ABSTRACT

Title of Document: EMPIRICAL ANALYSIS AND MODELING OF FREEWAY INCIDENT DURATION

Woon Kim, Master of Science, 2007

Directed By: Professor Gang-Len Chang,
Department of Civil and Environmental Engineering

This study presents a set of models for predicting incident duration and identifying variables associated with the incident duration in the state of Maryland. The incident database for years 2003 to 2005 from the Maryland State Highway (MDSHA) database is used for model development, and year 2006 for the model validation. This study, based on the preliminary analysis with the Classification Tree method, has employed the Rule-Based Tree Model to develop the primary prediction model. To enhance the prediction accuracy for some incidents with complex nature or limited samples, the study has also proposed and calibrated several supplemental components based on the Multinomial Logit and Regression methods. Although the prediction accuracy could still be improved if a data set with better quality is available, the developed set of models offers an effective tool for responsible agencies to estimate the approximate duration of a detected incident, which is crucial in projecting the potential impacts on the highway network.

EMPIRICAL ANALYSIS AND MODELING OF FREEWAY INCIDENT DURATION

By

Woon Kim

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2007

Advisory Committee:
Professor Gang-Len Chang, Chair
Professor Paul M. Schonfeld
Assistant Professor Cinzia Cirillo

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Traffic incidents have long been recognized as the main contributor of congestion in highway networks. Incidents defined in this study include vehicle disablements, fire, road debris, construction, police activities and vehicle accidents. On congested highways, any incident regardless of involving personal fatalities, injuries, or property damages will cause considerable reduction in roadway capacity due to lane closures or impediments. As reported in literature, one lane blockage on a three-lane road will reduce the capacity by 50% (TRB, 1994). The capacity reduction during the incident duration will inevitably result in heavy congestion, delay, and thus give birth to enormous socioeconomic loss. In the day-to-day traffic control and management, if some reliable way for predicting incident duration in real time is available, responsible agencies can convey information to travelers via the variable message signs (VMS), estimate the resulting queue length, 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 non-recurrent congestion in highway networks.

1.2 Definition of Incident Duration

According to Highway Capacity Manual (TRB, 1994), the entire duration of incidents consists of four phases as shown in Figure 1.1. The first phase is the detection time that represents the time elapsed from incident occurrence to its detection. The response time corresponds to the period of time between the incident detection and the arrival of any emergency or incident response unit. The 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 is the recovery time that measures the time required for the traffic condition to return back to its normal condition.

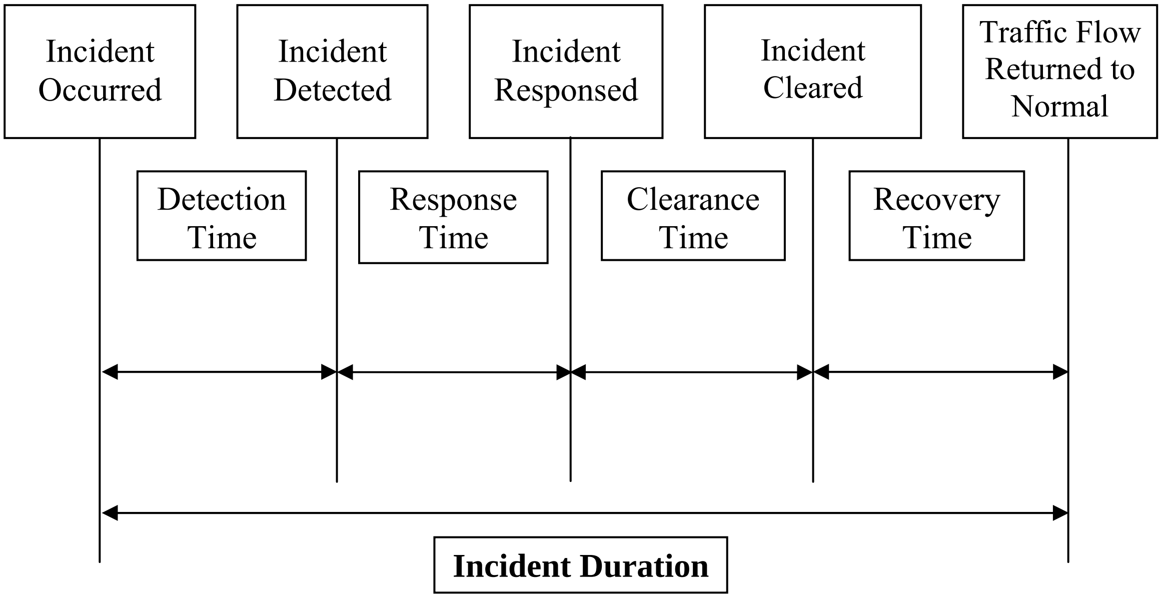


Figure 1.1 Phases of Traffic Incident Duration

In general, it is difficult to know the exact timestamp of incident occurrence, and the recovery time is usually regarded as being out of scope for the incident duration study. Moreover, the database used for this study includes only reliable records for response and clearance time. Thus, in this study, incident duration is defined as the time elapsed from incident detection to its clearance, which is the sum of 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 rigorous statistical models. Improvements in reporting techniques and incident information database have facilitated a detailed analysis of critical variables that influence incident duration and hence its prediction. Previous studies in this field have resulted in different prediction methods and models. However, it must be noted that these prediction models are developed based on the sets of data that are derived from different sources. Thus, information available for predicting the duration of an incident may vary between different databases. It is also observed that incident duration is influenced by various location-specific factors. Hence, to ensure reliable and efficient modeling of incident duration prediction for an area, one needs to calibrate the model from a well-designed database which includes all critical information of 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. The CHART (Coordinated Highways Action Response Team) database from Maryland State Highway Administration (MDSHA) is used in this study.

This study begins with a review of related literature in Chapter 2, including the most representative approaches for predicting incident duration – (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.

Chapter 3 is focused on the description of available data and the statistical analyses of interrelations between key variables. This chapter includes preliminary analysis for the distribution of incident duration, statistical tests for independent variables using 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 with Classification and Regression Tree (CART) Model. Based on the findings from CART, it further explores a new model, named the Rule-Based Tree Model. Detailed procedures for model development and its performance as well as validation are also included in the following subsections. Chapter 4 concludes with the overall findings from the Rule-Based Tree Model, 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 Models (MNL) and their performance with a test dataset. This is followed by the development of Multiple Linear Regression Models for some incident natures with small sample data and their performance. Potential applications of supplemental models are highlighted in the last section.

Chapter 6 summarizes 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 with 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, and (6) Discrete Choice Models. Although there are a variety of existing techniques with acceptable results, they cannot be directly applied to incidents that occurred 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 for the incident duration 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 that involved trucks, and found that the incident duration 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 duration. In 1999, Ozbay and Kachroo also found that the distribution of incident duration from their data set shows 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 follow 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 learned that their data set can 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 time 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 duration under certain given conditions; for example, the probability of incident duration 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 incident duration would be different, depending on how long the incident has lasted (known as duration dependence in Nam and Mannering (2000)), and the incident characteristics. One of the interesting approaches under this concept is the hazard-based duration model. This model allows researchers to formulate incident duration with conditional probability models. Such models have been widely used in biometrics and industrial engineering fields to determine causality from the duration data. Due to its similarity with the nature of traffic incident duration, their theoretical concepts and models have recently been applied in the transportation field. With such approach, researchers' interests have been expanded from simply estimating and predicting the incident duration to computing the likelihood that the incident will 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 set of two-year data from Washington State. Through their study, it is shown that each incident time (i.e. detection/reporting, response, and clearance times) is significantly affected by numerous factors, and different assumptions of distribution are recommended for different incident times. They also found that the estimated coefficients were unstable through the two-year data used in the model development. As concluded by Nam and Mannering, this approach is more useful to determine which variable has greater influence on incident duration, than to estimate or predict the incident duration for a set of given explanatory variables.

Another simple methodology to predict incident duration 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 most well-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 duration indicates that the incident duration in this data set follows 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 is the most significant factor in affecting the incident duration, which is followed by weather condition, peak hour, truck involvement, and the combined effect of 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 total set of variables will be more accurate and reliable, they are less practical for the real-time incident duration prediction because this full set of variables can only be available after the incident is cleared. Thus, they 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 developed by Khattak et al. (1995) 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, more variables are included progressively 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 is to introduce and demonstrate the time sequential model rather than proving 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 the incident duration models. One of the representative models is developed by Ozbay and Kachroo (1999) for the Northern Virginia region. They began with developing a model to predict clearance time using linear regression, based on a large size of samples. Unfortunately, they completed the analysis with a poor result ($R^2 \approx 0.35$), and learned that the incident duration follows neither a lognormal 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 an 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 which is estimated from the past records.

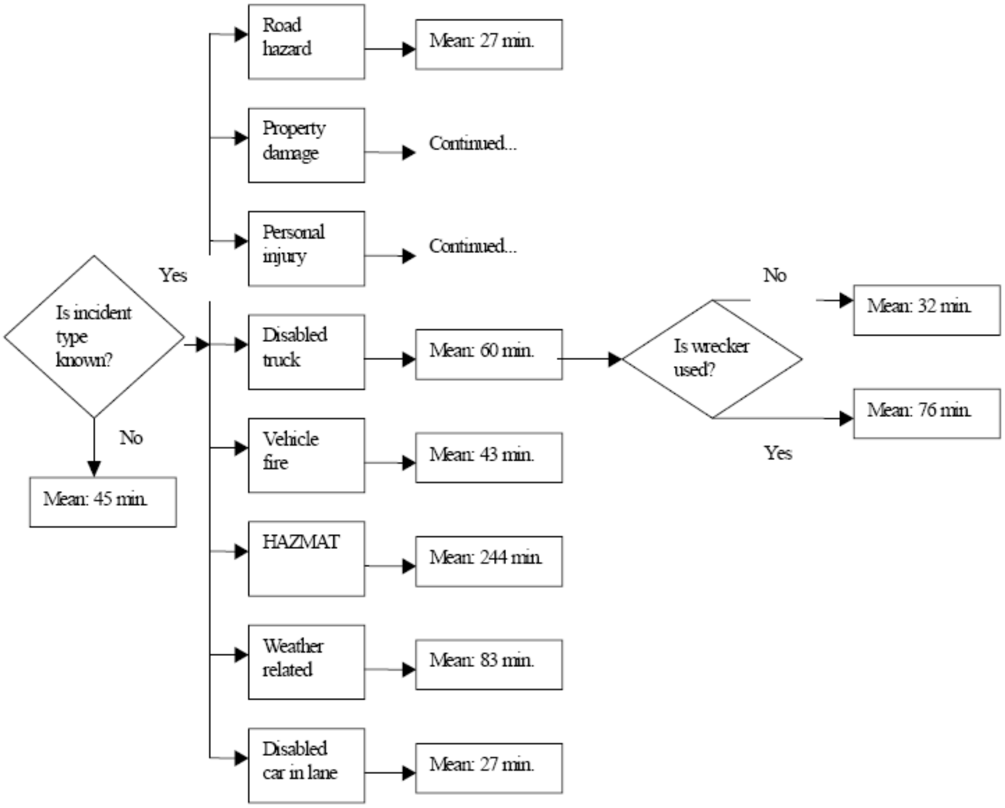


Figure 2.1 A Part of the Complete Decision Tree to Predict Clearance Time by Ozbay and Kachroo (1999)

Ozbay and Kachroo were satisfied with the new tree, based on the test results since about 57.14 % (44 out of 77) of tested incidents were predicted within 10 minutes of prediction error. They also found that the large differences between predicted and actual clearance time were caused by numerous outliers.

Smith and Smith (2001) who were inspired by the study of Ozbay and Kachroo 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.

The last 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 for predicting 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.

Grounded on the work by Lin et al. with an enriched dataset, this study has explored the integrated application of a set of new models, including a Rule-Based Tree Model, Discrete Choice Model and Multiple Regression Model. The proposed methodology will be discussed in more details in Chapter 4 and 5.

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 ANOVA test, Tukey test, and Multiple Correspondence Analyses. 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, Maryland State Highway Administration (MSHA) has developed an incident management database called CHART (Coordinated Highways Action Response Team), since 1996. CHART has collected major and minor incidents 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 year 2003 to year 2005 for model development, and year 2006 for the model validation. The data set from CHART-II for this research includes;

- Incident duration: detected, responded, and cleared timestamps;
- Incident characteristics: number of shoulder lane blockage, 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, debris, disabled vehicle, vehicle fire, police activities, off road activities, and emergency roadwork;
- Response team information: participation of MDSHA 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 hours in time when an incident was detected;
- Location information: county, road name, and exit no 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 includes a missing value for any information was excluded 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 response time and clearance time since the detection time is not available. In addition, records with duration below 5 minutes were excluded, since those seem unreasonable. After cleaning up the raw database, 6765 records are 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, it was found that incident duration follows several different but similar shapes of distributions. Golob et al. (1987) discovered that the total incident duration fits in the log normal distribution by using trucks involved accident data, while the incident duration can be illustrated by the log-logistic distribution according to Jones et al. (1991). The finding of Golob et al. has been supported by several researchers in the subsequent years (Giuliano, 1989, Garib et al, 1997, and Sullivan, 1997). Ozbay and Kachroo (1999) found out that the duration of incidents with similar type and severity shows a normal distribution, while Nam and Mannering (2000) suggested a Weibull distribution for incident duration. 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 the short duration as shown in the Figure 3.1 below (Smith and Smith, 2002).

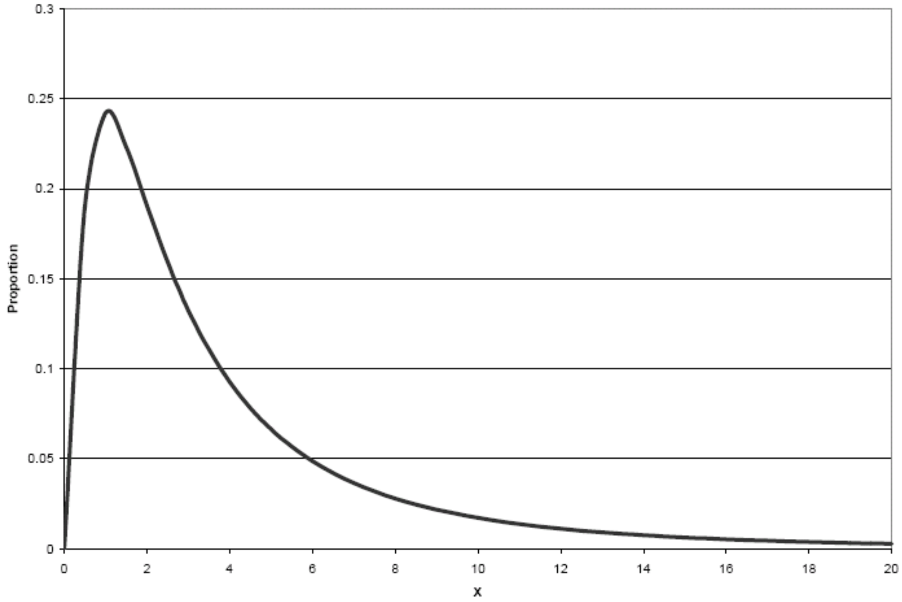


Figure 3.1 General Shape of Log Normal Distribution (Smith and Smith, 2002)

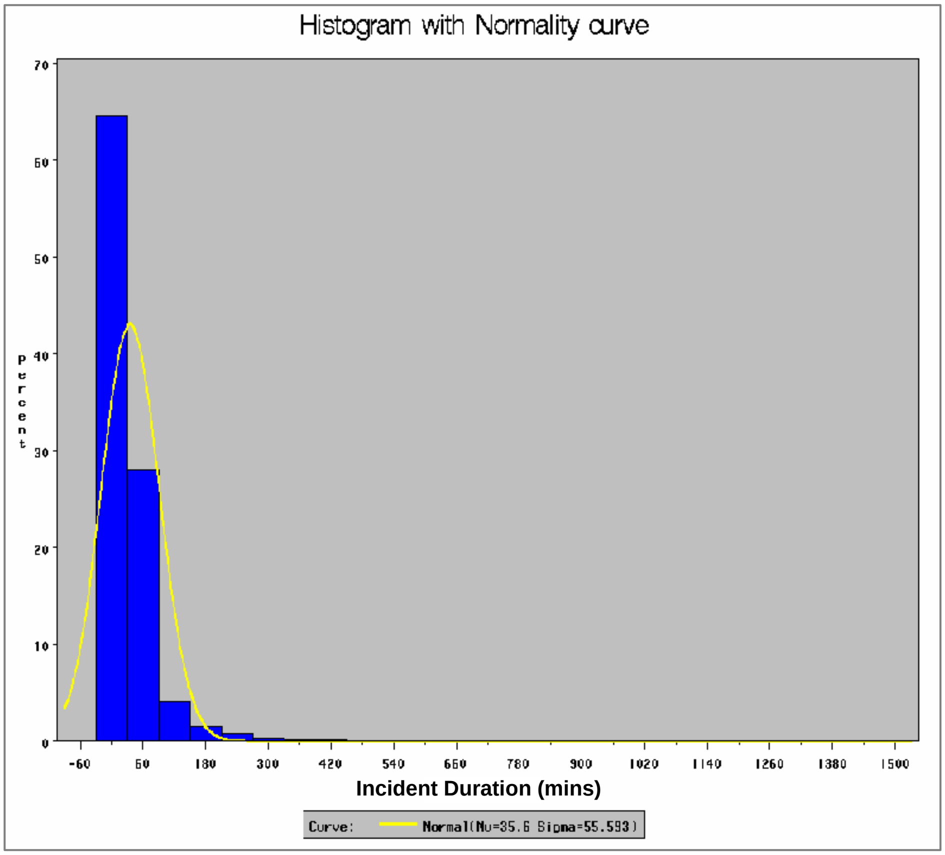


Figure 3.2 Histogram with a Normality Curve of the Incident Duration Used in This Study

To understand the distribution of incident duration, the entire available data set (including data with incident duration less than 5 minutes) is plotted in the histogram shown in Figure 3.2. It is clear that the available incident duration forms a similar shape of distribution as 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 seems closer to a log-normal distribution, but not for the Weibull distribution based on the resulting plots. However, the hypothesis tests such as Kolmogorov-Smirnov test, Anderson-Darling test and Chi-square test for distributions of log-normal, log-logistic, Weibull and so on all reject its normality at 0.01 and 0.05 significance levels.

Since the following statistical tests are performed under the assumption of normality of data set, it is essential to transform the original data for fitting 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 fit in a normal distribution as shown in the descriptive statistics (e.g. histograms, Q-Q plots or P-P plots). In a histogram, the overall shape of the distribution of transformed data set becomes nearly symmetry (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 by the diagonal dashed line (see Figure A1.3 and A1.4 in 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. The results of basic statistical measures and hypothesis tests by SAS are attached in Appendix 1.

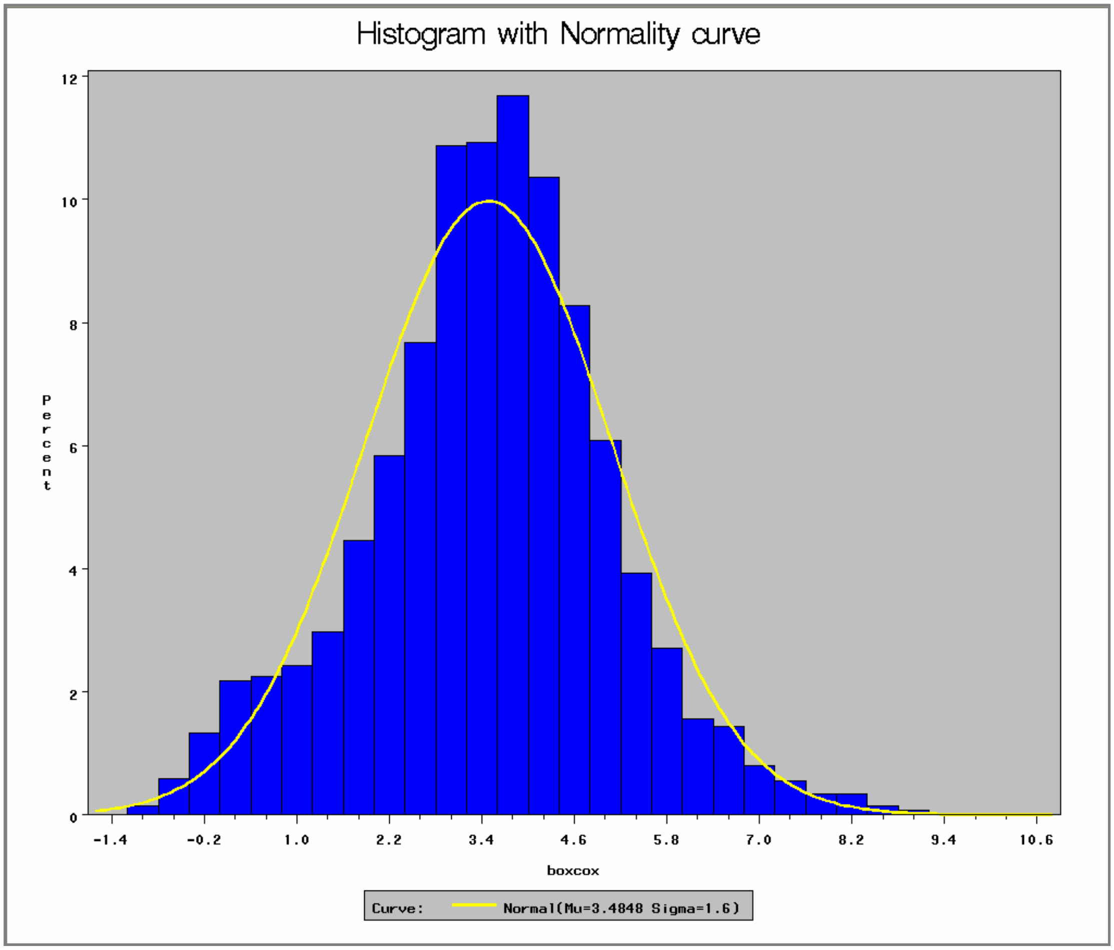


Figure 3.3 Histogram of the Box-Cox Power Transformed Data Set

The same procedure is performed with the data set which excludes incident duration less than 5 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, it is discovered that the statistics of tests become much smaller than those for the data including incident duration less than 5 minutes (see Table 3.1). This means that the truncated data set fits better 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 incident duration is continuous in nature, it is practically more useful to predict the duration by interval such as between 20~30 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 duration is categorized based on the cumulated percentage of the available samples. A category is defined by the range that covers approximately 15% of total samples, while the records with duration 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 Classification and Regression Tree (CART). Since CART is used to build a preliminary model in this research, a three-category variable similar to the one by Smith and Smith (2002) is also considered as one of options for classifying the dependent variable set. In addition, a four-category variable (short, middle, long, and very long) is explored as well. For the more detailed analysis with the primary model, the Rule-Based Tree Model, incident duration is also categorized for every 5 minutes up to 120 minutes. As in the first categorization, records with duration 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

Different from the previous study by Lin et al. (2004), this study specifies independent variables as a discrete variable, such as 0, 1, 2 and 3, 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. These independent variables are summarized along with other variables in Table 3.2.

