12
2	Rule 2-a	[5, 30]	66.67	4	6	493	N/A	N/A	0
3	Rule 2-b	[5, 40]	80.00	4	5	[5, 35]	42.86	3	7
4	Rule 3-a	[5, 30]	50.00	2	4	105	0.00	0	2
5	Rule 3-b	[5, 40]	80.00	4	5	(15, 40]	N/A	N/A	0
6	Rule 3-c	[5, 20]	50.00	1	2	(30, 50]	20.00	1	5
7	Rule 4-a	[5, 25]	33.33	1	3	(25, 40]	100.00	1	1
8	Rule 4-b	[5, 50]	100.00	4	4	(10, 40]	100.00	1	1
						607	N/A	N/A	0
9	Rule 4-c	(30, 40]	N/A	N/A	0	(10, 25]	N/A	N/A	0
10	Rule 5-a	[5, 25]	50.00	1	2	(20, 35]	N/A	N/A	0
11	Rule 5-b	[5, 40]	75.00	3	4	(45, 60]	N/A	N/A	0
12	Rule 5-c	(10, 40]	100.00	1	1	(90, 300]	N/A	N/A	0
13	Rule 6-a	[5, 25]	50.00	2	4	(60, 75]	N/A	N/A	0
14	Rule 6-b	(45, 60]	100.00	2	2	> 120	100.00	1	1
15	Rule 6-c	(15, 40]	100.00	1	1	(10, 20]	N/A	N/A	0
16	Rule 6-d	(80, 100]	N/A	N/A	0	(240, 300]	N/A	N/A	0
17	Rule 7-a	[5, 25]	50.00	2	4	> 120	N/A	N/A	0

Note: 1. Sample size is 18.
2. Highlighted cells are terminal nodes in the RBTM.
[1] Conf. stands for ***confidence***.

4.9 Overall Findings and Conclusions

This section summarizes the overall findings with the RBTMs.

1. For the categories of *Collision-Personal Injury*, *Collision-Property Damage*, *Disabled Vehicle* and *Others*, it turned out that the spatial factor, *County*, emerged as the second splitter. This implies that the durations for the same types of incident varied significantly among different jurisdictions.

2. The sequence of splitters varies significantly among different categories of incidents. This is likely due to the fact that incidents of different natures have different characteristics and are associated with different contributing factors.
3. RBTMs are more flexible for assigning an appropriate estimated incident duration range in given conditions (sub-data set or node) than CART. Unlike CART, this model includes a function to regroup categories of the dependent variable (i.e., 5-minute intervals of incident duration from 5 to 120 minutes), so as to determine the most appropriate range of incident duration for a selected subset.
4. As expected, heavy vehicle involvement tended to increase incident durations, due to its complexity to manage or the need for special equipment for clearance operations (e.g., wrecker).
5. Incidents occurring at nighttime or during off-peak hours generally had longer durations than those in daytime, due to the lack of sufficient response units for incident clearance operations.
6. When incidents resulted in *Collision-Fatality,* or *Property Damage*, the clearance operation was generally more efficient in the blocked shoulder lane scenarios than in those leaving it open. This finding implies that shoulder lane blockage helps reduce the duration of severe accidents, as it provides a wider space for emergency response units to work.
7. Similarly, during *Collision-Fatality* incidents, if the emergency response unit can close more lanes in the same direction, it generally results in a shorter duration.
8. The impact of wet pavement, a proxy variable for rainy days, on the efficiency of incident response operations was not definitive for the existing data records. It

showed a positive correlation with incident durations for those resulting in *Collision-Property Damage*, but a reverse relation for the category of *Collision-Fatality* incidents. For all other types of incidents, its impact on the resulting incident durations was not statistically significant.

Due to the complex nature of incidents and of response operations, one should not expect the above RBTM to capture all embedded relations and provide an operationally acceptable performance for real-world applications. Hence, based on the promising information generated from the RBTM, this study further developed some supplemental models to improve prediction accuracy for the duration of a detected incident. Depending on the available size of sample data, this study employed either the multinomial logit (MNL) model or the multiple linear regression model to develop supplemental components.

Lastly, RBTMs, illustrated in a tree shape, are included in Appendix 2.

Chapter 5: Supplemental Models

5.1 Introduction

This chapter presents two supplemental models for improving the accuracy of predicting incident durations. The first, the multinomial logit (MNL) model, is used for the sub-data sets with unsatisfactory results from the main (RBTM) models for incidents of type *CPI* and *CPD*. This model is proposed because samples in those subsets show a condensed distribution, and the sample size was large enough to compare to the number of categories in a dependent variable. The second model is the multiple linear regression model, used for data sets from incident natures of *CF* and *Others*, since those relatively small data sets show a scattered distribution.

Figures 5.1 and 5.2 illustrate the sub-data sets used for developing supplemental models for incident natures of *CPI* and *CPD*. Incident natures of *CF* and *Others* used the entire data set to develop their supplemental models. Brief descriptions of core concepts for these two methods, along with estimation and validation results, are presented below.

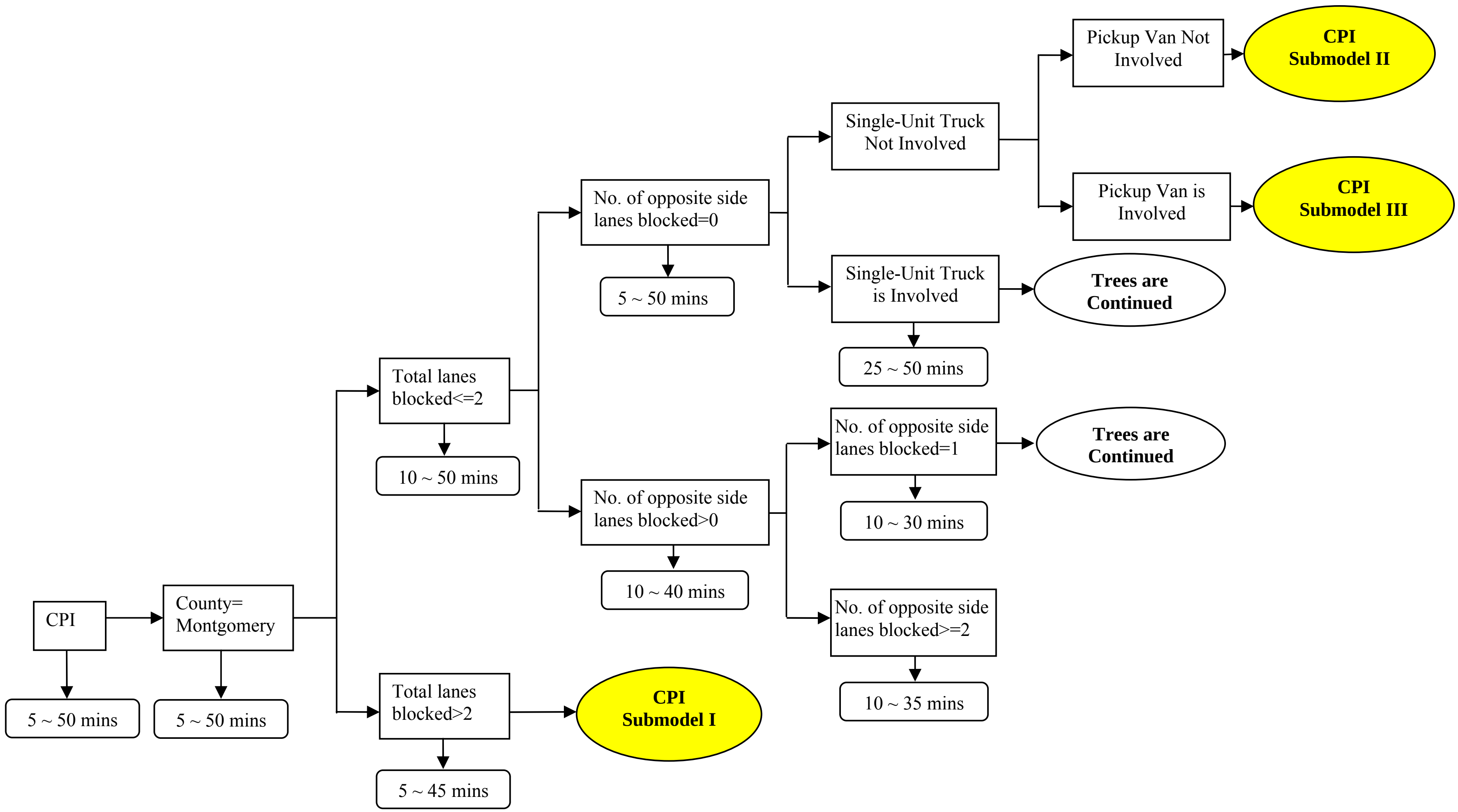


Figure 5.1 Sub-Data Sets Used for Developing Supplemental Models for Incidents Causing Collision-Personal Injury

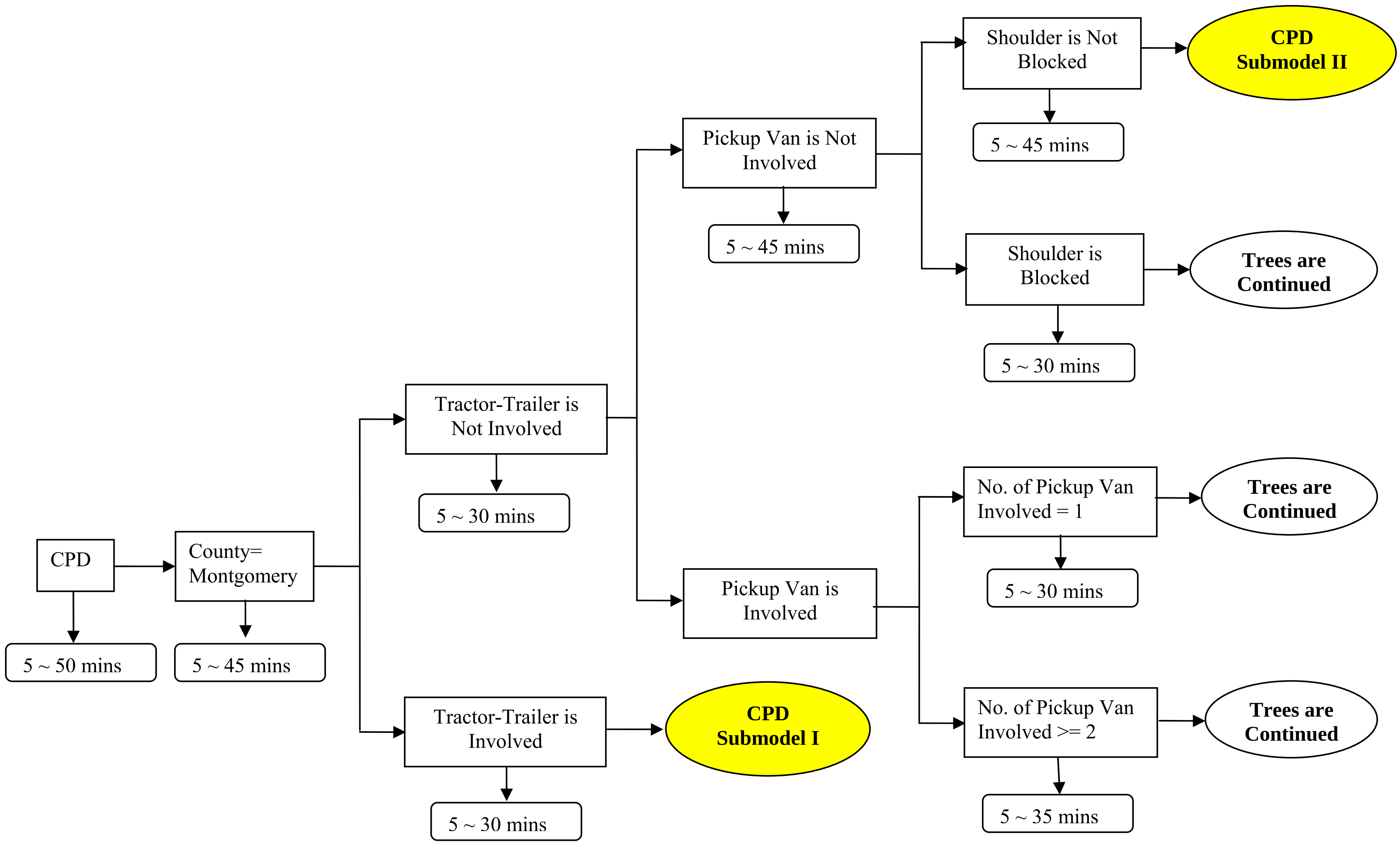


Figure 5.2 Sub-Data Sets Used for Developing Supplemental Models for Incidents Causing Collision-Property Damage

5.2 Supplemental Model – 1: Multinomial Logit (MNL) Models

5.2.1 Multinomial Logit Models

Analyses of discrete or nominal scale data are one of the major areas in transportation studies, as many interesting policy-sensitive analyses are implemented based on such data (Washington et al., 2003). Examples of discrete scale data in the transportation field are the travel mode (automobile, bus, metro), the type or class of vehicle owned, and the type of accident/injury severity (property damage only, personal injuries, fatalities). These types of data could be classified into two categories based on conceptual viewpoint — behavioral choice and description of discrete outcomes from a physical event (Washington et al., 2003). The travel mode choice and class of vehicle owned belong to the former category, i.e., these are behavioral choices; accident injury severity belongs to the latter category, since it merely explains discrete outcomes of a physical event. Similarly, intervals of incident durations can be treated as discrete outcomes from physical events.

Although these two conceptual perspectives are modeled by statistically identical methodologies, the fundamental theories used to derive those models show many differences (Washington et al., 2003). For instance, discrete scale models for a behavioral choice are derived from economic theories, while the model for the description of physical phenomena is based on simple probabilistic theories (Washington et al., 2003). In addition, though both discrete scale models for the two categories are derived from random utility theory (McFadden, 1974), different functions are used for determining a choice.

In a behavior model, the choice is made based on the utility function, and it is assumed that the decision maker will choose the alternative that has the greatest value of utility function among all available alternatives. However, for incidents, the individuals are no longer decision makers who make the best choice among alternatives. Rather, they are accident victims who have been injured or need responses from specialists. Thus, in the physical phenomenon model, a choice is made to the alternative with the highest value in ***propensity*** function (Khorashadi, 2003). Nonetheless, the possible forms of the two models are the same. The only difference is the interpretation of functional elements, such as utility or propensity (Khorashadi, 2003).

One of the most common models used for analyzing discrete data is the logit model. It has been widely used in mode choice and incident severity studies, but it is a relatively new approach to the study of incident durations.

For sub-data sets in *CPI* and *CPD* which showed unsatisfactory results in the RBTM, the MNL model was applied to estimate the relation between each category of incident durations and its associated factors. A well-calibrated model would allow its users to predict the duration category of a detected incident. The core concept of MNL, like that used in accident severity models (Khorashadi, 2003; Ulfarsson, 2001), is briefly described below:

The ***propensity*** function, R_{ni}, which represents the propensity of incident n towards interval i of incident durations, is defined as

$$R_{ni} = \beta_i X_{ni} + \varepsilon_{ni} \qquad \forall i \in I \qquad \text{(Eq. 5.1)}$$

where I is a set of preclassified incident durations (defined in an interval form), X_{ni} is a vector of observable characteristics (e.g., environmental conditions, geometric conditions, and so on) that determine the discrete outcome for observation n (incident n), β_i is a vector of estimated parameters, and ε_{ni} is an error term accounting for unobservable attributes and effects that influence the determination of discrete outcomes for observation n. Assuming that the disturbance terms of the ***propensity*** function are (1) independent, (2) identically distributed, and (3) follow the Gumbel distribution with a location parameter η=0 and a scale parameter μ=1, the MNL model is derived as

$$P_n(i) = \frac{e^{\mu R_{ni}}}{\sum_{j \in C_n} e^{\mu R_{nj}}} = \frac{e^{\beta_i X_{ni}}}{\sum_{j \in C_n} e^{\beta_i X_{nj}}} \qquad \text{(Eq. 5.2)}$$

where, β_i is a vector of coefficients, and X_{ni} and X_{nj} are vectors of attributes for alternatives i and j. A detailed discussion regarding MNL models can be found in the literature (Ben-Akiva and Lerman, 1985; Koppelman and Bhat, 2006; Washington et al., 2003).

The initial specification of the ***propensity*** functions is set as follows:

$$\begin{aligned} R_i &= \beta_0^i + \beta_{NoTT} \cdot NoTT + \beta_{NoVehInv} \cdot NoVehInv + \beta_{I270} \cdot I270 + \beta_{I495} \cdot I495 \\ &+ \beta_{Night} \cdot Night + \beta_{rt\_ttlbl} \cdot RtTTLBL + \beta_{\mathrm{Re}spTime} \cdot \mathrm{Re}spTime \\ &+ \beta_{NoLnBl(S)} \cdot NoLnBl(S) + \beta_{Pave\_SI} \cdot Pave\_SI \end{aligned}$$

for $\forall i \in I$ but the last alternative, i_L

$R_{i_L} = 0$ (Base) for the last alternative, i_L

where:

- β_0^i is an alternative specific constant for each alternative.
- *NoTT* is the number of tractor-trailers involved.
- *NoPUV* is the number of pickup vans involved.
- *NoSUT* is the number of single-unit trucks involved.

- *NoVehInv* is the number of vehicles involved.
- *I270* is 1 if the incident occurred on Interstate I-270; 0 otherwise.
- *I495* is 1 if the incident occurred on Interstate I-495; 0 otherwise.
- *Night* is 1 if the incident occurred at night; 0 otherwise.
- *RtTTLBL* is the ratio of total number of blocked lanes over the total number of lanes.
- *RespTime* is the response time in minutes.
- *NoLnBl(S)* is the number of lanes blocked in the same direction.
- *Pave_SI* is 1 if the pavement condition is snowy/icy; 0 otherwise.

Since there are too many variables included, the model development is initialized with all coefficients being set as generic, except for alternative specific constants. First, variables showing insignificance at the 0.10 significance level were removed from the ***propensity*** functions (for a two-tailed test, the critical values of t-statistic are ±1.65 for the 0.10 significance level). Then, variables not included at the initial stage were included to test their significance in ***propensity*** functions. After filtering out insignificant variables, all coefficients were set as alternative specific to test if all variables are significant. Insignificant variables were removed from the corresponding ***propensity*** function. Lastly, variables previously removed from the model were included one by one, again with their coefficients being set as alternative specific, to verify whether any significant variable was left out.

5.2.2 Estimation Results with MNL

As shown in Figure 5.1, for *CPI*, three MNL models were needed, while two MNL models were required for *CPD* (see Figure 5.2), since each sub-data set needed a different model to result in the best performance. The categories (intervals) of incident durations were defined differently for each MNL model, since their distributions of incident durations differed from one another. Table 5.1, following, summarizes the categories of incident durations for each MNL model.

Table 5.1 Categories of Incident Durations (minutes) for Each MNL Model

	Submodel I	Submodel II	Submodel III
CPI[1]	[5, 25] (25, 45] > 45	[5, 25] (25, 50] > 50	[5, 25] (25, 45] > 45
CPD[2]	[5, 30] [3] > 30 [3]	[5, 25] (25, 45] > 45	N/A

[1] CPI stands for *Collision-Personal Injury*

[2] CPD stands for *Collision-Property Damage*

[3] Since this submodel includes only two categories for a dependent variable, a binary logit model was applied instead of MNL. However, the theoretical concepts and background for binary logit models are the same as for MNL.

Developed MNL models are presented in Tables 5.2(a) to 5.3(b), and the estimated and validated probabilities for incident durations for each MNL model are summarized in Table 5.4. All of the estimated coefficients, except for the alternative specific constant in the ***propensity*** function for incident durations of 5 to 25 minutes for CPI-Submodel I, show significance at the 90 percent level (an absolute value of the *t*-statistic should be above 1.65). The insignificance of alternative specific constants is irrelevant, because they reflect the average effects of variables which were not included in the model. Thus, they should always be included, even though they are not well understood in the behavioral interpretation (Koppelman and Bhat, 2006).

In general, the signs and magnitudes of coefficients for all variables were as expected. In previous chapters, an increase in the number of heavy vehicles (single-unit trucks, pick up vans, or tractor-trailers) involved was found to cause an increase in incident duration. This observation is reflected as the negative sign of the coefficients for variables *NoTT, NoSUT,* and *NoPUV* in the short incident duration alternatives, e.g., 5 to 25 or 25 to 45 minutes, of the MNL models. The observation that incident duration increases as the number of vehicles involved increases is reflected in the same way. The negative coefficient for *Night* in alternatives 5 to 25 and 25 to 45 minutes reflects the observation that when an incident occurs at night, it is likely to last longer than one occurring in the daytime. Models also show a positive effect of *I-495* in reducing incident duration by having a positive coefficient in those short incident duration alternatives. In other words, incidents occurring on Interstate I-495 are more likely to be cleared earlier than the other cases. Some noticeable outcomes for each explanatory variable are summarized below.

1. In MNL models for *CPI,* the pavement condition showed different effects on each submodel. In Submodel I, the pavement condition *Dry* is likely to shorten incident durations, as it has a positive coefficient for the alternative of 5 to 25 minutes. But, in Submodel III, this variable tends to increase incident durations, since it has a negative coefficient for the alternative of 25 to 45 minutes. Meanwhile, incidents occurring in the pavement condition *Snow/Ice* tend to increase the durations in Submodel II, which is reflected by this variable having a negative coefficient for the incident duration alternatives of 5 to 25 and 25 to 50 minutes.

2. Interstate I-270 has different effects in the submodels for *CPI* and *CPD*. In CPI-Submodel II, the variable *I-270*, shows a tendency to decrease incident durations, which is reflected by the larger positive coefficient for the 5 to 25 minutes alternative than for the 25 to 45 minutes alternative. On the other hand, *I-270* shows a negative effect on shortening incident durations in CPD-Submodel I.
3. Particular locations (exits) on I-495 and I-270 cause longer incident durations. This is reflected in several MNL models with negative coefficients for the related variables in short incident duration alternatives, e.g., 5 to 25 or 25 to 45 minutes. Exits that commonly appeared to have this kind of effect were 27, 33, 36, and 39 on I-495 and 1, 4, 9, and 18 on I-270. The reason for this can be found in the complexity of geometric configurations around these exit areas or in their long distances from traffic operation centers. Especially when incidents occurred in the areas around exits 33, 36, and 39 on I-495 and exits 1 and 4 on I-270, the response and clearance times for the incidents were longer, due to the difficulty in access caused by complex geometric configurations and heavy traffic at those locations. I-495 splits from I-270 at exits 34 and 35, and merges again with I-270 at exit 38. I-270 splits into two directions at exit 2 and merges with I-495 at exits 34 and 38. Such features around this area cause heavy weaving traffic, which can interrupt the main stream.
4. Response times were proportional to the incident durations in CPD-Submodels, and this relation exhibits a negative coefficient for the shortest incident duration alternative in CPD-Submodels I and II.

5. In CPD-Submodel II, *Incident Hour* represents the hour in time when the incident occurred; this showed a strong relationship with incident durations. The format of *Incident Hour* is defined as numbers from 0 to 23 without AM or PM. Developed submodels implied that incidents occurring in the evening (approximately 20 ~ 23) were likely to last longer than those occurred in any other time. This effect is similar to the one from *Night* factor, but with more sensitivity to each hour.

As shown in Table 5.4, the probabilities for the three categories of incident durations do not show large discrepancies from one another in the submodels for *CPI*. For example, for two categories (25 to 45 minutes and > 45 minutes) in CPI-Submodel I, the difference in probability was only about 2 percent. Similar phenomena can also be found in CPI-Submodels II and III for the first two categories of incident durations. In MNL models for *CPD*, the difference in probability between alternatives was larger, but still no alternative dominated the entire data set (i.e., had over 70 percent probability). For this reason, such probabilistic models as MNL models must be applied to those subsets for which it is hard to find any short range of incident durations with a high probability of satisfying given conditions.

Developed MNL models were validated with year 2006 data set. By using this data set, predicted probabilities for each incident duration category in each model were found and summarized in Table 5.4. The difference between the estimated and validated probability was within 10 percent.