Statistical Tests for Independent Variables

In this study, 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 multi-categorical 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, Multiple Correspondence Analysis (MCA) is implemented to determine a set of most significant variables which can explain most parts of the entire dataset.

1. ANOVA Test

ANOVA tests are performed to test if any of the descriptive variables has significant effects on the incident duration. Each of the descriptive variables is 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 ANOVA test for this variable is 0.094 so that the null hypothesis, the mean of incident durations involved with pick up vans is equal to the one not 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 Truck Involved	0, 1, 2, 3, 4	0, 1, >=2
No of Pick-Up Van Involved	0, 1, 2, 3, 4, 5, 6, 8	(0 or 1), >=2
No of Tractor-Trailer 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 Blockage	0, 1, 2, 3, 4, 5, 6, 7	0, 1, 2, >=3
No of Opposite Direction Lane Blockage	0, 1, 2, 3, 4, 5	0, 1, >=2
No of Shoulder Blockage	0, 1, 2, 3, 4	0, 1, >=2
Shoulder Blockage Indicator	0, 1	N/A
Total Lane Blockage	0, 1, 2, 3, 4, 5, 6, 7, 8, 12	0, 1, 2, >=3
Ratio of Same Direction Lane Blockage	0.00 ~ 1.00	N/A
Ratio of Opposit Direction Lane Blockage	0.00 ~ 1.00	N/A
Ratio of Total Direction Lane Blockage	0.00 ~ 1.00	N/A
No of Lane (One Direction)	2, 4, 8	N/A
Hour Incident Occurred	1, 2, 3, …., 23, 24	Day : 6 ~ 20 Night : Else
Response Time (minute)	> 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 applies 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 with sorting the means of groups in the 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 sample 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 $\mathrm{M_S}$ 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 due to its simplicity to understand and compute. The adjusted error rate by Bonferroni inequality is α/c, where c is the number of comparisons. The regrouped independent variables using Tukey test with Bonferroni inequality adjustment on α are summarized in Table 3.2 along with the original categories.

Initially, the incident nature was categorized into 7 classes. Tukey test shows 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*) can be grouped as one large. The number of data having single unit trucks and tractor-trailers is recategorized into three groups (*0*, *1*, and *>=2*), whereas the number of pick up vans is recategorized into two groups (*0* or *1*, and *>=2*).

3. Variables Selection Using Multiple Correspondence Analyses (MCA)

The Correspondence Analysis originally was developed by Jean-Paul Benzécri in France in the early 1970's (Benzécri, 1973). It has the same function as the factor analysis but 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 from other countries with different names, such as optimal scaling, quantification method, or homogeneity analysis (Hill and Lewicki, 2005). As the first step to perform the Correspondence Analysis, one needs to compute the relative frequencies for the frequency table of two variables, such that the sum of all entries of the frequency table equals 1.0. The row or column totals in the relative frequency table is referred to as the row *mass* or column *mass*, respectively (Greenacre, 1984). In the table that rows and columns are completely independent, the entries of the rows and columns can be recreated by the totals of rows and columns, which is referred as row and column *profiles* in the Correspondence Analysis (Hill and Lewicki, 2005).

Under the condition that rows and columns of the frequency table are completely independent to 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, the 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 frequency in the CA (Greenacre, 1984), by expressing a small number of dimensions that represent the deviations from the expected values.

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

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

Since this study includes more than two categorical predictors, the Multiple Correspondence Analysis is performed to find the most significant independent variables that can explain deviations from the expected values. Thus, regrouped variables are input to MCA, and 32 dimensions, which contain all information in the input table, are 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 in that dimension and dominates that dimension (Jolliffe, 1972 and 1973). Table A1.1 in Appendix 1 summarizes the largest coefficients value and the corresponding variables for these 32 dimensions. As shown in the table, the most significant factor in the first and second dimensions, which is also the most significant factor for the entire study, is the number of blocked lanes for the opposite direction that is greater than or equal to two. This result reflects that the incidents involving more than one lane blockage in the opposite direction are 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 of the variables repeatedly appear in different dimensions. The categories which make the variable significant in MCA are indicated in the 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 is computed to investigate its relationships with explanatory variables. Tables 3.3(a)-3.3(c) summarize the statistical results of incident duration under different classifications. As shown in Table 3.3(a), the incident duration increases 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. The incident durations on weekends and at night are generally longer than the durations on weekdays and in the daytime due to the longer response and clearance times.

It is noticeable that incidents occurred in the four major freeways, I-495, I-95, I-695, and I-270, have relatively shorter duration than others. It can be explained by the location of operations centers which determine the accessibility of the response units. In Maryland, there are 6 operations centers – one statewide operations center, and 5 traffic operations centers. Among them, 5 operations centers are located near those four major roads, because they are primary roads around the two metropolitan areas – Washington D.C. and Baltimore area – in Maryland.

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

Incidents with property damage also show a similar shape of distribution, and 90.2% of such incidents take between 5 minutes to 70 minutes. However, unlike the incidents with disabled vehicles, they are quite evenly distributed up to 30 minutes. Incidents causing personal injuries and fatalities are more likely to have longer duration. For example, 94.2 percent of incidents resulting in fatalities last over 70 minutes, and 78.6 percent of them last over 120 minutes. Note that 80.8 percent of incidents causing personal injuries result in the duration longer than 20 minutes, while 60.9 percent of the entire personal injury incidents lie between 20 minutes and 70 minutes. In the category of incidents classified as *Others*, its incident durations distribute quite evenly across all intervals. These results are consistent with the observations that the distribution of incident durations varies with its 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 Peakhour	45.3	4058
Peakhour	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

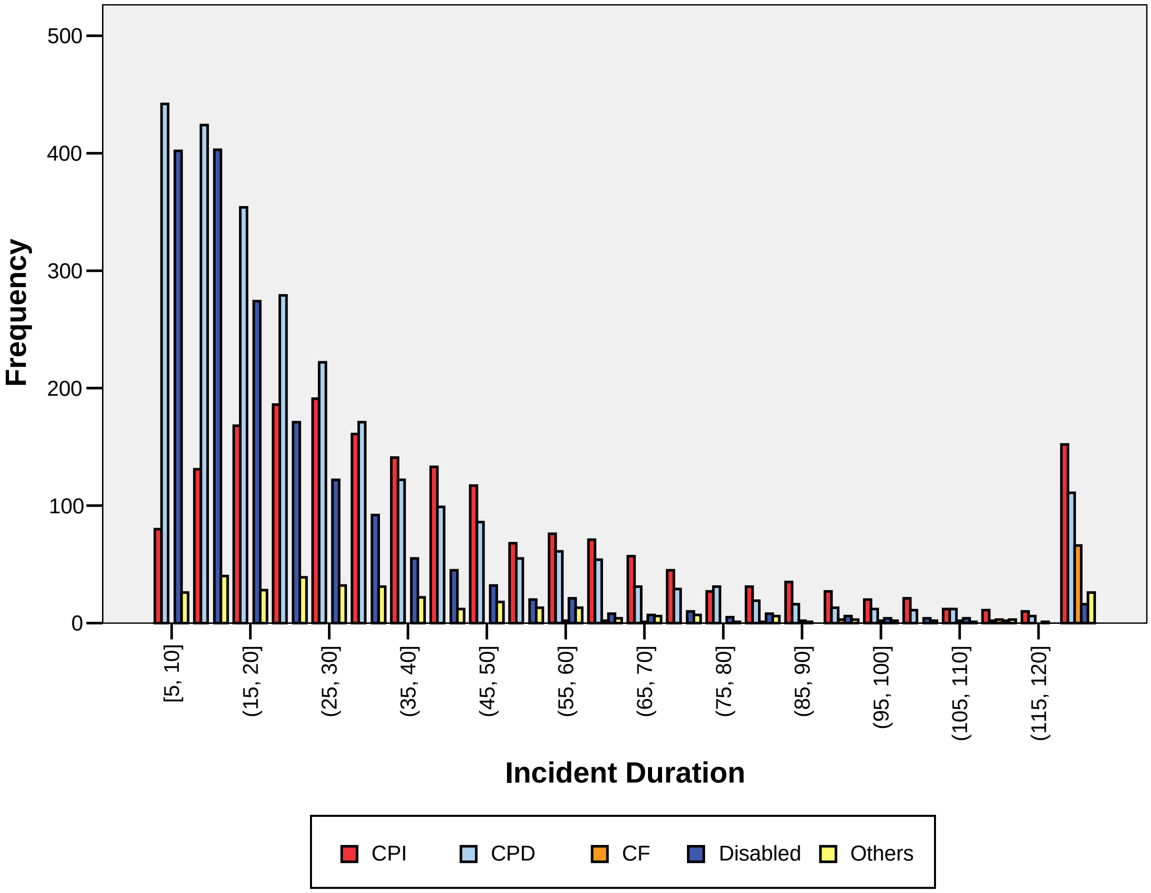


Figure 3.4 Distribution of Incident Duration Frequency by Each Incident Nature

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 discussion of the preliminary analyses with Classification and Regression Tree (CART). Based on the findings from CART, this study has further developed a Rule-Based Tree Model in Section 4.3 along with its calibration procedures. Presentation of the entire model structures, its performances, and validations are illustrated in Section 4.4 to Section 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

Classification and Regression Tree, known as CART or C&RT as well, 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 that was developed based on 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 variable.

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

In the tree building step, first, the root node, including all data set, is 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 cost of classifying *i* as *j*,

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

N_i is number of category *i* in dataset,

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

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

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

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

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

(3) Reaches the maximum tree level that is externally set by users.

Usually, a tree created by aforementioned procedures is likely to be over fit. That may result in difficulties for uses to read and interpret and, so the process of tree pruning is recommended. To prune the over-fit tree, the method of "cost-complexity" is used in general. In this method, the complexity parameter, α, is gradually increasing 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 is increasing, the tree is getting 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 dataset is well fit but not overfit (Lewis, 2000). A detailed discussion regarding CART is available in the literature (Breiman et al, 1984, Lewis, 2000, Yohannes and Hoddinott, 1999, and Lemon et al., 2003).

4.2.2 Results and Findings from CART

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

1. Among 25 independent variables, the incident nature 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* are relatively short since about 53% of these incidents exhibit the duration between 5 minutes and 20 minutes. On the other hand, incident durations for Collision-Personal Injury, Fatality, and Others are more likely to be longer because about 59% of these incidents distribute between 20 minutes and 70 minutes. These relations are consistent with the frequency distribution of incident duration (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. Among 25 independent variables, the incident nature 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* are relatively short since about 53% of these incidents exhibit the duration between 5 minutes and 20 minutes. On the other hand, incident durations for Collision-Personal Injury, Fatality, and Others are more likely to be longer because about 59% of these incidents distribute between 20 minutes and 70 minutes. These relations are consistent with the frequency distribution of incident duration (see Figure 3.4 in Chapter 3).

3. 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).

4. Based on the experimental results, the difference of tree performance between using the original independent variables and regrouped independent variables is trivial. Also, the CART algorithm itself has an ability to choose the most significant variable as the best splitter, and it can also find the best regrouped categories within the selected variable.
5. Tables 4.2 to 4.4 summarize the prediction result for each tree. Table 4.2 shows that the overall percentage of the correct prediction with the tree developed for 9-categorized (*Basic*) dependent variable, called Tree 1, is 30.2 %. About 71% of the incidents that have the duration between 5 and 10 minutes have been predicted correctly. But the tree at this level could not predict correctly for incidents having durations for 70~90 minutes and 90~120 minutes. Trees developed for 3-categorized (*RCDV1*) dependent variable (Tree 2) and for 4-categorized (*RCDV2*) dependent variable (Tree 3) reflect the similar trend, but achieve a better level of performance, where the overall percents of correct prediction are 63.5 % and 63.1% for Tree 2 and Tree 3, respectively. Both trees, however, are not sufficient for use in predicting incident duration exceeding 70 minutes. For example, Tree 2 predicted 22.8 % correctly for incidents lasting longer than 70 minutes. In Tree 3, incident duration for 70~120 minutes was not predicted correctly at all, whereas 31.1% of incidents lasting for more than 2 hours was predicted correctly.

Overall, CART performs quite well for short or middle ranges of incident duration, especially, for these between 5 to 10 minutes. However, it does not provide satisfactory results for incidents of long duration (e.g., longer than 1 hour). Similar results are also found from the research implemented by Smith and Smith (2001), although their tree is developed to forecast the clearance time. The overall prediction accuracy of their classification tree was 58.47%, and they concluded that this accuracy level is 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 duration. Based on this finding along with other analysis results from CART discussed previously, this study has redesigned a classification tree, named a Rule-Based Tree Model (RBTM), using the following procedures. Note that incident duration, which was grouped into 5-minute intervals, is 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 to develop a separate model. Thus, the regrouped incident nature was considered as the more appropriate splitter than the original one.

Step 2: Determine the next splitter for each node.

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

Step 3: Split the node based on the determined splitter in each category.

The focus of this step is to convert each splitting node in *If-then; Else-then* statement, which will constitute the set of rules for determining the incident duration for the node.

Step 4: Assign the predicted incident duration range for each split node.

This is to determine the best representative range of incident duration for each node. To achieve this, one shall first search the interval less than or equal to 30 minutes which covers at least 70% of all cases within a node. If no such interval exists within the node, then one can assign the shortest interval covering at least 60% of all cases within the node as the predicted incident duration for that node.

Step 5: Repeat Step 2 to Step 4 for all nodes until satisfy the predetermined criteria for stopping the tree growth.

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. There is only one observation left in a node.

To evaluate the performance of rules for each node, this study adopts 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 that satisfies 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 c***onfidence*** 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 dataset by *County*, one can repeat the same procedures to complete the Rule-Based Tree Model for each *County* of each incident nature. Due to the constraints of samples, the study has analyzed only the data from the Montgomery County. Figure 4.1 shows the structure of the Rule-Based Tree Model.

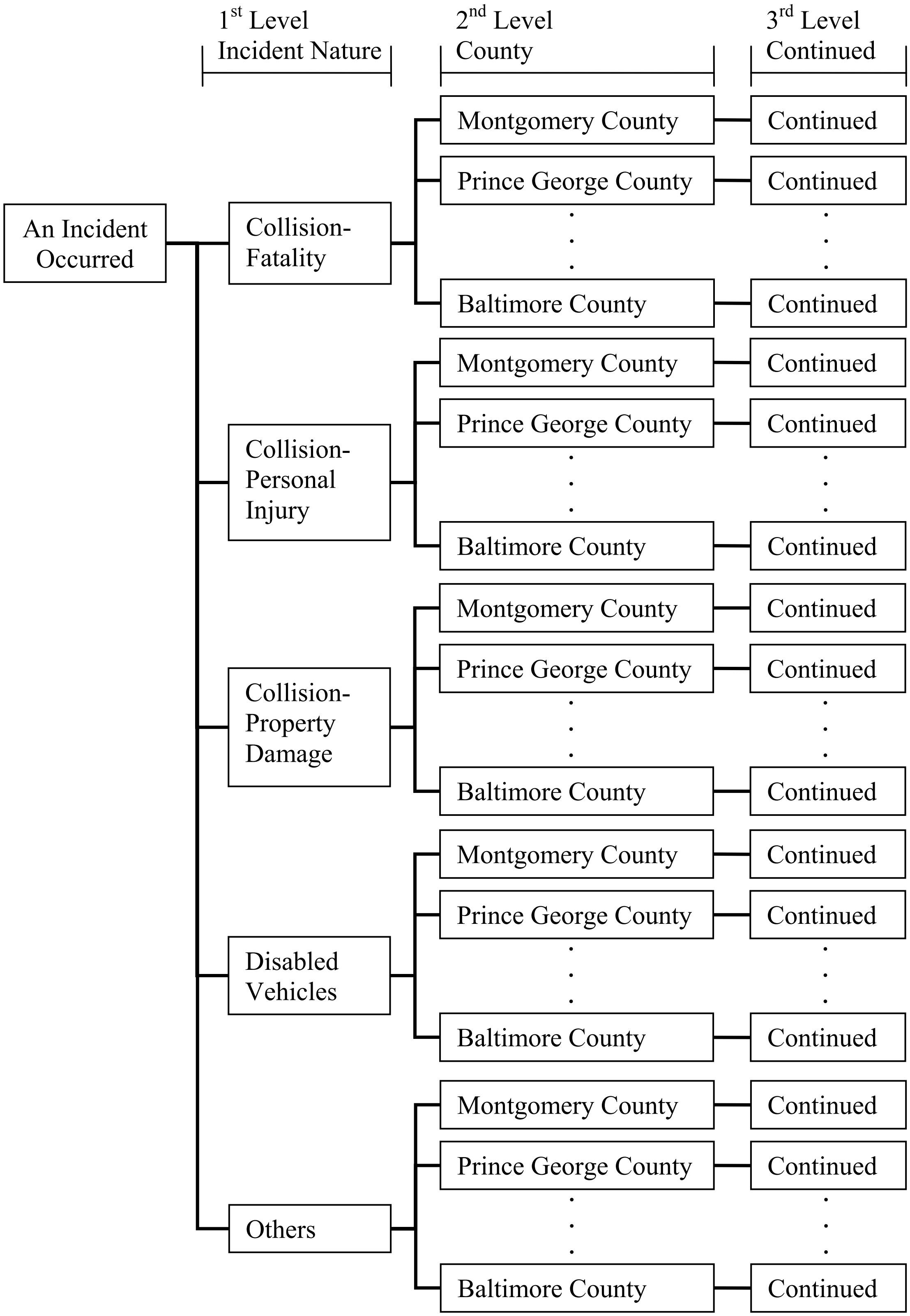


Figure 4.1 The Structure of Rule-Based Tree Models

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

4.4.1 The Tree Structure

For these incidents resulting in *Collision-Fatality*, their distributions over 300 minutes are scattered over a wide duration range (300 ~ 1500 minute), while the distribution in the range of 60~300 minutes is condensed and nearly symmetric (see Figure 4.2). Most of those cases last over 300 minutes occurred on roads which are out of our scopes, and about 78% of those cases show the *Ratio of Blocked Lanes in the Same Direction* is greater than or equal to 0.5. This means that those incidents resulted in an extreme level of severity. In addition, about 68% of these occurred between midnight and 6 AM. One extreme case was involved with 73 vehicles, including 5 tractor-trailers, and it resulted in the longest duration of 1501 minutes. Since these cases require special response and operational efforts, this study has excluded them from the model development.

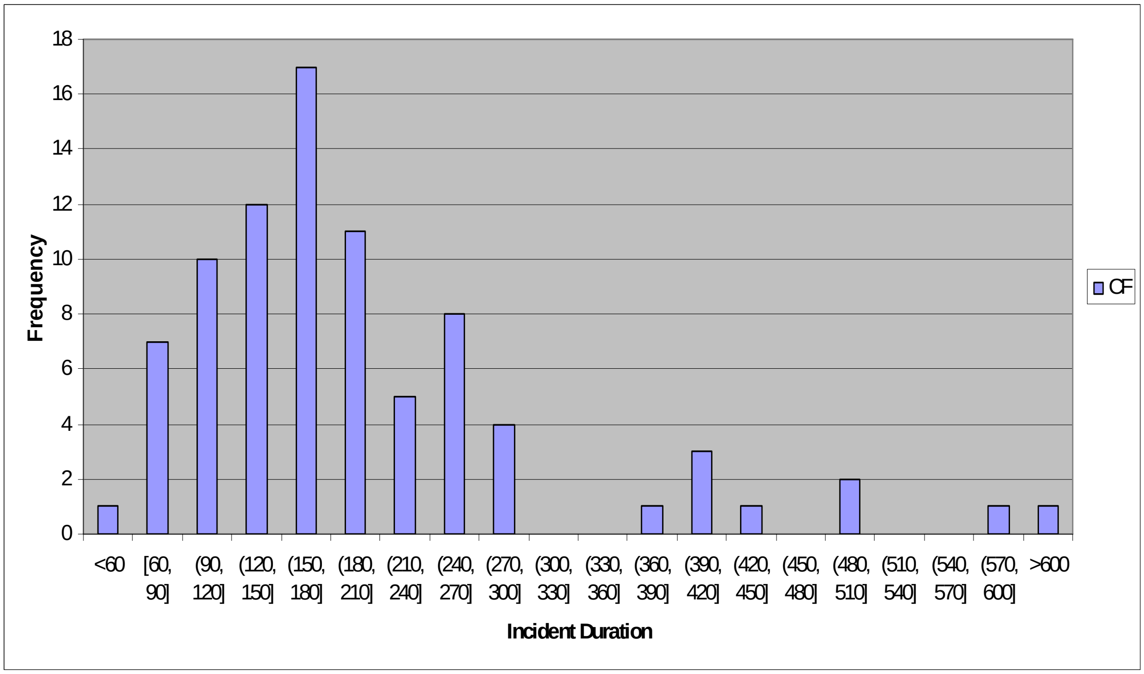


Figure 4.2 Distribution of Frequencies for Incidents with Collision-Fatality (CF)

The Rule-Based Tree Model 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. In all depictions 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, it is observed that heavy vehicles such as pickup vans, single unit trucks and tractor-trailers show a strong effect on the resulting duration 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** occurred 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 duration of incidents that incur fatalities. When a shoulder lane is blocked, the incident duration is more likely to be shorter than that 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)). It 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.