Table 5.2(a) CPI-Submodel I: Estimated ***Propensity*** Functions for the MNL Model

$$R_{5\text{-}25} = \underset{(0.9)}{0.910} \underset{(-2.9)}{-3.550*NoTT} \underset{(-2.4)}{-2.140*Night} \underset{(-2.4)}{-0.536*NoVehInv} \underset{(3.2)}{+2.434*I495} \underset{(-3.3)}{-3.053*NoSUT} \underset{(-2.3)}{-0.971*NoPUV} \underset{(1.6)}{+1.053*Pave\_Dry}$$

$$R_{25\text{-}45} = \underset{(2.9)}{2.131} \underset{(-2.0)}{-1.241*NoTT} \underset{(-3.2)}{-2.678*Night} \underset{(-2.4)}{-0.536*NoVehInv} \underset{(1.9)}{+1.253*I495} \underset{(-3.3)}{-3.053*NoSUT}$$

R_{gt45} = 0 (Base)

The number of observations used : 98
Likelihood with zero coefficients = -106.5654
Likelihood with constants only = -105.5362
Final value of Likelihood = -76.2511

Note : Numbers in parentheses are *t*-statistic values

<Legend>
I495 : 1 if an incident occurred on Interstate I-495; 0 otherwise
Night : Binary variable for incident time (Night=1, otherwise=0)
NoTT: Number of Tractor-trailers involved
NoPUV : Number of Pickup Vans involved
NoVehInv : Number of vehicles involved
NoSUT : Number of Single-Unit Trucks involved
Pave_Dry : 1 if Pavement Condition is Dry; 0 otherwise

Table 5.2(b) CPI-Submodel II: Estimated ***Propensity*** Functions for the MNL Model

$$R_{5\text{-}25} = \underset{(2.5)}{1.952} + \underset{(2.0)}{1.827}*I270 \underset{(-3.1)}{-0.655}*NoVehInv + \underset{(2.3)}{2.663}*I495 \underset{(-2.7)}{-2.776}*Pave\_SI \underset{(-2.1)}{-2.050}*Ex495$$

$$R_{25\text{-}50} = \underset{(2.0)}{1.576} + \underset{(1.8)}{1.568}*I270 \underset{(-2.2)}{-0.422}*NoVehInv + \underset{(2.1)}{2.471}*I495 \underset{(-2.7)}{-3.626}*Pave\_SI \underset{(-2.3)}{-2.253}*Ex495$$

$R_{gt50} = 0$ (Base)

The number of observations used : 189
Likelihood with zero coefficients = -206.5391
Likelihood with constants only = -179.5752
Final value of Likelihood = -167.4129

Note : Numbers in parentheses are *t*-statistic values

<Legend>
I495 : 1 if an incident occurred on Interstate I-495; 0 otherwise
I270 : 1 if an incident occurred on Interstate I-270; 0 otherwise
NoVehInv : Number of vehicles involved
Ex495 : Binary variable to indicate the specific locations on I-495 (exit no. 27, 28, 33, 34, 36, 38, 39)
Pave_SI : 1 if Pavement Condition is Snow/Ice; 0 otherwise

Table 5.2(c) CPI-Submodel III: Estimated ***Propensity*** Functions for the MNL Model

$R_{5\text{-}25}$ = 1.868 -3.346*NoTT -2.773*Night -2.509*PEAKHR -3.874*Ex270
(2.8) (-3.2) (-2.1) (-2.2) (-3.6)

$R_{25\text{-}45}$ = 3.031 -3.346*NoTT -1.603*Night -2.095* PeakHR -2.727* Ex270 -0.865*Ex495 -1.099*Pave_Dry
(3.8) (-3.2) (-1.7) (-1.9) (-3.1) (-1.5) (-2.1)

R_{gt45} = 0 (Base)

The number of observations used : 82
Likelihood with zero coefficients = -90.0862
Likelihood with constants only = -85.9470
Final value of Likelihood = -65.3223

Note : Numbers in parentheses are *t*-statistic values

<Legend>
Ex495 : Binary variable to indicate the specific locations on I-495 (exit nos. 27, 28, 33, 34, 36, 38, 39)
Ex270 : Binary variable to indicate the specific locations on I-270 (exit nos. 1, 4, 9, 13, 15, 18, 22)
Night : Binary variable for incident time (Night=1, otherwise=0)
NoTT : Number of Tractor-trailers involved
PeakHR : 1 if an incident occurred during peak hours; 0 otherwise
Pave_Dry : 1 if Pavement Condition is Snow/Ice; 0 otherwise

Table 5.3(a) CPD-Submodel I: Estimated ***Propensity*** Functions for the MNL Model

==

$$R_{5\text{-}30} = \underset{(3.4)}{8.517} \underset{(-3.3)}{-4.610}*NoTT \underset{(-1.8)}{-2.390}*NoPUV \underset{(-1.9)}{-0.136}*RespTm \underset{(-2.5)}{-3.804}*I270$$

$R_{gt30} = 0$ (Base)

==

The number of observations used : 46
Likelihood with zero coefficients = -31.8848
Likelihood with constants only = -30.7891
Final value of Likelihood = -13.7119

Note : Numbers in parentheses are *t*-statistic values

<Legend>
NoTT : Number of Tractor-trailers involved
NoPUV : Number of Pickup Vans involved
RespTm : Response Time in minutes
I270 : 1 if an incident occurred on Road I-270; 0 otherwise

Table 5.3(b) CPD-Submodel II: Estimated ***Propensity*** Functions for the MNL Model

==

$R_{5\text{-}25}$ = 6.772 -0.169 *IncHR -0.782*NoVehInv -3.078*Ex495 -3.333*Ex270 +1.228*Pave_Dry -0.089*RespTm
(4.1) (-2.4) (-2.0) (-3.6) (-3.1) (1.7) (-3.2)

$R_{25\text{-}45}$ = 5.155 -0.171*IncHR -0.948*NoVehInv -2.654*Ex495 -2.883*Ex270 +1.572*SUT_Ind +1.349*Pave_Dry
(3.1) (-2.3) (-2.2) (-3.0) (-2.4) (2.4) (1.8)

R_{gt45} = 0 (Base)

==

The number of observations used : 109
Likelihood with zero coefficients = -119.7487
Likelihood with constants only = -107.2160
Final value of Likelihood = -79.9817

Note : Numbers in parentheses are *t*-statistic values

<Legend>
IncHR : Hour in time incident occurred (0 ~ 23)
NoVehInv : Number of vehicles involved
Ex495 : Binary variable to indicate the specific locations on I-495 (exit nos. 27, 33, 36, 39, 41)
Ex270 : Binary variable to indicate the specific locations on I-270 (exit nos. 1, 4, 9, 18)
SUT_Ind : 1 if Single-Unit Trucks involved; 0 otherwise
Pave_Dry : 1 if Pavement Condition is Dry; 0 otherwise
RespTm : Response Time in minutes

Table 5.4 Summary of Incident Duration Probabilities Estimated and Validated by MNL Submodels

	Submodel I				Submodel II				Submodel III			
	Incident Duration (mins)	Obs. Prob.	Est. Prob.	Val. Prob.	Incident Duration (mins)	Obs. Prob.	Est. Prob.	Val. Prob.	Incident Duration (mins)	Obs. Prob.	Est. Prob.	Val. Prob.
CPI	[5, 25] (25, 45] > 45	0.276 0.378 0.346	0.265 0.378 0.357	0.328 0.388 0.284	[5, 25] (25, 50] > 50	0.481 0.408 0.111	0.483 0.408 0.108	0.494 0.428 0.078	[5, 25] (25, 45] > 45	0.366 0.439 0.195	0.366 0.439 0.195	0.461 0.379 0.160
CPD	[5, 30] > 30	0.609 0.391	0.609 0.391	0.576 0.424	[5, 25] (25, 45] > 45	0.550 0.285 0.165	0.550 0.285 0.165	0.609 0.235 0.156	N/A	N/A	N/A	N/A

Note: Val. Prob. stands for ‘Validated Probability’.

5.3 Supplemental Model – 2: Multiple Linear Regression Models

5.3.1 Multiple Linear Regression Models

Linear regression is one of the most widely studied and used statistical and econometric techniques for its usefulness in modeling various relationships between variables. Moreover, numerical estimation, interpretation, and application of regression models are relatively easy, since such models can be solved by a number of non-specialty commercial statistical software.

Multiple linear regression models include two or more independent variables, assuming that the dependent variable is a linear function of a series of independent variables and an error term. In general, the multiple linear regression models can be mathematically expressed as

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad \text{(Eq. 5.3)}$$

where, Y_i is the dependent variable, X_{ki} is the ith observation on independent variable X_k, ε_i is the error term, and β_k is the estimated coefficient for independent variable X_k. β_k is estimated in a way to minimize the error sum of squares (known as a *least-squares* procedure), defined as

$$ESS = \sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y}_i)^2 \qquad \text{(Eq. 5.4)}$$

where, $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$, and $\hat{\beta}_k$ is the slope estimate. Since there are numerous references and much literature regarding this estimation technique (Washington et al., 2003; Pindyck and Rubinfeld, 1998), this report will not discuss it in detail.

5.3.2 Estimation Results with Multiple Linear Regression Models

The estimated multiple linear regression models for *Collision-Fatality* and *Others* incidents are presented in Tables 5.6 and 5.7, respectively. Generally, the sign and magnitude of estimated coefficients for variables included in models were as expected. Independent variables were tested at the 90 percent significance level, which means that the absolute value of the *t*-statistic should be greater than or equal to 1.65 for that variable to be considered as significant. The estimated models for *Collision-Fatality* and *Others* were valid at the 90 percent significance level, since the p-values for both models were less than 0.0001. Specific discussions are summarized below for each model.

Collision-Fatality (CF)

1. As shown in Table 5.5, the involvement of heavy vehicles (tractor-trailers and single-unit trucks) increased fatality incident durations, and this result is similar to the result from RBTMs. The results also confirmed that the increase in the number of blocked lanes in the same direction, including shoulder lanes, contributed to the reduction of incident durations. This observation was reflected in the term *Ratio_sdbl*SHDBK* having a negative coefficient and a high *t*-statistic value (i.e., -2.87).
2. As mentioned in Chapter 4, one interesting finding from RBTMs regarding *Collision-Fatality* incidents was the decrease of incident durations in the wet pavement condition. This finding was also reflected in this estimated linear regression model as a negative coefficient and a high *t*-statistic value (i.e., -2.11) for the wet pavement condition.

3. This model, also reflecting that durations increased for *Collision-Fatality* incidents occurring on Interstates I-68 or MD/I-295, gave a positive coefficient to this binary variable.
4. The observation from RBTMs, that the durations of fatality incidents occurring at night are likely to be longer than in the daytime, was reflected in this estimated model by a positive coefficient for the binary variable *Night*.
5. As shown in Table 5.7, the overall percentage of correct estimation result for durations of fatality incidents using the estimated regression model was 74.7 percent. Incident durations greater than or equal to 120 minutes were well estimated, while other categories for incident durations were not estimated correctly at all.
6. The model was tested using the validation data set — incidents occurring in year 2006 — and the results are shown in Table 5.8. The overall percentage of correct prediction result was 78.1 percent, which was slightly higher than the one for estimation. Similarly, the predictions for incident durations greater than or equal to 120 minutes were satisfactory, while the predictions for other categories of incident durations were unsatisfactory.
7. Absolute error, defined as the absolute value of the difference between observed and estimated/predicted value, was also computed as a reference to evaluate the estimated incident durations model. In the model estimation results, 50.7 percent of records showed an absolute error within 30 minutes, while 81.3 percent of records had an absolute error within 60 minutes. For model results with the

validation data set, 40.6 percent and 59.3 percent of records showed an absolute error within 30 minutes and 60 minutes, respectively.

Others

1. Unlike the linear regression model for *CF* incidents, the dependent variable in the model for *Others* is the logarithm of the incident duration. This transformation of the dependent variable was applied to identify linear relationships between the dependent and independent variables, which is a requirement of the regression modeling framework (Washington et al., 2003).
2. According to the estimated linear regression model, heavy vehicle (tractor-trailers and single-unit trucks) involvement was likely to increase incident durations. Tractor-trailer involvement (*TT_Ind*) showed an especially strong positive relationship with incident durations, as the *t*-statistic value for this was very high (i.e., 4.64). This relation was not found in the RBTMs.
3. The model, reflecting the observation that incident durations for *Others* increased as the number of blocked lanes in the same direction increases, assigned a positive coefficient to that variable. Response times also showed a strong positive relationship with incident durations in the estimated model.
4. Among other incident natures, *Debris* showed a negative relationship with incident durations, while *Emergency Road Work* had a positive relationship. Other events in incident type *Others* did not show any significance with incident durations. That is, durations of incidents caused by debris were likely to be shorter than those caused by any other incident type in *Others*. On the other hand,

emergency road work caused longer incident durations than any other type of incident in *Others*.

5. The overall percentage of correct estimation was 66 percent, as shown in Table 5.9. For relatively short (i.e., 5 to 30 minutes) and long (i.e., >=120 minutes) incident durations, the model performed well. However, for incident durations between 30 and 120 minutes, the model did not give a good estimation. Especially for incidents with durations between 60 and 120 minutes, this model did not give any correct estimations.
6. For predicted results based on the validation data set, as summarized in Table 5.10, the overall percentage of correct prediction percentage was slightly lower (i.e., 61.1 percent) than the estimation results. This table also shows that the model predicted incident durations between 5 and 60 minutes quite well, while incident durations longer than 60 minutes were not predicted correctly at all.
7. An absolute error was also computed for each record in the model development and validation data set. In the data set used for model development, 61.7 percent of records showed an absolute error within 15 minutes, while 80.9 percent of them showed it within 30 minutes. In the validation data set, the results for absolute errors were similar to these in the model development data set; 61.1 percent of them were within 15 minutes, while 77.8 percent of them were within 30 minutes.

Table 5.5 Estimated Multiple Linear Regression Model for Incident Nature-CF

Incident Duration (mins) = 162.95	- 31.94*Pave_Wet	+ 32.05*NoSUT	+ 42.03*NoTT	+ 29.50*Night	+ 59.10*Rd68_295	- 42.03*Ratio_sdbl*SHDBK
(13.64)	(-2.11)	(2.02)	(3.17)	(2.33)	(2.47)	(-2.87)

Number of observations used : 75
$R^2 = 0.3730$
F-value for Model = 6.74
P-value for Model = < 0.0001

Note : Numbers in parentheses are *t*-statistic values

<Legend>
Pave_Wet : 1 if Pavement Condition is Wet; 0 otherwise
NoSUT : Number of Single-Unit Trucks involved
NoTT: Number of Tractor-trailers involved
Night : Binary variable for incident time (Night=1, otherwise=0)
Rd68_295 : 1 if an incident occurred on Road I-68 or MD/I-295
Ratio_sdbl : Number of lanes blocked in same direction/Number of lanes in that direction
SHDBK : 1 if Shoulder lane is blocked; 0 otherwise

Table 5.6 Estimated Multiple Linear Regression Model for Incident Nature-Others

$$\text{Log(Incident Duration)} = \underset{(13.03)}{2.67} + \underset{(2.28)}{0.96}*\text{SUT\_Ind} + \underset{(4.64)}{1.73}*\text{TT\_Ind} + \underset{(2.38)}{0.23}*\text{No\_sdbl} + \underset{(2.31)}{0.04}*\text{RespTm} - \underset{(-1.93)}{0.72}*\text{Debris}$$

$$+ \underset{(2.00)}{1.83}*\text{EmgRdWk}$$

Number of observations used : 47
$R^2 = 0.6017$
F-value for Model = 10.07
P-value for Model = < 0.0001

Note : Numbers in parentheses are *t*-statistic values

<Legend>
SUT_Ind : 1 if Single-Unit Trucks are involved; 0 otherwise
TT_Ind : 1 if Tractor-trailers are involved; 0 otherwise
No_sdbl : Number of lanes blocked in same direction
RespTm : Response Time in minutes
Debris : 1 if Incident Nature is Debris; 0 otherwise
EmgRdWk : 1 if Incident Nature is Emergency Road Work; 0 otherwise

Table 5.7 Estimated Results of Multiple Linear Regression Model for Incident Nature *CF*

Incident Duration (mins)	Estimated				
Observed	< 60	[60, 90)	[90, 120)	>=120	Correct Percent
< 60	0	0	0	1	0.0%
[60, 90)	0	0	2	5	0.0%
[90, 120)	0	1	0	9	0.0%
>=120	0	1	0	56	98.2%
Overall Correct Percent	N/A	0.0%	0.0%	78.9%	74.7%

Note: sample size is 75.

Table 5.8 Predicted Results of Multiple Linear Regression Model for Incident Nature *CF*

Incident Duration (mins)	Predicted				
Observed	< 60	[60, 90)	[90, 120)	>=120	Correct Percent
< 60	0	0	0	0	N/A
[60, 90)	0	0	0	5	0.0%
[90, 120)	0	0	0	7	0.0%
>=120	0	0	2	50	96.2%
Overall Correct Percent	N/A	N/A	0.0%	80.6%	78.1%

Note: sample size is 64.

Table 5.9 Estimated Results of Multiple Linear Regression Model for Incident Nature *Others*

Incident Duration (mins)	Estimated					
Observed	[5, 30)	[30, 60)	[60, 90)	[90, 120)	>=120	Correct Percent
[5, 30)	21	2	1	0	0	87.5%
[30, 60)	9	5	0	0	0	35.7%
[60, 90)	0	1	0	0	0	0.0%
[90, 120)	0	0	2	0	0	0.0%
>=120	0	1	0	0	5	83.3%
Overall Correct Percent	70.0%	55.6%	0.0%	N/A	100.0%	66.0%

Note: sample size is 47.

Table 5.10 Predicted Results of Multiple Linear Regression Model for Incident Nature – *Others*

Incident Duration (mins)	Predicted					
Observed	[5, 30)	[30, 60)	[60, 90)	[90, 120)	>=120	Correct Percent
[5, 30)	8	1	0	0	0	88.9%
[30, 60)	4	3	0	0	0	42.9%
[60, 90)	0	0	0	1	0	0.0%
[90, 120)	0	0	0	0	1	0.0%
>=120	0	0	0	0	0	N/A
Overall Correct Percent	66.7%	75.0%	N/A	0.0%	0.0%	61.1%

Note: sample size is 18.

To sum up, linear regression models are suitable for finding the relationships between incident duration and its factors. In the estimated regression models, several findings discovered from RBTMs are confirmed. Given these estimation/prediction results and absolute errors, further research is recommended for more reliable models, especially for *CF* incidents. This recommendation is also supported by Figures 5.3 to 5.6, since they show that the incident durations between observed and estimated/predicted for *CF* are quite different, while those for *Others* are close. In general, fatality incidents cause longer incident durations, and they require a more specific and systematic incident management strategy based on well-predicted incident durations to soothe their impact (i.e., traffic congestion or delay). To achieve this, the first thing to accomplish is to collect additional incident records with additional information, e.g., the number of pedestrians, drivers, and/or occupants injured or killed and the collision type (head-on, rear-end, etc).

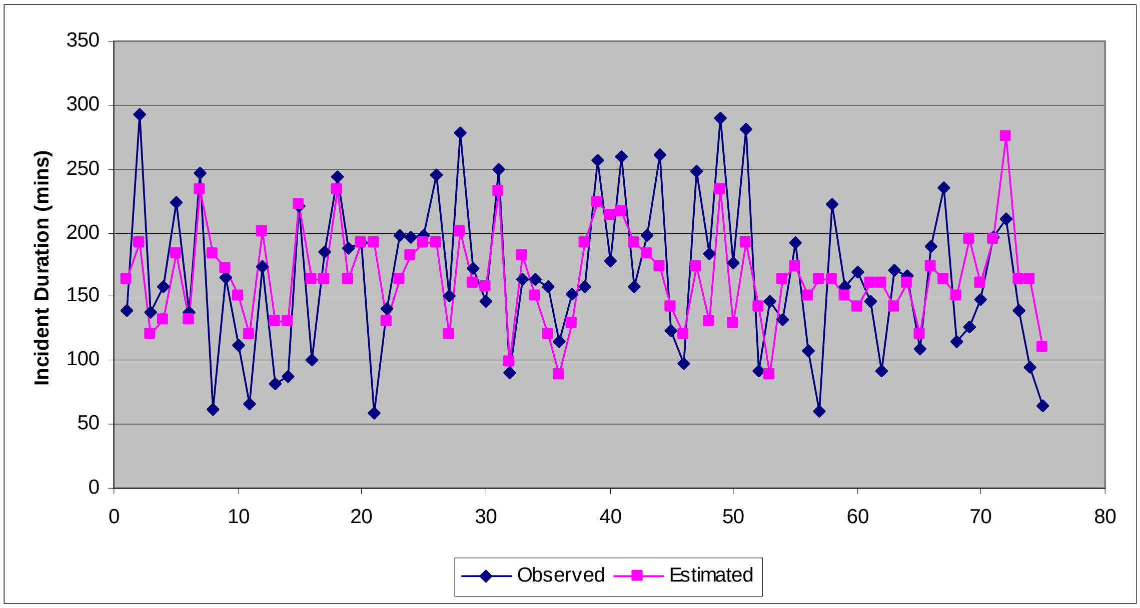


Figure 5.3 Comparisons between Observed and Estimated Incident Durations Using Developed Multiple Linear Regression Model for Incident Nature-*CF*

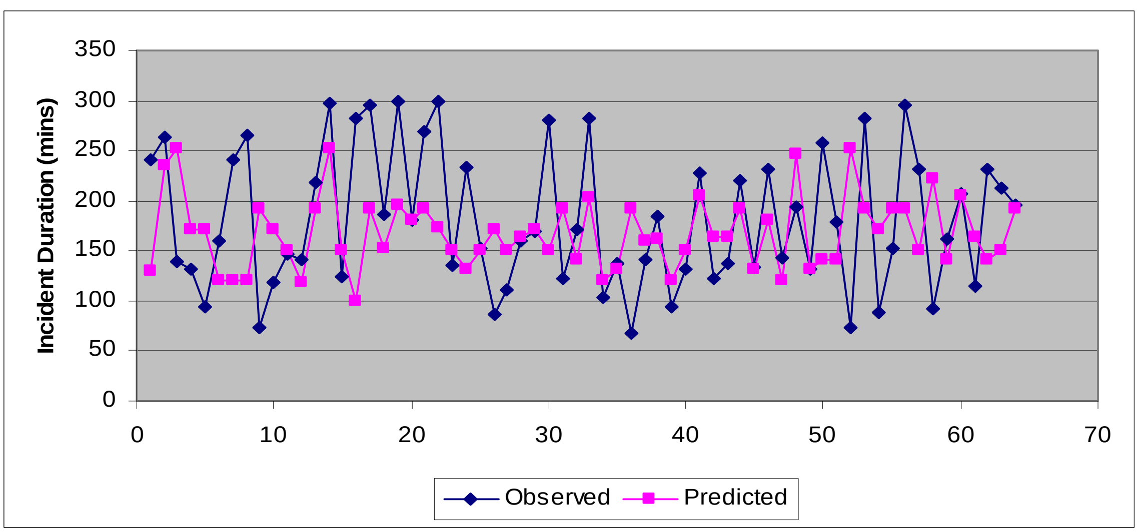


Figure 5.4 Comparisons between Observed and Predicted Incident Durations Using Developed Multiple Linear Regression Model for Incident Nature-*CF*

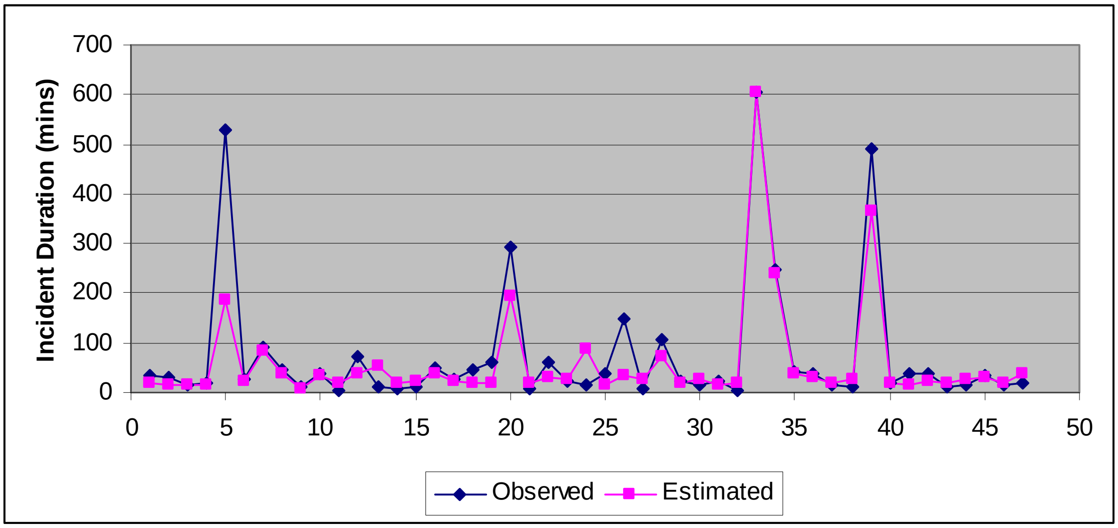


Figure 5.5 Comparisons between Observed and Estimated Incident Durations Using Developed Multiple Linear Regression Model for Incident Nature-*Others*

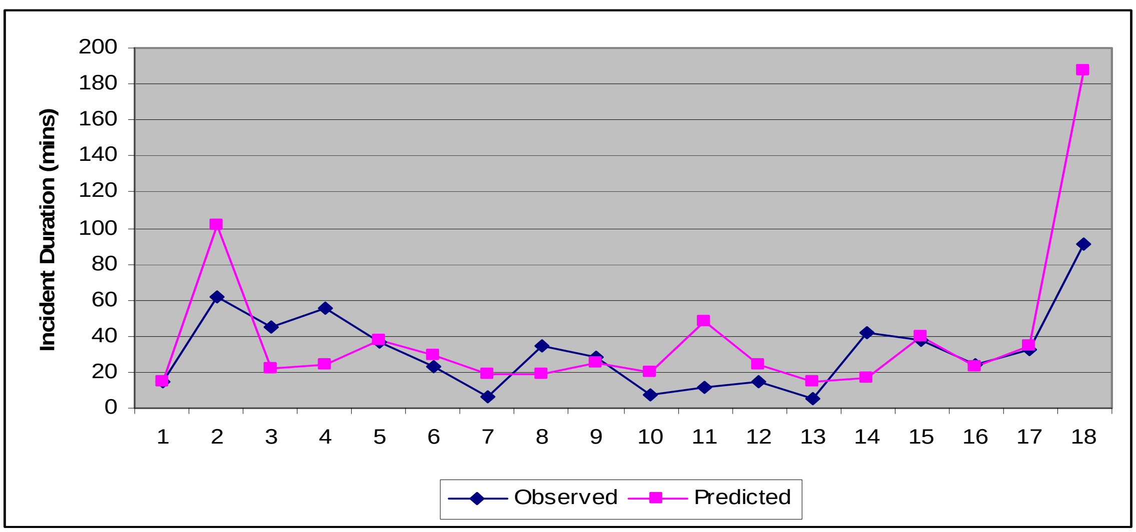


Figure 5.6 Comparisons between Observed and Predicted Incident Durations Using Developed Multiple Linear Regression Model for Incident Nature-*Others*

5.4 Illustrative Description of the Application of the Developed Model

To apply the developed model in this study to real-time incident management operations, reliable information about a detected incident should first be acquired promptly from dispatched response units. Then, using that information, operators can employ the following steps to predict the approximate range of incident duration.

Step 1: Identify the detected incident nature and location of its jurisdiction to select the appropriate RBTM.

Step 2: Trace the selected RBTM from its root to the corresponding terminal node using the traffic incident information provided by dispatched units.

Step 3: At the corresponding terminal node, take the predicted incident duration if the predicted outcome satisfies the evaluation criteria based on its historical data set.

Step 4: Otherwise, trace back node by node until the node satisfying the evaluation criteria is found.

Step 5: If one cannot find a satisfactory node in the RBTMs, then use a supplemental model to predict the incident duration or the probability distribution of the target incident durations.

This whole process can be expedited if the models (RBTMs and supplemental models), along with evaluation criteria, are programmed with a user-friendly interface.

Table 5.11 provides actual examples of traffic incident information from dispatched units and the predicted incident duration using the RBTMs and supplemental models. Variable names appearing in Table 5.11 are described in Table 5.12. The first example concerns a fatality incident occurring in Prince George County in 2006. The

RBTM predicted incident duration of 80 to 100 minutes with 33.33 percent ***confidence*** based on the data set collected from years 2003 to 2005. Since it did not satisfy one of our criteria (i.e., ***confidence*** should be greater than 70 percent), we used a supplemental model (the multiple linear regression model presented in Table 5.5) to obtain a more reliable prediction of incident duration. The model predicted that the incident duration would be approximately 121 minutes, and this prediction was closer to the observed incident duration, 144 minutes, than the one from the RBTM. The same phenomenon was observed in the fifth example.

For the second example, the RBTM predicted incident duration of 10 to 35 minutes with 75 percent ***confidence***. Since this outcome satisfied our criteria, application of a supplemental model to this case was not needed. A similar explanation can be applied to the fourth example for the disabled vehicle incident. On the other hand, the third example showed only 60 percent ***confidence*** with the RBTM, so a supplemental model was required. The supplemental model (the MNL model presented in Table 5.3(b)) predicted an incident duration of 5 to 25 minutes with 0.84 probability, and this was taken as the predicted duration of the detected incident. Note that the RBTM predicted a duration of 30 to 45 minutes, which was quite far from the observed incident duration of 12.82 minutes.

Table 5.11 Traffic Incident Information Examples and Their Predicted Incident Duration

<table>
<tr><td>Example No.</td><td>1</td><td>2</td><td colspan="2">3</td><td>4</td><td>5</td></tr>
<tr><td>Event Open
Date & Time</td><td>2006-03-25
14:14:33</td><td>2003-07-02
18:33:52</td><td colspan="2">2006-11-06
19:47:01</td><td>2004-05-26
08:19:00</td><td>2006-02-02
08:57:28</td></tr>
<tr><td>County</td><td>Prince George</td><td>Montgomery</td><td colspan="2">Montgomery</td><td>Montgomery</td><td>Montgomery</td></tr>
<tr><td>Incident Nature</td><td>CF</td><td>CPI</td><td colspan="2">CPD</td><td>DISABLED</td><td>OTHERS
(Vehicle Fire)</td></tr>
<tr><td>Pavement Condition</td><td>Dry</td><td>Dry</td><td colspan="2">Dry</td><td>Dry</td><td>Dry</td></tr>
<tr><td>Road Info</td><td>I-495 IL</td><td>I-495 IL</td><td colspan="2">I-270 N</td><td>I-270 S</td><td>I-495 OL</td></tr>
<tr><td>Exit No.</td><td>23</td><td>41</td><td colspan="2">11</td><td>5</td><td>38</td></tr>
<tr><td>CHART Involvement</td><td>1</td><td>1</td><td colspan="2">1</td><td>1</td><td>1</td></tr>
<tr><td>SUT_Ind</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>PUV_Ind</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>TT_Ind</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>No_TT</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>No_SUT</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>No_PUV</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>No_Veh_Inv</td><td>3</td><td>1</td><td colspan="2">2</td><td>1</td><td>1</td></tr>
<tr><td>Weekend</td><td>1</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>Peak Hour</td><td>0</td><td>0</td><td colspan="2">0</td><td>1</td><td>1</td></tr>
<tr><td>no_sd_lane_bl</td><td>4</td><td>1</td><td colspan="2">1</td><td>0</td><td>1</td></tr>
<tr><td>no_od_lane_bl</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>no_shd_bl</td><td>2</td><td>1</td><td colspan="2">0</td><td>0</td><td>1</td></tr>
<tr><td>Shoulder Blockage</td><td>1</td><td>1</td><td colspan="2">0</td><td>0</td><td>1</td></tr>
<tr><td>total_lane_bl</td><td>4</td><td>1</td><td colspan="2">1</td><td>0</td><td>1</td></tr>
<tr><td>ratio_sd_bl</td><td>1</td><td>0.25</td><td colspan="2">0.125</td><td>0</td><td>0.25</td></tr>
<tr><td>ratio_od_bl</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>ratio_total_bl</td><td>1</td><td>0.12</td><td colspan="2">0.125</td><td>0</td><td>0.25</td></tr>
<tr><td>no_lane_one</td><td>4</td><td>4</td><td colspan="2">8</td><td>4</td><td>4</td></tr>
<tr><td>Incident Hour</td><td>14</td><td>18</td><td colspan="2">19</td><td>8</td><td>8</td></tr>
<tr><td>Night</td><td>0</td><td>0</td><td colspan="2">0</td><td>0</td><td>0</td></tr>
<tr><td>Response Time
(minutes)</td><td>0.38</td><td>23.91</td><td colspan="2">0.17</td><td>0.78</td><td>2.05</td></tr>
<tr><td>Clearance Time
(minutes)</td><td>143.15</td><td>8.81</td><td colspan="2">12.65</td><td>5.6</td><td>5.02</td></tr>
<tr><td>Observed-INCDm[1]</td><td>143.53</td><td>32.73</td><td colspan="2">12.82</td><td>6.38</td><td>7.07</td></tr>
<tr><td>Predicted-RBTM[2]</td><td>(80, 100]</td><td>(10, 35]</td><td colspan="2">(30, 45]</td><td>[5, 30]</td><td>(30~50]</td></tr>
<tr><td rowspan="4">Predicted-SM[3]</td><td rowspan="4">120.93</td><td>SM II[4]</td><td>SM II[5]</td><td>[5, 25]</td><td rowspan="4">N/A</td><td rowspan="4">19.78</td></tr>
<tr><td rowspan="3">N/A</td><td>[5, 25]</td><td>0.84</td></tr>
<tr><td>(25, 45]</td><td>0.13</td></tr>
<tr><td>> 45</td><td>0.03</td></tr>
<tr><td>Confidence
in RBTM[2]</td><td>33.33%</td><td>75.00%</td><td colspan="2">60.00%</td><td>78.13%</td><td>44.44%</td></tr>
</table>

[1] Observed incident duration in minutes
[2] Predicted incident duration in minutes based on RBTM
[3] Predicted incident duration in minutes based on Supplemental Models (SM)
[4] CPI-Submodel II presented in Table 5.2(b)
[5] CPD-Submodel II presented in Table 5.3(b)

Table 5.12 Descriptions of Variable Names

Example No.	Example number
Event Open Date & Time	Date and time of incident occurred
County	County
Incident Nature	Incident nature
Pavement Condition	Pavement condition
Road Info	Road information (Road name and direction)
Exit No.	Exit number for I-495, I-95, I-695 and I-270
CHART Involvement	1 if CHART is involved; 0 otherwise
SUT_Ind	1 if any single-unit truck is involved; 0 otherwise
PUV_Ind	1 if any pick up van is involved; 0 otherwise
TT_Ind	1 if any tractor-trailer is involved; 0 otherwise
No_TT	Number of tractor-trailers involved
No_SUT	Number of single-unit trucks involved
No_PUV	Number of pick up vans involved
No_Veh_Inv	Number of vehicles involved
Weekend	1 if the incident occurred day is weekend; 0 otherwise
Peak Hour	1 if the incident occurred time is peak hours; 0 otherwise
no_sd_lane_bl	Number of blocked lanes in the same direction
no_od_lane_bl	Number of blocked lanes in the opposite direction
no_shd_bl	Number of blocked shoulder lanes
Shoulder Blockage	1 if any shoulder lane is blocked; 0 otherwise
total_lane_bl	Total number of blocked lanes in same and opposite direction
ratio_sd_bl	= no_sd_lane_bl / no_lane_one
ratio_od_bl	= no_od_lane_bl /no_lane_one
ratio_total_bl	= total_lane_bl / (2×no_lane_one)
no_lane_one	Number of lanes in same direction
Incident Hour	Hour in time that an incident is detected (occurred) (0 to 23)
Night	0 if 6 <= Incident Hour < 20; 0 otherwise
Response Time (minutes)	Response Time in minutes
Clearance Time (minutes)	Clearance Time in minutes
Observed-INCDm	Observed incident duration in minutes
Predicted-RBTM	Predicted incident duration in minutes based on RBTM
Predicted-SM	Predicted incident duration in minutes based on supplemental models
Confidence in RBTM	***Confidence*** based on RBTMs

Chapter 6: Advanced Study for Collision-Fatality Incidents

6.1 Introduction

Traffic congestion has long been considered as one of the primary operational problems in urban highway networks. Although incidents on highways, regardless of their resulting severities, will cause significant delays due to reductions in capacity, those incidents involving fatalities, personal injuries, or property damages generally require much longer clearance times and, consequently, long traffic queues, as well as potential secondary incidents.

The durations of fatality incidents are mostly distributed in the wide range of time from about 60 minutes up to 400 minutes. This implies that it takes traffic at least 60 minutes to recover from an incident which results in fatalities, and most fatality incidents can last up to three or four hours, or even longer. Their uniquely long durations highlight the need to develop a separate prediction model for such incidents. In addition, the annual frequency of fatality incidents is much smaller than the frequency of other types of incident. Thus, fatality incidents are highly likely to be considered as outliers when developing a generalized incident duration models.

Chapters 4 and 5 developed an individual model for *CF* incidents using RBTMs and regression models. Although the developed methodology works quite well for other kinds of incidents, it has performed unsatisfactorily in *CF* incidents.

To enhance model performance in predicting durations of fatality incidents required the inclusion of more detailed information, especially information related to the severity of collisions, which, in the CHART-II database, is usually reflected by the number of vehicles involved. However, to fully capture the complex characteristics of fatality incidents, it is essential to include additional factors, such as the number of fatalities or injuries, to capture the majority of the duration variances of fatality-incidents. To accomplish that aim, this study integrated the Accident Report database, archived in the Maryland State Police Department, with CHART-II to explore more information. A detailed description of the database utilized for this study is provided in the next section.

6.2 Integrated Data Descriptions

This study was conducted with two different databases, namely the CHART-II database and the Accident Report database, managed by MDSHA and the Maryland State Police Department, respectively.

While CHART-II includes many kinds of measurable and observable information for characterizing responded incidents regardless of severity, the Accident Report database is the primary source for providing more comprehensive information, particularly for severe incidents resulting in fatalities and personal injuries. Those two databases were integrated manually in order to generate a high-quality database.

The additional variables retrieved from the Accident Report DB included:

- Severity information: the number of fatalities and injuries for pedestrians or driver/occupants;
- Collision information: collision types and vehicle movements; and
- Others: weather, light and road division conditions.

Figure 1, created based on the fatality incidents in the integrated database, shows that the distribution of incident durations for fatalities used in this study is right-skewed. The fatality incidents hardly appear in the short and middle ranges of durations, that is, less than 60 minutes, while about 95 percent of them are distributed in the range between 60 minutes and 360 minutes. This range is considerably wider than those for other incident natures. For instance, about 95 percent of incidents causing only property damage are distributed in the range of five to 90 minutes in CHART-II.

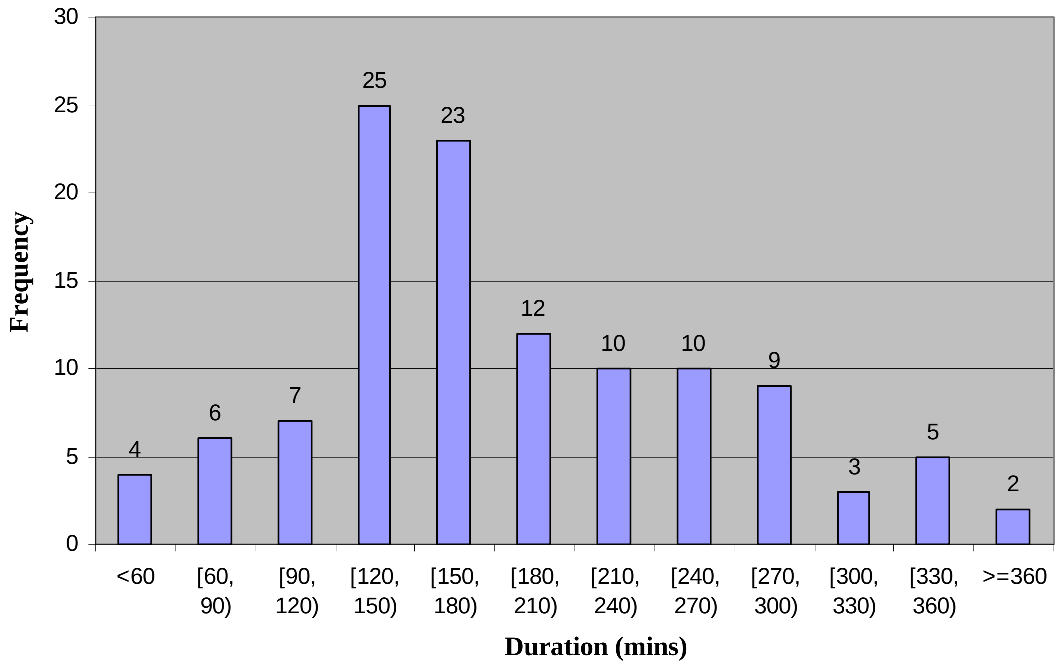


Figure 6.1 Distribution of Durations of Fatality Incidents in the Integrated Database

Another feature of fatality-incidents, along with distribution over a wide range, is small sample size. Basically, the frequency of fatality incidents is much smaller than the frequency of other incidents of different types. In CHART-II, the fatality incidents constitute approximately one percent of the total incident records. Since some of those records had information missing, only several dozens of fatality incidents with complete information were available annually for the analysis. For this study, using the integrated databases, the number of fatality incidents with complete information that could be extracted totaled 116 for years 2003 to 2006. Due to the expected small sample size, about two-thirds of them (76) were randomly selected for model estimation; the remaining 40 cases were used for model validation. The statistical summary of each data set is briefly described in Table 6.1.

Table 6.1 Statistical Summary for Estimation and Validation Data Sets

	Estimation Set	Validation Set
Average Incident Duration	191.40	182.21
Standard Deviation in Durations	83.73	65.62
Median in Durations	171.63	164.10
Average number of fatalities	1.08	1.05
Average number of personal injuries	1.00	0.76

Since this study adopts the Naïve Bayesian Classifier (NBC) as the primary methodology for model development, incident durations initially recorded as continuous values were converted to several predefined categories of time interval. Based on the distribution of fatality incident durations presented in Figure 6.1, this study has classified incidents lasting less than 60 minutes as one category, while incidents lasting over 300 minutes are classified as another category. For incidents lasting between 60 and 300 minutes, we further divided them into four intervals, each covering a range of 60 minutes, as shown in Table 6.2.

Table 6.2 Incident Duration Categories Converted for Model Development

Category	Corresponding Incident Duration
1	< 60 minutes
2	[60 ~ 120) minutes
3	[120 ~ 180) minutes
4	[180 ~ 240) minutes
5	[240 ~ 300) minutes
6	>=300 minutes

6.3 Methodology

6.3.1 Naïve Bayesian Classifier (NBC)

The NBC assigns the object I to one of the discrete categories, $D_1, D_2, \cdots, D_m$, based on its attributes, $X_1, X_2, \cdots, X_m$. To do so, the NBC first calculates probabilities that I belongs to each category, conditioned on the observed and measurable attributes. I is assigned to the category with the greatest probability. This classifier is modeled based on Bayes' theorem, with the strong assumption that the available attributes are mutually independent. In other words, an NBC assumes that the presence of a specific attribute is unrelated to the presence of any other attributes.

The probability that I belongs to each category is calculated based on conditions of the observed attributes, that is, $P(I \in D_i | X_1, X_2, \cdots, X_n)$. Applying Bayes' Theorem, this can be rewritten as

$$P(I \in D_i | X_1, X_2, \cdots, X_n) = \frac{P(I \in D_i) P(X_1, X_2, \cdots, X_n | I \in D_i)}{P(X_1, X_2, \cdots, X_n)}$$

Assuming that X_is are mutually independent, this reduces to

$$P(I \in D_i | X_1, X_2, \cdots, X_n) = \frac{P(I \in D_i) \prod_{j=1}^{n} P(X_j | I \in D_i)}{P(X_1, X_2, \cdots, X_n)}$$

for each category D_i. Due to the fact that the denominator will be the same for all categories, the only thing to be done is to compute the numerator for each category i. Then, we choose i satisfying

$$i^* \in \arg\max P(I \in D_i) \prod_{j=1}^{n} P(X_j | I \in D_i)$$

and assign I to category D_{i^*} (Boyles, 2007).

Applying NBCs to the incident duration model, the attribute X_i corresponds to an observable or measurable incident characteristic, such as pavement conditions, locations of incidents, number of vehicles involved, number of lanes blocked, and so on. Non-discrete features of incidents needed to be discretized first. Also, continuous values of incident durations were redefined as discrete categories of time interval, as discussed in the previous section. Once an incident occurs, the NBC would calculate the probability that the incident duration would fall into each category and choose the category of the incident duration with the highest probability. The advantage of doing so is that it can still compute the probability, even if the information of some attributes is unavailable.

Parameters (i.e., the probabilities $P(I \in D_i)$ and $P(X_j \mid I \in D_i)$) in this model can be estimated with relative frequencies from the training set. Thus, if given category and attribute values never occurred together in the training set, then the probability of the given conditions would be estimated as zero. This is problematic, since it would wipe out all information in the other probabilities when they are multiplied. To avoid this, zero probabilities were replaced by a small positive value when calculating these products.

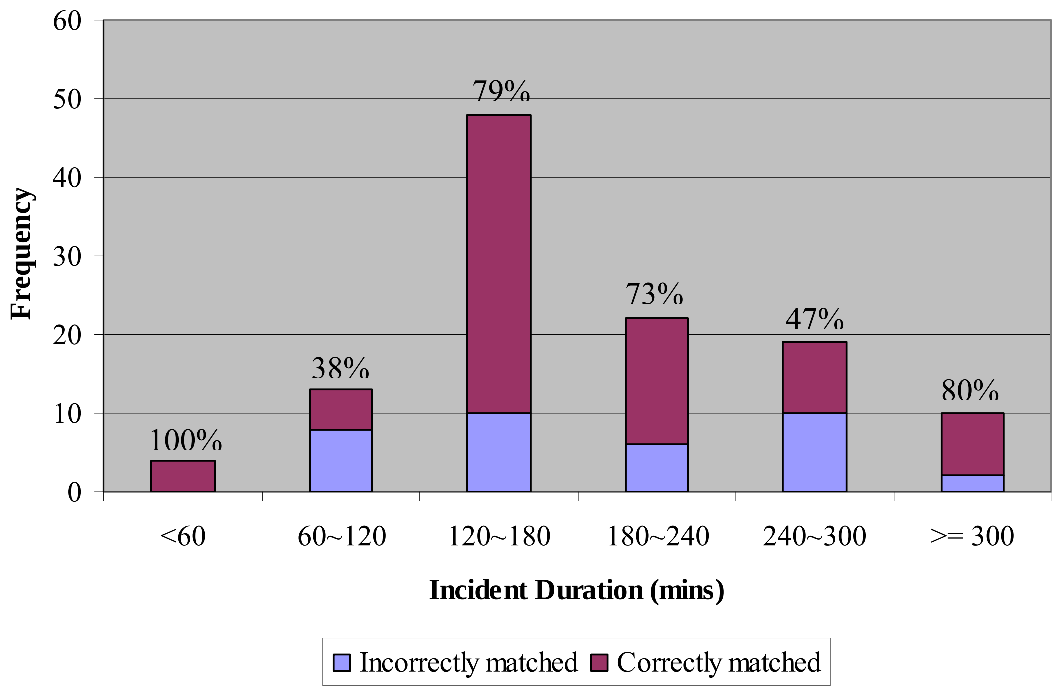
6.4 Model Development and Results

Since not all attributes positively correlate to the incident durations, it is necessary to identify the critical attributes that can improve the NBC model. To find the best attribute set, a simple NBC model for each attribute alone was first developed separately. By comparing all estimated results, attributes with best results were chosen as the initial set of attributes. Among those selected attributes, we selected any one attribute and made another set adding one more attribute. Then we compared results from those two attribute sets and took the best results. We repeated this process of adding attributes one by one until we acquired the best set of attributes. Additionally, attributes consisting of multiple categories were re-created as dummy variables for each category to investigate if any of those categories had a stronger impact on the model. Here is the selected list of the best set of attributes for the model:

- Counties;
- Pavement Conditions: Unknown, Dry, Wet or Snow/Ice;
- Number of Tractor-Trailers;
- Number of Pick-Up Vans;
- PM Peak Hour Indicator: 1 if occurred in 4 PM to 6:30 PM; 0 otherwise;
- Night Indicator: 1 if occurred in 8 PM to 6 AM; 0 otherwise;
- Number of Blocked Shoulder Lanes;
- Number of Drivers/Occupants Injured;
- Number of Drivers/Occupants Killed;
- Lighting Conditions: Daylight, Dawn/Dusk, Dark-Lights On or Dark-No lights;
- Collision Type — Head-On indicator;
- Collision Type — Head-On Left Turn indicator; and
- Road-795: 1 if an incident occurred on I-795; 0 otherwise.

To test the performance of the calibrated model, the estimated categories of incident durations were compared with the observed categories of incident durations. The estimated categories matched exactly with the observed categories in 60 out of 76 cases in the estimation set. The calibrated model was then applied to the validation set to test its prediction performance. Overall, 24 out of 40 cases in the validation set were correctly predicted by the calibrated model. This result implies that the developed model was accurate in approximately 78.9 percent and 60.0 percent of cases in the estimation and validation sample sets, respectively.

Figure 6.2 illustrates the overall model performance, including estimation and validation sets. Considering the sample size, the developed model performed satisfactorily for incident durations of 120 to 180 minutes and 180 to 240 minutes. The model also performed well for the durations of <60 and >=300 minutes, though the number of incidents for these categories were relatively small. On the other hand, the model performed worst in durations of 60 to 120 minutes.



*Note: Percentages represent the proportion of correctly estimated/predicted incidents based on the developed NBC model.

Figure 6.2 Distribution of Overall Model Performance

To verify the power of the developed NBC model, the research team tried to develop other classification models, such as discrete choice models, to compare their performances. Unfortunately, parameters included in the calibrated MNL model hardly showed statistical significance, except for three attributes: Number of Blocked Shoulder Lanes, Road Division Conditions (Not divided, One-way, Divided without Barrier, or Divided with Barrier) and Fixed Object (Culvert-Ditch, Curb, Guardrail-Barrier, Embankment, Fence, Light Support Pole, Tree-Shrubbery, or Others).

Alternately, a multiple linear regression model was developed using the incident duration as a continuous variable. Similar to the MNL model, few attributes appeared to be statistically significant on the calibrated model at the 90 percent significance level. The developed regression model is summarized in Table 6.3. Observed incident durations and estimated/predicted incident durations by the regression model were categorized on the basis of Table 6.2 and compared to each other to check the performance of the developed regression model. The regression model correctly estimated about 29 percent of incidents in the estimation set and correctly predicted approximately 28 percent of incidents in the validation data set. Although it is hard to conclude that the developed NBC model is the best approach among existing methodologies, these comparisons in model performance at least prove that the NBC model would be one of the most promising approaches for the target analysis, especially under the limited available sample size.

In the NBC model, it is noticeable that collision types were one of the significant attributes contributing to the durations of fatality incidents. Even though it is intuitively reasonable, the developed model statistically proved that collisions associated with head-on and head-on-left-turn, rather than other collision types, are one of the primary determinants which cause the fatality incident durations longer. The regression model developed for the model performance comparisons also led to the same conclusion. As matter of fact, the average incident duration for head-on and head-on-left-turn collisions was 255 and 249 minutes, respectively. In contrast, the average incident duration for a sideswipe collision, which is also one of the most common collision types, was 166 minutes in this study sample.

Another notable finding is the contribution of road division conditions. In the MNL and multiple linear regression models, the road division condition exhibited as a statistically significant factor. The average incident durations based on the study sample also showed the significant difference of this attribute, since incidents on not-divided roads (14 cases) lasted 244 minutes, on average, but lasted 181 minutes on divided roads (59 cases). This is likely due to the fact that incidents on not-divided roads tend to be more severe than those on divided roads. This is supported by the statistical results that the percentage of head-on or head-on-left-turn collisions among all types of collisions on not-divided roads is nine percent higher than that on divided roads.

Table 6.3 Estimated Multiple Linear Regression Model for Fatality Incidents

$$\begin{aligned} \textit{Incident Duration (mins)} = {} & \underset{(6.86)}{216.36} + \underset{(3.29)}{88.31}*\textit{Ratio\_Total\_Bk} - \underset{(-2.68)}{22.50}*\textit{Rd\_Div} \\ & + \underset{(2.96)}{76.18}*\textit{Coll\_Type\_HOLT} - \underset{(-1.81)}{33.69}*\textit{Rd\_Others} \end{aligned}$$

Note : Numbers in parentheses are *t*-statistic values

Number of observations used : 76
$R^2 = 0.26$
F-value for Model = 6.32
P-value for Model = 0.0002

<Legend>
Incident Duration : Fatality incident duration in minutes
Ratio_Total_Bk : Number of lanes blocked in both directions/Total Number of lanes in both directions
Rd_Div : Road Division Conditions
1 if not divided
2 if one-way road or street
3 if divided with no barrier
4 if divided with barrier
Coll_Type_ HOLT : 1 if a collision type is Head-on or Head-on-Left-Turn; 0 otherwise
Rd_Others : 1 if an incident occurred on non-major roads; 0 otherwise

6.5 Closure

In this advanced study, the NBC was used to identify the major variables which have a significant influence on fatality incident duration and to predict the most plausible incident durations under given conditions.

Despite the small available sample size, the analysis results clearly indicate that incidents causing fatalities usually have much longer durations than those of other natures and that their durations are distributed over a wide range. On the other hand, the frequency of fatality incidents is far less than those of other types. Thus, it is quite difficult to collect a sufficient sample size for model development and validation.

The NBC model developed from this study showed satisfactory performance, achieving 79 percent and 60 percent accuracy in the model calibration and validation processes, respectively. Although the model was developed based on a relatively small sample, it outperformed some conventional models developed using the same data set.

Due to the complex nature of incident durations and associated factors, the research team fully recognizes that much remains to be done to produce a more reliable and generalized model for predicting fatality incident durations. This study, however, provides a preliminary model and results to enable state highway agencies to estimate the approximate delay from, and the potential impact area of, an identified fatality incident, which should help when assessing the need to implement a detour plan.

Chapter 7: Application of Incident Duration Models: Estimations of Incident-Induced Delay and Queue

7.1 Introduction

Freeway incidents, regardless of severity, are the primary contributors of traffic congestion in most urbanized areas. Unexpected delays caused by freeway incidents affect the daily travel time variability, and consequently may significantly influence a driver's decision about route choices, departure time, and mode choices. The unpredicted delay of travel will inevitably increase the commuting cost of drivers.

In highway traffic management, the most significant indicators to measure the impacts of incidents are the delay and the resulting queue length. Information associated with delay and queue can also be used for transportation planning at different levels. Transportation agencies responsible for incident management can use such indicators to implement the most cost-effective strategies. In fact, the delay data have evolved as the most critical factor for estimating traffic and economic impacts due to incidents, such as increased fuel consumption, increased emissions, and increased air pollution.

This chapter presents the estimation of incident-induced delay and its queue length using a data set generated from CORSIM. We explore factors contributing to delay and queue, and present preliminary models for estimating the total vehicle delay and the resulting queue to the incident.

7.2 Methodology

7.2.1 Experimental Design

This study simulated the I-495 network with CORSIM. The entire network, a four-lane loop format highway which includes no lane drops, grades or any other local bottlenecks, was chosen in order to generate delay and queue solely due to incidents. Delay and queue are computed based on the congestion caused by one isolated incident. In addition, the delay and queue in the opposite direction due to the rubbernecking factor were not considered in this study.

7.2.2 Description of Variables

To identify factors contributing to the delay and queue induced by incidents, the experiment explored a number of related variables and their possible range of variation. A list of these variables is presented below.

- Incident duration (in minutes)
    - Minor (less than 30 minutes): 5, 15, 25
    - Moderate (30 mins to 60 mins): 35, 45, 55
    - Severe (> 60 mins): 65, 75, 85

- One main lane closure scenario
    - lane 1, lane 2, lane 3, or lane 4

- Two main lane closure scenario
    - (lane 1, lane 2), (lane 1, lane 3), (lane 1, lane 4), (lane 2, lane 3), (lane 2, lane 4), or (lane 3, lane 4)

- Three main lane closure scenario
    - (lane 1, lane 2, lane 3), (lane 1, lane 2, lane 4), (lane 1, lane 3, lane 4), (lane 2, lane 3, lane 4)

- Main lane volume: 1000, 3000, 5000, 7000, or 9000 vph

- On-ramp volume: 1000, 1200, 1400, 1600, 1800, 2000 vphpl

- Off-ramp volume ratio: 5 percent, 10 percent, 15 percent, 20 percent, or 25 percent

- Heavy vehicle factors: 5 percent, 10 percent, 15 percent, or 20 percent

- Rubbernecking proportion: 0 percent, 20 percent, 40 percent, 60 percent, 80 percent

- Incident impact length (in ft)
    - 50: minor incident (e.g., disabled vehicles)
    - 100: moderate incident (e.g., two-vehicle collisions)
    - 150: semi-severe incident
    - 200: severe incident (e.g., several injuries or fatalities)

- Incident location
    - Before passing off-ramp (1/3 mile away from the off-ramp)
    - Near off-ramp (±0.05 mile from ramps)
    - Between on- and off-ramps
    - Near on-ramp (±0.05 mile from ramps)
    - After passing on-ramp (1, 2/3, or 1/3 miles away from the on-ramp)

As described in the above list, all lane blockages were excluded from this study due to their unusual impact. Lane blockage was also specified in terms of the lane ID; that is, scenarios were discerned on the basis of which lanes out of the four lanes available were blocked. Lane IDs are illustrated in Figure 7.1. Scenarios generated for lane blockages did not regard shoulder blockages, because CORSIM does not have a function to reliably capture a shoulder blockage.