4th Level

Rule 4-a: **If** occurred 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

5th Level

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** occurred 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

It is noticeable at this level that there exists a relation between *Road* (i.e., the highway segment) and incident duration.

<u>6th Level</u>

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** occurred in the *Daytime*, **then** 220<*IncD* <=240; **Else** 280<*IncD* <=300

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

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

<u>7th Level</u>

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

Rule 7-b: **If** occurred 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 increases, the incident duration generally decreases. It can be explained by the fact that more blocked lanes during the operations may provide a wider working space for incident response units to efficiently clear an incident.

8th Level

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 shown at this level is about the pavement conditions. In general, the wet pavement condition reflects an inclement weather, which tends to increase the number of incidents and the incident duration. However, in the study dataset for *CF*, the relationship between wet pavement and incident duration was opposite to what is expected. This dataset shows that incident durations in the wet pavement condition are 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 Duration for Different Pavement Conditions. It can be explained by the fact that in inclement weather incident response units are on alert and more staffs are available for emergency medical services (EMS).

Table 4.5 Average Fatality Incident Duration 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 Rule-Based Tree Models using the dataset collected from year 2003 to 2005.

While most samples for other incident natures are distributed within 2 hours (i.e., 5 ~ 120 minutes), samples for *CF* are scattered between 60 and 300 minutes. In addition, the sample size is very small (i.e., 84), although they have been collected for three years. Thus, the ranges of incident durations assigned at many of terminal nodes (highlighted cells) in Rule-Based Tree Models (*Then* statement in rules) are more 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 fall in a relatively wide range, the ***confidences*** for most of rules are acceptable.

Table 4.6 Summary of Estimation Results for the Rule-Based Tree Model for Collision-Fatality Incidents Occurred 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 Rule-Based Tree Model for Collision-Fatality Incidents Occurred 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 Rule-Based Tree Model.

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

However, the overall validation results shown in Tables 4.7 and 4.7(a) using a dataset collected in year 2006 (sample size is 70) indicate that only two nodes show the ***confidence*** over 70%. Many validation results of the nodes that appear close to terminal nodes show a low ***confidence***. Some of terminal nodes (highlighted cells) are not even able to be validated, since in the validation dataset there are no records satisfying *If* conditions given in those nodes. Models for *Collision-Fatality* show unsatisfactory performance even with the dataset 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 Rule-Based Tree Model for Collision-Fatality Incidents Occurred 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 Rule-Based Tree Model for Collision-Fatality Incidents Occurred 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 Rule-Based Tree Model.

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

4.5 The Rule-Based Tree Model for Incident Nature of Collision - Personal Injury (CPI)

4.5.1 The Tree Structure

The following rules construct the Rule-Based Tree Model 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 Unite 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, heavy vehicles (e.g., single unit trucks, pickup vans, and tractor-trailers) involvement shows a strong relation to determine incident duration.

4th Level

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 truck* =1, **then** Rue 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** occurred in the *Daytime* 15< *IncD* <=40

At this level, it is observed that the duration of incidents occurred on I-495 shows a different range of incident duration from those on other roads.

5th Level

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** *Pick up van* is not involved, **then** Rule 6-e; **Else** 25 <*IncD*<=50

Ruel 5-d: **If** *Pick up 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

In overall, this level selects the information regarding pickup van involved 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*= 270 N, **then** 40 <*IncD*<=65

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

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

Rule 6-e: **If** occurred 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** occurred 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** occurred 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** 2hrs <*IncD*<=3.5hrs

It is observed at this level that as the number of vehicles involved with an incident increases, the incident duration is likely to increase.

8th Level

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** occurred during *Off Peak hours,* **then** 40 <*IncD*<=65

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

Rule 8-f: **If** occurred 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 occurred in the dry pavement condition are likely to be shorter than those in other conditions as expected. Also, noticeable is that the effect of *Off Peak Hours* on incident duration is different with the subsets. For example, with the Rule 8-e, incidents occurred during off peak hours result in a shorter duration, while with the Rule 8-f, they results in a longer duration.

9th Level

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 becomes a significant factor in determining the incident duration.

10th Level

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** occurred during *Off Peak hours*, **then** 5<=*IncD*<=30; **Else** 10<*IncD*<=35

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

It has been observed at this level that the *Peak Hour* factor shows a different degree of influence in different subsets. With Rule 10-c, the duration of incidents occurred during peak hours is likely to be longer than that during off peak hours, and vice versa with Rule 10-d.

To complete the Rule-Based Tree Model 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 duration for 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 dataset.

However, with Rules 3-d and 4-d, the predicted range of incident duration is over 30 minutes with unsatisfactory ***confidences***, which are lower than 70%. Therefore, a supplement model is needed. Due to the limit of sample size, the supplemental model will be developed with the sub-dataset that was used for developing the 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-dataset, including all these cases (i.e., a subset for the Rule 5-a) will be used to develop a separate supplemental model.

Lastly, another supplemental model will be 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 Rule-Based Tree Model for Collision-Personal Injury Incidents Occurred 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 Rule-Based Tree Model for Collision-Personal Injury Incidents Occurred 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 Rule-Based Tree Model.

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

Table 4.9 Summary of Validation Results for the Rule-Based Tree Model for Collision-Personal Injury Incidents Occurred 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 Rule-Based Tree Model for Collision-Personal Injury Incidents Occurred 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 Rule-Based Tree Model.

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

4.6 The Rule-Based Tree Model for Incident Nature of Collision - Property Damage (CPD)

4.6.1 The Tree Structure

The rules, constituting the Rule-Based Tree Model 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

The 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 to determine the resulting incident duration.

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** occurred during *Off Peak Hours*, **then** Rule 6-a; **Else** 6-b

Rule 5-b: **If** occurred 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 is likely to be shorter in the dry pavement condition than the one in other pavement conditions.

6th Level

Rule 6-a: **If** *Pavement* is Dry, **then** Rule 7-a; **Else** Rle 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** occurred during *Off Peak Hours*, **then** 5 <=*IncD* <=25

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

Rule 6-j: **If** occurred 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 for pavement conditions and blocked lanes play a significant role to determine the duration of incidents resulting in property damage. It is observable that the incident duration increases as the lane-blockage ratio increases. In addition, it is found that the time that an incident occurred has 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** occurred 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 Rule 7-f and Rule 7-g, incidents occurred on I-495 are more likely to be shorter than those same types of incidents but on I-270. The same relations have also been observed in developing Rule 3-c at the 3rd level.

8th Level

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) shows 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 duration and the ***confidence***. With Rule 1, the performance with *If* condition itself demonstrates satisfactory results without any additional splitter. The predicted range of incident duration is less than 30 minutes, and the probability (***confidence***) is greater than 0.7 (70%). 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 shows the unsatisfactory performance results. Even with additional splitters, the performance for this sub-dataset is 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 Rule-Based Tree Model can not yield definitive results. This suggests the need to calibrate a supplement model.

In addition, since Rule 5-c, Rule 7-b and Rule 7-d cannot perform up to the expected level, a supplement model is also needed. 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 demonstrate satisfactory results in the validation dataset, while many rules at lower levels do not perform as expected due either to the need of additional factors or the lack of sufficient sample data.

Table 4.10 Summary of Estimation Results for the Rule-Based Tree Model for Collision-Property Damage Incidents Occurred 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 Rule-Based Tree Model for Collision-Property Damage Incidents Occurred 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 Rule-Based Tree Model.

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

Table 4.11 Summary of Validation Results for the Rule-Based Tree Model for Collision-Property Damage Incidents Occurred 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 Rule-Based Tree Model for Collision-Property Damage Incidents Occurred 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 Rule-Based Tree Model.

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

4.7 The Rule-Based Tree Model for Incident Nature of Disabled Vehicles (DV)

4.7.1 The Tree Structure

The following rules are used to construct the Rule-Based Tree Model for incidents with disabled vehicles occurred in Montgomery County.

1st Level

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

2nd Level

Rule 2-a: **If** occurred 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 occurred during peak hours are more likely to be cleared in a shorter duration than those during off-peak hours. Also, when any shoulder lane is blocked at peak hours due to a disabled vehicle, its average duration is slightly shorter than that for incidents without a shoulder blockage.

When a disabled vehicle is a pickup van, the estimated range for incident duration is 5 ~ 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 of incidents due to *Disabled Vehicle* (83.3% for Montgomery County only) lie in a relatively short range of 5~30 minutes, one can use a simple rule to predict their resulting duration. It is also found that even after applying a series of additional splitters to subdivide the dataset, 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 single vehicle.

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

Table 4.12 Summary of Estimation Results for the Rule-Based Tree Model for Disabled Vehicles Incidents Occurred 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 Rule-Based Tree Model.

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

Table 4.13 Summary of Validation Results for the Rule-Based Tree Model for Disabled Vehicles Incidents Occurred 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 Rule-Based Tree Model.
[1] Conf. stands for the ***confidence***.

4.8 The Rule-Based Tree Model for Incident Nature of Others

4.8.1 The Tree Structure

The rules used to construct the Rule-Based Tree Model 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** occurred 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 result in longer incident duration 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 rules at this level are used to collectively capture the fact that the number of lanes being blocked during the response operation is positively correlated with the resulting incident duration. So is the relation between the incident duration and the heavy vehicles or trucks involved.

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** occurred 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 detection time is one of the critical factors that contribute to the resulting incident duration. In general, the duration of incidents occurred in the daytime is 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 incurred at some locations may take a longer duration than those of the some types but at other locations. For instance, Rule 6-c indicates that the incidents occurred at Exit 1 on I-270 are 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 is relatively small, it was difficult to develop a reliable Rule-Based Tree Model. It was even more challenging to validate this model, because the validation dataset has only 18 records of such incidents. As a result, more than 50% of rules were unable to 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 is thus developed and presented in the next chapter.

Table 4.14 Summary of Estimation Results for the Rule-Based Tree Model for Incident Nature – Others Occurred 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 Rule-Based Tree Model.

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

Table 4.15 Summary of Validation Results for the Rule-Based Tree Model for Incident Nature – Others Occurred 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 Rule-Based Tree Model.

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

4.9 Overall Findings and Conclusions

This section summarizes the following overall findings with the Rule-Based Tree Models.

1. For the categories of *Collision-Personal Injury*, *Collision-Property Damage*, *Disabled Vehicle* and *Others*, it turned out that the spatial factor, *County*, has emerged as the second splitter. It implies that the duration for the same type of incidents varies 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. Rule-Based Tree Models are more flexible for assigning an appropriate estimated incident duration range in given conditions (sub-dataset or node) than Classification and Regression Tree Models (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 minutes to 120 minutes), so as to determine the most appropriate range of incident duration for a selected subset.
4. As expected, heavy vehicles involvement tends to increase the incident duration due to its complexity to manage or the need of special equipment for clearance operations (e.g., wrecker).
5. Incidents occurring at night time or during off-peak hours generally take a longer duration than those in daytime, due to the lack of sufficient response units for incident clearance operations.

6. When incidents resulting in *Collision-Fatality,* or *Property Damage*, the clearance operation is generally more efficient in the shoulder-lane blocked scenarios than 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 do the work.
7. Similarly, during the *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 is not definitive for the existing data records. It shows a positive correlation with the incident duration 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 impacts on the resulting incident duration are not statistically significant.

Due to the complex nature of incidents and response operations, one shall not expect the above Rule-Based Tree Model to capture all embedded relations and provide the operationally acceptable performance for real-world applications. Hence, grounded on the promising information generated from the Rule-Based Tree Model, this study has furthered developed some supplemental models for improving the prediction accuracy for the duration of a detected incident. Depending on the available size of sample data, this study has employed either the Multinomial Logit Model or the Regression Model in the development of supplemental components.

Lastly, Rule-Based Tree Models illustrated in a tree shape are included in the Appendix 2.

Chapter 5: Supplemental Models

5.1 Introduction

This chapter presents two supplemental models for improving the prediction accuracy of incident duration. The first is the Multinomial Logit Model (MNL) that is used for the sub-datasets with unsatisfactory results from the main model (Rule-Based Tree Models) for incidents causing *Collision-Personal Injury (CPI)* and *Property Damage (CPD)*. This model is proposed because samples in those subsets show a condensed distribution and have a large enough size that is comparable to the number of categories in a dependent variable. The second model is the regression model that is used for datasets from incident natures of *Collision-Fatality* and *Others*, since those datasets show a scattered distribution with a relatively small size.

Figures 5.1 and 5.2 illustrate the sub-datasets used for developing supplemental models for incident natures of *Collision-Personal Injury* and *Property Damage*. Incident natures of *Collision-Fatality* and *Others* use the entire dataset to develop their supplemental models. Brief descriptions of core concepts for these two methods along with estimation and validation results are presented below.

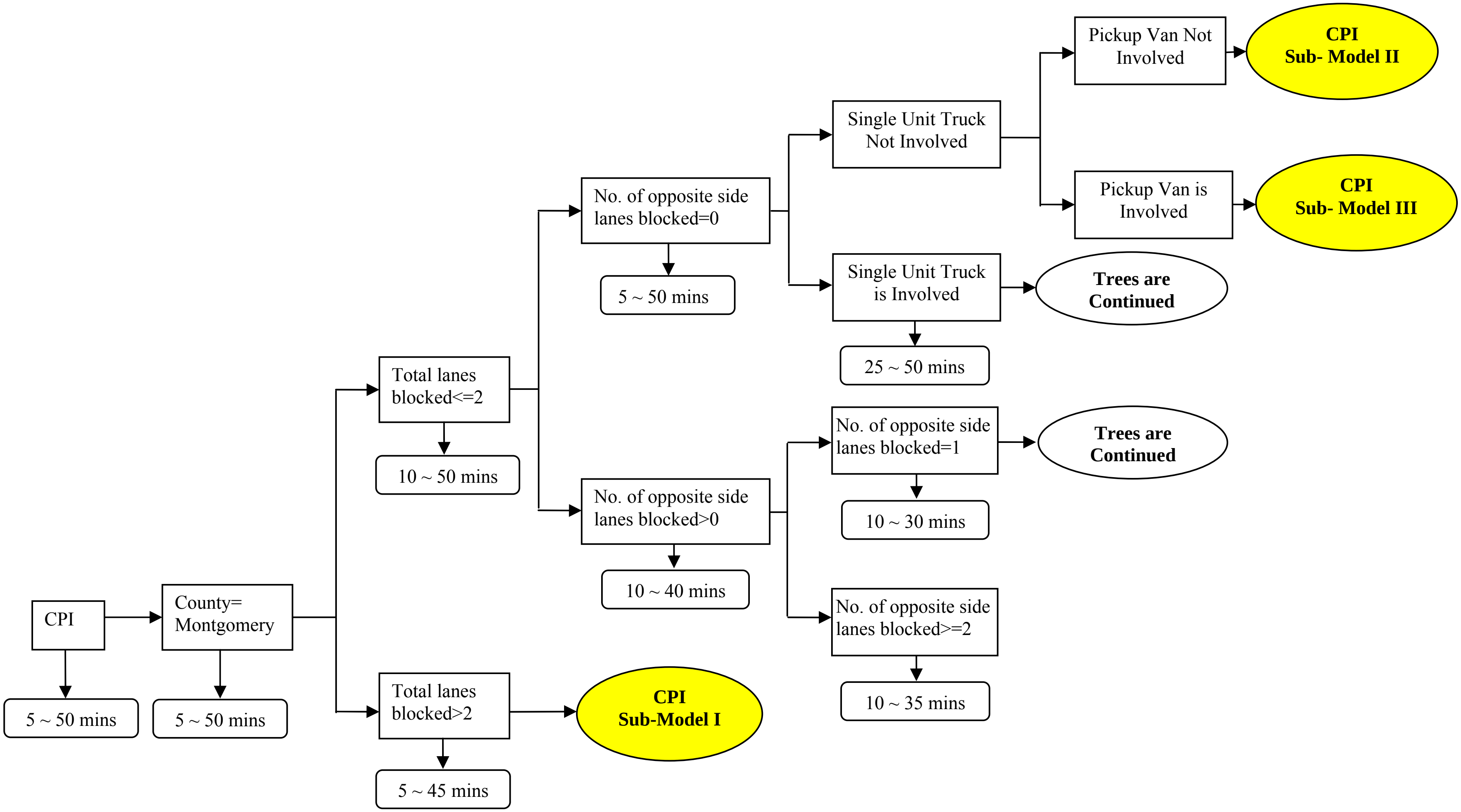


Figure 5.1 Sub-Datasets Used for Developing Supplemental Models for Incidents Causing Collision-Personal Injury

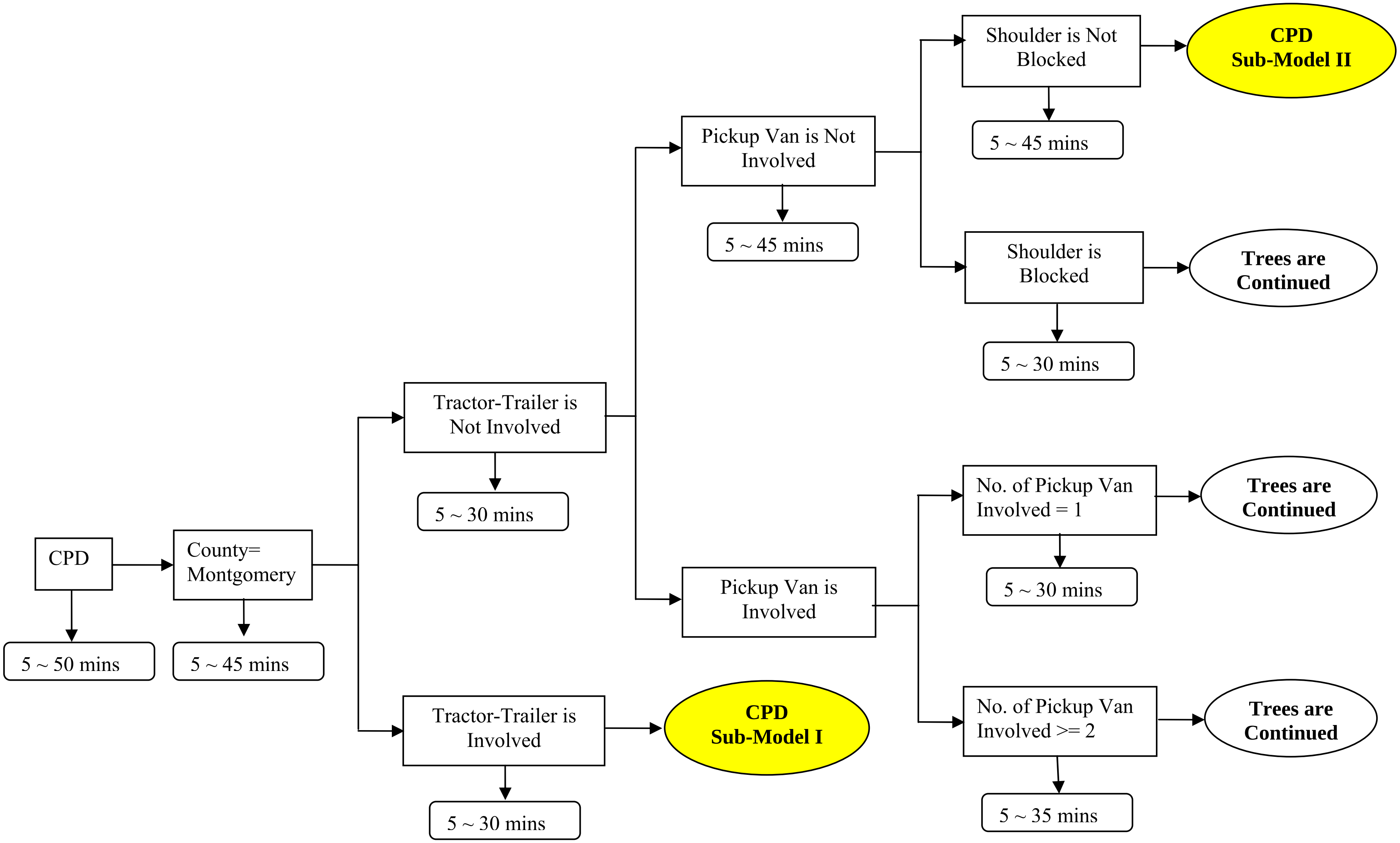


Figure 5.2 Sub-Datasets Used for Developing Supplemental Models for Incidents Causing Collision-Property Damage

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

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 for these discrete scale data in 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 a conceptual viewpoint – a behavioral choice and a 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 - a behavioral choice, while accident injury severity belongs to the latter category since it is merely explaining discrete outcomes of a physical event. Similarly, intervals of incident duration 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 a lot of differences (Washington et al., 2003). For instance, discrete choice 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 the discrete choice models for the two categories are derived based on the 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 assumed that the decision maker will choose an 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 become accident victims that got injured or need responses from specialists. Thus, in the physical phenomenon model, a choice is made to an alternative with the highest value in ***propensity*** function (Khorashadi, 2003). Nonetheless, the possible forms of two functions 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, although it is a relatively new approach in the incident duration study.

For sub-datasets in *Collision-Personal Injury* and *Collision-Property Damage* which show unsatisfactory results in the Rule-Based Tree Model, the Multinomial Logit Model is applied to estimate the relation between each category of incident duration and its associated factors. A well calibrated model will allow its users to predict the duration category of a detected incident. The core concept of MNL, same as that used in accident severity model (Khorashadi, 2003 and Ulfarsson, 2001), is briefed below:

The ***propensity*** function, R_{ni}, which represents the propensity of incident *n* towards interval *i* of incident duration 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 pre-classified incident duration (defined in an interval form), $\boldsymbol{X_{ni}}$ 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*), $\boldsymbol{\beta_i}$ 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* (incident *n*). Assuming that the disturbance terms of the ***propensity*** functions 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, $\boldsymbol{\beta_i}$ is a vector of coefficients, and $\boldsymbol{X_{ni}}$ and $\boldsymbol{X_{nj}}$ are vectors of attributes for alternative *i* and *j*. The detailed discussion regarding MNL models would 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:

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

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 the interstate road I-270; 0 otherwise.
- I495 is 1 if the incident occurred on the interstate road 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 are 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). And then, variables not included at the initial stage are included to test their significance in ***propensity*** functions. After filtering out insignificant variables, all coefficients are set as alternative specific to test if all variables are significant. If not, the insignificant variable is removed from the corresponding ***propensity*** function. Lastly, variables previously removed from the model are included one by one again with their coefficients being set as alternative specific to verify whether any significant variable is left out.