The incident impact length was estimated based on the number of vehicles involved and the incident management vehicles. For instance, if a police car, an ambulance, and two wreckers were dispatched to manage the incident of a two-vehicle collision, the impact of the incident would be represented approximately by 140 feet in length, considering that each vehicle is 20 feet long. To examine the contribution of incident locations on delay, scenarios were also distinguished by locations where the incidents occur.

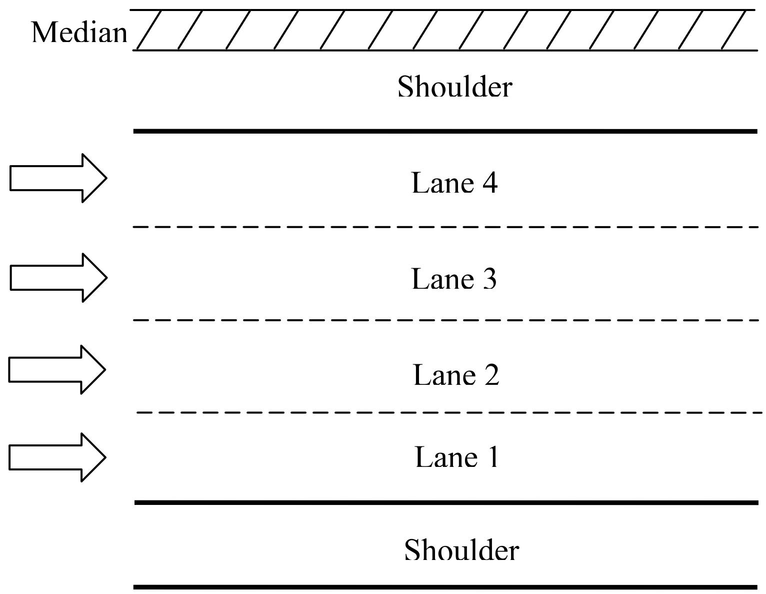


Figure 7.1 Illustration of Lane IDs

7.2.3 Model Development and Results

Total Delay Estimation

In CORSIM, to develop a model for computing the total delay, one first selects a set of variables of different ranges using the random sampling procedure. Delay calculated from the CORSIM output is the sum of delays that a vehicle has experienced due to an incident. It is computed based on the definition that delay is the difference between the free-flow and incident-incurred travel times over the same highway segment.

Based on 380 samples acquired from the CORSIM output, a regression model was developed to estimate/predict the total delay. Table 7.1 summarizes the estimated results.

All parameters of the proposed model of 13 variables were significant at the 5 percent significance level with reasonable signs. The measure of the goodness of fit of the model (R^2) was approximately 0.83, which indicates that the developed model explained about 83 percent of samples used for the model calibration. Since simulated total delays follow a log-normal distribution, a natural logarithm form of the total delay in minutes was used as the dependent variable.

As expected, the model shows that the most significant variable was the incident duration that had the *t*-statistic value 27.69, representing the significance of this variable. These estimation results justified the need to have reliable estimation of incident durations so as to increase the robustness of estimating the incident-induced delay.

As revealed in the model, traffic volume and related factors exhibited significant impacts on the total delay. For instance, an increase in high-heavy vehicles and a high ratio of through traffic would directly increase the resulting delay during any incident.

One interesting finding in this regard is associated with where the incident occurred. As presented in the previous section, different scenarios were generated in terms of incident locations. For example, considering that the segment between two exits is 1.5 miles long on average, the incident scenarios were generated at five different locations; (1) before passing the off-ramp, (2) near the off-ramp, (3) between the off-ramp and the on-ramp, (4) near the on-ramp, and (5) after passing the on-ramp. Figure 7.2 illustrates the relative locations of these five scenarios defined in this study.

Table 7.1 Estimated Regression Model for Incident Induced Delay

$$
\begin{aligned}
\text{Log(delay(mins))} = &- \underset{(-4.81)}{3.4204} + \underset{(7.20)}{0.0723}*\text{HeavyVeh} + \underset{(27.01)}{0.0008}*\text{Main\_Vol} + \underset{(3.84)}{0.0297}*\text{Thru\_Rt} + \underset{(27.69)}{0.0596}*\text{Inc\_Dur} + \underset{(5.56)}{0.6124}*\text{Away\_On} \\
&+ \underset{(14.26)}{1.9102}*\text{LnB1} + \underset{(12.61)}{1.6571}*\text{LnB2} + \underset{(9.89)}{1.3205}*\text{LnB3} + \underset{(7.97)}{1.0366}*\text{LnB4} + \underset{(3.24)}{0.4903}*\text{LnR1} + \underset{(2.12)}{0.3225}*\text{LnR2} + \underset{(3.24)}{0.5105}*\text{LnR3} \\
&+ \underset{(2.22)}{0.3402}*\text{LnR4}
\end{aligned}
$$

Number of observations used : 380
$R^2 = 0.8330$
F-value for Model = 140.39
P-value for Model = < 0.0001

Note : Numbers in parentheses are *t*-statistic values

<Legend>
HeavyVeh : Heavy vehicle factor
Main_Vol : Volume on four main lanes (vph)
Thru_Rt : Ratio of through traffic at an off ramp (100-(off-ramp volume ratio))
Inc_Dur : Incident duration in minutes
Away_On : 1 if an incident occurred at least 1/3 miles away after passing an on-ramp; 0 otherwise
LnB1 : 1 if Lane 1 is blocked due to the incident; 0 otherwise
LnB2 : 1 if Lane 2 is blocked due to the incident; 0 otherwise
LnB3 : 1 if Lane 3 is blocked due to the incident; 0 otherwise
LnB4 : 1 if Lane 4 is blocked due to the incident; 0 otherwise
LnR1 : 1 if Lane 1 has a rubbernecking effect due to the incident; 0 otherwise
LnR2 : 1 if Lane 2 has a rubbernecking effect due to the incident; 0 otherwise
LnR3 : 1 if Lane 3 has a rubbernecking effect due to the incident; 0 otherwise
LnR4 : 1 if Lane 4 has a rubbernecking effect due to the incident; 0 otherwise

The *t*-statistic result stated that the location factor was meaningful only when an incident occured somewhere in the downstream of an on-ramp, which is defined as Area 5 in Figure 7.2. The positive sign of its estimated parameter implied that the total delay caused by an incident was highly likely to increase when the incident occured somewhere after passing the on-ramp and before reaching the next off-ramp. This may be attributed to the fact that the traffic volume after passing the on-ramp generally increases by the in-flow volume from the on-ramp.

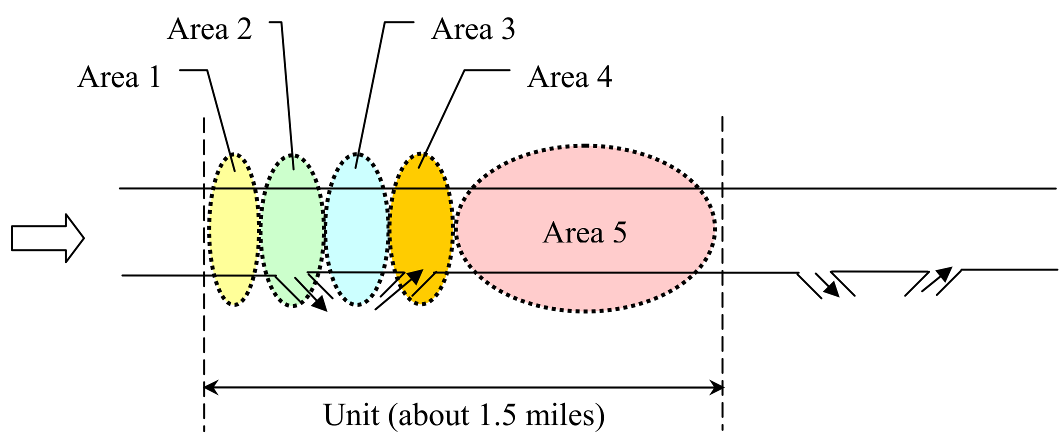


Figure 7.2	Illustration of the Network Unit Used for This Study

The model also reflected an interesting statistical result regarding the lane blockage and rubbernecking effects. It revealed that, all other conditions being equal, the actual impact of an incident may vary with its lane location. As shown in Table 7.1, the closure of Lane 1 had the most significant influence on delay, whereas Lane 4 had the least impact, as it is close to the median. In most conditions, Lane 1 is the slowest lane — since most conservative drivers take this lane for better safety, and it interacts with on- and off-ramps via the weaving areas — while Lane 4 is the fastest lane occupied by aggressive drivers. Conservative drivers generally take a longer time in changing lanes and demand wider gaps. The impact of such behavioral patterns becomes evident when Lane 1 is unexpectedly closed due to an incident: these conservative drivers generally exhibit some difficulty in finding acceptable lane-changing gaps.

Similarly, the most significant rubbernecking effects were those occurring in Lane 1. However, unlike the lane blockage factor, no systematic pattern can be observed among different lanes.

Queue Length Estimation

Queue, in this study, is defined as the length of the maximum spillback due to incidents. Based on the shockwave theory, the maximum spillback would occur at the moment that all closed or restricted lanes caused by incidents are recovered to normal conditions. Vehicles showing a mean speed lower than 20 mph are viewed as being part of the incident-induced queue, which is used as the criterion to determine the maximum queue length.

To acquire simulated data for the maximum queue lengths from CORSIM, detectors were placed on the network to detect the breakpoint where the traffic flow speed was greater than 20 mph. Also, the distance from the incident site to the breakpoint was measured as the incident-induced maximum queue length at the moment that all restrictions on the roads by incidents were cleared.

Based on 285 samples acquired from the CORSIM output, a regression model was developed to estimate the maximum queue length. Table 7.2 summarizes the estimated results.

Table 7.2 Estimated Regression Model for Incident Induced Queue

$$
\begin{aligned}
\text{Log(queue(ft))} = {} & \underset{(51.07)}{6.6736} + \underset{(3.92)}{0.0191}\text{*HeavyVeh} + \underset{(15.79)}{0.0002}\text{*Main\_Vol} + \underset{(13.53)}{0.0149}\text{*Inc\_Dur} + \underset{(3.32)}{0.1930}\text{*LnB2} + \underset{(1.97)}{0.1147}\text{*LnB3} \\
& + \underset{(2.71)}{0.1528}\text{*LnB4} + \underset{(7.63)}{1.0079}\text{*Away\_Off\_1/3} + \underset{(6.82)}{0.8094}\text{*Near\_Off\_Bf} + \underset{(9.23)}{1.0020}\text{*Near\_Off\_Af} + \underset{(6.18)}{0.8100}\text{*Btw\_On\_Off} \\
& + \underset{(5.51)}{0.6371}\text{*Near\_On\_Bf} + \underset{(5.66)}{0.6284}\text{*Near\_On\_Af} + \underset{(5.31)}{0.5501}\text{*Away\_On\_1/3} + \underset{(1.68)}{0.1604}\text{*Away\_On\_2/3}
\end{aligned}
$$

Number of observations used : 285
$R^2 = 0.7360$
F-value for Model = 53.76
P-value for Model = < 0.0001

Note : Numbers in parentheses are *t*-statistic values

<Legend>
HeavyVeh : Heavy vehicle factor
Main_Vol : Volume on four main lanes (vph)
Inc_Dur : Incident duration in minutes
LnB2 : 1 if Lane 2 is blocked due to the incident; 0 otherwise
LnB3 : 1 if Lane 3 is blocked due to the incident; 0 otherwise
LnB4 : 1 if Lane 4 is blocked due to the incident; 0 otherwise
Away_Off_1/3 : 1 if an incident occurred about 1/3 miles away before passing the nearest off-ramp; 0 otherwise (Area 1 in Figure 7.2)
Near_Off_Bf : 1 if an incident occurred near an off-ramp (within 500 ft) but before passing an off-ramp; 0 otherwise (Area 2 in Figure 7.2)
Near_Off_Af : 1 if an incident occurred near an off-ramp (within 500 ft) but after passing an off-ramp; 0 otherwise (Area 2 in Figure 7.2)
Btw_On_Off : 1 if an incident occurred somewhere between on-ramp and off-ramp; 0 otherwise (Area 3 in Figure 7.2)
Near_On_Bf : 1 if an incident occurred near an on-ramp (within 500 ft) but before passing an on-ramp; 0 otherwise (Area 4 in Figure 7.2)
Near_Off_Af : 1 if an incident occurred near an on-ramp (within 500 ft) but after passing an on-ramp; 0 otherwise (Area 4 in Figure 7.2)
Away_On_1/3 : 1 if an incident occurred about 1/3 miles away after passing an on-ramp; 0 otherwise (Area 5 in Figure 7.2)
Away_On_2/3 : 1 if an incident occurred about 2/3 miles away after passing an on-ramp; 0 otherwise (Area 5 in Figure 7.2)

All 14 variables included in the proposed queue model showed reasonable parameter signs. They were all significant at the 10 percent significance level. Similar to the delay model, a natural logarithm form of the maximum queues was used as the dependent variable, since the simulated maximum queues approximately follow a log-normal distribution.

Unlike the delay model, the most significant variable in this model was the volume in the main lanes. Incident durations also showed great significance for queue length estimation. The estimation results also showed that, as expected, the queue length would increase with the traffic volume and incident durations.

Lane closures for Lane 2, Lane 3, and Lane 4 all had statistically significant impacts on the maximum queue, while rubbernecking effects had little effect.

Interestingly, the queue model is highly sensitive to the location of the incident sites. Most variables defined to capture the nature of the incident location (in Table 7.2) had considerable significance for the model, except for the variable *Away_On_1*, which is defined to be 1 if an incident occurred about 1 mile away after passing an on-ramp and 0 otherwise. One can also notice that the variable *Away_On_2/3* (defined in Table 7.2) was much less significant than other incident-location variables. In addition, variables indicating incident locations before reaching the next on-ramp (e.g., *Away_Off_1/3*, *Near_Off_Bf*, *Near_Off_Af*, and *Btw_On_Off* in Table 7.2) showed greater significances with higher estimated coefficients. This implies that incidents that occur before reaching the next on-ramp are more likely to increase the queue.

7.3 Applications of Developed Models

Note that one can apply the delay and queue models if the incident duration and other traffic information are available from any source. The CHART-II database would be the main source for the information needed to estimate the incident duration and to perform the delay and queue estimations. Additional data required for executing the models are the heavy vehicle factor, traffic volume in the main lanes and at ramps, incident lane location, and rubbernecking effects on each opened lane.

Table 7.3 presents three examples of incidents occurring on I-495 with different incident natures. Incident durations were estimated based on the RBTM and supplemental models. When the estimated/predicted incident duration is an interval output, the value of the middle point of the range would represent the incident duration and be used as an input for delay and queue models. For instance, in the second example, the estimated incident duration was 10 to 35 minutes. The middle point of the incident duration interval — that is, 22.5 minutes — would be used as the input for the delay and queue estimation models.

The first example shows how severe the impact of the fatality incident would be in terms of delay and queue. The total delay experienced by all vehicles on the highway due to the fatality incident would have been much larger than the delays from other natures of incidents, as shown in the table. Similarly, the maximum queue length of the fatality incident would be approximately four times larger than that from the *CPI*. The relative scales of the predicted results from these examples seem consistent with field observations.

Table 7.3 Model Applications of Incident Duration, Delay and Queue

Example No	1	2	3
Event Open Date & Time	2006-03-25 14:14:33	2003-07-02 18:33:52	2006-02-02 08:57:28
County	Prince George	Montgomery	Montgomery
Incident Nature	CF	CPI	OTHERS (Vehicle Fire)
Road Info	I-495 IL	I-495 IL	I-495 OL
Exit No	23	41	38
Weekend	1	0	0
Peak Hour	0	0	1
Heavy Vehicle Factor	11%	12%	10%
Volume in Main Lanes	6200	7000	7600
Thru Traffic Ratio	90%	90%	90%
Lane Blockage Info	Ln 1 and 2	Ln 1	None (Shoulder Blockage)
Lane Restriction Info	Ln 3: 80% Ln 4: 60%	Ln 2: 80% Ln 3: 60% Ln 4: 40%	Ln 1: 20%
Incident Location Info	Btw_On_Off	Near_Off_Bf	Away_On_1
Observed Incident Duration (mins)	143.53	32.73	7.07
Estimated Incident Duration (mins)	120.93	(10, 35)	19.78
Predicted Total Delay (hrs)*	301,046	463	76
Predicted Max. Queue (miles)*	12.6	2.9	2.4

* Total delay and maximum queue length are predicted based on the estimated incident durations

Chapter 8: Conclusions

8.1 Summary of Research Results

This study has presented a set of models for estimating incident durations using the incident data from years 2003 to 2005 available in the Maryland SHA CHART II database. The proposed models consist of primary components developed with the Rule-Based Tree Model and supplemental components calibrated with either multinomial logit or linear regression models. In conducting this study, it has been found that *Incident Nature* is the most influential factor associated with the duration of an incident; *County* emerges as the second most critical factor. The proposed RBTM has also been found to be quite flexible in assigning appropriate estimated incident duration ranges to nodes in the decision tree.

Additionally, studies for estimating delay and queue, based on estimated incident durations, were conducted as a primary part of the model application using data simulated in CORSIM. Theses studies are essential, since the impacts of incidents must be identified in order to promptly implement an effective and efficient incident management system. This can help motorists reduce the uncertainty of travel times, diminishing the ensuing economic losses due to unexpected delays. As expected, incident durations turned out to be the most significant factor for determining incident-induced delays. Incident durations also played a vital role in estimating the consequent queues.

This chapter next presents a summary of the incident duration research findings from both primary and supplemental models, ordered by incident nature; this is followed by a summary of research findings from delay and queue studies.

<u>*Collision-Fatality (CF)*</u>

- The range of predicted incident durations with the RBTM for fatality-related incidents is likely to be wider (e.g., about 60 minutes on average) than for other incident types (e.g., about 25 minutes in *Collision-Personal Injury*). However, the ***confidences*** for most of the rules were acceptable, since most of them were greater than or equal to 70 percent.
- For example, with the data set from years 2003 to 2005, the RBTM predicted with 75 percent ***confidence*** that the durations of incidents occurring on weekdays without the involvement of tractor-trailers would be between 100 and 200 minutes**.** It also predicted with 94 percent ***confidence*** that when fatality-related incidents occurred on weekends, their durations would be between 80 and 200 minutes.
- The multiple linear regression model, which is the supplemental model for predicting durations of incidents causing fatalities, can achieve about 75 percent accuracy.
- The clearance operation is generally more efficient in the scenarios where more lanes in the same direction (including shoulder lanes) were blocked than in scenarios leaving them open. The impact of wet pavement, a proxy variable for rainy days, was also found to have a negative correlation with the durations of incidents resulting in fatalities.
- Advanced research was carried out to enhance the model by integrating the CHART database with the Accident Report database managed by the Maryland State Police Department. The integrated database includes more specific

information regarding the severity of fatality-collisions, such as the number of fatalities and injuries for pedestrians or driver/occupants.

- The NBC was adopted in the advanced research, and it demonstrated accuracy of 78.9 percent and 60.0 percent for predicting incident durations in the estimation and validation sample sets, respectively. Although it was developed from a fairly small sample, this model has shown promise of being more suitable for incident duration data featuring wide distribution ranges and small sample sizes, e.g., fatality incident durations.

Collision-Personal Injury (CPI)

- Most rules having terminal nodes in RBTM can predict the range of incident duration within 30 minutes with a ***confidence*** exceeding 70 percent.
- RBTM can predict incidents occurring in Montgomery County causing less than three blocked lanes (including one blocked lane in the opposite direction) within a range of 10 to 30 minutes with approximately 85 percent ***confidence***. For incidents without lane blockage in the opposite direction but involving single-unit trucks, the predicted duration of 25 to 50 minutes can be achieved with about 81 percent ***confidence***.
- The probabilistic distribution of incident durations predicted with multinomial logit (MNL) models differs from the observed data by less than 10 percent .

Collision-Property Damage (CPD)

- Most rules in RBTM can achieve satisfactory results, such that the interval of predicted incident duration is within the range of 30 minutes and with 70 percent ***confidence***.
- Incidents not involving tractor-trailers and resulting in only property damage have been predicted to last up to 30 minutes with 75 percent ***confidence***.
- The predicted probabilistic distribution of incident duration with MNL models as the supplemental component is within 5 percent difference from the observed data.

Disabled Vehicles

- Most of the incidents caused by disabled vehicles (83.3 percent for Montgomery County alone) lie in a relatively short range of 5 to 30 minutes.
- Since about 84 percent of incidents in Montgomery County due to disabled vehicles had durations in the range of 5 to 30 minutes, one can use this simple rule to predict their resulting duration. Furthermore, based on the rules in RBTM, the durations of disabled vehicle-related incidents occurring on weekends in Montgomery County would be in a range of 5 to 25 minutes with 82 percent ***confidence***.

Others (Debris, Fire, Police Activity, Emergency Road Work, or Off-Road Work)

- Due to the limited sample data in this category, the development of a reliable RBTM was particularly challenging. In addition, more than 50 percent of rules could not be validated. Nevertheless, except for some rules with small sample sizes, the overall performance of RBTM was promising.

- Using the multiple linear regression model as a supplemental component for this category worked quite well. It predicted that the durations of incidents caused by *Other* types of incidents would fall in the range of 30 minutes with an 81 percent accuracy level.

Delay and Queue

- To estimate incident-induced delay and queue, this study successfully developed regression models with R^2 (the goodness of fit of a model) of 0.83 and 0.74 for delay and queue models, respectively. In both models, the incident duration and traffic volume in main lanes emerged as the most dominant factors.
- In the delay model, detailed information related to lane blockages and restrictions (rubbernecking effects) identified themselves as significant contributing factors. Also, if the incident occurs at locations downstream from an on-ramp, the total delay is likely to increase.
- On the other hand, the maximum queue length was relatively less sensitive to lane restrictions, while detailed information associated with incident locations had a great influence on the resulting queue length.
- Even though these proposed models are preliminary in nature, they offer a potentially useful set of tools for SHA to effectively manage incidents and to inform drivers in a timely manner.

8.2 Future Research

Developing reliable models for prediction based on field data is always a challenging task. It generally takes time to collect sufficient high-quality samples for model calibration. Besides, identifying outliers of samples requires in-depth knowledge about the environment of data collection and about the fundamental relationship between factors and predicted variables. To contend with the complex nature of incident duration prediction, this study has proposed the integrated application of three different models — the RBTM, the multinomial logit model and the multiple linear regression model — which, based on the available data, seem to yield quite promising results.

Moreover, advanced research for fatality incidents using the integrated database was been conducted to enhance model performance. To illustrate the potential model applications, this study further calibrated delay and queue length models for a detected incident with an estimated duration.

However, due to the variety of factors that may contribute to the resulting duration of a detected incident, much remains to be done to produce a reliable and generalized prediction model for use in practice. Some further research needs are summarized below.

- An alternative approach with additional data is needed to develop a more reliable model for predicting incident durations, since the CHART database contains information mainly about operations but not other safety factors.
- For incidents resulting in fatalities, it is essential to integrate the CHART database with police accident reports to construct an integrated data set with better quality for calibrating models of incidents involving fatalities.

- To implement integrated models of incident duration and the resulting delay/queue, having the database archive traffic conditions at the moment that an incident has occurred is recommended.

Appendix 1

Table A1.1 Summary of Results of MCA

Dimension	Largest Coeff.	Value	Meaning of Variables (Categories)
1	noodlb2+	2.03692	No of Lane Blockage for Opposite Direction(>=2)
2	noodlb2+	1.45660	
3	nosut2+	2.18925	No of Single Unit Trucks Involved(>=2)
4	nosut1	1.97307	No of Single Unit Trucks Involved(=1)
5	nosdlb3+	-1.51015	No of Lane Blockage for Same Direction(>=3)
6	nosdlb2	1.92873	No of Lane Blockage for Same Direction(=2)
7	extranr1	-2.11342	Incident Nature-Extra
8	road5	1.92107	Regrouped Road : Group 5 (I-68)
9	road5	-1.66567	
10	cf1	2.63923	Incident Nature-Collision_Fatality
11	road5	-4.29491	
12	road5	7.04303	
13	road5	5.55222	
14	nosut2+	3.68769	
15	road5	-3.78040	
16	road5	-7.71557	
17	road5	2.77578	
18	road5	8.69242	
19	cf1	-2.16741	
20	noodlb2+	4.21256	
21	road5	-2.83608	
22	road5	-3.62245	
23	extranr1	1.37016	
24	cf1	3.72513	
25	road5	2.95051	
26	nshdlb2+	-2.05738	No of Shoulder Blockage(>=5)
27	nopuv2+	2.15205	No of PickUp Van Involved(>=2)
28	novi1	-0.48775	No of Vehicles Involved(=1)
29	shdb0	0.41488	Shoulder Blockage Indicator(=0)
30	nosdlb2	-0.55401	
31	nosdlb3+	0.27334	
32	nottlb3+	-0.14306	No of Total Lane Blockage(>=3)

Figure A1.1 The Quantile-Quantile Plot (Q-Q Plot) of the Original Incident Duration Data Set – Log-normal Distribution

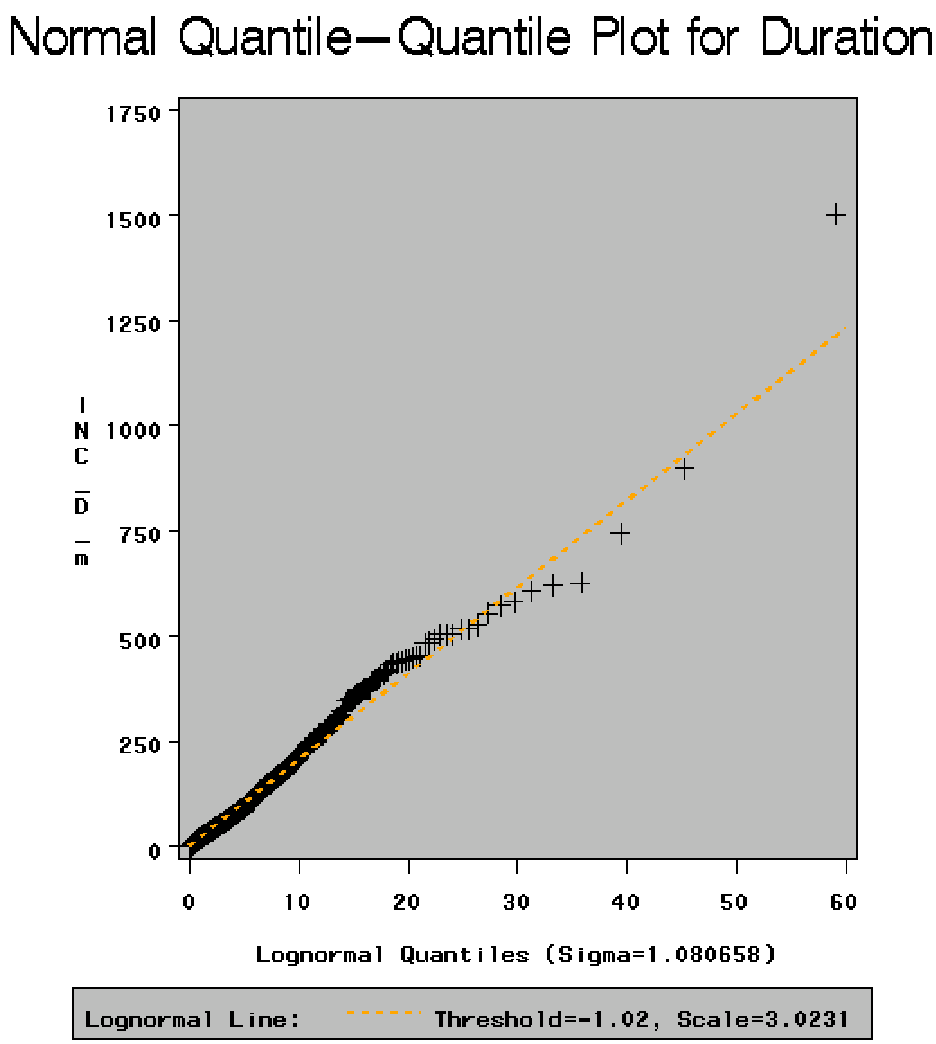


Figure A1.2 The Quantile-Quantile Plot (Q-Q Plot) of the Original Incident Duration Data Set – Weibull Distribution

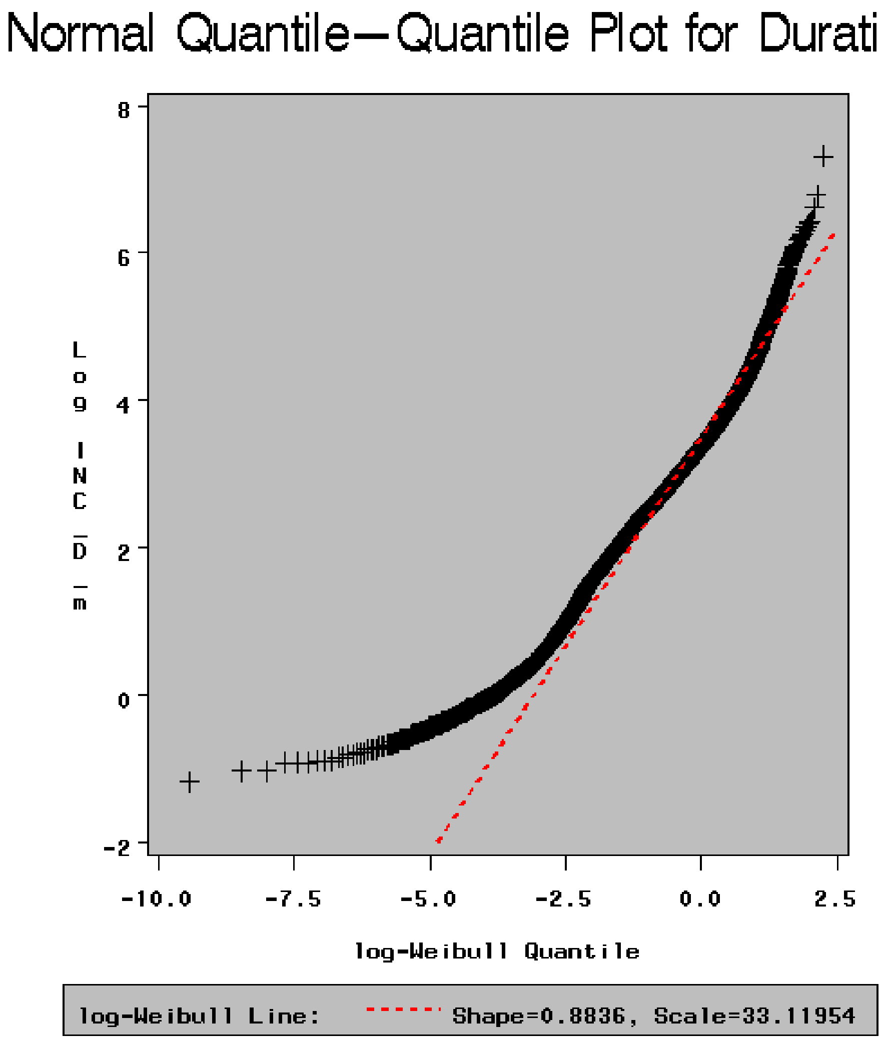


Figure A1.3 The Quantile-Quantile Plot (Q-Q Plot) of the Box-Cox Power Transformed Data Set

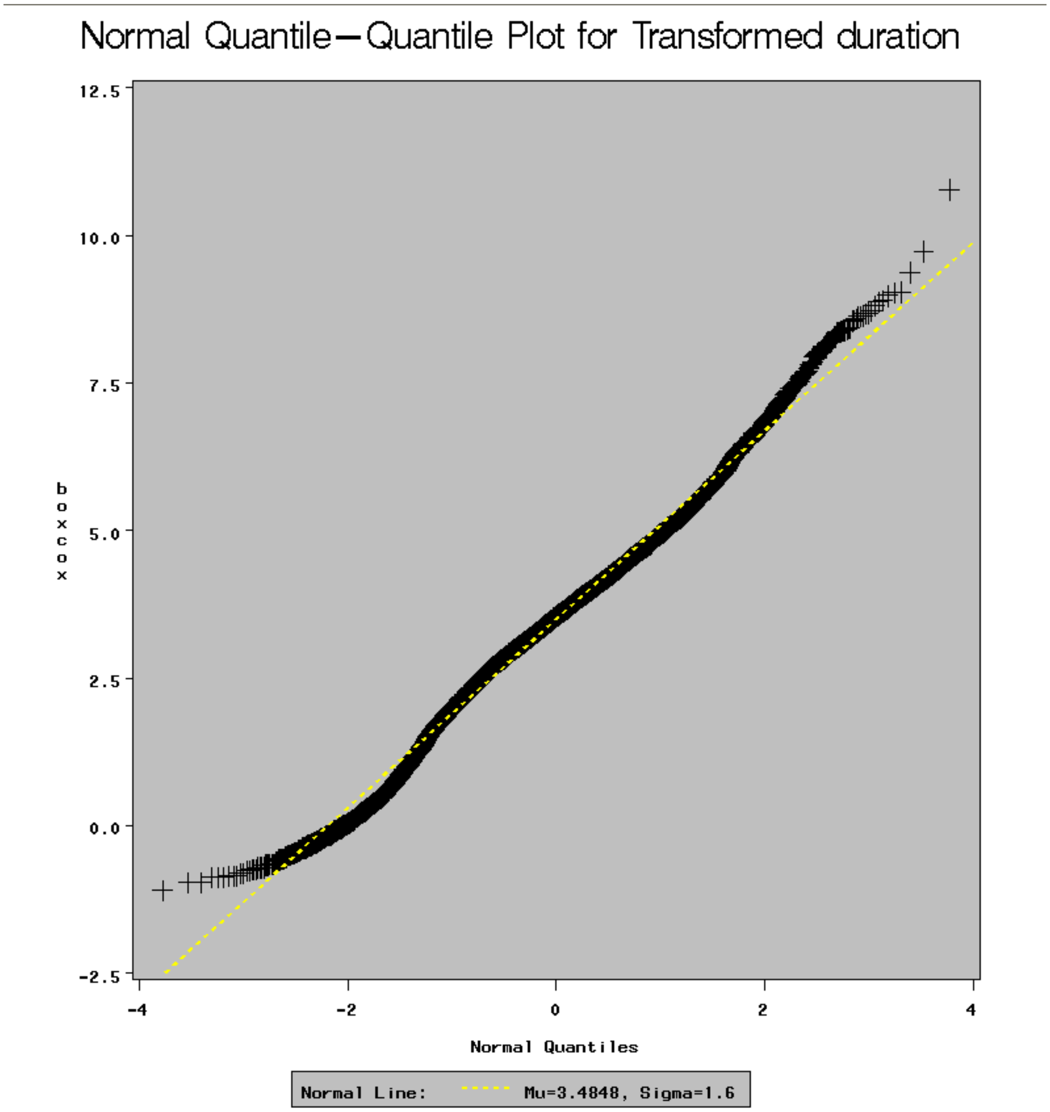


Figure A1.4 The Probability Plot (P-P Plot) of the Box-Cox Power Transformed Data Set

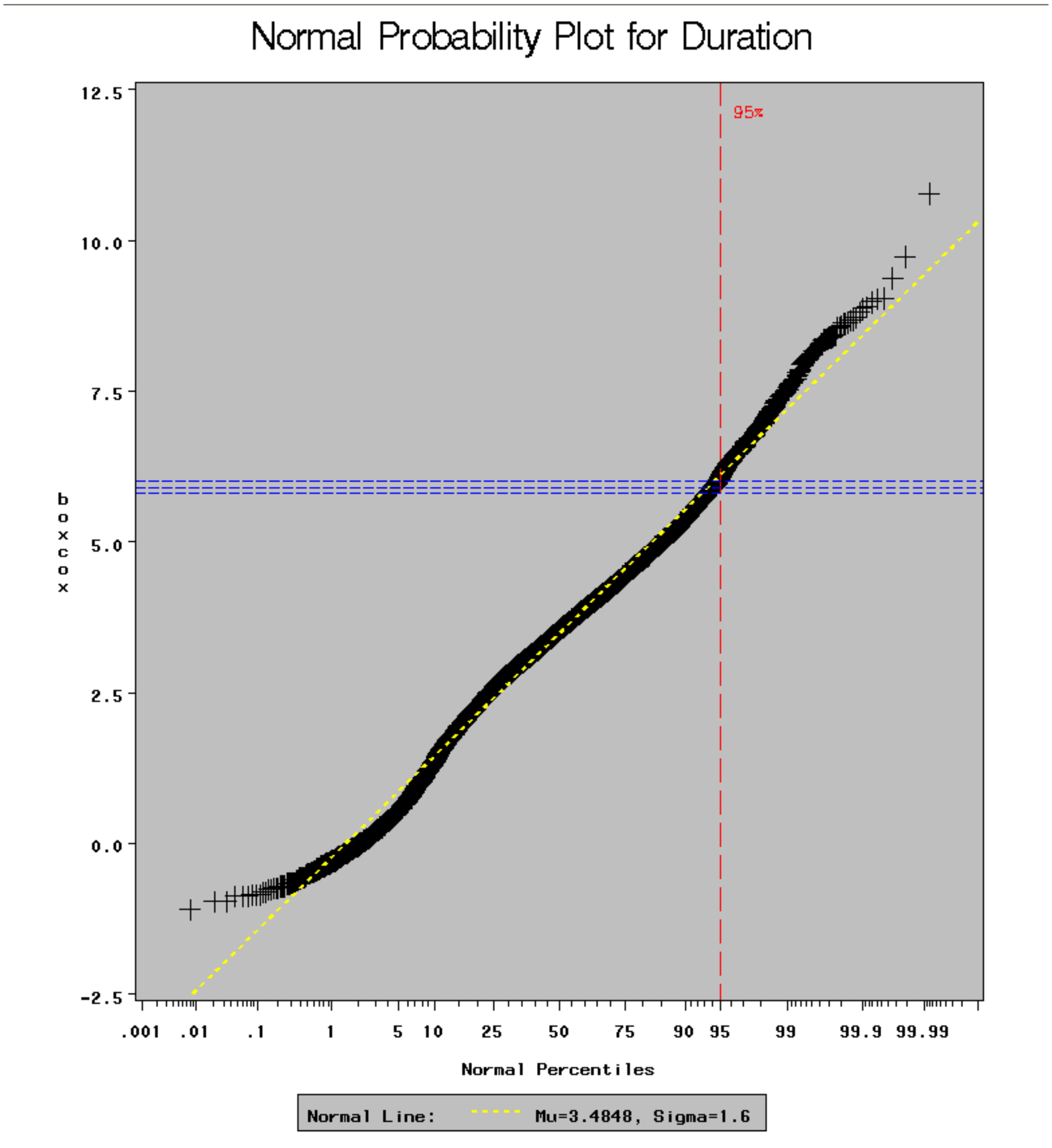


Figure A1.5 The Quantile-Quantile Plot (Q-Q Plot) of the Power Transformed Data from the Truncated Data Set (Incident Duration >= 5 minutes)

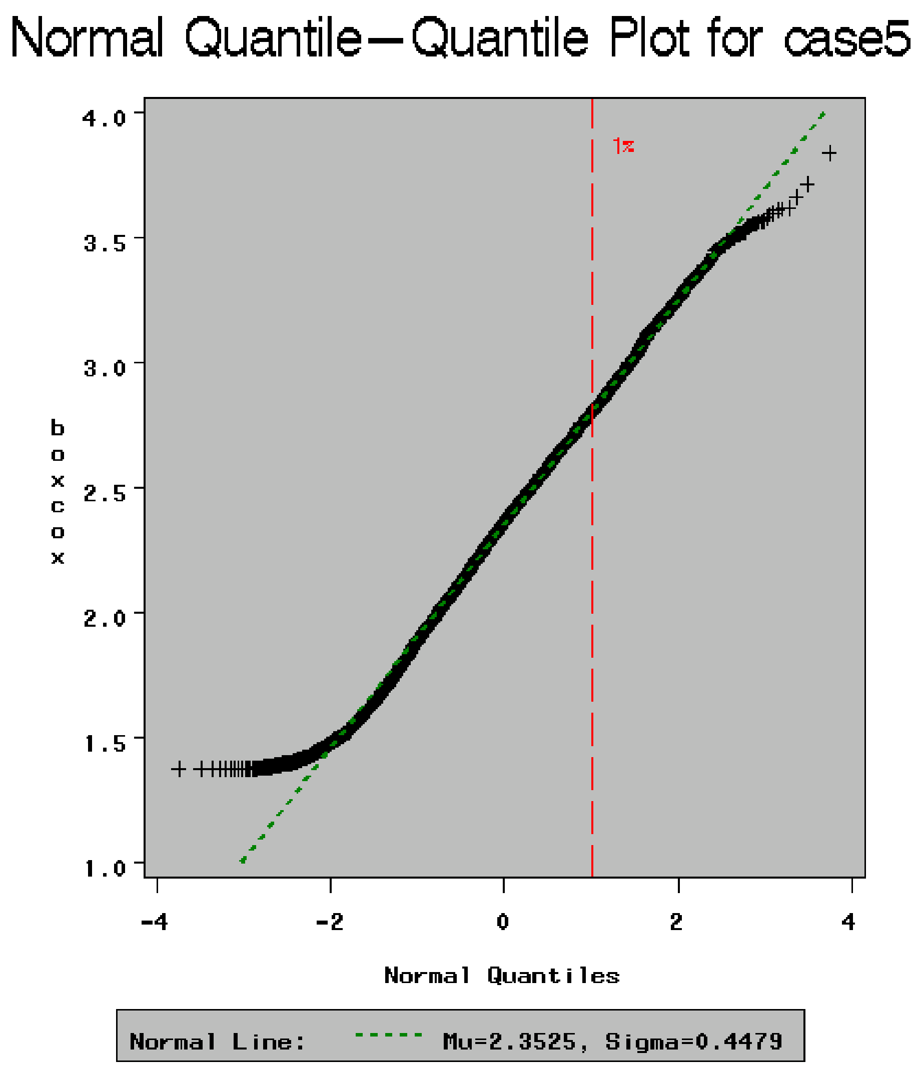
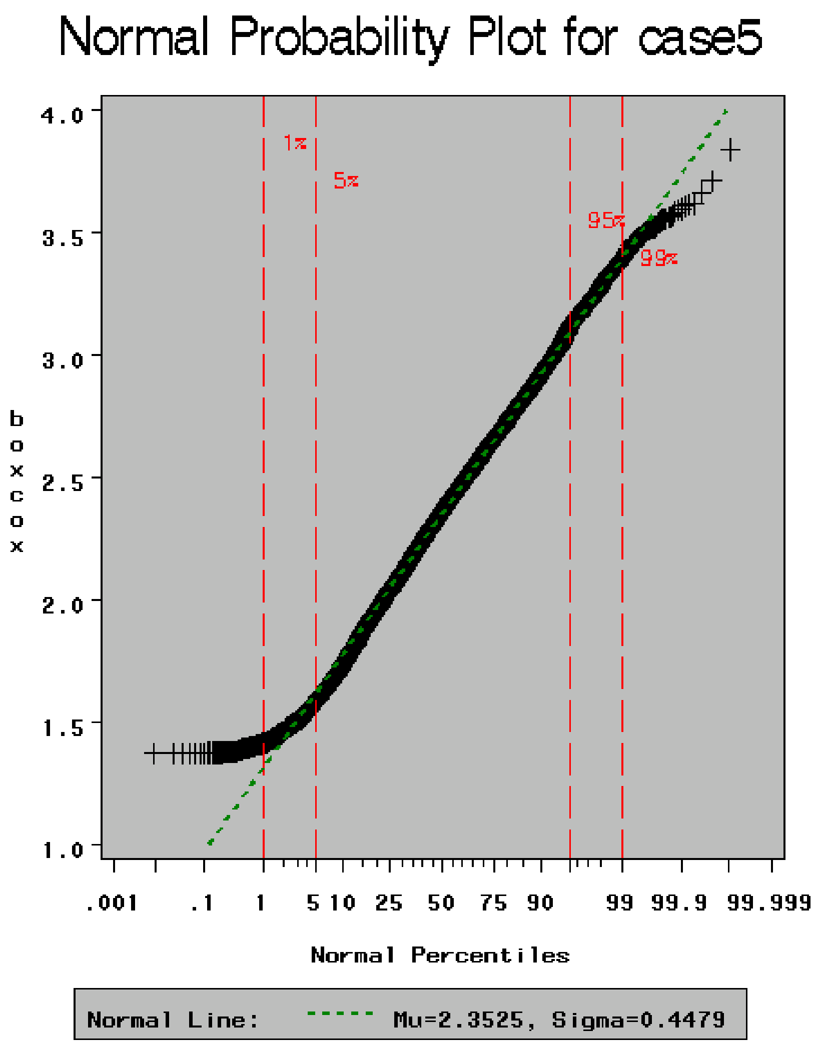


Figure A1.6 The Probability Plot (P-P Plot) of the Power Transformed Data from the Truncated Data Set (Incident Duration >= 5 minutes)



< SAS Output for the Basic Statistical Measures Using the Power Transformed Data Set>

Histogram with Normality curve

The CAPABILITY Procedure
Variable: boxcox

Moments

N	7798	Sum Weights	7798
Mean	3.4847733	Sum Observations	27174.2622
Std Deviation	1.60004234	Variance	2.5601355
Skewness	-0.0177011	Kurtosis	0.42382116
Uncorrected SS	114657.52	Corrected SS	19961.3765
Coeff Variation	45.9152491	Std Error Mean	0.01811924

Basic Statistical Measures

Location		Variability	
Mean	3.484773	Std Deviation	1.60004
Median	3.545526	Variance	2.56014
Mode	2.975670	Range	11.88588
		Interquartile Range	1.91399

< SAS Output for the Hypothesis Tests Using the Power Transformed Data Set >

Histogram with Normality curve

The CAPABILITY Procedure
Fitted Normal Distribution for boxcox

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	3.484773
Std Dev	Sigma	1.600042

Goodness-of-Fit Tests for Normal Distribution

Test	----Statistic-----		DF	------p Value------	
Kolmogorov-Smirnov	D	0.037447		Pr > D	<0.010
Cramer-von Mises	W-Sq	3.047849		Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	19.235050		Pr > A-Sq	<0.005
Chi-Square	Chi-Sq	393.636177	27	Pr > Chi-Sq	<0.001

<SAS Output for the Basic Statistical Measures and Hypothesis Test Statistics Using the Power Transformed Data from the Truncated Data Set (Incident Duration >= 5 minutes)>

Histogram with Normality curve-case5

The CAPABILITY Procedure
Variable: boxcox

Moments

N	6770	Sum Weights	6770
Mean	2.35245655	Sum Observations	15926.1309
Std Deviation	0.44787173	Variance	0.20058909
Skewness	0.04870144	Kurtosis	-0.3624111
Uncorrected SS	38823.3184	Corrected SS	1357.78755
Coeff Variation	19.03847	Std Error Mean	0.00544326

Basic Statistical Measures

Location		Variability	
Mean	2.352457	Std Deviation	0.44787
Median	2.364128	Variance	0.20059
Mode	2.030315	Range	2.46605
		Interquartile Range	0.62082

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	432.1776	Pr > |t|	<.0001
Sign	M	3385	Pr >= |M|	<.0001
Signed Rank	S	11459918	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value-----	
Kolmogorov-Smirnov	D	0.016158	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.341464	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	3.606629	Pr > A-Sq	<0.005

Appendix 2

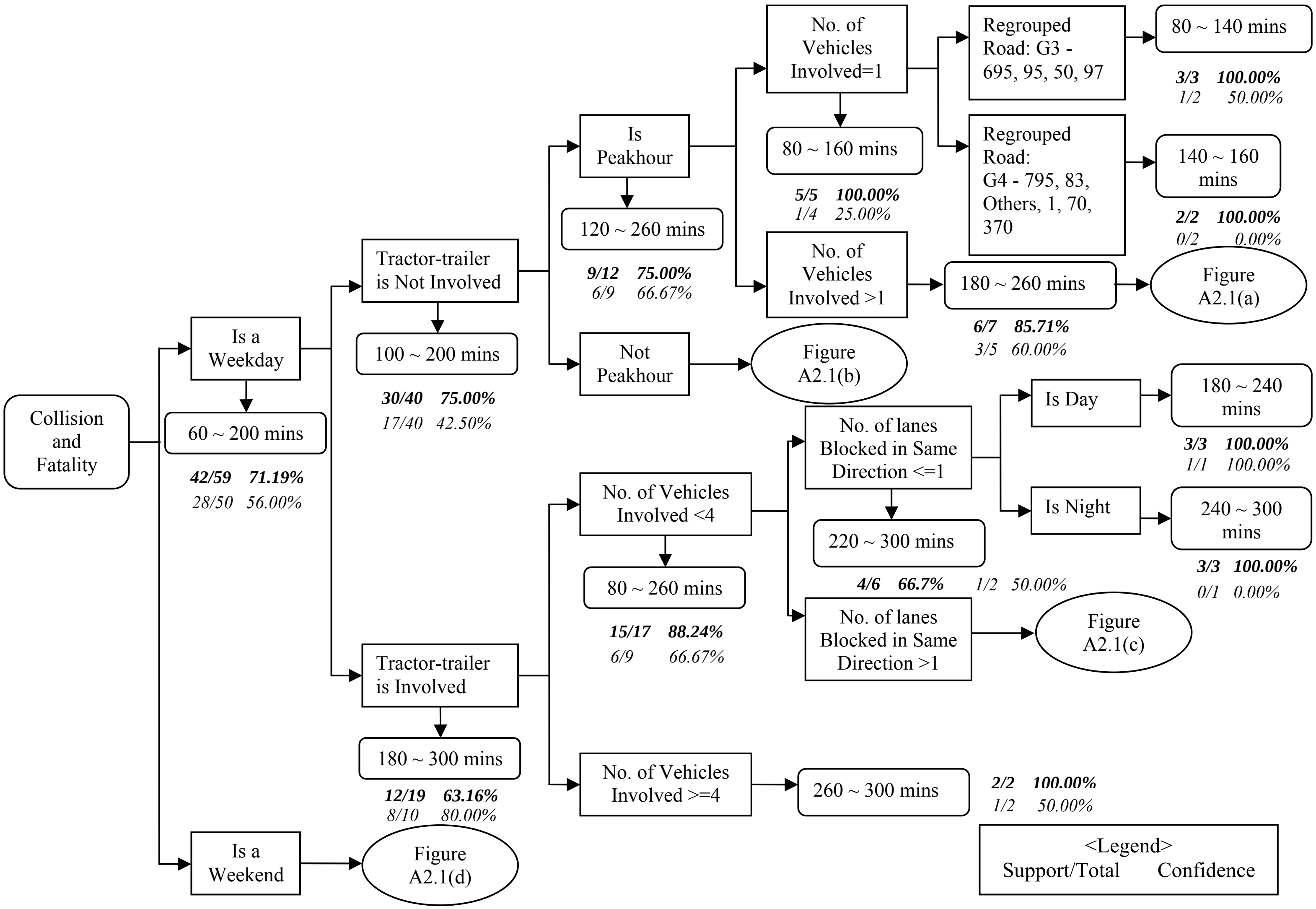


Figure A2.1 Rule Based Tree Model for Collision-Fatality in Montgomery County

* Numbers in bold italics are based on the dataset from year 2003 to year 2005, while numbers in italics are based on dataset from year 2006.