5.2.2 Estimation Results with MNL

As shown in Figure 5.1, for *Collision-Personal Injury (CPI)*, three MNL models are needed, while two MNL models are required for *Collision-Property Damage (CPD)* (see Figure 5.2) since each sub-dataset needs a different model to result in the best performance. The categories (intervals) of incident duration are defined differently for each MNL model, since the distribution of incident duration is different from one another. The following Table 5.1 summarizes the categories of incident duration for each MNL model.

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

	Sub-Model I	Sub-Model II	Sub-Model 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 sub-model includes only two categories for a dependent variable, a binary logit model is applied instead of MNL. But, the theoretical concept and background of binary logit models remain same as those of MNL.

Developed MNL models are presented in Tables 5.2(a) to 5.3(b), and the estimated and validated probabilities for incident duration 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 duration 5~25 minutes of CPI-Sub-Model I, show a significance at the 90% level (an absolute value of *t*-statistic should be above 1.65). The insignificance of alternative specific constants is irrelevant because they reflect the average effects of variables which are 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 sign and magnitude of coefficients for all variables are as expected. In previous chapters, it was found that the increase in the number of heavy vehicles (single unit trucks, pick up vans, or tractor-trailers) involved causes an increase in incident duration. This observation is reflected as the negative sign of the coefficients for variables *NoTT, NoSUT,* and *NoPUV* in short incident duration alternatives, e.g., 5~25 or 25~45 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, 5~25 and 25~45 minutes, reflects the observation that when an incident occurs at night, the incident is likely to last longer than that 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 road 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 shows different effects on each sub-model. In *Sub-Model I*, the pavement condition-*Dry* is likely to shorten the incident duration as it has a positive coefficient for the alternative of 5~25 minutes. But, in *Sub-Model III*, this variable shows a tendency to increase the incident duration for having a negative coefficient for the alternative of 25~45 minutes. Meanwhile, incidents occurring in the pavement condition-*Snow/Ice*

show an effect to increase the duration in *Sub-Model II*, and this is reflected by its negative coefficient for the incident duration alternatives of 5~25 and 25~50 minutes.

2. The interstate road I-270 shows different effects in the sub-models for *CPI* and *CPD*. In *CPI-Sub-Model II*, the variable, *I-270*, shows an effect to decrease incident duration, which is reflected by the positive and larger coefficient in the 5~25 minutes alternative than the 25~45 minutes alternative. On the other hand, *I-270* shows a negative effect on shortening incident duration in *CPD-Sub-Model I*.
3. Particular locations (exits) on I-495 and I-270 cause longer incident duration. This is reflected in several MNL models with negative coefficients of the related variables in short incident duration alternatives, e.g., 5~25 or 25~45 minutes. Exits that are commonly appeared to have this kind of observations are 27, 33, 36, 39 on I-495 and 1, 4, 9, 18 on I-270. The reason for this can be found in the complexity of geometric configuration around these exit areas or for their long distance from the traffic operation centers. When incidents occur especially in the area around exits 33, 36, 39 on I-495 and 1, 4 on I-270, the response and clearance time for the incidents will be longer due to the difficulty in access caused by complex geometric configuration and heavy traffic of those locations. I-495 is split into I-270 at exits 34 and 35, and merged with I-270 at exit 38 again. I-270 is split into two directions at exit 2 to be merged with I-495 at exits 34 and 38. Such features around this area cause heavy weaving traffic to interrupt the main stream.

4. Response time is proportional to the incident duration in *CPD-Sub-Models*, and this relation exhibits a negative coefficient for the shortest incident duration alternative in *CPD-Sub Model I* and *II*.
5. In *CPD-Sub-Model II*, *Incident Hour* which represents the hour in time that the incident occurred shows a strong relationship with incident duration. The format of *Incident Hour* is defined as numbers from 0 to 23 without AM or PM. As the value of in *Incident Hour* increases, the incident duration is likely to increase, which implies that incidents occurred in the evening are likely to last longer than those occurred in the early morning. This effect is similar to the one from *Night* factor, but more sensitive to each hour.

As shown in Table 5.4, the probabilities for the three categories of incident duration do not show large discrepancies from one another in the sub models for *CPI*. For example, for two categories (25~45 minutes and > 45 minutes) in *CPI-Sub-Model I*, the difference in probability is only about 2%. Similar phenomena can also be found in *CPI-Sub-Model II* and *III* for the first two categories of incident duration. In MNL models for *CPD*, the difference in probability between alternatives is larger, but still no alternative dominates the entire dataset (i.e., over 70% probability). For this reason, probabilistic models, such as MNL models, are required to be applied for those subsets in which it is hard to find any short range of incident duration with high probability to satisfy given conditions.

Developed MNL models are validated with year 2006 dataset. By using this dataset, predicted probabilities for each incident duration category in each model are found and summarized in Table 5.4. The difference between the estimated and validated probability is within 10%.

Table 5.2(a) CPI-Sub-Model I: Estimated Propensity Functions for the Multinomial Logit 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 Road 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-Sub-Model II: Estimated Propensity Functions for the Multinomial Logit Model

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

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

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 Road I-495; 0 otherwise
I270 : 1 if an incident occurred on Road 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-Sub-Model III: Estimated Propensity Functions for the Multinomial Logit Model

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

$$\begin{array}{llllllll}
R_{25\text{-}45} = 3.031 & -3.346*\text{NoTT} & -1.603*\text{Night} & -2.095*\ \text{PEAKHR} & -2.727*\ \text{Ex270} & -0.865*\text{Ex495} & -1.099*\text{Pave\_Dry} \\
\quad\quad (3.8) & (-3.2) & (-1.7) & (-1.9) & (-3.1) & (-1.5) & (-2.1)
\end{array}$$

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 no. 27, 28, 33, 34, 36, 38, 39)
Ex270 : Binary variable to indicate the specific locations on I-270 (exit no. 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-Sub-Model I: Estimated Propensity Functions for the Multinomial Logit 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-Sub-Model II: Estimated Propensity Functions for the Multinomial Logit Model

---

$$\begin{array}{lllllll} R_{5\text{-}25} = 6.772 & -0.169\ *\text{IncHR} & -0.782*\text{NoVehInv} & -3.078*\text{Ex495} & -3.333*\text{Ex270} & +1.228*\text{Pave\_Dry} & -0.089*\text{RespTm} \\ \quad\quad\ (4.1) & (-2.4) & (-2.0) & (-3.6) & (-3.1) & (1.7) & (-3.2) \end{array}$$

$$\begin{array}{lllllll} R_{25\text{-}45} = 5.155 & -0.171*\text{IncHR} & -0.948*\text{NoVehInv} & -2.654*\text{Ex495} & -2.883*\text{Ex270} & +1.572*\text{SUT\_Ind} & +1.349*\text{Pave\_Dry} \\ \quad\quad\quad (3.1) & (-2.3) & (-2.2) & (-3.0) & (-2.4) & (2.4) & (1.8) \end{array}$$

$R_{gt45} = 0$ (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 no. 27, 33, 36, 39, 41)
Ex270 : Binary variable to indicate the specific locations on I-270
(exit no. 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 Sub-Models

	Sub-Model I				Sub-Model II				Sub-Model 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]	0.276	0.265	0.328	[5, 25]	0.481	0.483	0.494	[5, 25]	0.366	0.366	0.461
	(25, 45]	0.378	0.378	0.388	(25, 50]	0.408	0.408	0.428	(25, 45]	0.439	0.439	0.379
	> 45	0.346	0.357	0.284	> 50	0.111	0.108	0.078	> 45	0.195	0.195	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 it 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 *i*th 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 *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 literature regarding this estimation technique (Washington et al., 2003 and Pindyck and Rubinfeld, 1998), it is not be discussed in detail here.

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 are as expected. Independent variables are tested at the 90% significance level which means that the absolute value of *t*-statistic should be greater than or equals to 1.65 for that variable to be considered as significant. The estimated models for *Collision-Fatality* and *Others* are valid at the 90% significance level, since the p-values for both models are less than 0.0001. Specific discussions are summarized below for each model.

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

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

3. This model also reflects the effect on increasing incident duration for *Collision-Fatality* incidents, when an incident occurs on interstate roads, I-68 or MD/I-295 as a positive coefficient for this binary variable.
4. The observation from Rule-Based Tree Models that the duration of fatality incidents occurring at night is more likely to be longer than that in the daytime is reflected in this estimated model by a positive coefficient of the binary variable *Night*.
5. As shown in Table 5.7, the overall correct estimation result for duration of fatality incidents using the estimated regression model is 74.7%. Incident duration greater than or equal to 120 minutes is well estimated, while other categories for incident duration are not estimated correctly at all.
6. The model is tested using the validation dataset – incidents occurred in year 2006 – and the results are shown in Table 5.8. The overall correct predicted result is 78.1%, which is slightly higher than the one for estimation. Similarly, the predictions for incident duration greater than or equal to 120 minutes are satisfactory, while the predictions for other categories of incident duration are unsatisfactory.
7. Absolute error, defined as an absolute value of the difference between observed and estimated/predicted value, is also computed as a reference to evaluate the estimated model. In the model estimation results, 50.7% of records show an absolute error within 30 minutes, while 81.3% of records have an absolute error within 60 minutes. For model results with the validation dataset, 40.6% and

59.3% of records show an absolute error within 30 minutes and 60 minutes, respectively.

<u>*Others*</u>

1. Unlike the linear regression model for *Collision-Fatality*, the dependent variable in the model for *Others* is the logarithm of incident duration. This transformation of dependent variable is 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, the heavy vehicle (tractor-trailers and single unit trucks) involvement is likely to increase incident duration. Especially, the tractor-trailer involvement (*TT_Ind*) shows a strong positive relationship with incident duration as the *t*-statistic value for this is very high (i.e. 4.64). This relation is not found in Rule-Based Tree Models.
3. The model reflects the observation that incident duration for *Others* increases as the number of blocked lanes in the same direction increases by a positive coefficient for that variable. Response time also shows a strong positive relationship with incident duration in the estimated model.
4. Among other incident natures, *Debris* shows a negative relationship with incident duration, while *Emergency Road Work* has a positive relationship. Other events in *Others* incident natures do not show any significance with incident duration. That is, duration of incidents caused by debris is more likely to be shorter than the one

with any other natures in *Others*. On the other hand, emergency road work causes longer incident duration than any other natures in *Others*.

5. The overall percentage for correct estimation is 66% as shown in Table 5.9. For relatively short (i.e., 5~30 minutes) and long (i.e., >=120 minutes) incident duration, the model performs well. However, for incident duration between 30 and 120 minutes, the model does not give a good estimation. Especially, for incidents with duration between 60 and 120 minutes, this model does not give any correct estimation.
6. For predicted results based on the validation dataset as summarized in Table 5.10, the overall correct prediction percentage is slightly lower (i.e., 61.1%) when compared with the estimation results. The table also shows that the model predicts the incident duration between 5 and 60 minutes quite well, while incident duration longer than 60 minutes is not predicted correctly at all.
7. An absolute error is also computed for each record in the model development and validation dataset. In the dataset used for model development, 61.7% of records show an absolute error within 15 minutes, while 80.9% of them show it within 30 minutes. In the validation dataset, the results for absolute errors are similar to these in the model development dataset; 61.1% of them are within 15 minutest, while 77.8% of them are within 30 minutes.

Table 5.5 Estimated Multiple Linear Regression Model for Incident Nature-*Collision-Fatality*

$$\text{Incident Duration (mins)} = \underset{(13.64)}{162.95} - \underset{(-2.11)}{31.94}*\text{Pave\_Wet} + \underset{(2.02)}{32.05}*\text{NoSUT} + \underset{(3.17)}{42.03}*\text{NoTT} + \underset{(2.33)}{29.50}*\text{Night} + \underset{(2.47)}{59.10}*\text{Rd68\_295}$$

$$- \underset{(-2.87)}{42.03}*\text{Ratio\_sdbl}*\text{SHDBK}$$

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 is involved; 0 otherwise
TT_Ind: 1 if Tractor-trailers is 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 – *Collision-Fatality*

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 – *Collision-Fatality*

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 to find the relationships between incident duration and its factors. In the estimated regression models, several findings discovered from Rule-Based Tree Models are reflected. According to their estimation/prediction results and absolute error, further research is recommended for more reliable models, especially for *Collision-Fatality*. It is also supported by Figures 5.3 to 5.6, since the incident duration between observed and estimated/predicted for *Collision-Fatality* is quite different, while the one for *Others* is close as shown in those figures. In general, fatality incidents cause longer incident duration, and this requires more specific and systematic incident management strategy based on the well predicted incident duration to soothe their impact (e.g., traffic congestion or delay). To achieve this, the first thing to accomplish is to collect additional incident records with additional information for them, e.g., number of pedestrians/drivers/occupants injured or killed, collision type (head on, rear end, etc).

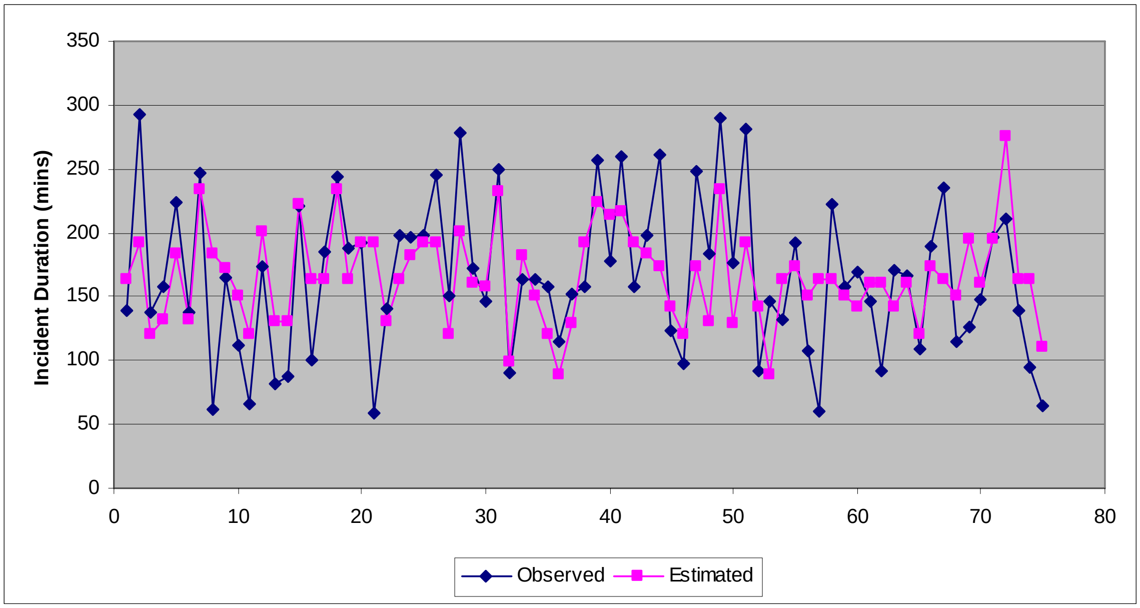


Figure 5.3 Comparisons between Observed and Estimated Incident Duration Using Developed Multiple Linear Regression Model for Incident Nature - *Collision-Fatality*

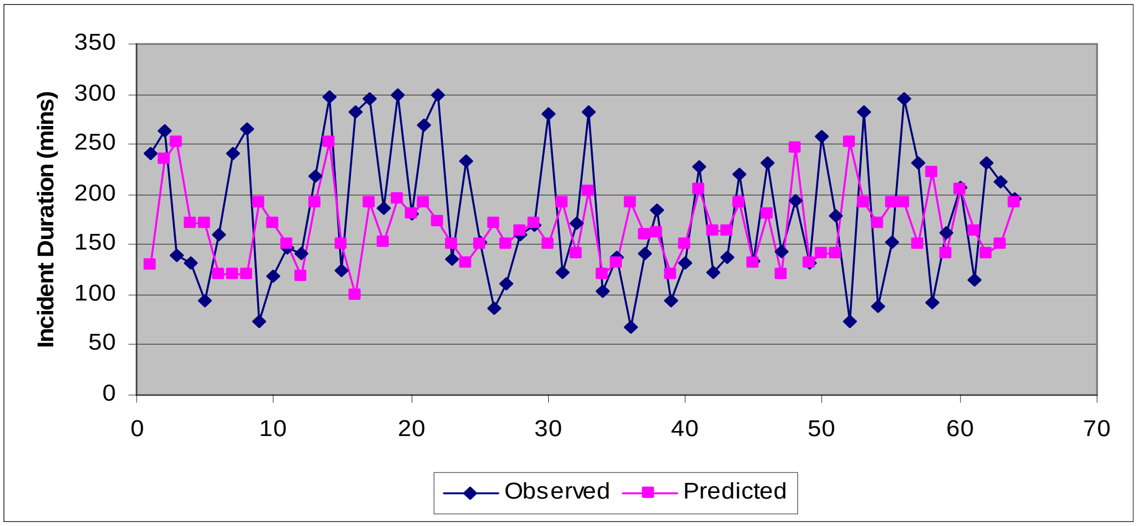


Figure 5.4 Comparisons between Observed and Predicted Incident Duration Using Developed Multiple Linear Regression Model for Incident Nature - *Collision-Fatality*

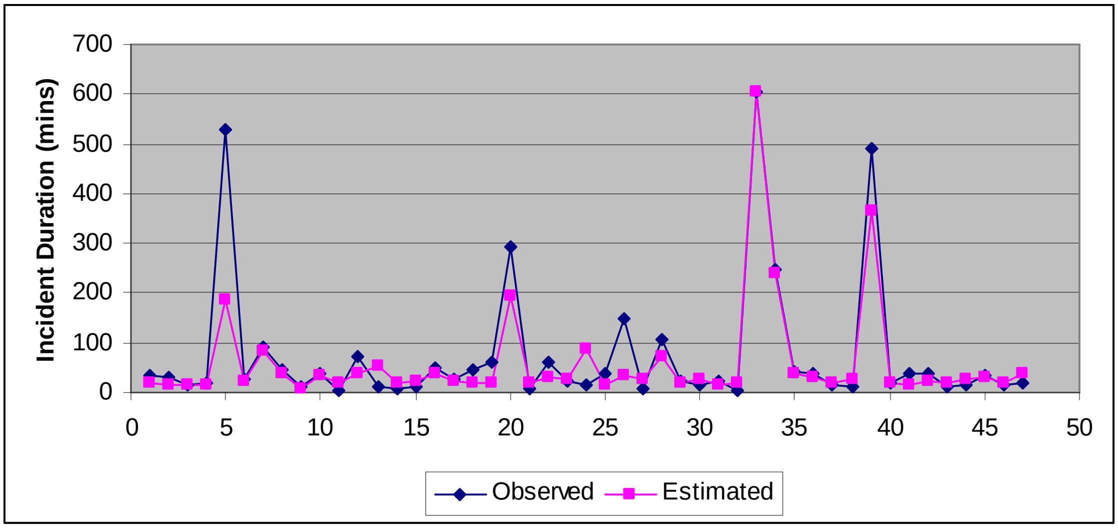


Figure 5.5 Comparisons between Observed and Estimated Incident Duration Using Developed Multiple Linear Regression Model for Incident Nature – *Others*

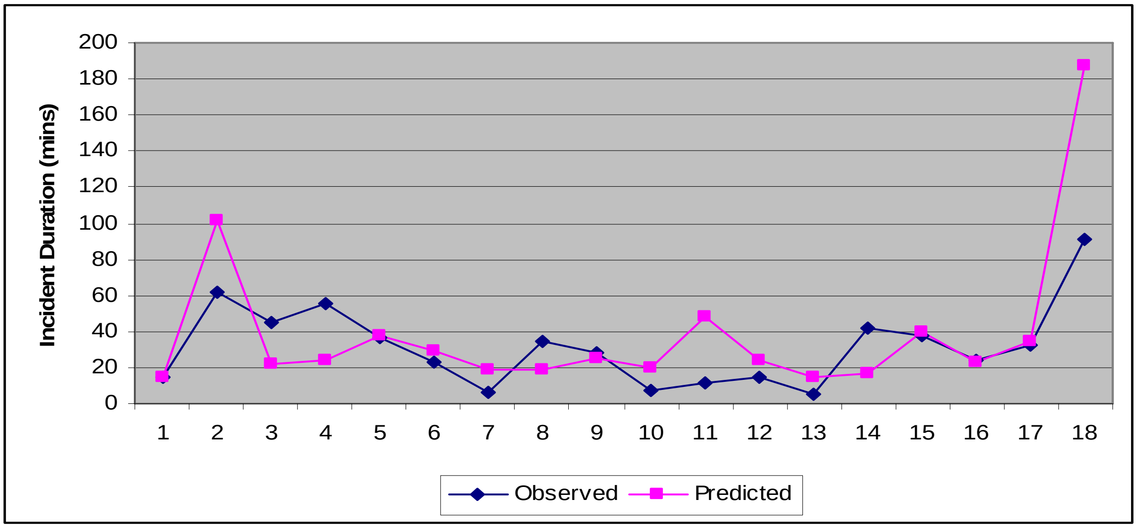


Figure 5.6 Comparisons between Observed and Predicted Incident Duration Using Developed Multiple Linear Regression Model for Incident Nature – *Others*

5.4 Illustrative Description of the Developed Model Application

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

Step 1: Identify the detected incident nature and location of its jurisdiction to select the appropriate Rule-Based Tree Model.

Step 2: Trace the selected Rule-Based Tree Model 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 dataset.

Step 4: Otherwise, trace back one node by one node until meeting the node satisfying the evaluation criteria.

Step 5: If one can not find the satisfactory node in Rule-Based Tree Models, then use a supplemental model to predict the incident duration or the probability distribution of the target incident duration.

This whole process can be expedited if those models (Rule-Based Tree Models and supplemental models) along with evaluation criteria are programmed on a user-friendly interface.