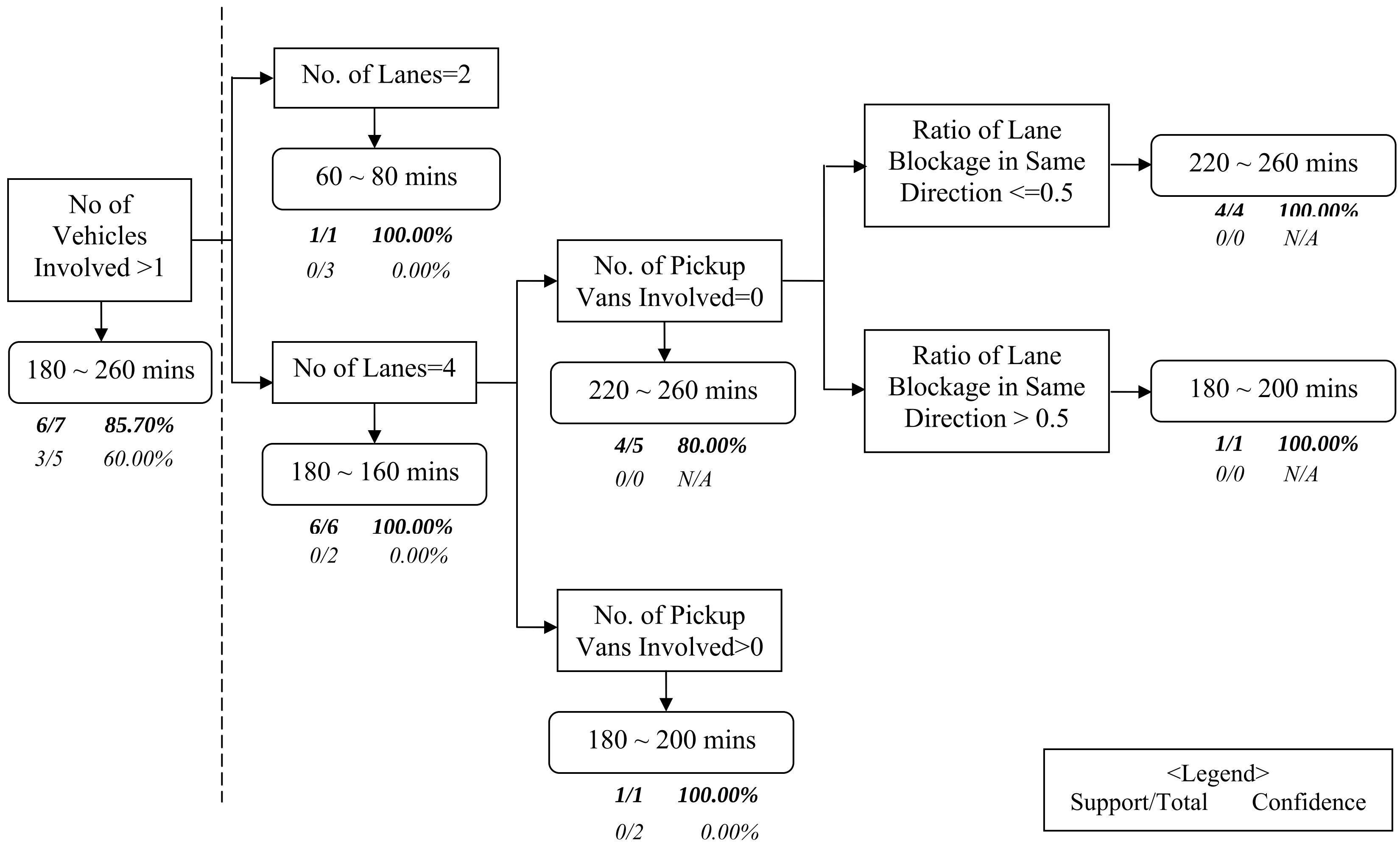


Figure A2.1(a) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

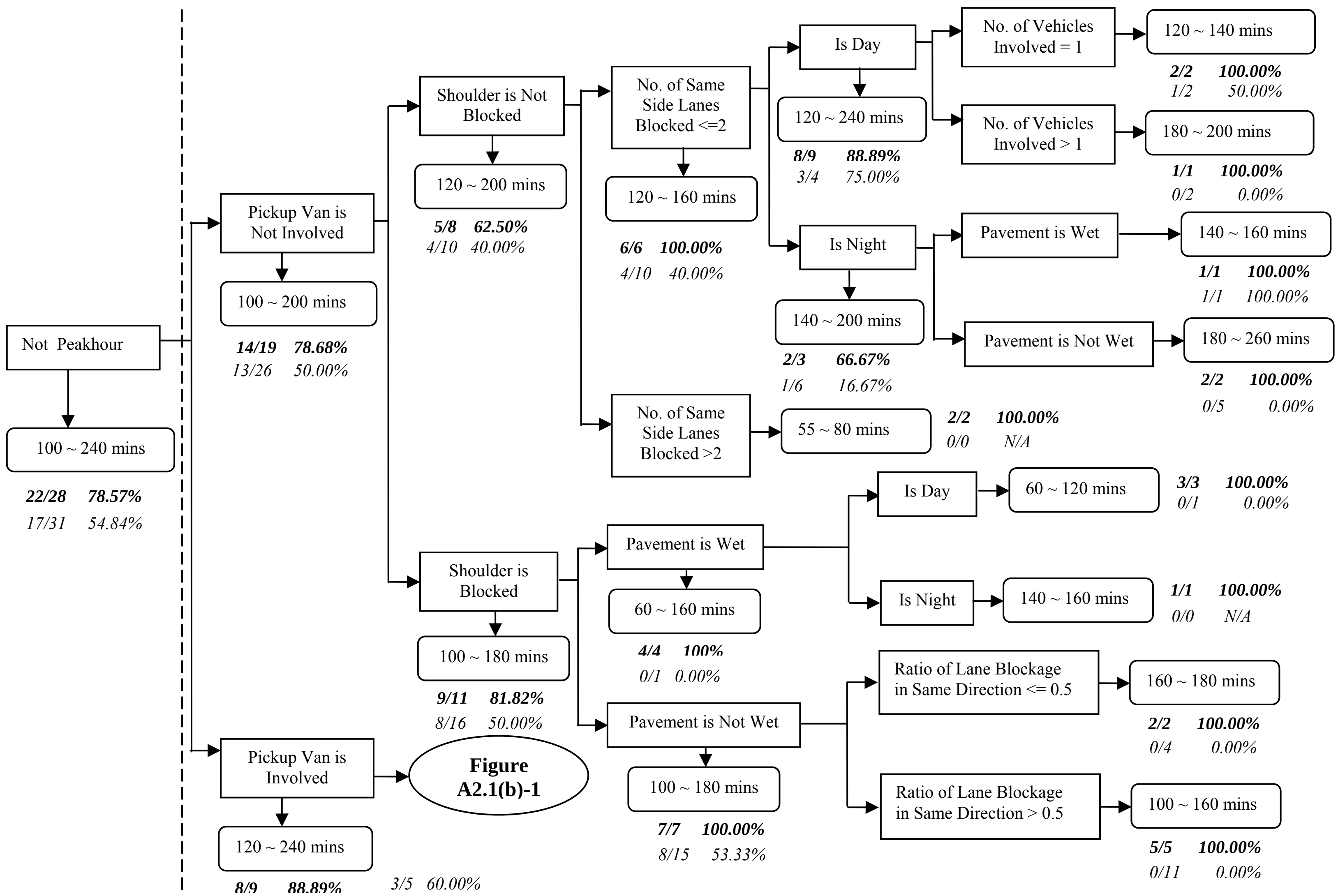


Figure A2.1(b) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

<Legend>
Support/Total Confidence

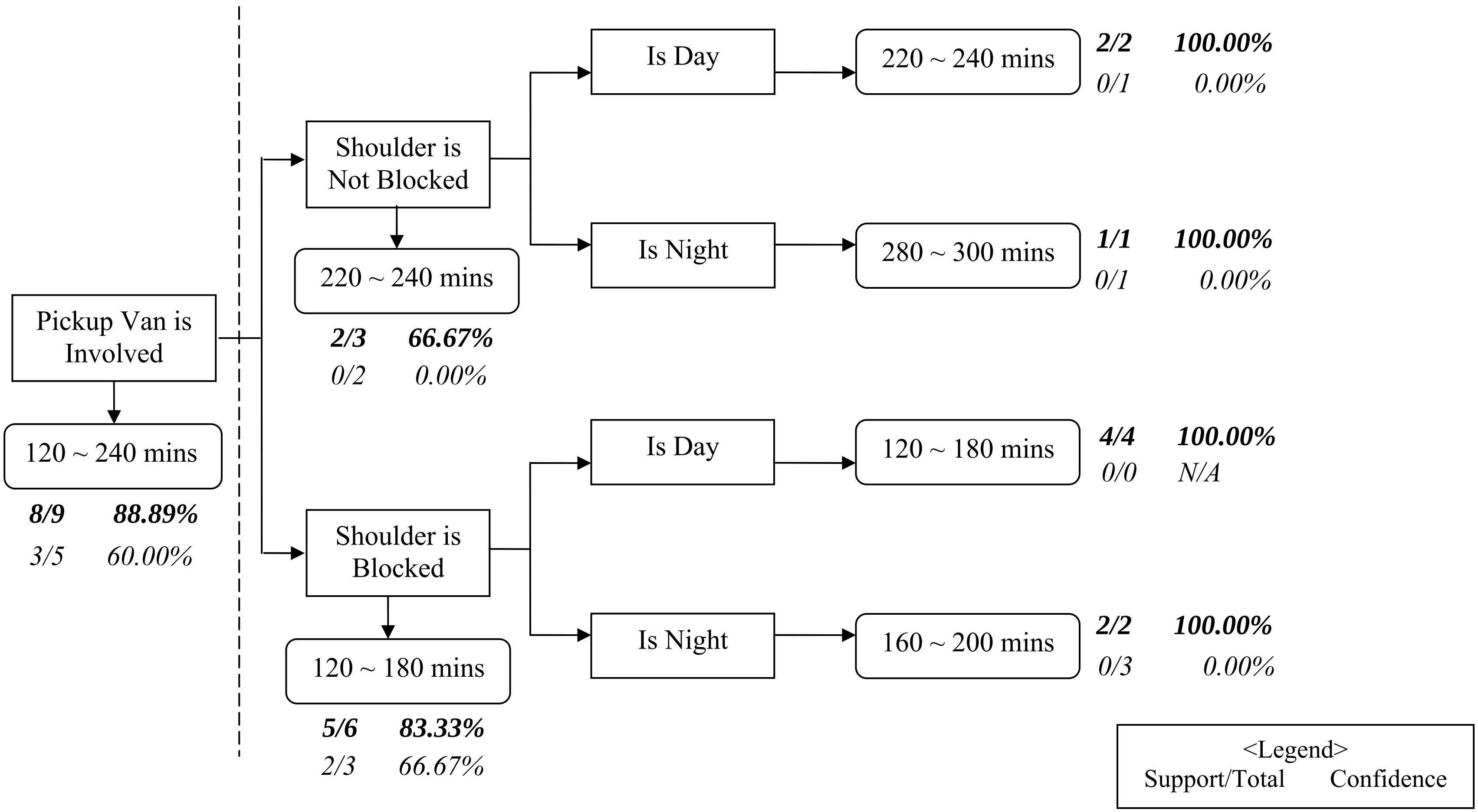


Figure A2.1(b)-1 Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

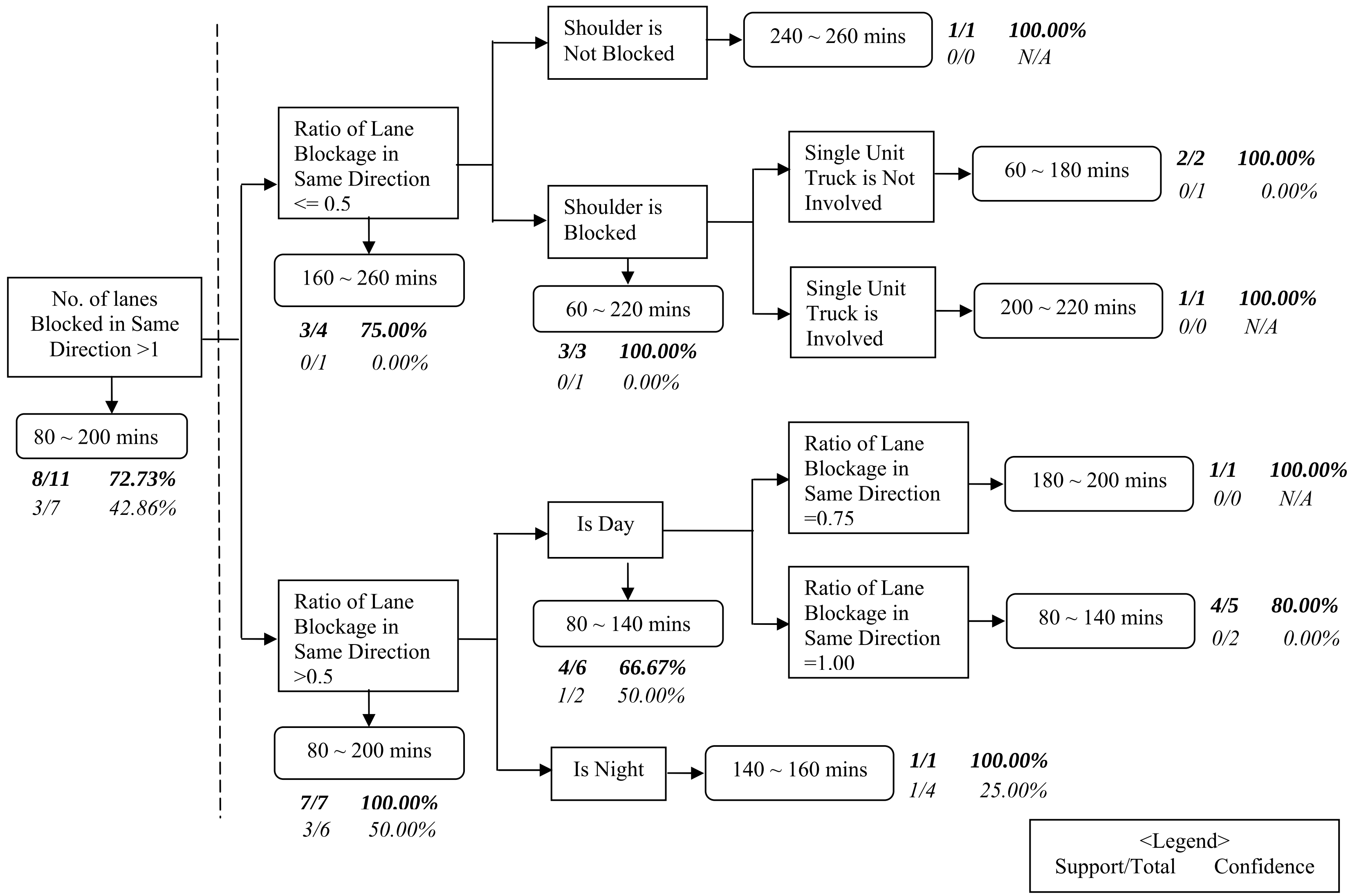


Figure A2.1(c) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

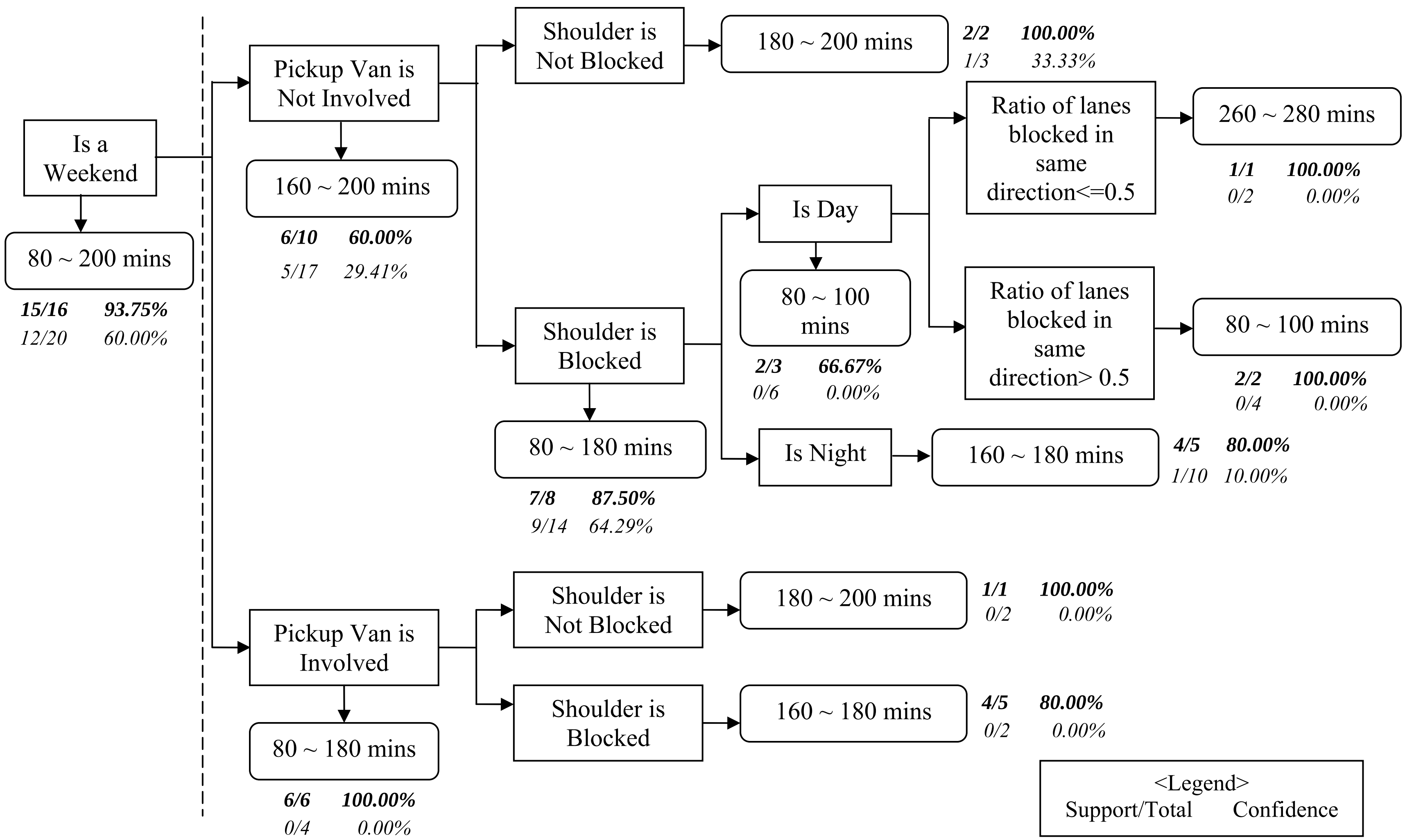


Figure A2.1(d) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

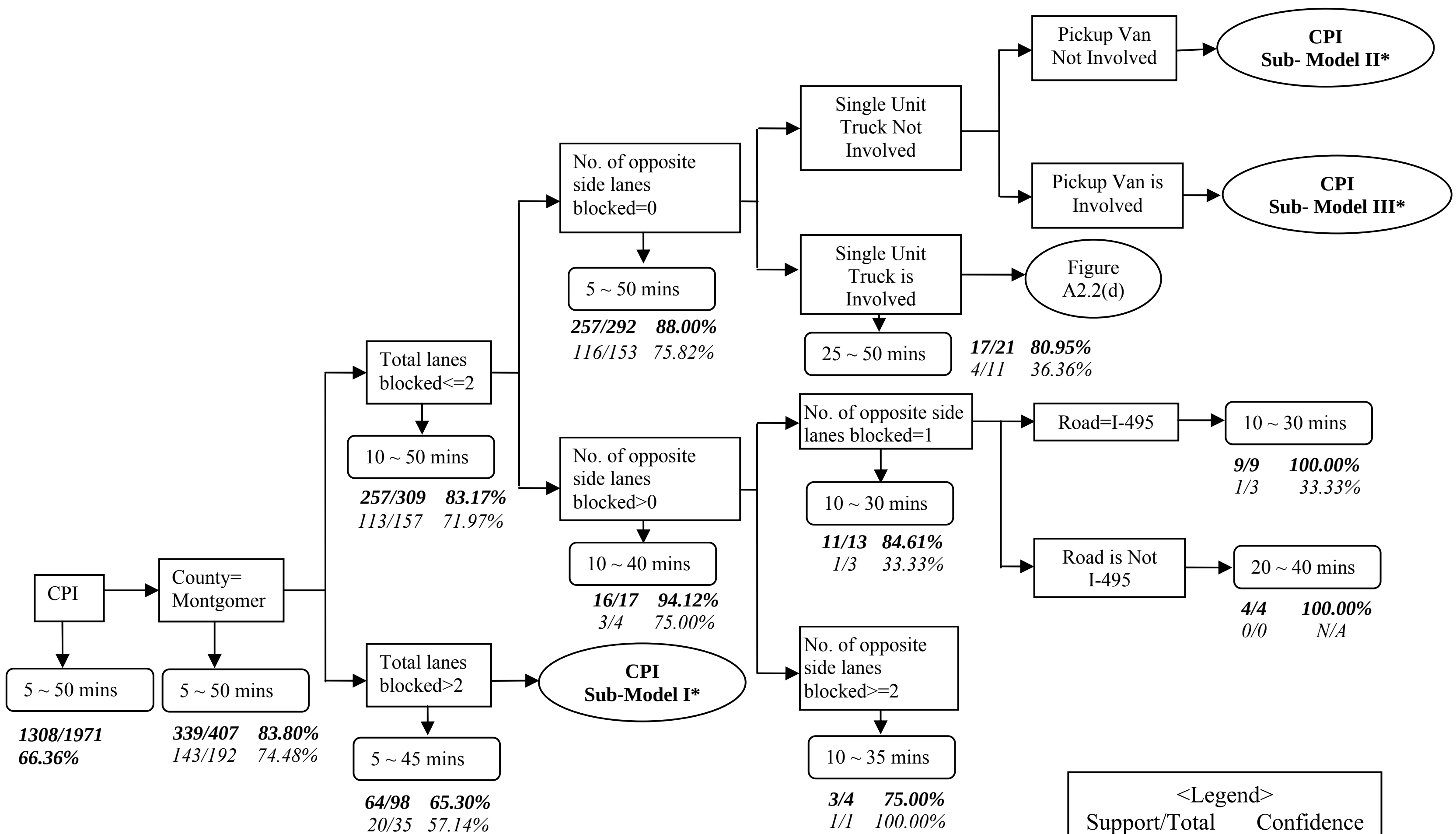
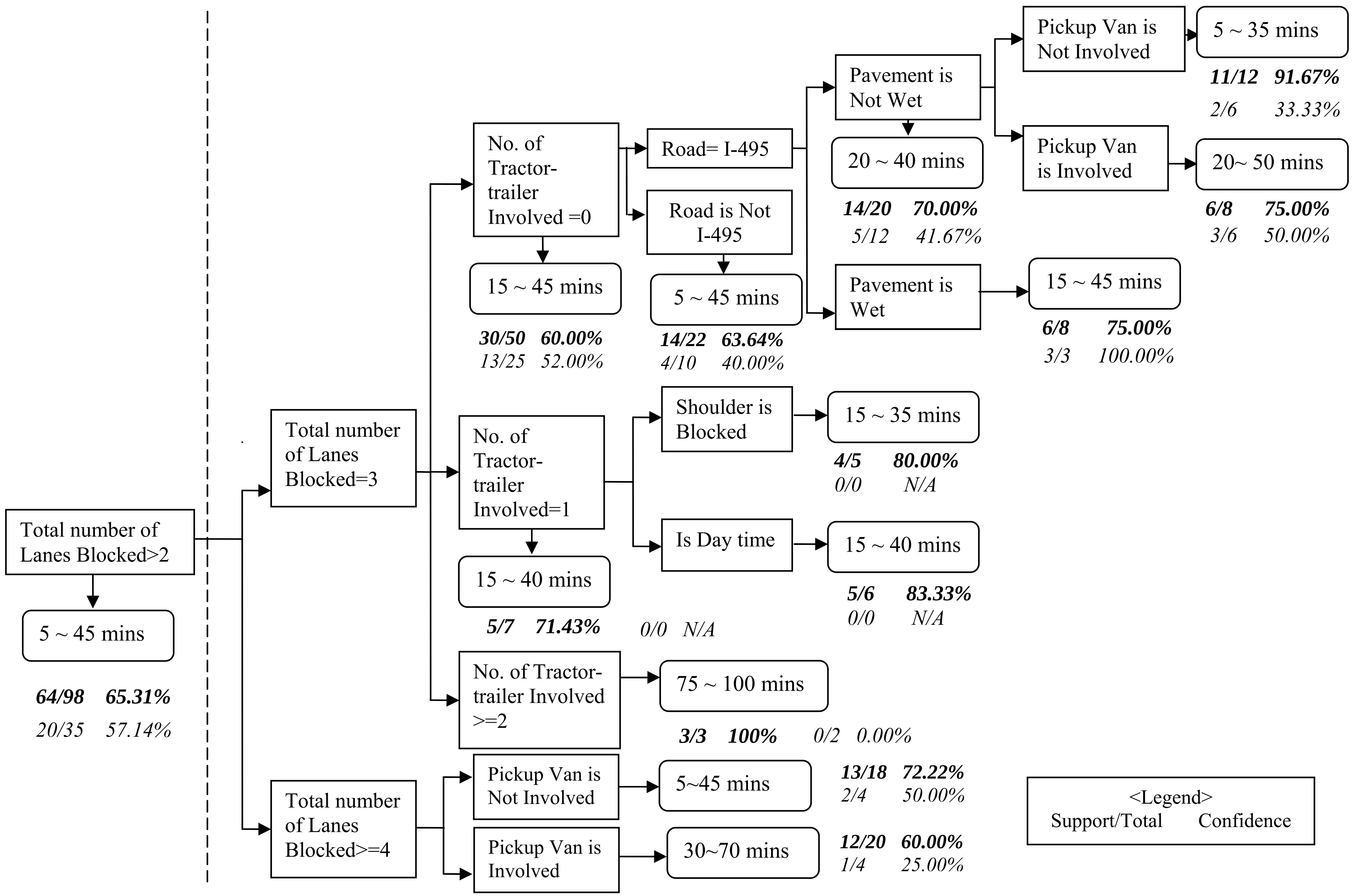


Figure A2.2 Rule Based Tree Model for Collision-Personal Injury in Montgomery County

* Detailed trees for subsets for CPI-Sub-Model-I, II and III could be found in Table A2.2(a), A2.2(b) and A2.2(c), respectively, in Appendix 2.



Total number of Lanes Blocked>2
5 ~ 45 mins
64/98 65.31%
20/35 57.14%
Total number of Lanes Blocked=3
No. of Tractor-trailer Involved =0
15 ~ 45 mins
30/50 60.00%
13/25 52.00%
Road= I-495
Road is Not I-495
5 ~ 45 mins
14/22 63.64%
4/10 40.00%
Pavement is Not Wet
20 ~ 40 mins
14/20 70.00%
5/12 41.67%
Pickup Van is Not Involved
5 ~ 35 mins
11/12 91.67%
2/6 33.33%
Pickup Van is Involved
20~ 50 mins
6/8 75.00%
3/6 50.00%
Pavement is Wet
15 ~ 45 mins
6/8 75.00%
3/3 100.00%
No. of Tractor-trailer Involved=1
15 ~ 40 mins
5/7 71.43%
0/0 N/A
Shoulder is Blocked
15 ~ 35 mins
4/5 80.00%
0/0 N/A
Is Day time
15 ~ 40 mins
5/6 83.33%
0/0 N/A
No. of Tractor-trailer Involved >=2
75 ~ 100 mins
3/3 100%
0/2 0.00%
Total number of Lanes Blocked>=4
Pickup Van is Not Involved
5~45 mins
13/18 72.22%
2/4 50.00%
Pickup Van is Involved
30~70 mins
12/20 60.00%
1/4 25.00%
<Legend>
Support/Total Confidence

Figure A2.2(a) Rule Based Tree Model for Subsets for CPI-Sub-Model I

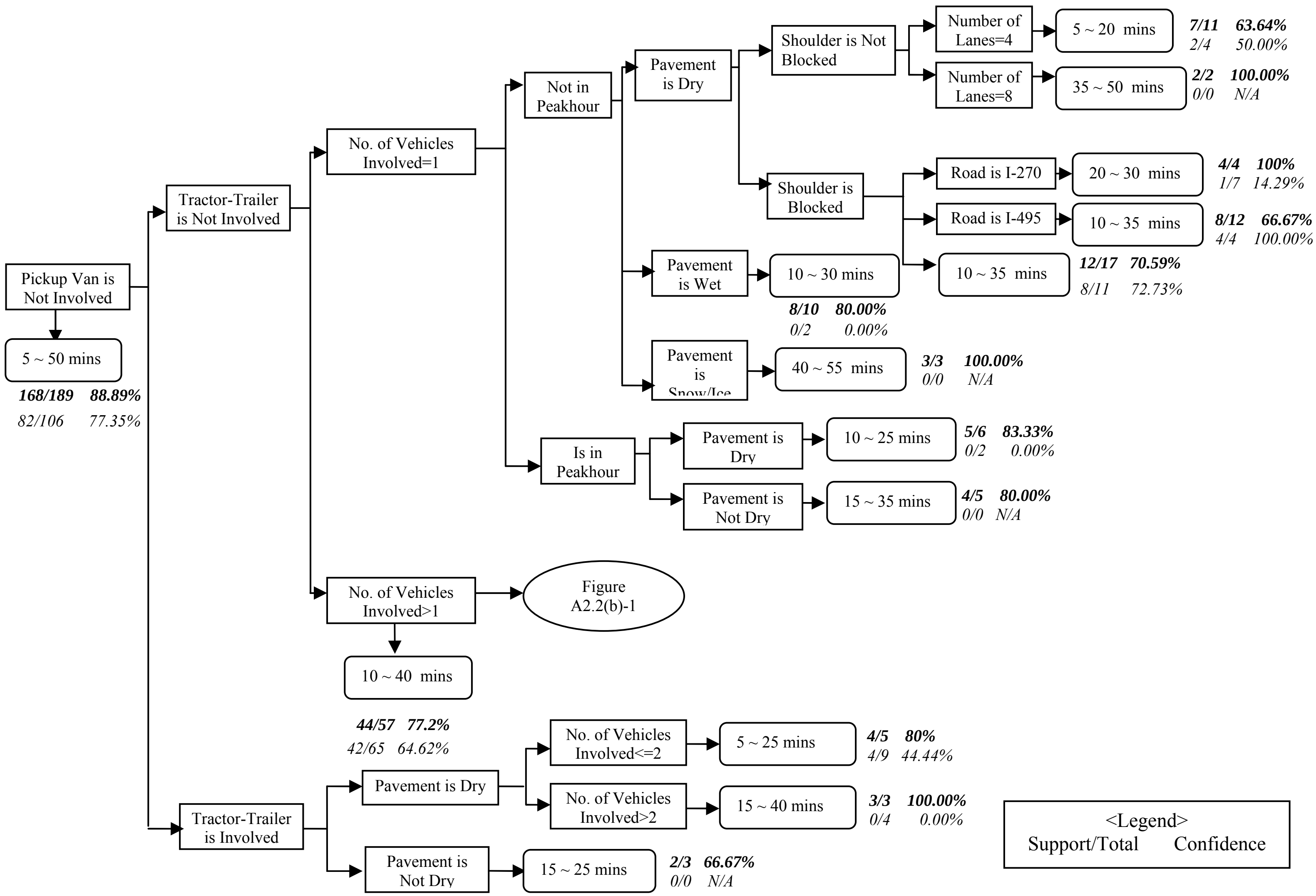


Figure A2.2(b) Rule Based Tree Model for Subsets for CPI-Sub-Model II

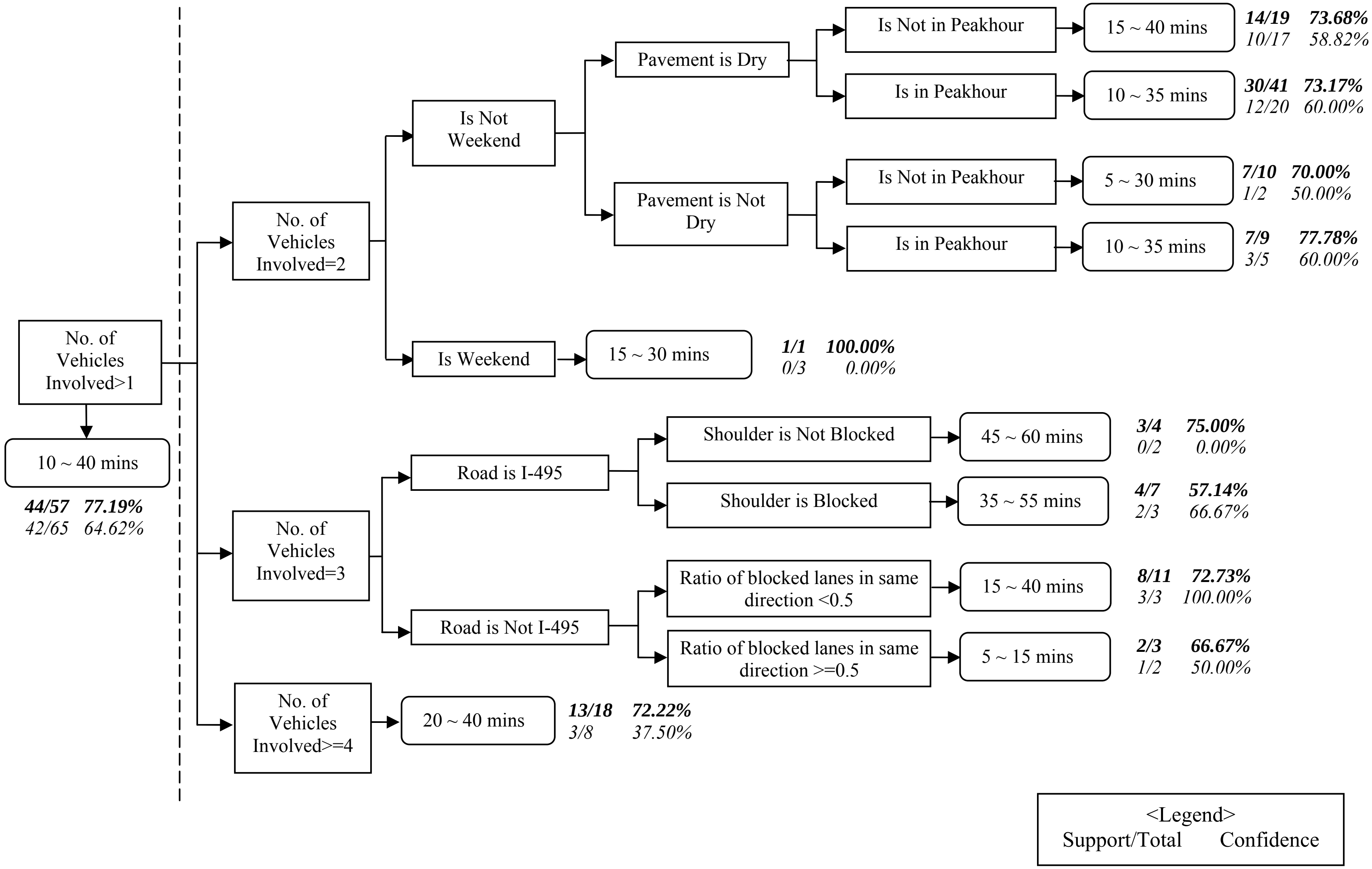


Figure A2.2(b)-1 Rule Based Tree Model for Collision-Personal Injury in Montgomery County

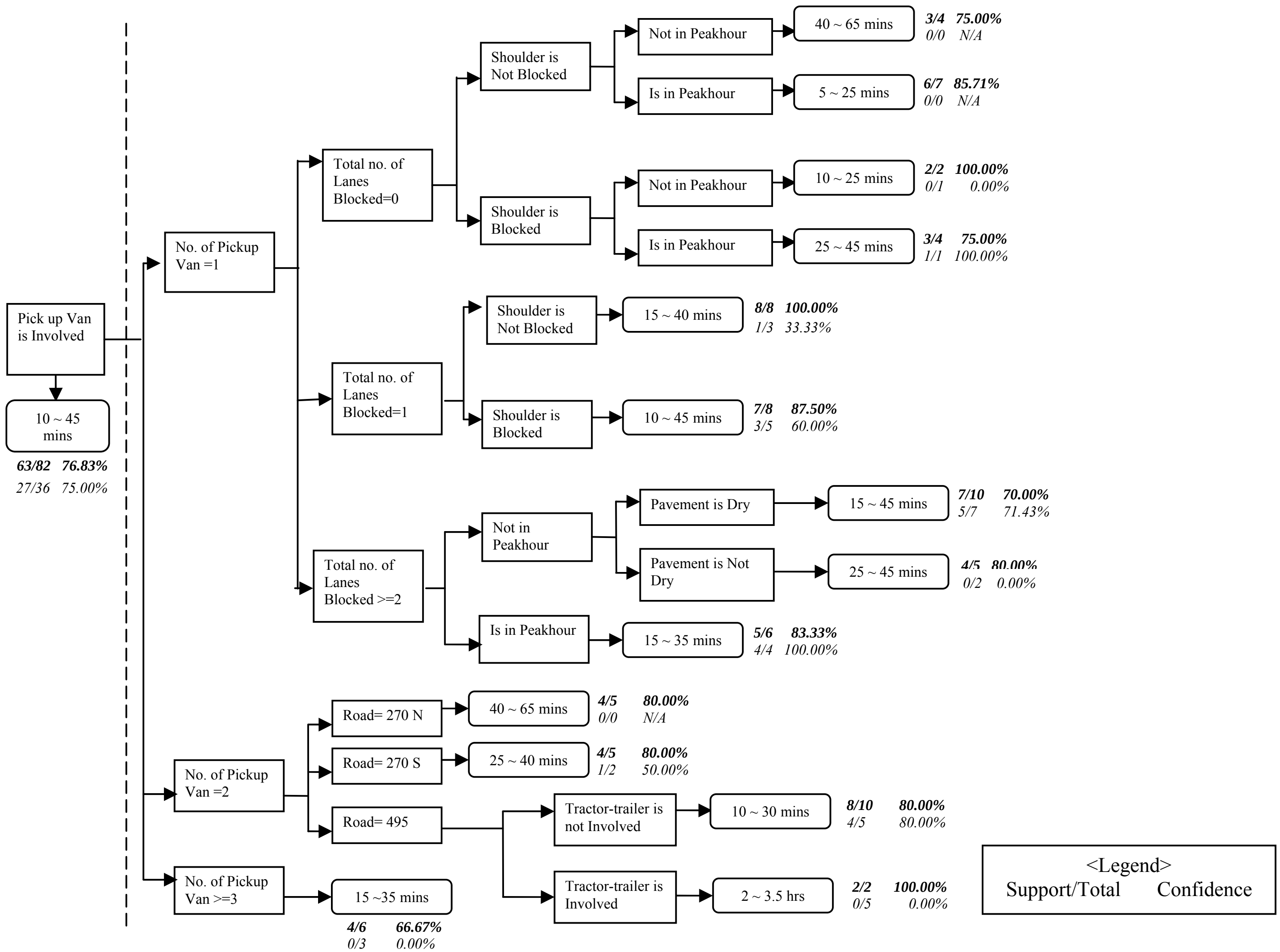


Figure A2.2(c) Rule Based Tree Model for Subsets for CPI-Sub-Model III

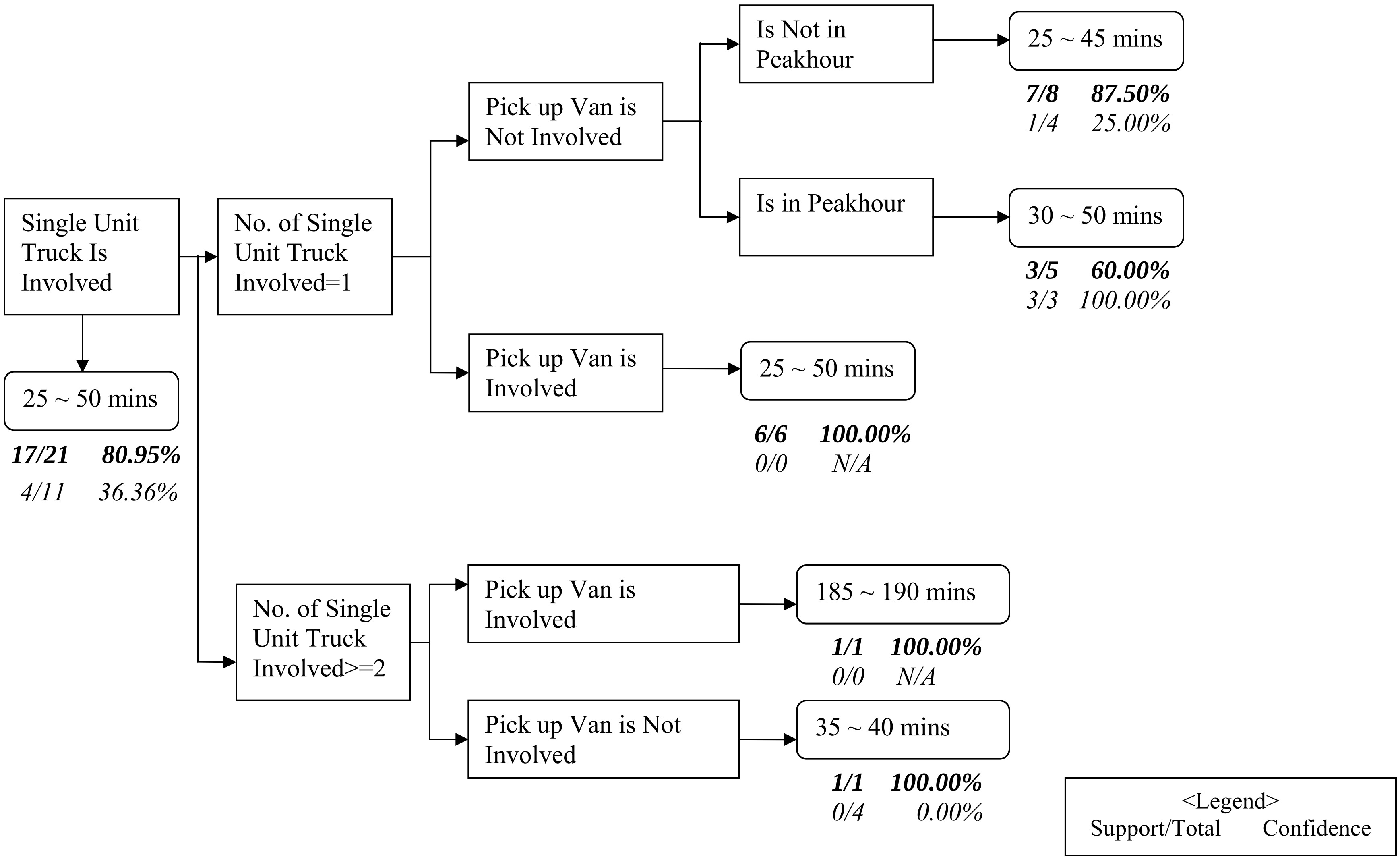


Figure A2.2(d) Rule Based Tree Model for Collision-Personal Injury in Montgomery County

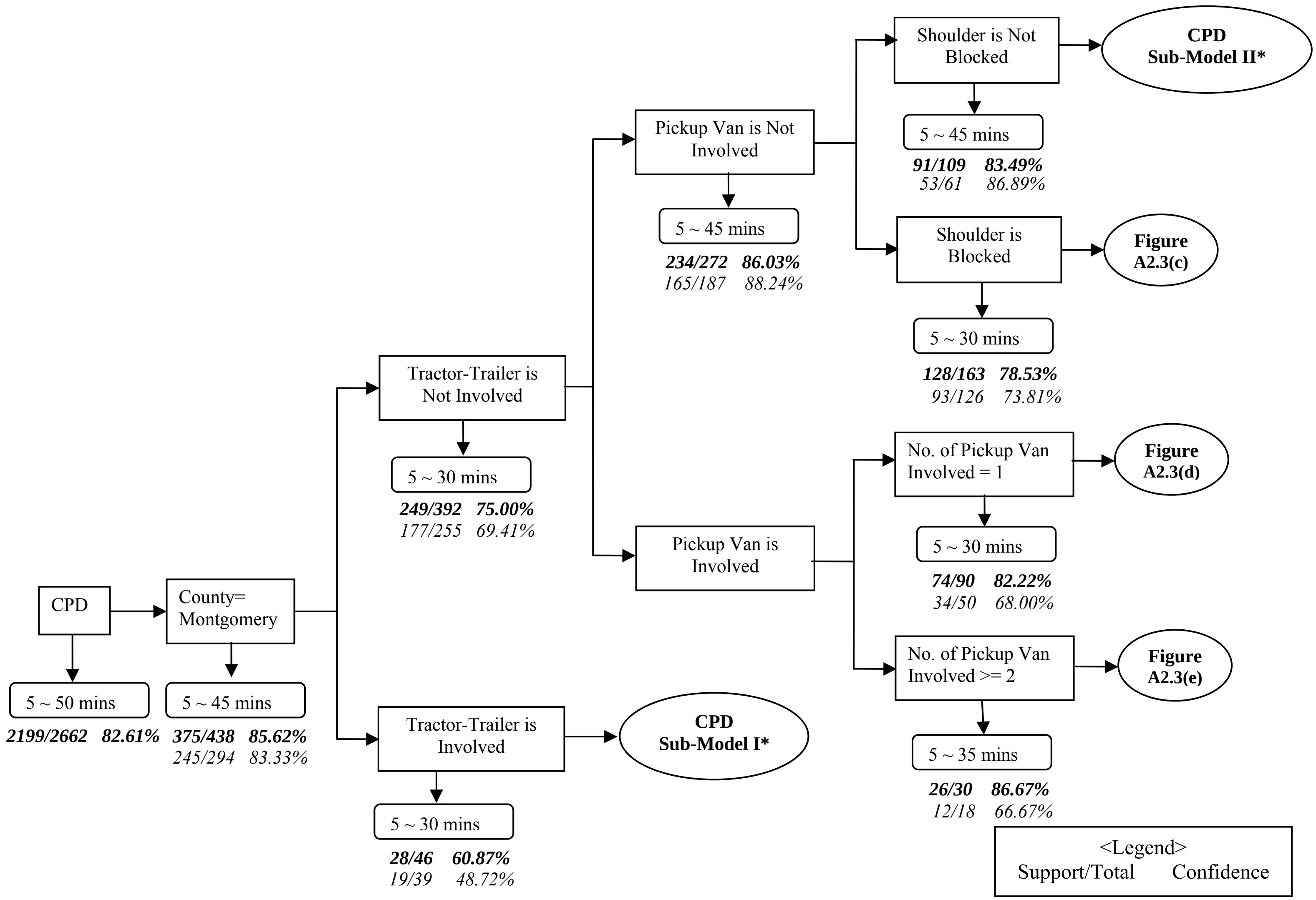


Figure A2.3 Rule Based Tree Model for Collision-Property Damage in Montgomery County

* Detailed trees for subsets for CPD-Sub-Model-I and II could be found in Table A2.3(a) and A2.3(b), respectively, in Appendix 2

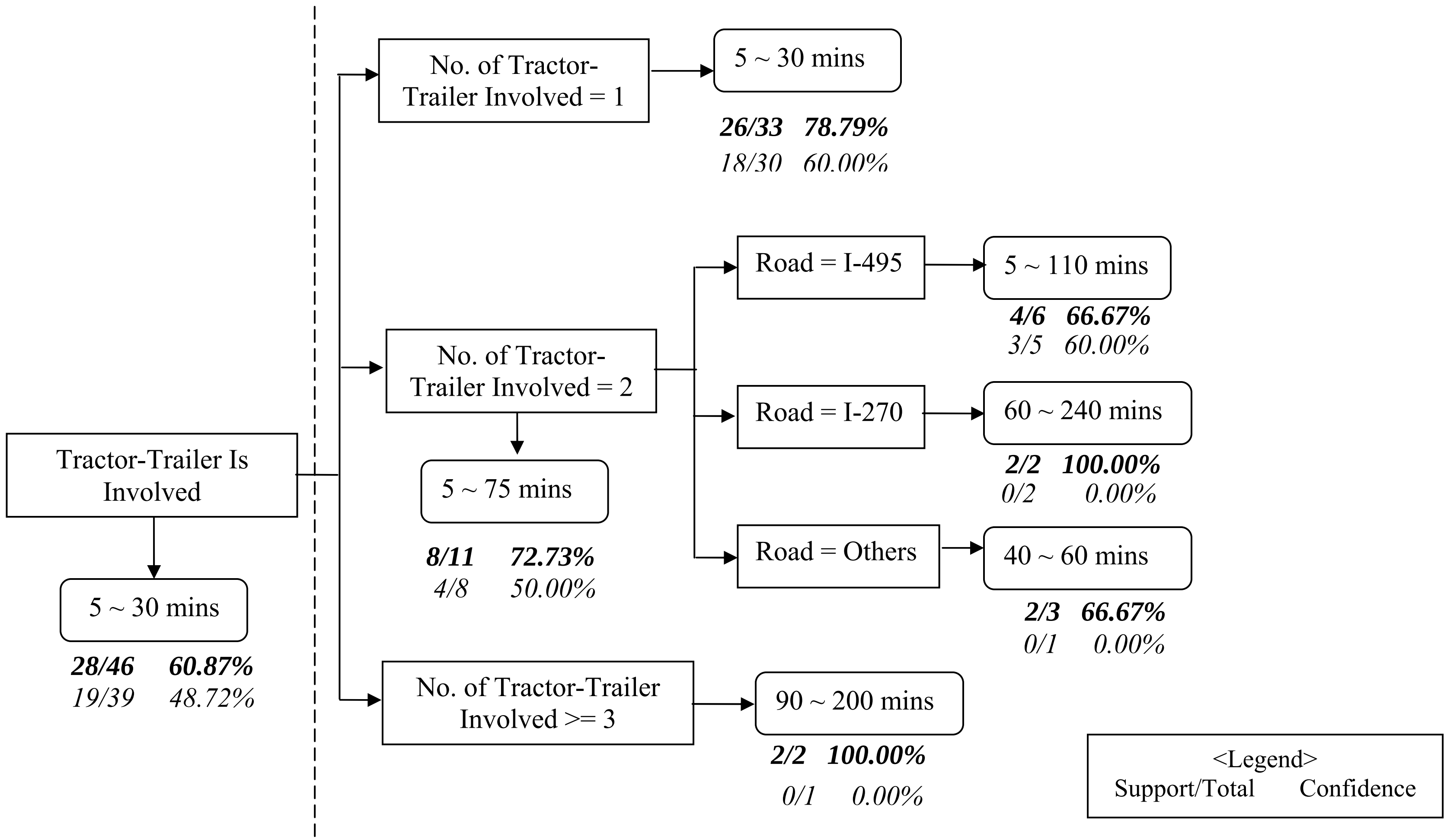


Figure A2.3(a) Rule Based Tree Model for Subsets for CPD-Sub-Model I

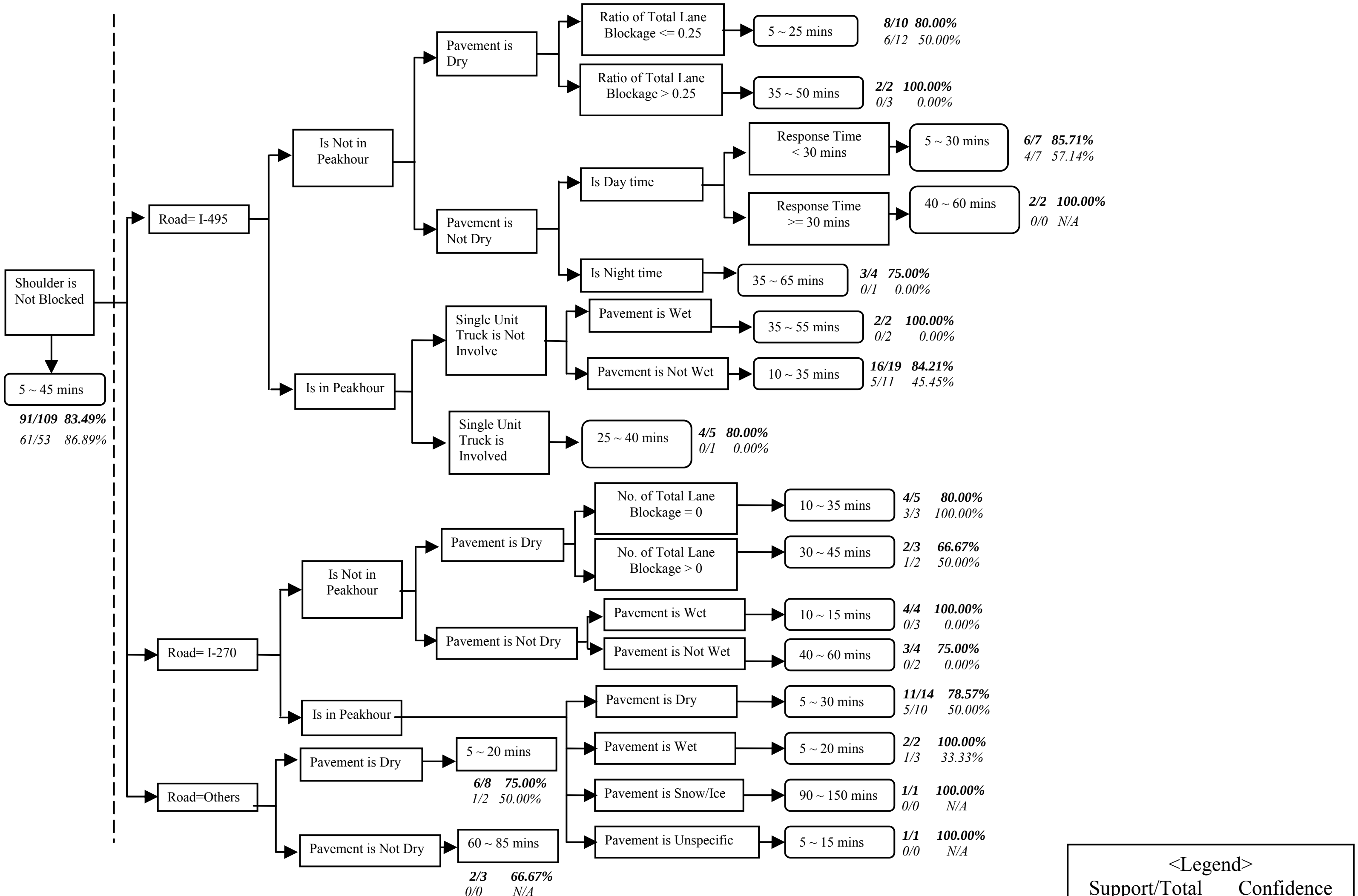


Figure A2.3(b) Rule Based Tree Model for Subsets for CPD-Sub-Model II

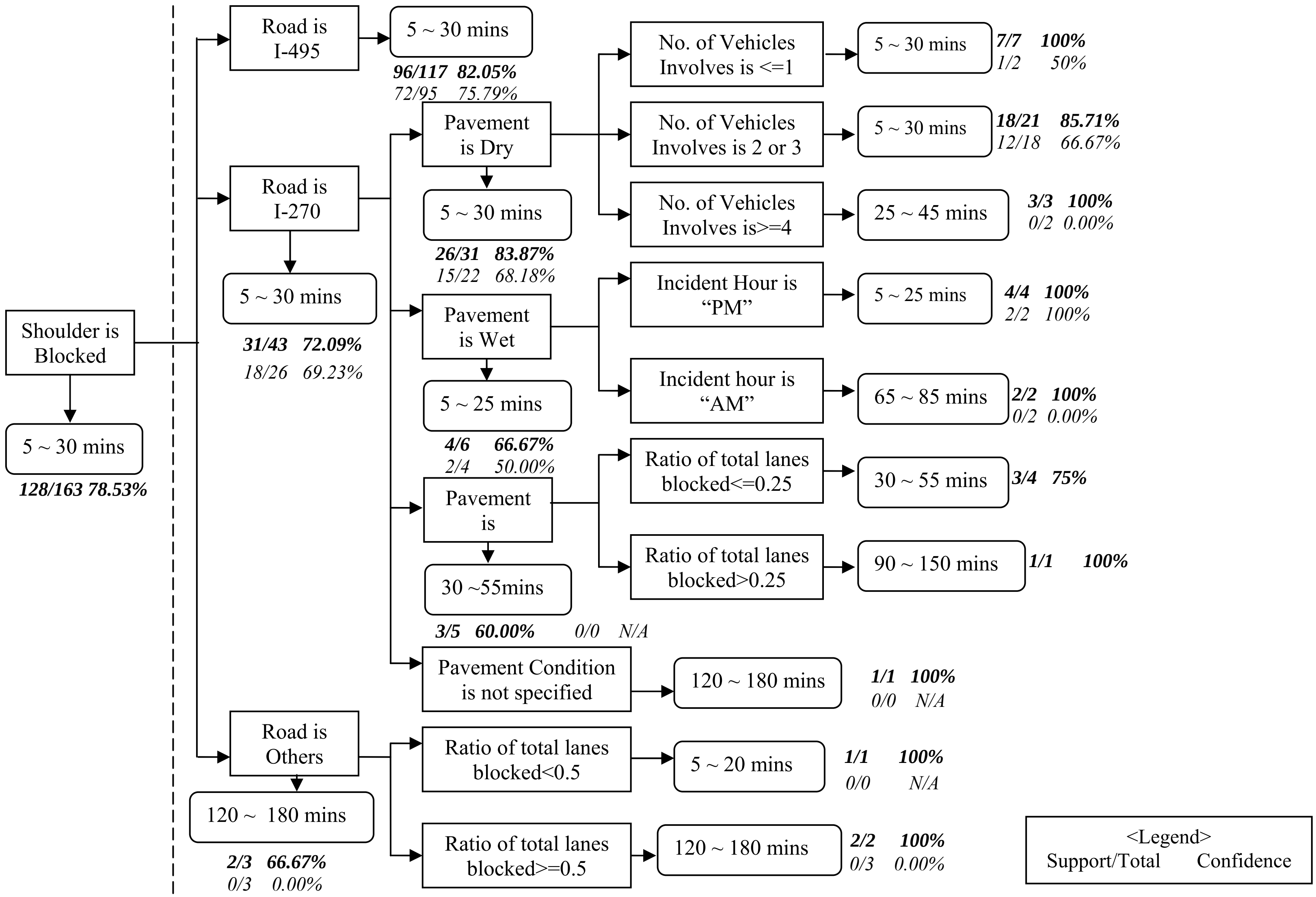


Figure A2.3(c) Rule Based Tree Model for Collision-Property Damage in Montgomery County