Table 5.11 provides the actual examples of traffic incident information from dispatched units and the predicted incident duration using the Rule-Based Tree Model and supplemental models. Variable names appearing in Table 5.11 are described in Table 5.12. The first example is about a fatality incident occurring in Prince George County in 2006. The Rule-Based Tree Model predicts the incident duration of 80~100 minutes with 33.33% ***confidence*** based on dataset collected from year 2003 to 2006. Since it does not satisfy one of our criteria (i.e., ***confidence*** should be greater than 70%), we use a supplemental model (Multiple Linear Regression Model presented in Table 5.5) to obtain a more reliable prediction of incident duration prediction. The model predicts that the incident duration will be approximately 121 minutes, and this prediction is closer to the observed incident duration, 144 minutes, than the one from the Rule-Based Tree Model. The same phenomenon is observed in the fifth example.

For the second example, the Rule-Based Tree Model predicts the incident duration of 10~35 minutes with 75% ***confidence***. Since this outcome satisfies our criteria, one does not need to apply any supplemental model to this case. Similar explanation can be applied to the fourth example for the disabled vehicle incident. On the other hand, the third example shows only 60% ***confidence*** with the Rule-Based Tree Model, so a supplemental model is required. The supplemental model (MNL model presented in Table 5.3(b)) predicts the incident duration of 5~25 minutes with 0.84 probability and we take this as the predicted duration of the detected incident. Note that the Rule-Based Tree Model predicts it to be 30~45 minutes.

Table 5.11 Traffic Incident Information Examples and Their Predicted Incident Duration

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

[1] Observed incident duration in minutes

[2] Predicted incident duration in minutes based on Rule-Based Tree Model (RBTM)

[3] Predicted incident duration in minutes based on Supplemental Models (SM)

[4] CPI-Sub-Model II presented in Table 5.2(b)

[5] CPD-Sub-Model 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 ~ 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 Rule-Based Tree Models
Predicted-SM	Predicted incident duration in minutes based on Supplemental Models
Confidence in RBTM	***Confidence*** based on Rule-Based Tree Models

Chapter 6: Conclusions

6.1 Summary of Research Results

This study has presented a set of models for estimating the incident duration using the incident data from year 2003 to 2005 available in the MDSHA CHART II Database. The proposed models consist of a primary component 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, whereas *County* emerges as the second most critical factor. It has also been found that the proposed Rule-Based Tree Model is quite flexible for assigning an appropriate estimated incident duration range for nodes in the decision tree.

A summary of research findings by incident nature from both primary and supplemental models is presented below.

Collision-Fatality (CF)

- The range of predicted incident durations with the Rule-Based Tree Model (RBTM) for fatality-related incidents is more likely to be wider (e.g., about 60 minutes on average) than that for other incident natures (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%.
- For example, with the dataset from year 2003 to 2005, RBTM predicted that the duration of incidents occurring on weekdays without tractor-trailers involved would be between 100 and 200 minutes with 75 percent ***confidence.*** It also

predicted that when fatality-related incidents occurred on weekends, the duration for those would be between 80 and 200 minutes with 94 percent ***confidence***.

- The multiple linear regression model, which is the supplemental model, for predicting durations of incidents causing fatalities can achieve about 75 percent of accuracy.
- The clearance operation is generally more efficient in the scenarios of blocking more lanes in the same direction (including shoulder lanes) than those leaving them open. It has also been found that the impact of wet pavement, a proxy variable for rainy days, showed a negative correlation with the duration of incidents resulting in fatalities.

Collision-Personal Injury (CPI)

- Most rules having terminal nodes in RBTM can predict the range of incident duration within 30 minutes with their ***confidence*** exceeding 70 percent.
- RBTM can predict incidents occurring in Montgomery County causing less than 3 blocked lanes (including 1 lane blockage in the opposite direction) within the range of 10~30 minutes at about 85 percent ***confidence***. For the incidents without lane blockage in the opposite direction but involving single unit trucks, the predicted duration of 25 to 50 minutes can be achieved at about 81 percent ***confidence***.
- The predicted probabilistic distribution of incident duration with multinomial logit models (MNL) is within 10 percent difference from the observed data.

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 multinomial logit models (MNL) 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% for Montgomery County only) lie in a relatively short range of 5~30 minutes.
- Since about 84 % of incidents in Montgomery County due to disabled vehicles have duration lied in a 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 duration of disabled vehicle-related incidents occurring on weekends in Montgomery County would be in a range of 5 to 25 minutes with 82% ***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 the reliable RBTM was particularly challenging. In addition, more than 50% of rules were unable to be validated. Nevertheless, the overall performance of RBTM was promising, except for some rules with small sample size.

- The multiple linear regression model as a supplemental component for its category performed quite well. It can predict the duration of incidents caused by other natures in the range of 30 minutes at 81 percent level of accuracy.

6.2 Future Research

Developing reliable models for prediction based on field data is always a challenging task. It generally takes time to collect sufficient quality samples for model calibration. Besides, identifying outliers of sample necessitates the in-depth knowledge about the environment of data-collection and 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 - Rule Based Tree Model, Multinomial Logit Model and Multiple Linear Regression Model, which seem to yield quite promising results based on the available data.

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.

- For incidents resulting in fatalities, an alternative approach with additional data is needed to develop a more reliable model for predicting incident duration, since the data from SHA-CHART contains mainly operation-related information, but not other safety factors such as the number of fatalities, severity of injuries.

- It is essential to integrate the CHART database with police accident records to construct a dataset with better quality for calibrating models of fatalities involved incidents.
- For the incidents type *Others*, it is necessary to recalibrate the model with a larger sample of data, since the data from year 2003 to 2005 for these incident types is relatively small.

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

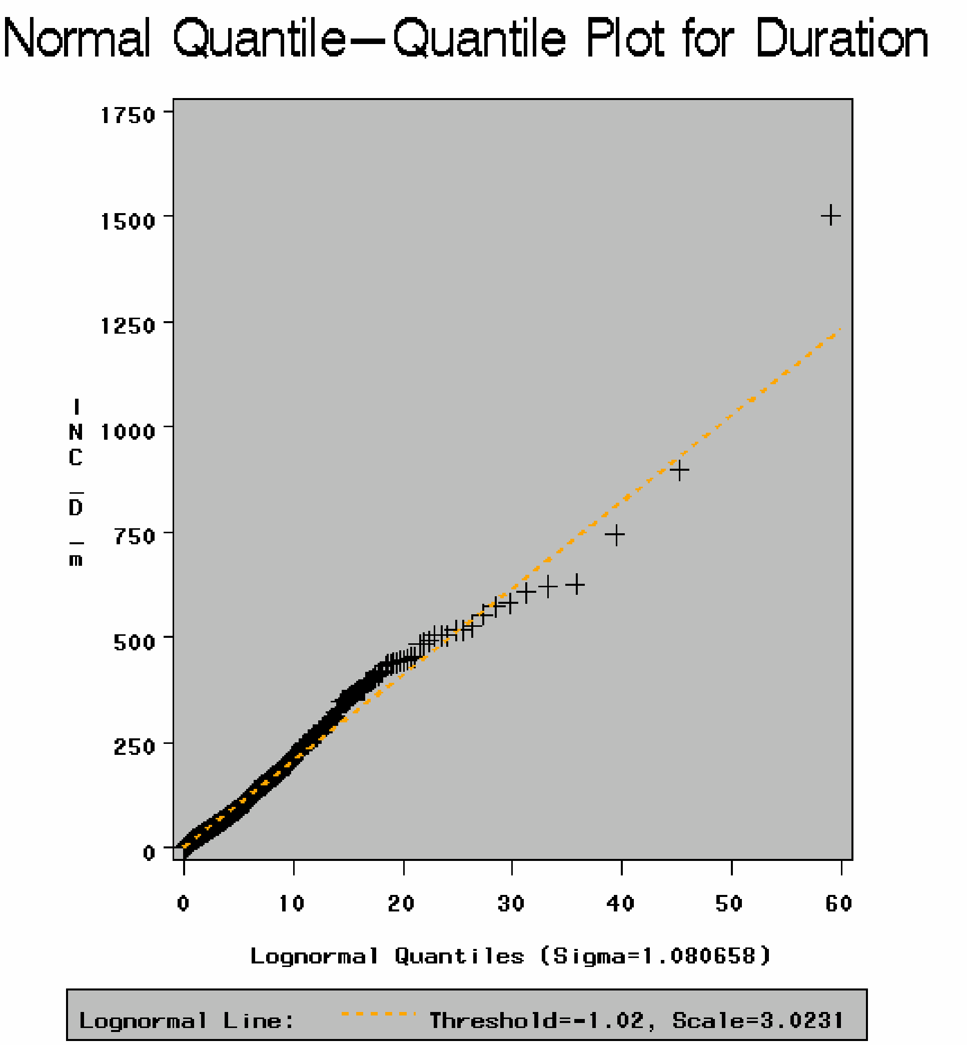


Figure A1.2 The Quantile-Quantile Plot (Q-Q Plot) of the Original Incident Duration Data Set – Weibull Distribution

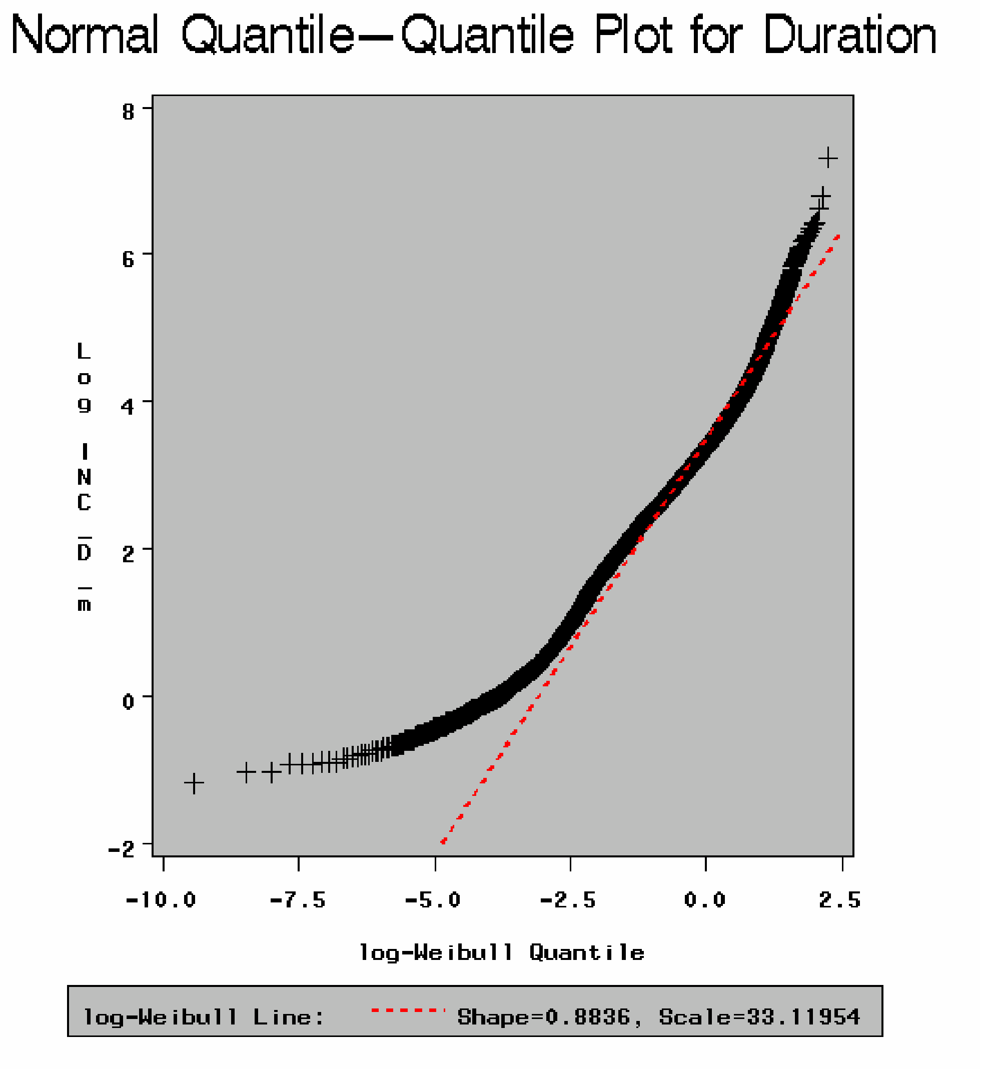


Figure A1.3 The Quantile-Quantile Plot (Q-Q Plot) of the Box-Cox Power Transformed Data Set

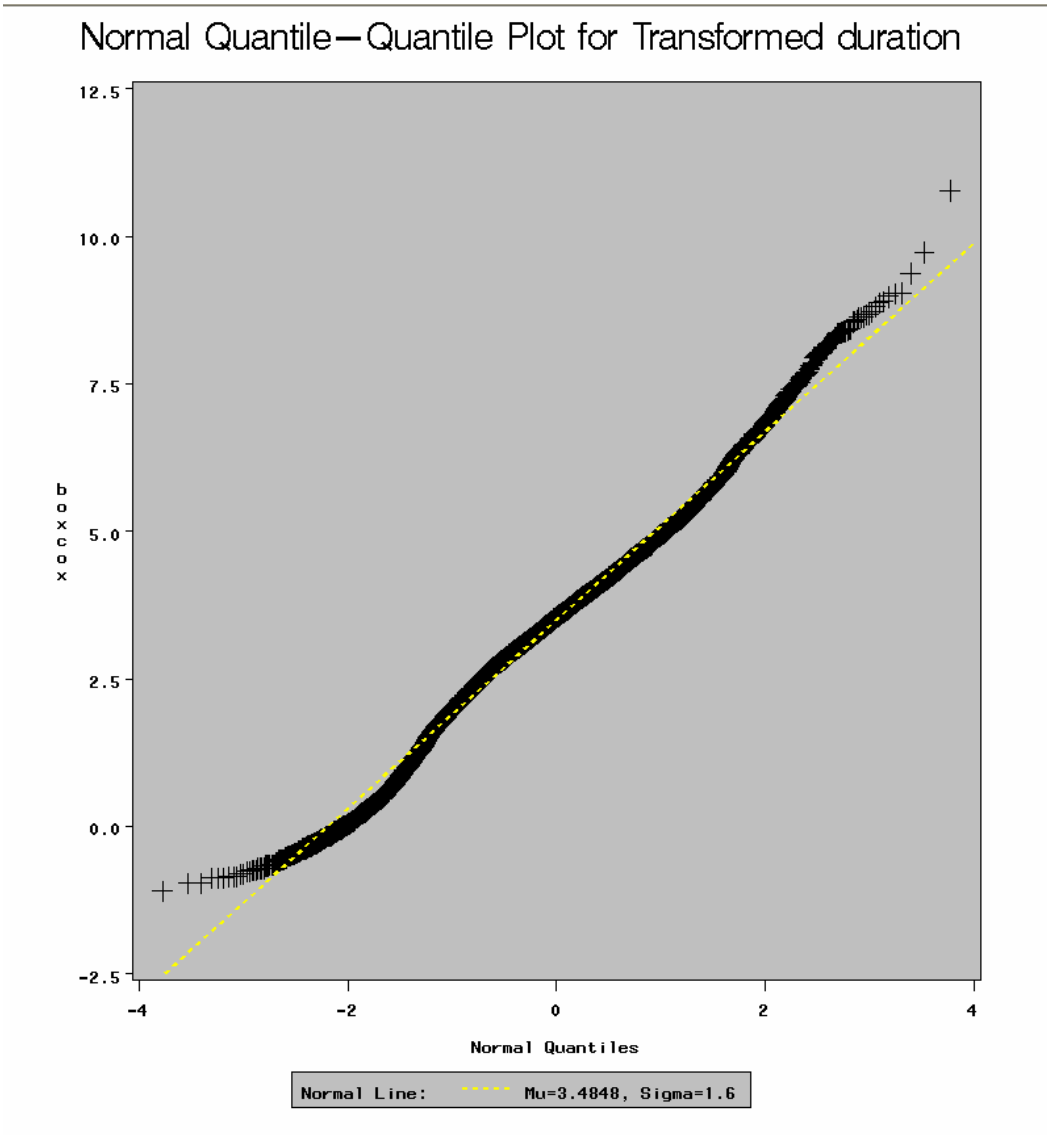


Figure A1.4 The Probability Plot (P-P Plot) of the Box-Cox Power Transformed Data Set

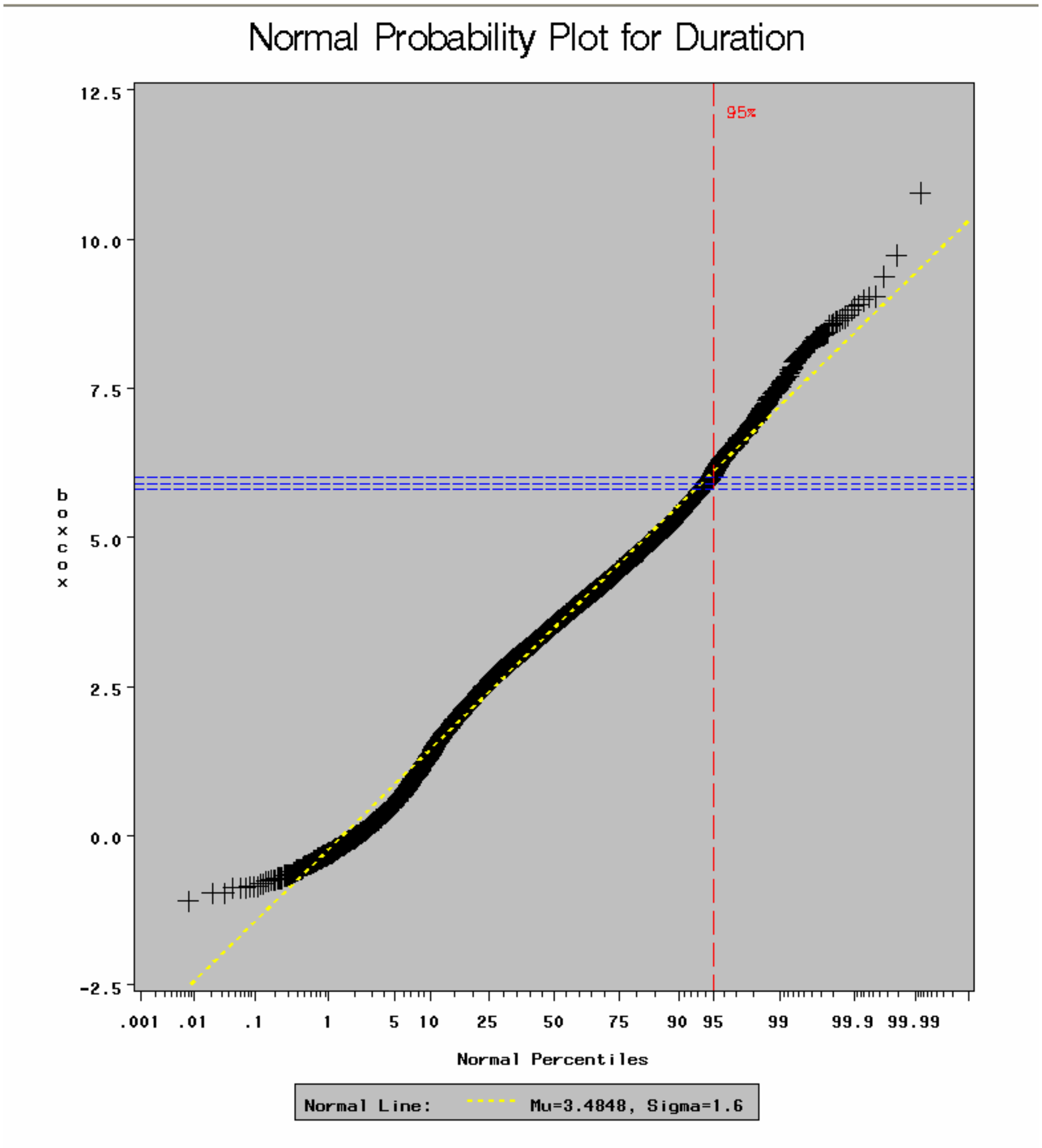


Figure A1.5 The Quantile-Quantile Plot (Q-Q Plot) of the Power Transformed Data from the Truncated Data Set (Incident Duration >= 5 minutes)

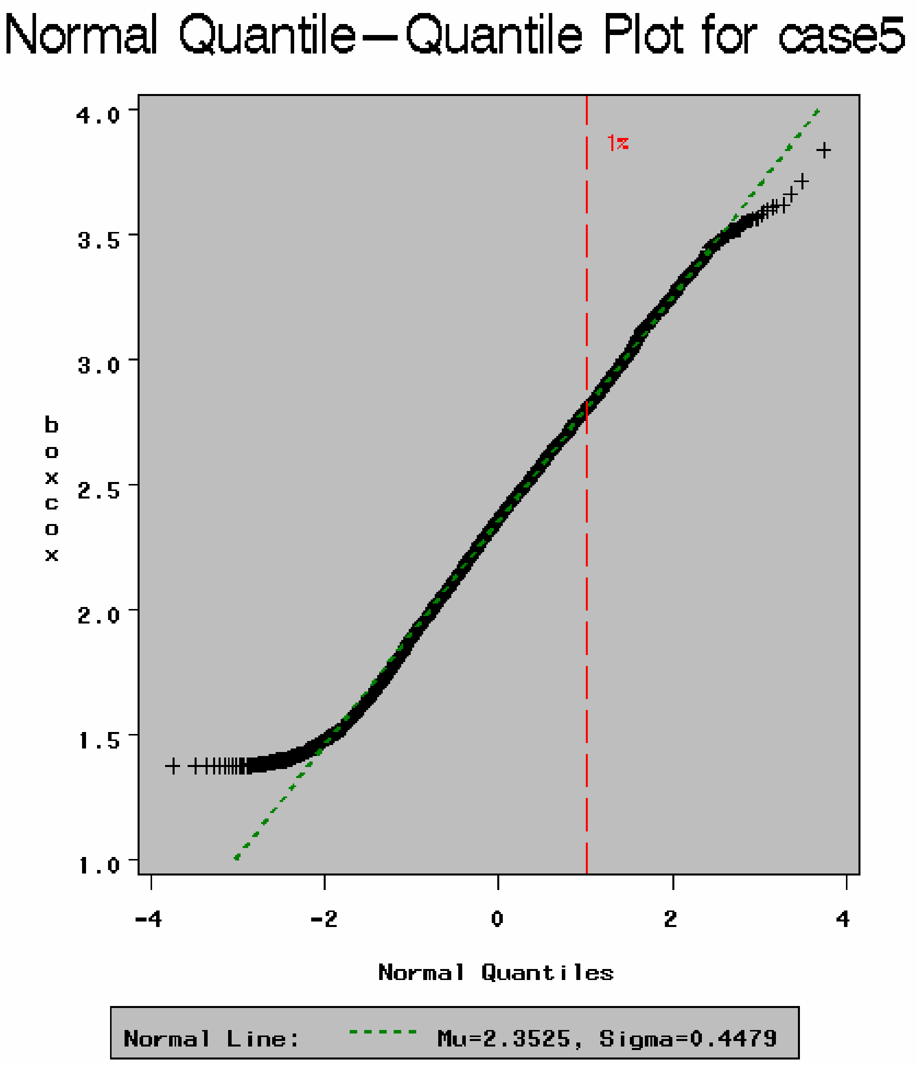
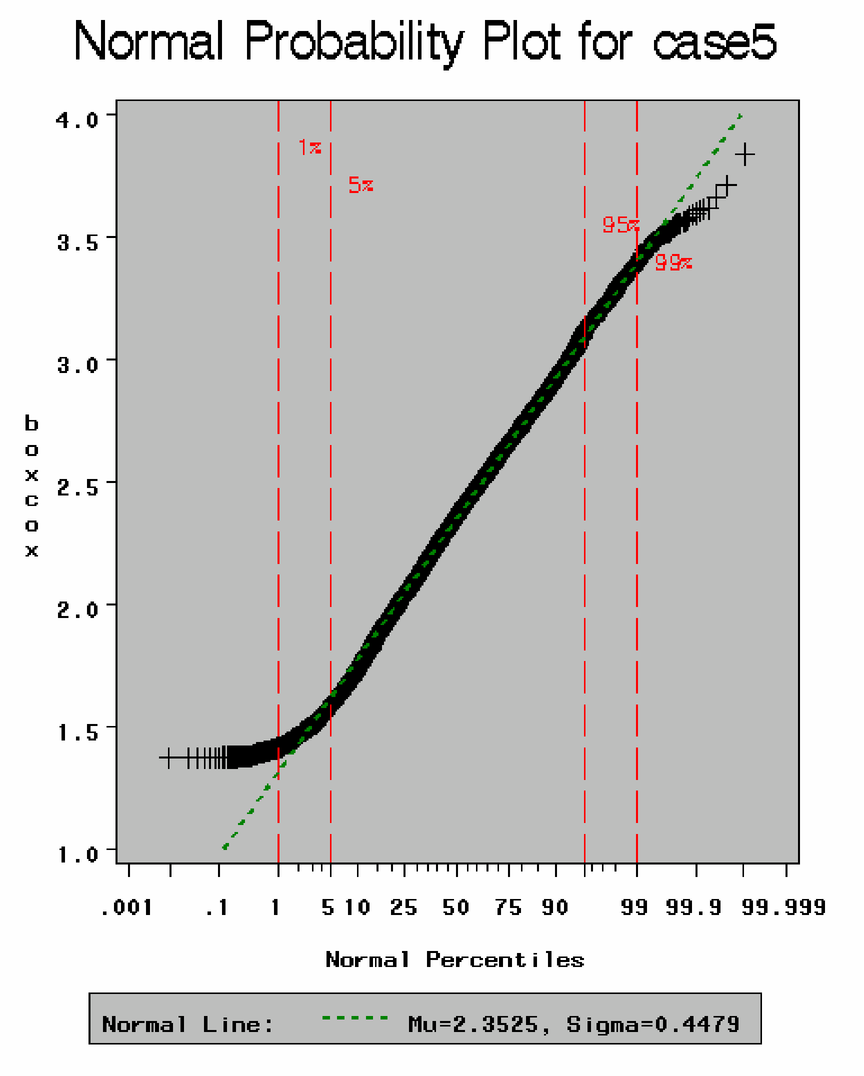
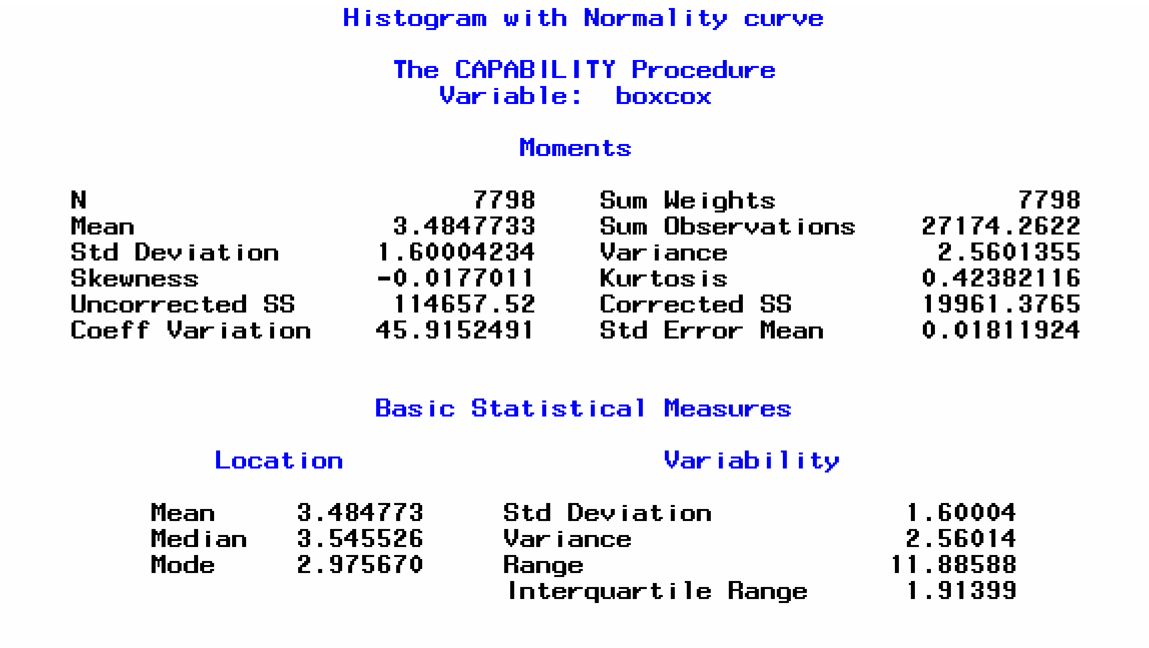


Figure A1.6 The Probability Plot (P-P Plot) of the Power Transformed Data from the Truncated Data Set (Incident Duration >= 5 minutes)

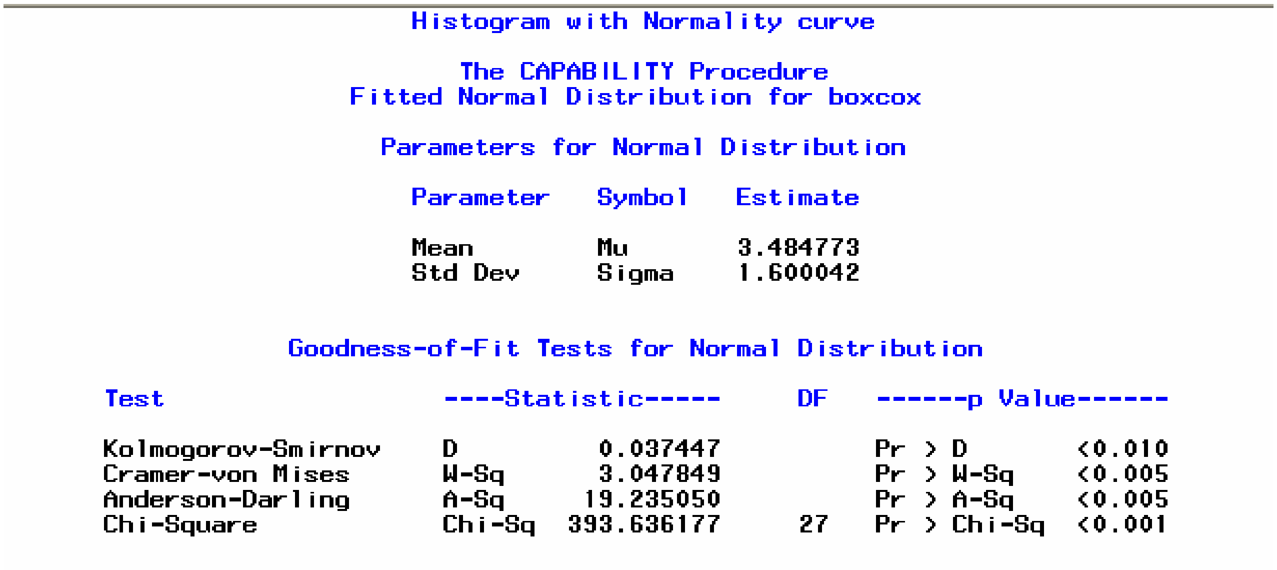


< SAS Output for the Basic Statistical Measures Using the Power Transformed Data Set>