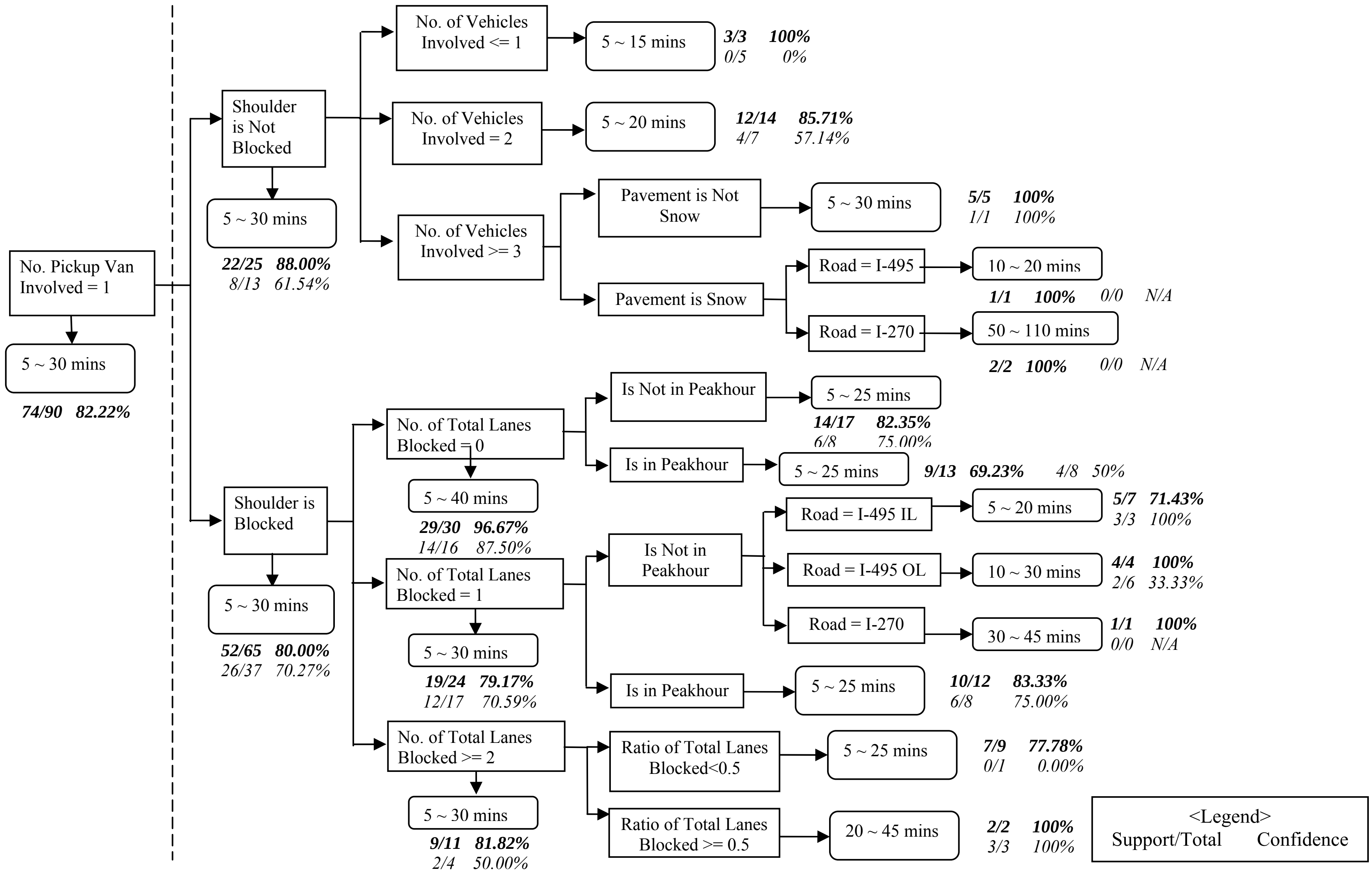


Figure A2.3(d) Rule Based Tree Model for Collision-Property Damage in Montgomery County

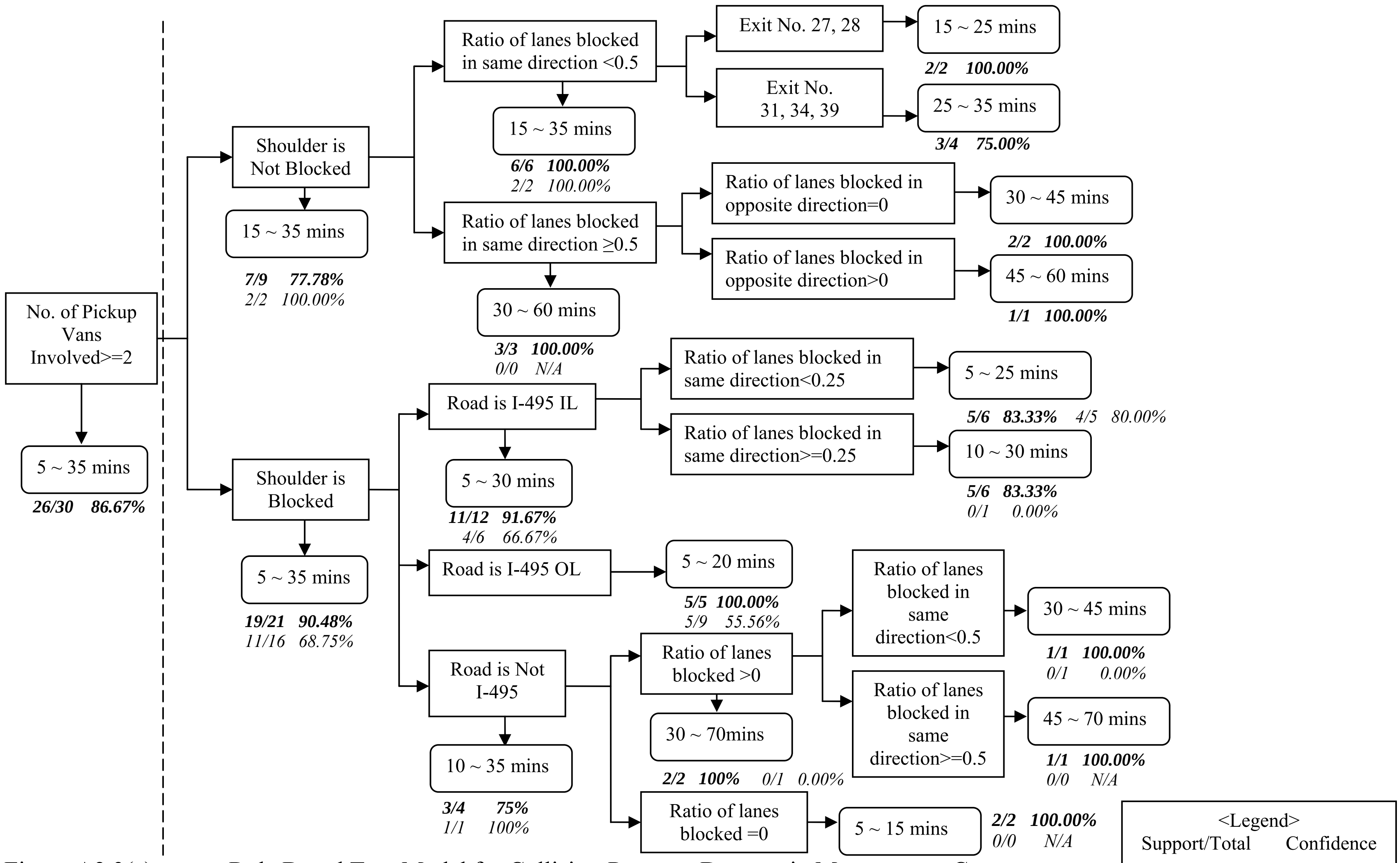


Figure A2.3(e) Rule Based Tree Model for Collision-Property Damage in Montgomery County

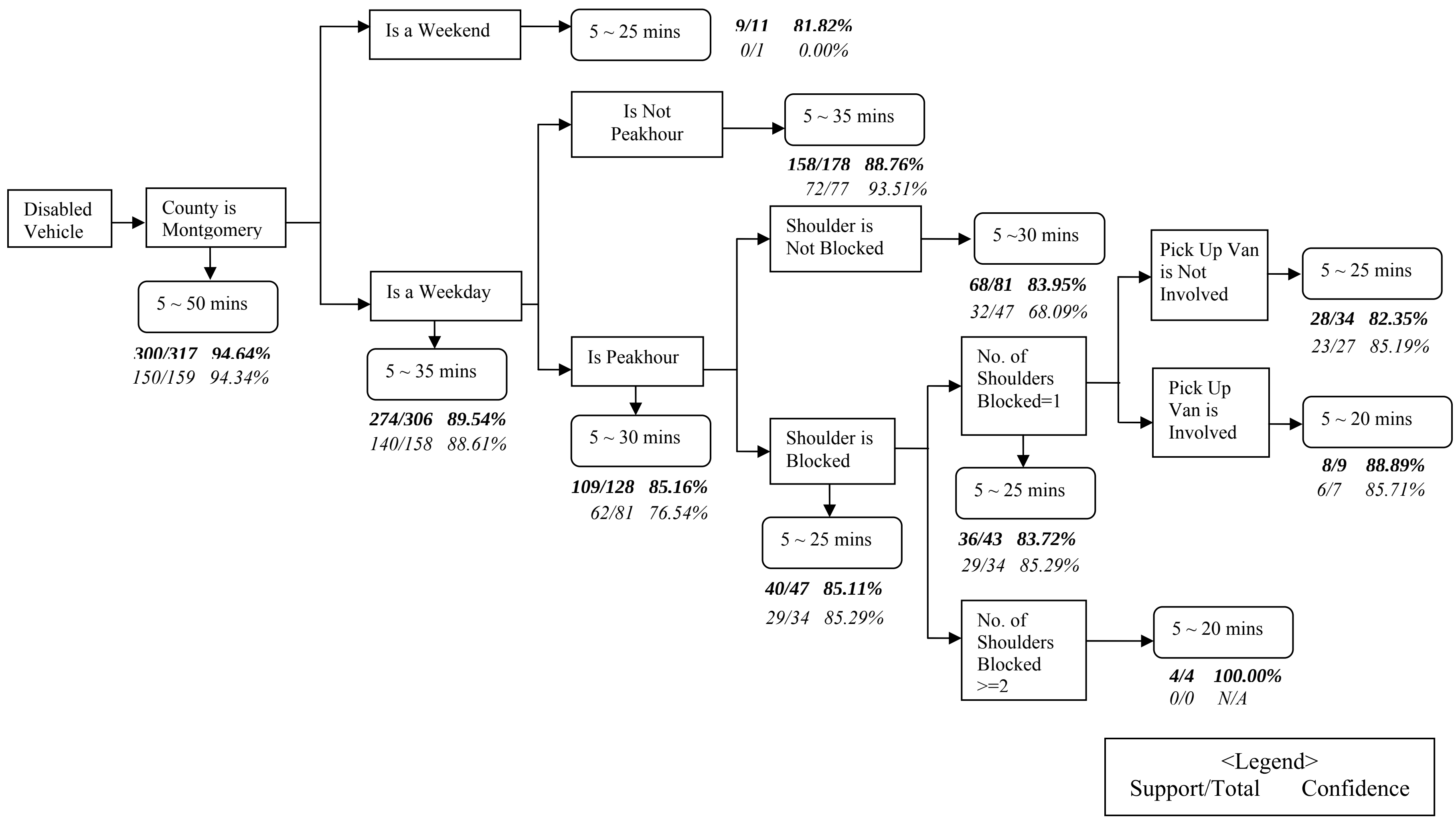


Figure A2.4 Rule Based Tree Model for Disabled Vehicles in Montgomery County

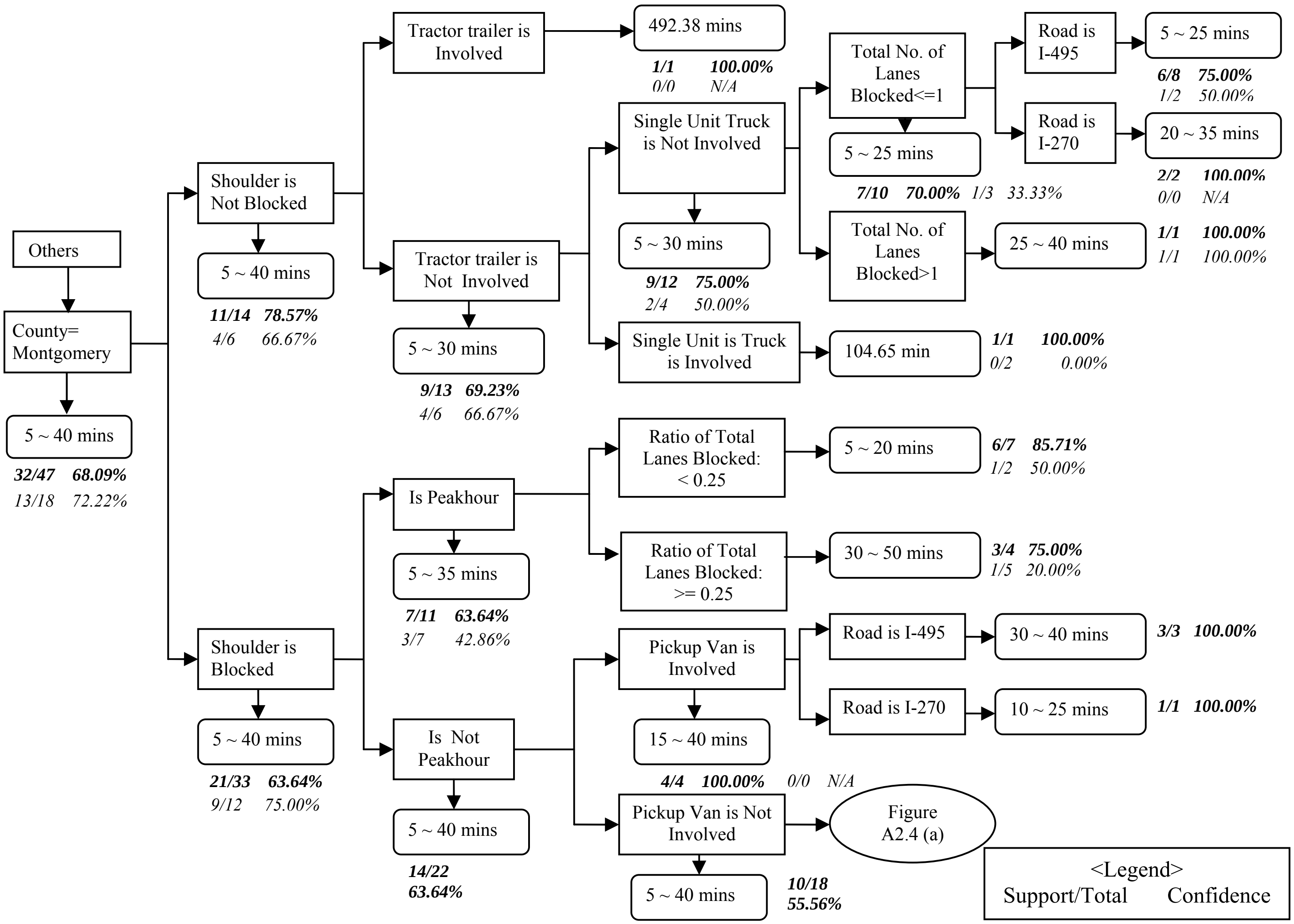


Figure A2.4 Rule Based Tree Model for Other Incident Natures in Montgomery County

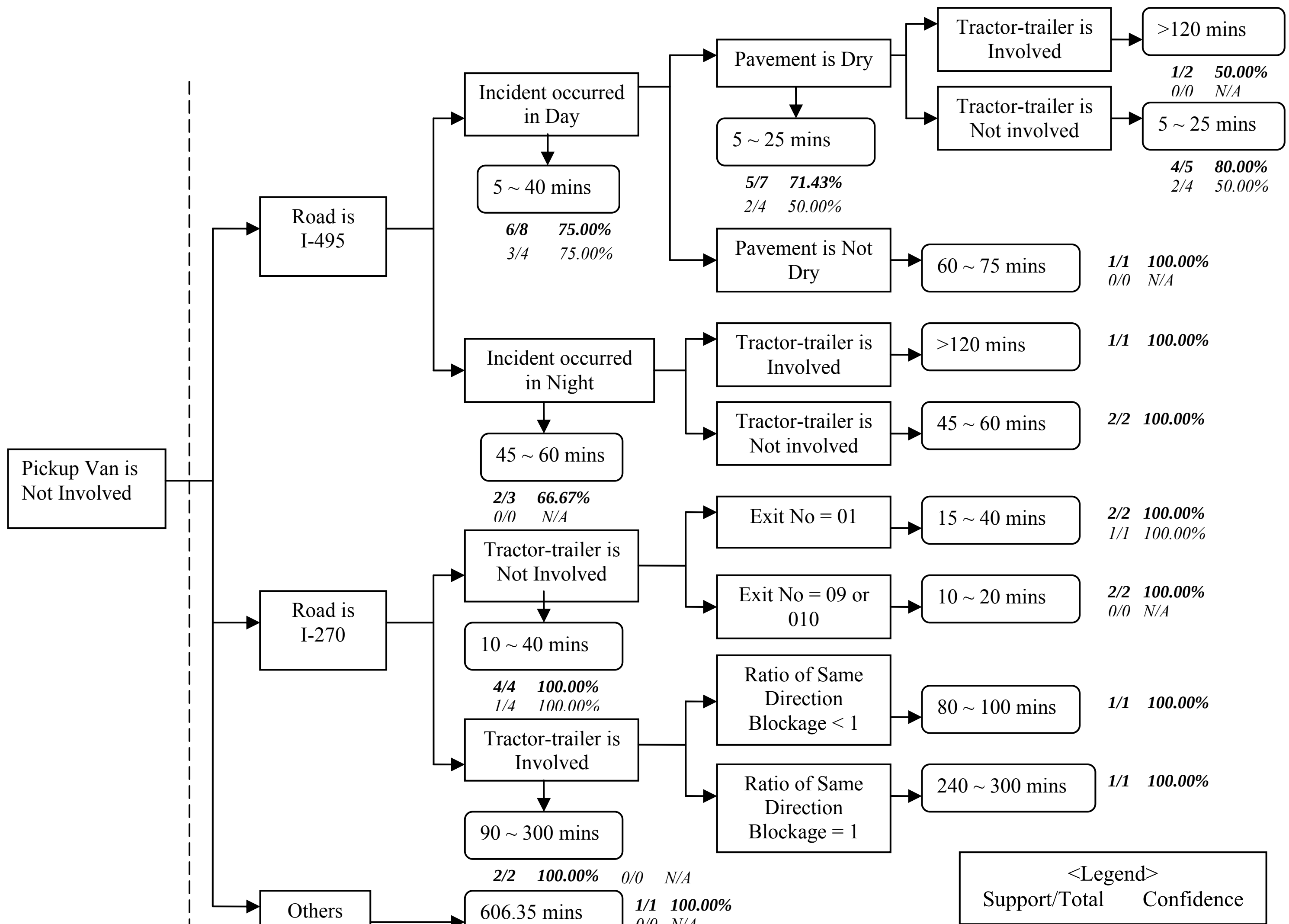


Figure A2.4(a) Rule Based Tree Model for Other Incident Natures in Montgomery County (Cont'd)

References

Ben-Akiva, M. and Lerman , S.R. (1985). "Discrete choice analysis: Theory and application to travel demand." MIT Press, Cambridge, Massachusetts.

Benzécri, J. P. (1973). "L'Analyse des Données: T. 2, I' Analyse des correspondances." Paris: Dunod.

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformations." *Journal of the Royal Statistical Society*, Series B, 26, 211–252.

Boyles, S., Fajardo, D., and Waller, S. T. (2007). "A Naïve Bayesian Classifier for Incident Duration Prediction." CD-ROM of Proceedings of the TRB 2007 Annual Meeting, Washington, D.C.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). "Classification and regression trees." Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Carrol, J. D., Green, P. E. and Schaffer, C. M. (1986). "Interpoint distance comparisons in correspondence analysis." *Journal of Marketing Research,* 23, 271-280.

Dimakos, Ioannis C. "Power Transformation Using SAS/IML Software." Syracuse University, Computing & Media Services.

Garib, A., Radwan, A. E. and Al-Deek, H. (1997). "Estimating Magnitude and Duration of Incident Delays." *Journal of Transportation Engineering*, Vol. 123, No. 6, November/December, 459-466.

Giuliano, G. (1989). "Incident Characteristics, Frequency, and Duration on a High Volume Urban Freeway." *Transportation Research – A*, Vol. 23A, No. 5, 387-396.

Golob, T. F., Recker, W. W. and Leonard, J. D. (1987). "An Analysis of the Severity and Incident Duration of Truck-Involved Freeway Accidents." *Accident Analysis and Prevention,* Vol. 19, No. 4, August, 375-395.

Greenacre, M. J. (1984). "Theory and applications of correspondence analysis." New York: Academic Press.

Hill, T. and Lewicki, P. (2005). "Statistics: Methods and Applications", StatSoft, Inc.

Hoffman, D. L. and Franke, G. R. (1986). "Correspondence analysis: Graphical representation of categorical data in marketing research." *Journal of Marketing Research*, *13*, 213-227.

Johnson, R. A. and Wichern, D. W. (1993). “Applied Multivariate Statistical Analysis (3rd ed).” Englewood Cliffs, NJ: Prentice-Hall.

Jolliffe, I. T. (1972). “Discarding Variables in a Principal Component Analysis, I: Artificail Data.” *Applied Statistics*, 21, 160-173.

Jolliffe, I. T. (1973). “Discarding Variables in a Principal Component Analysis, II: Real Data.” *Applied Statistics*, 22, 21-31.

Jones, B., Janssen, L. and Mannering, F. “Analysis of the Frequency and Duration of Freeway Accidents in Seattle.” *Accident Analysis and Prevention,* Vol. 23, No. 4, August 1991, 239-255.

Khattak, A. J., Schofer, J. L. and Wang, M-H. (1995) “A Simple Time Sequential Procedure for Predicting Freeway Incident Duration.” *IVHS Journal*, Vol. 2, No. 2, 113-138.

Khorashadi, P. E. (2003). “Analysis of Driver Injury Severity: Logit Models of Truck Involvement/Truck Causation.” Ph.D. dissertation.

Kononenko, I. (1993). “Inductive and Bayesian Learning in Medical Diagnosis.” *Applied Artificial Intelligence*, Volume 7, 317-337.

Koppelman, F. S. and Bhat, C. (2006). “A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models.”

Lemon, S., Roy, J., Clark, M. A., Friedmann, P. D. and Rakowski, W. (2003). “Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression.” *Annals of Behavioral Medicine*, Volume 26, 172-181.

Lewis, Roger J. (2000). “An Introduction to Classification and Regression Tree (CART) Analysis.” Presented in the Annual Meeting of the Society for Academic Emergency Medicine.

Lin, P-W., Zou, N. and Chang, G-L. (2004). “Integration of a Discrete Choice Model and a Rule-Based System for Estimation of Incident Duration: a Case Study in Maryland.” CD-ROM of Proceedings of the 83rd TRB Annual Meeting, Washington, D.C.

MacFadden, D. (1974). “Conditional Logit Analysis of Qualitative Choice Behavior.” In Frontiers in Econometrics. Zarembka, P., ed. Academic Press, NY.

Maron, M. (1961). “Automatic indexing: an experimental inquiry.” *Journal of the ACM*, Volume 8, 404-417.

Nam, Doohee and Mannering, F. (2000). “An Exploratory Hazard-Based Analysis of

Highway Incident Duration.” *Transportation Research – A*. Vol. 34A, No. 2, 85-102.

Ozbay, K. and Kachroo, P. (1999). “Incident Management in Intelligent Transportation Systems.” Artech House, Boston, MA.

Pindyck, Robert S. and Rubinfeld, D. L. (1998). “Econometric Models and Economic Forecasts.” Forth Edition, McGraw-Hill International Editions.

Rencher, Alvin C., (2002). “Methods of Multivariate Analysis.” Second Edition, USA, Wiley-Interscience.

Resampling Stats – XLMiner User Guide, Resampling Stats, co-published with On Demand Manual a division of Trafford Holding Ltd.

Powers, R., Goldszmidt, M. and Cohen, I. (2005). “Short Term Performance Forecasting in Enterprise Systems.” Publication HPL-2005-50. HP Labs.

Smith, K. and Smith, B. (2001). “Forecasting the Clearance Time of Freeway Accidents.” Research Report, STL-2001-01. Center for Transportation Studies, University of Virginia, Charlottesville, VA.

Sullivan, E. C. (1997). “New Model for Predicting Incidents and Incident Delay.” *ASCE Journal of Transportation Engineering*, Vol. 123, July/August, 267-275.

Transportation Research Board (TRB). (2000). “Highway Capacity Manual 2000.” National Research Council, Washington, D.C.

Tukey, J. W. (1952). “Allowances for Various Types of Error Rates.”

Tukey, J. W. (1953). “The Problem of Multiple Comparisons”

Ulfarsson, G. F. (2001). “Injury severity analysis for passenger car, pickup, sport utility vehicle and minivan drivers: Male and female differences.” Ph.D. dissertation.

Washington, S., Karlaftis, M. and Mannering, F. (2003). “Statistical and Econometric Methods for Transportation Data Analysis.” CRC Press, Boca Raton.

Yohannes, Y. and Hoddinott, J. (1999). “Classification and Regression Trees: An Introduction” Technical Guide #3, International Food Policy Research Institute, Washington, D.C.