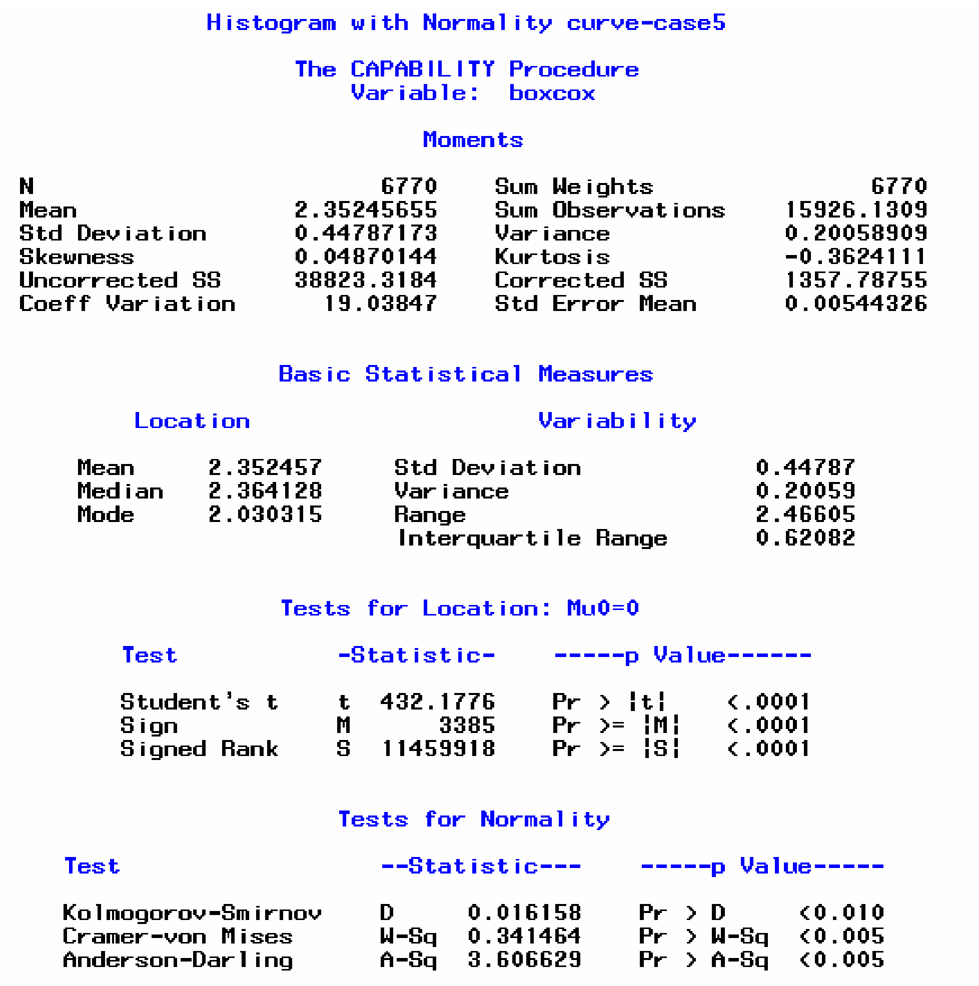
```
                          Histogram with Normality curve

                              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

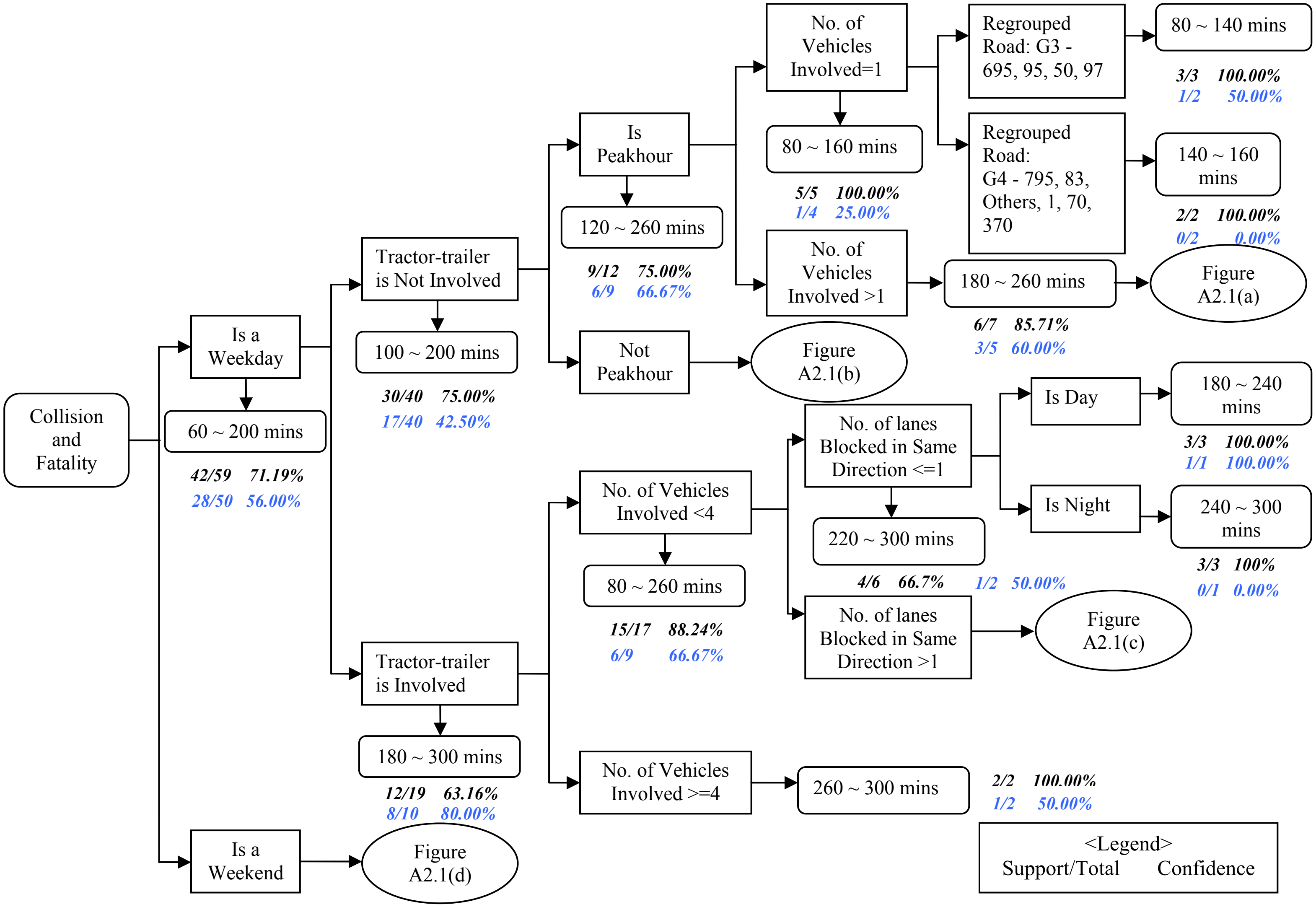


Figure A2.1 Rule Based Tree Model for Collision-Fatality in Montgomery County

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

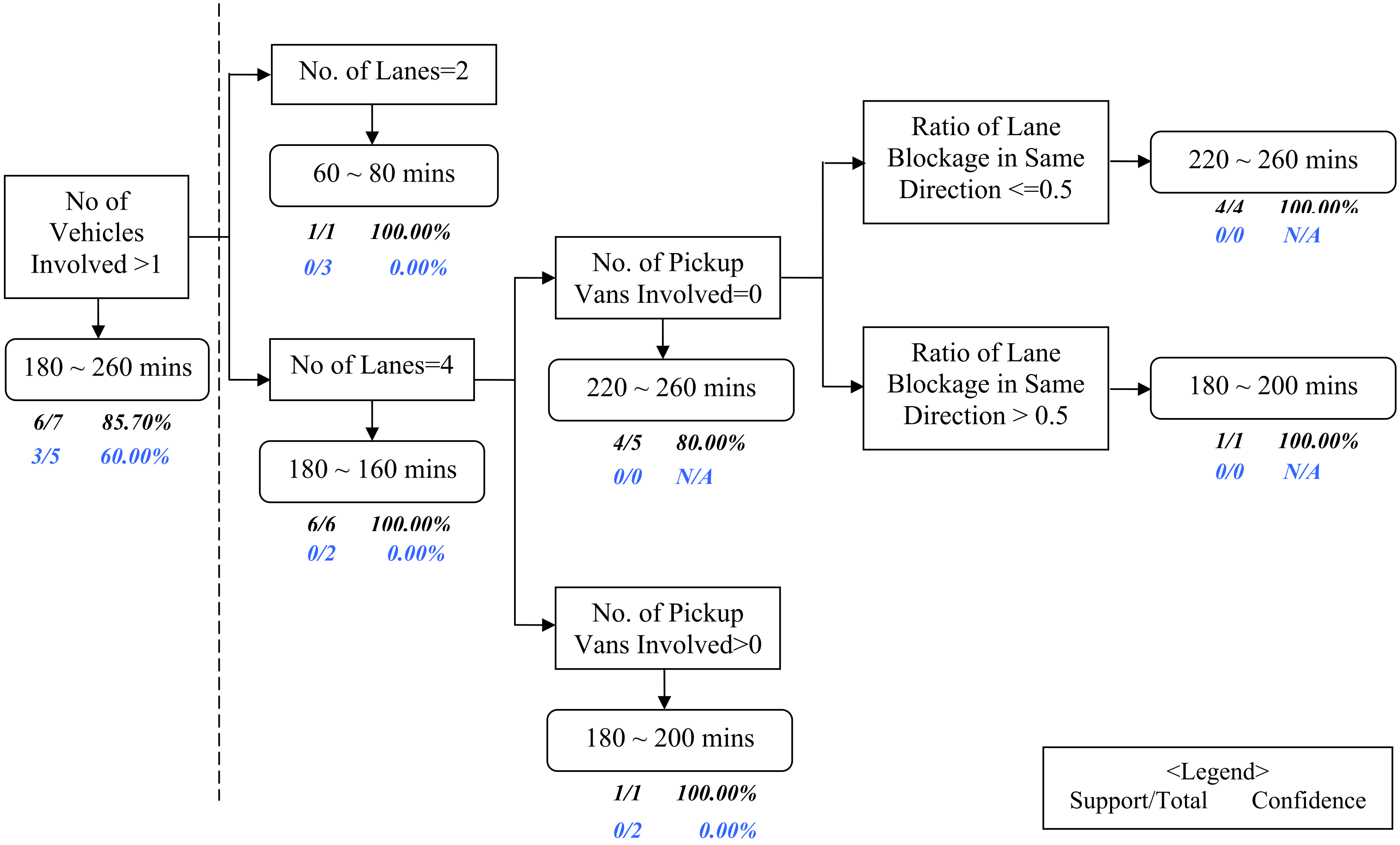


Figure A2.1(a) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

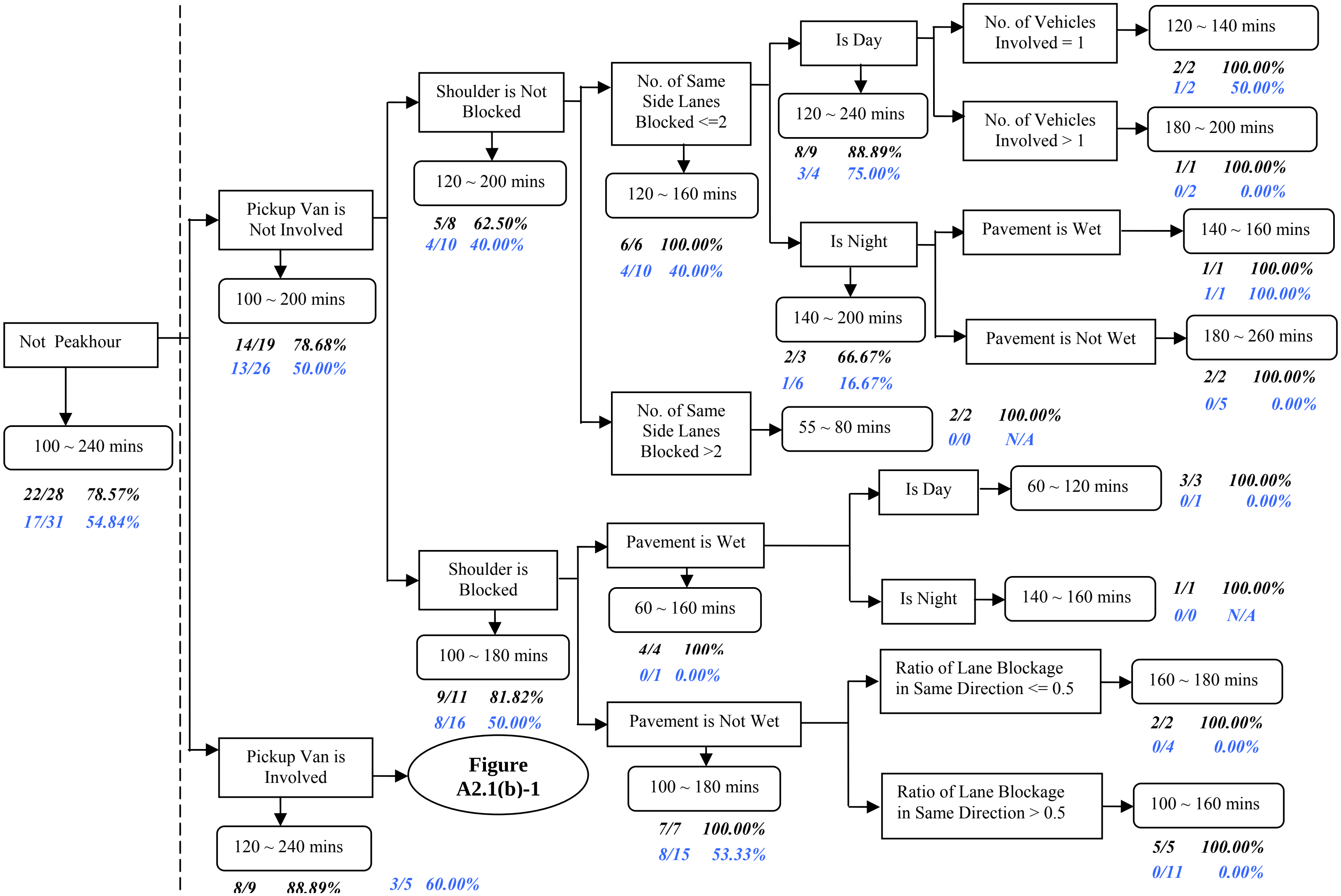


Figure A2.1(b) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

<Legend>
Support/Total Confidence

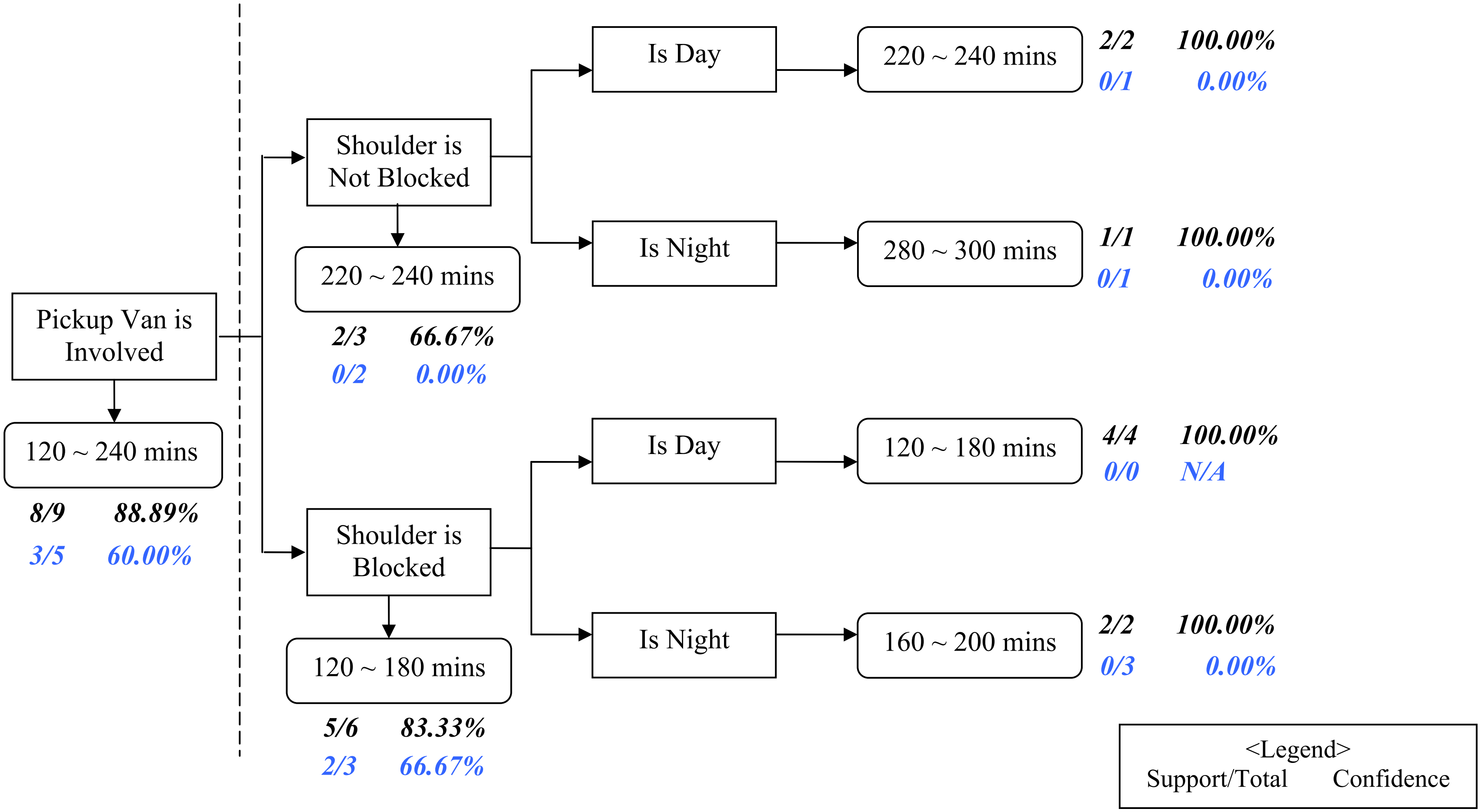


Figure A2.1(b)-1 Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

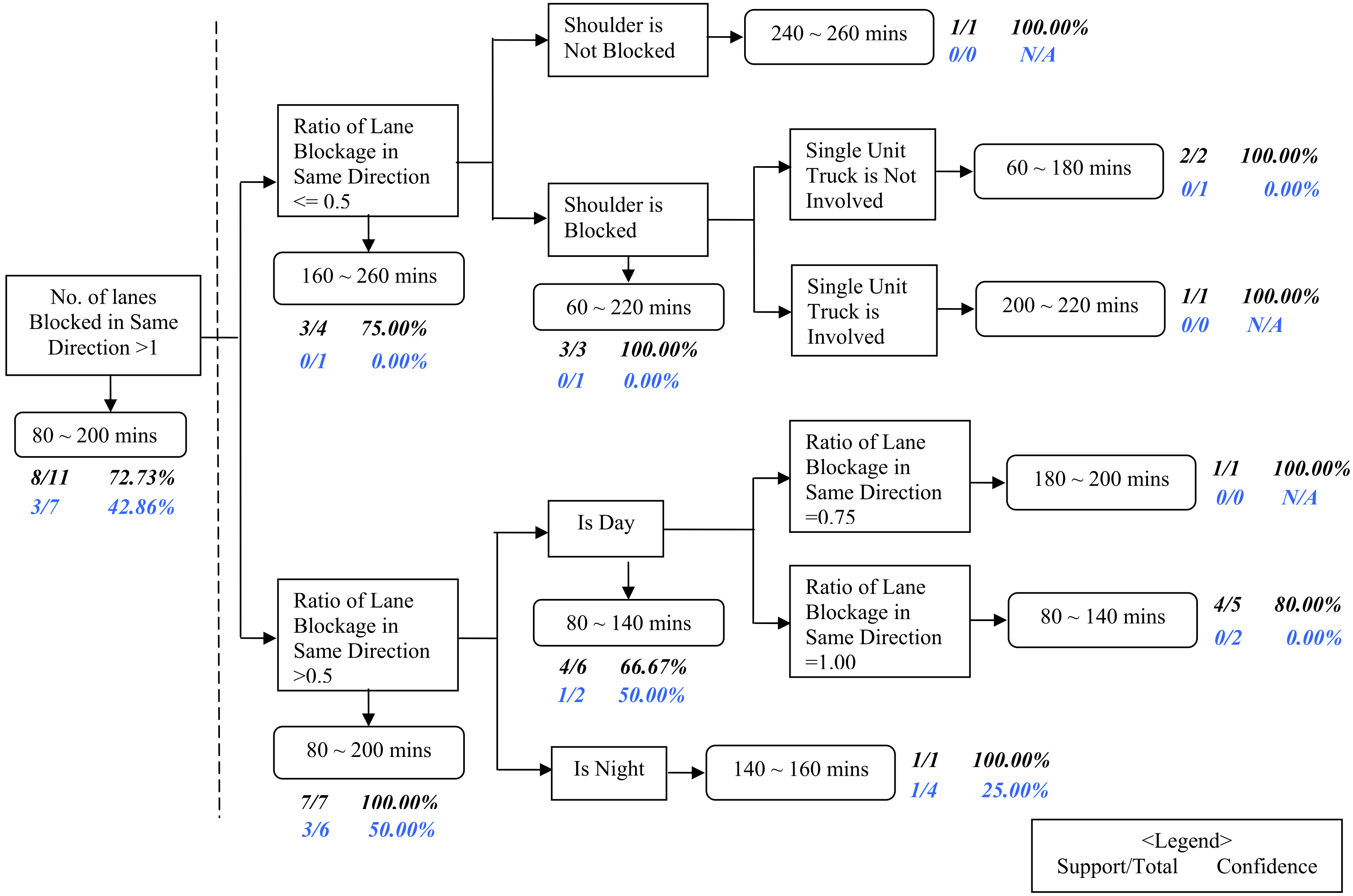


Figure A2.1(c) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

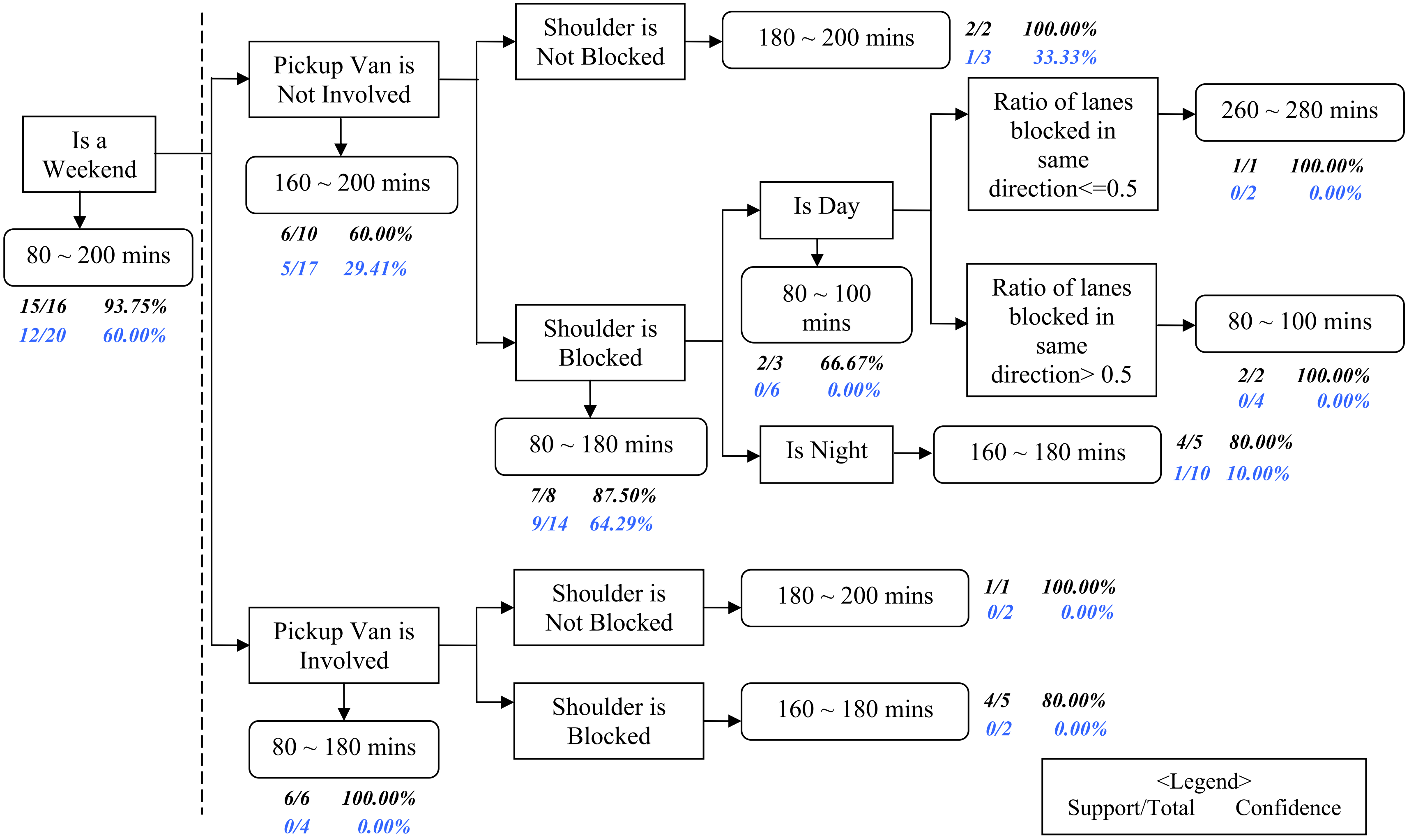


Figure A2.1(d) Rule Based Tree Model for Collision-Fatality in Montgomery County (Cont'd)

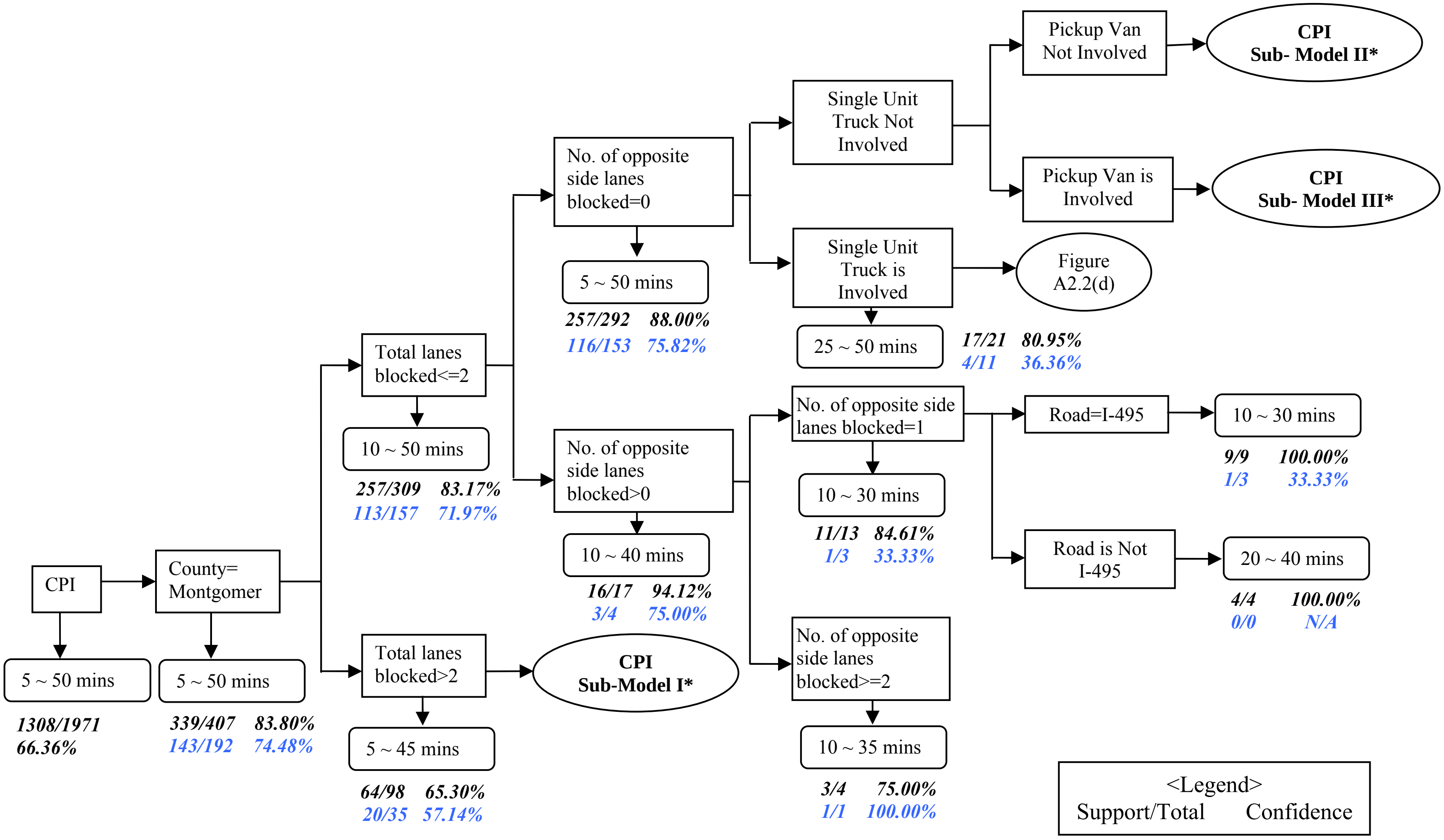


Figure A2.2 Rule Based Tree Model for Collision-Personal Injury in Montgomery County

* 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.

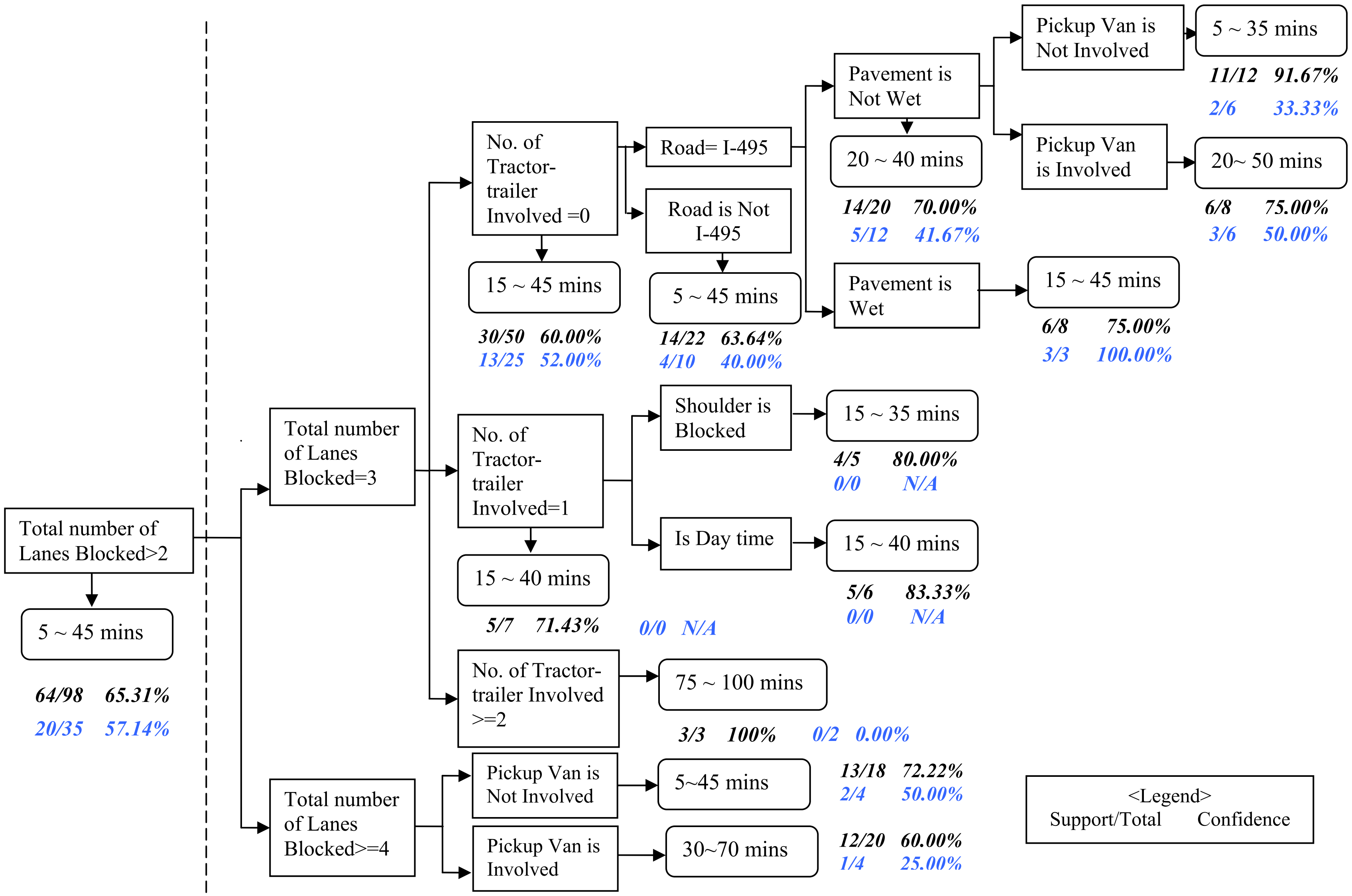


Figure A2.2(a) Rule Based Tree Model for Subsets for CPI-Sub-Model I

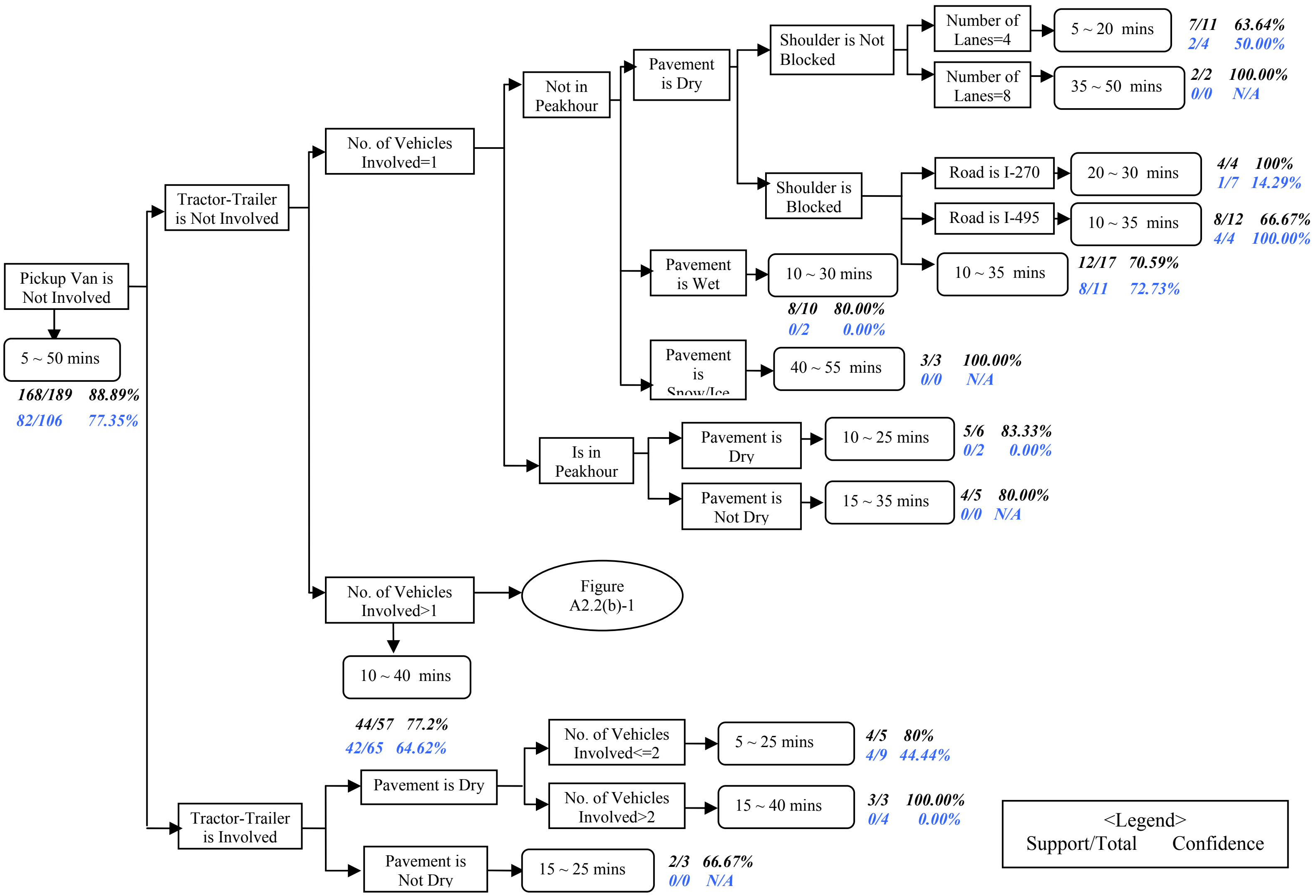


Figure A2.2(b) Rule Based Tree Model for Subsets for CPI-Sub-Model II

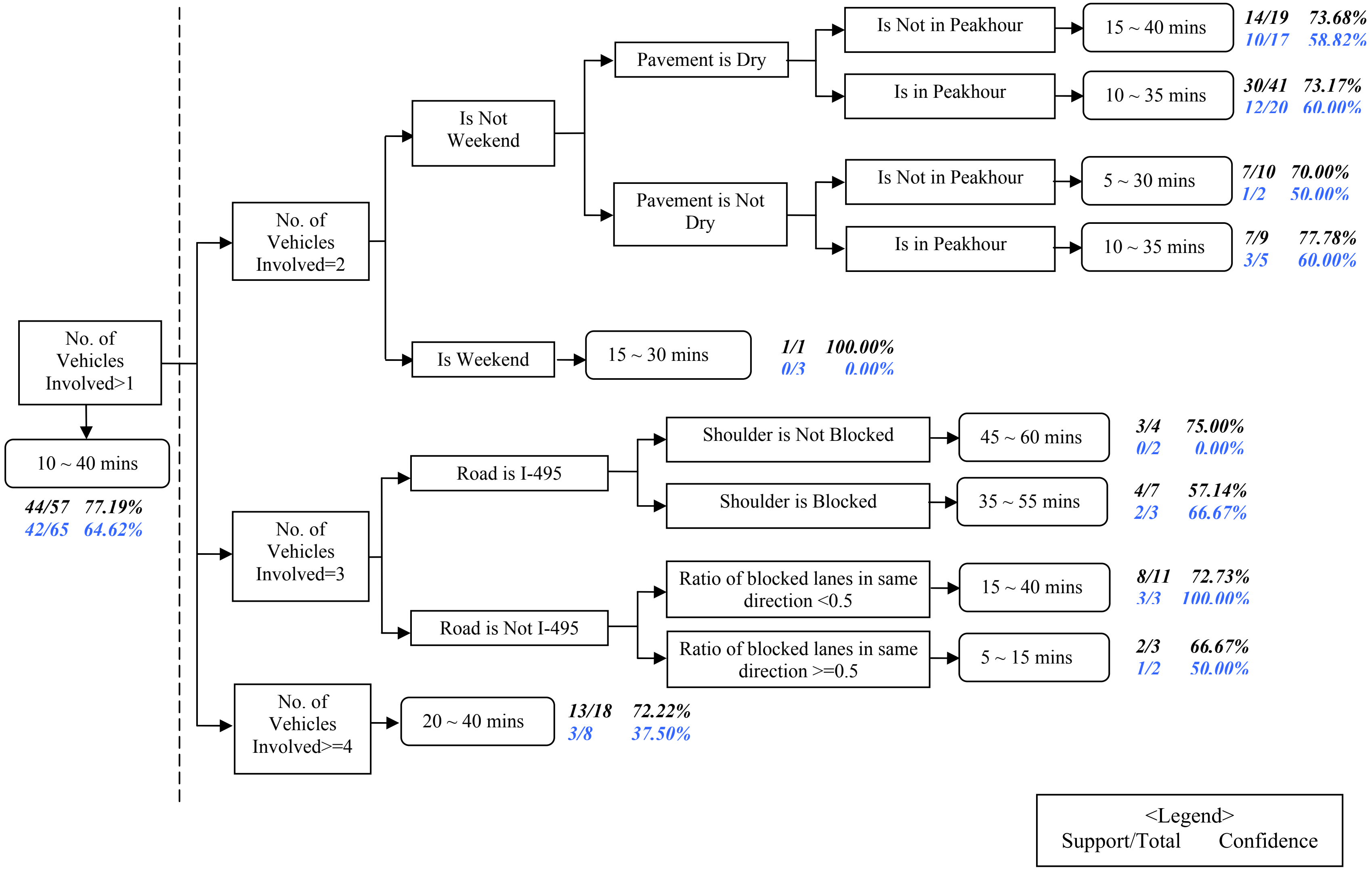


Figure A2.2(b)-1 Rule Based Tree Model for Collision-Personal Injury in Montgomery County

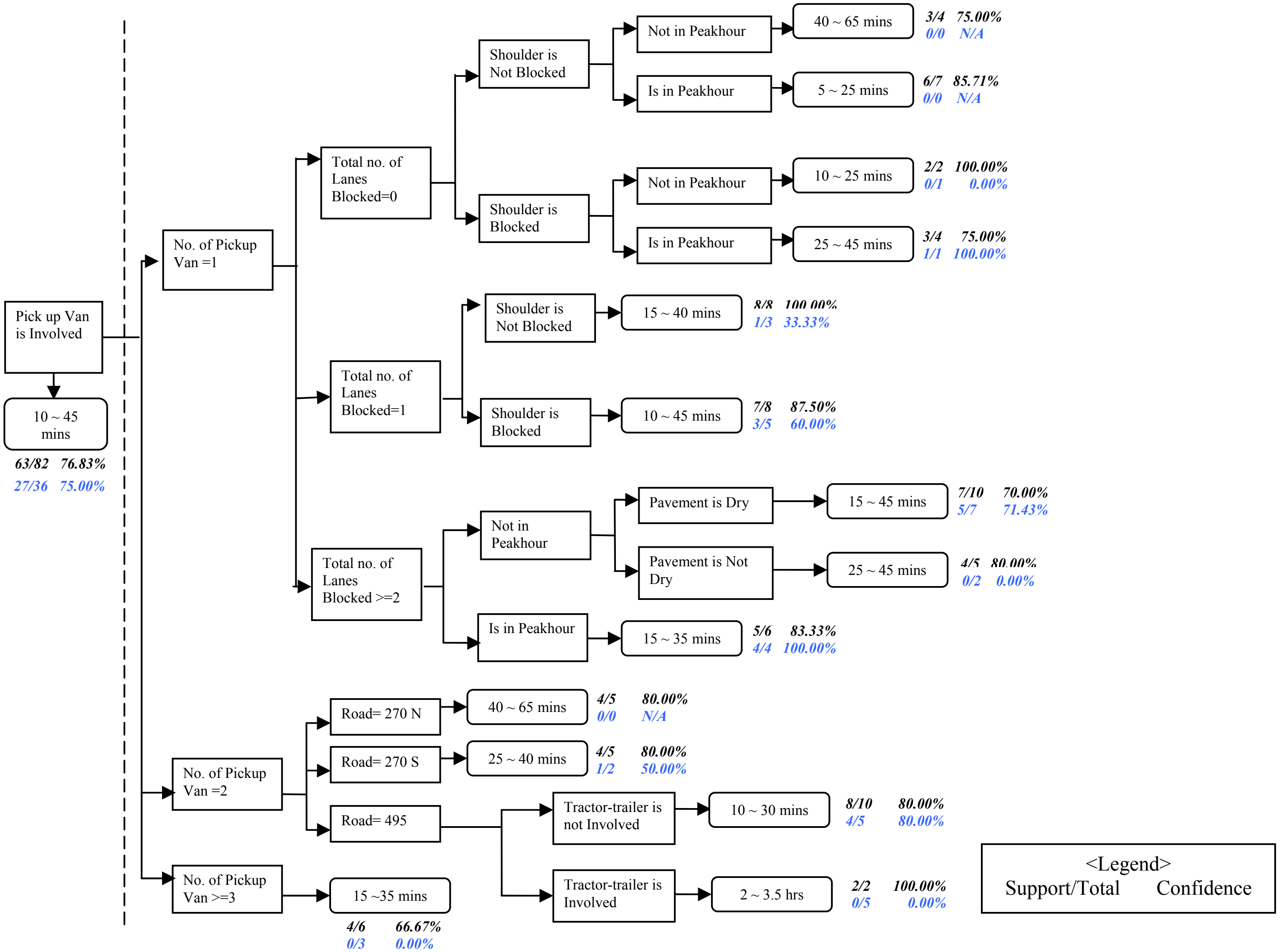


Figure A2.2(c) Rule Based Tree Model for Subsets for CPI-Sub-Model III

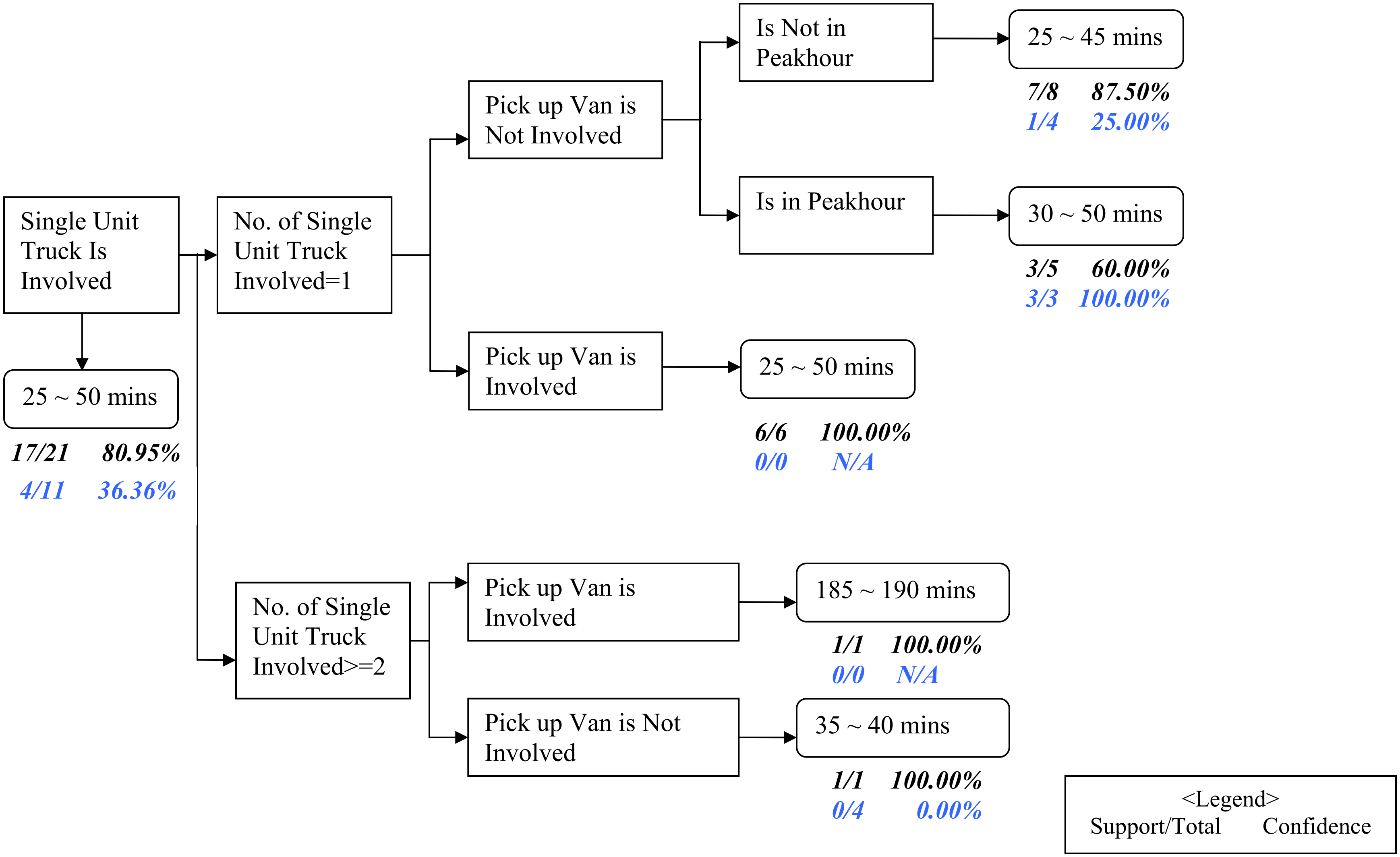


Figure A2.2(d) Rule Based Tree Model for Collision-Personal Injury in Montgomery County

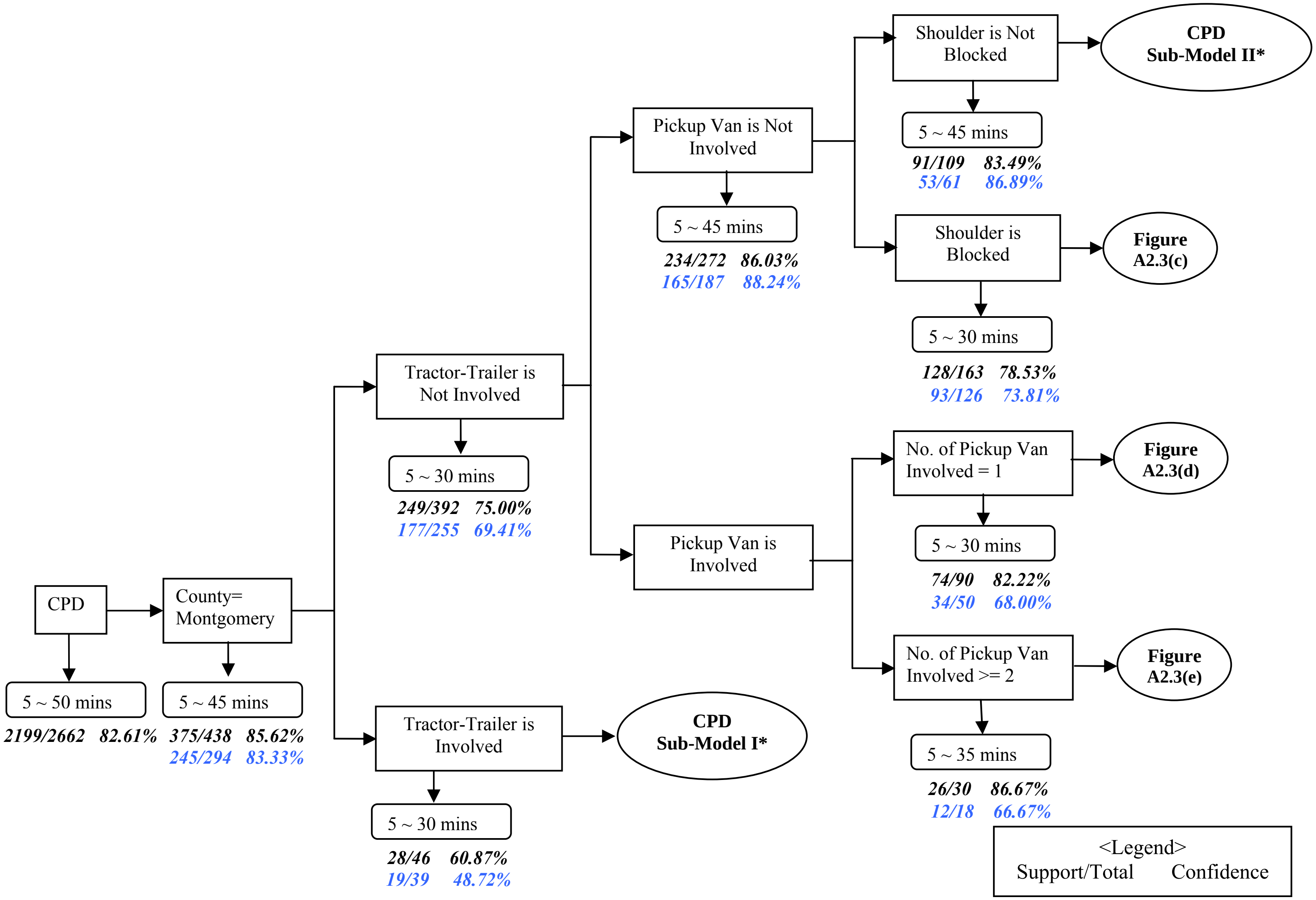


Figure A2.3 Rule Based Tree Model for Collision-Property Damage in Montgomery County
* Detailed trees for subsets for CPD-Sub-Model-I and II could be found in Table A2.3(a) and A2.3(b), respectively, in Appendix 2

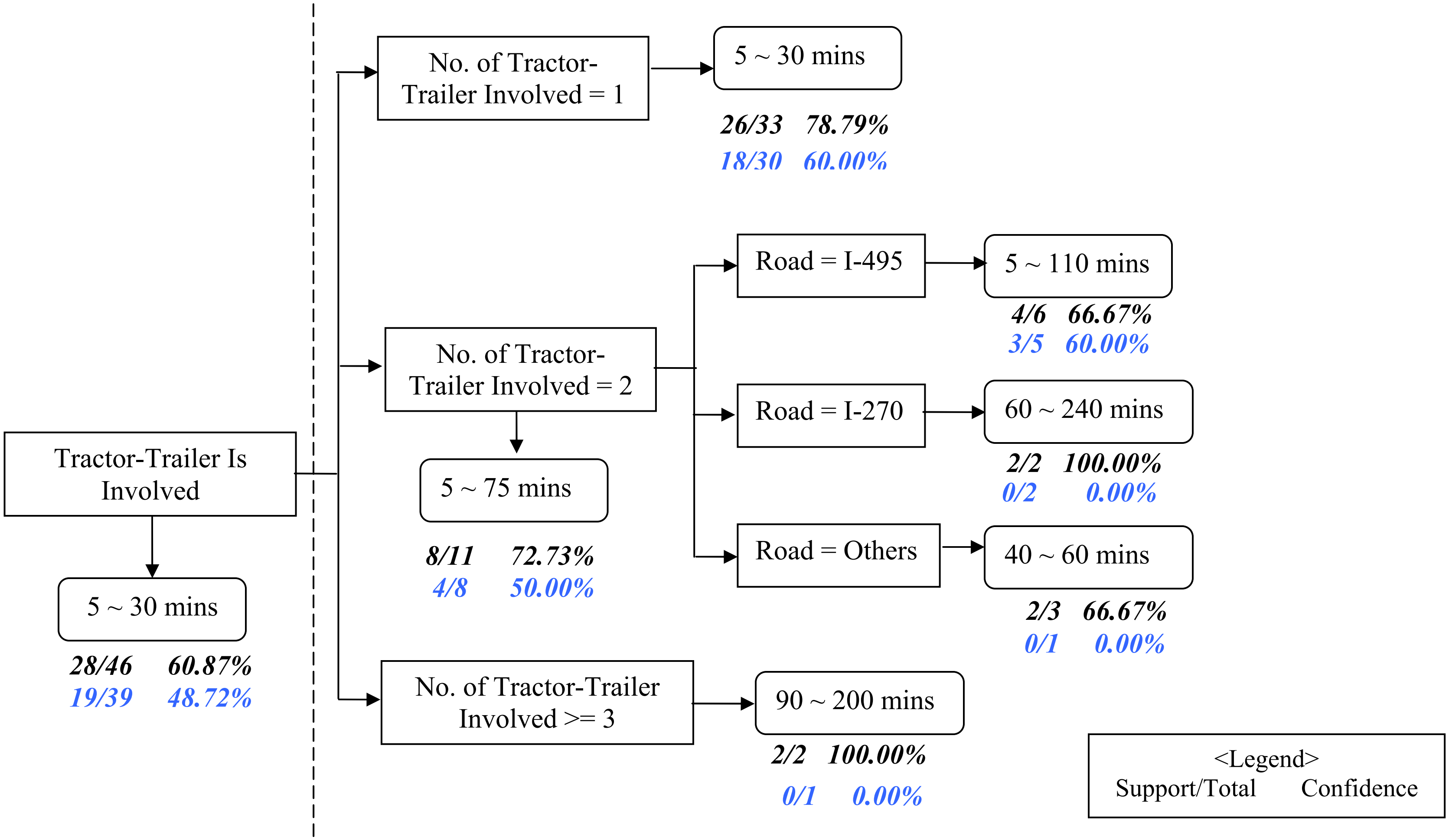


Figure A2.3(a) Rule Based Tree Model for Subsets for CPD-Sub-Model I

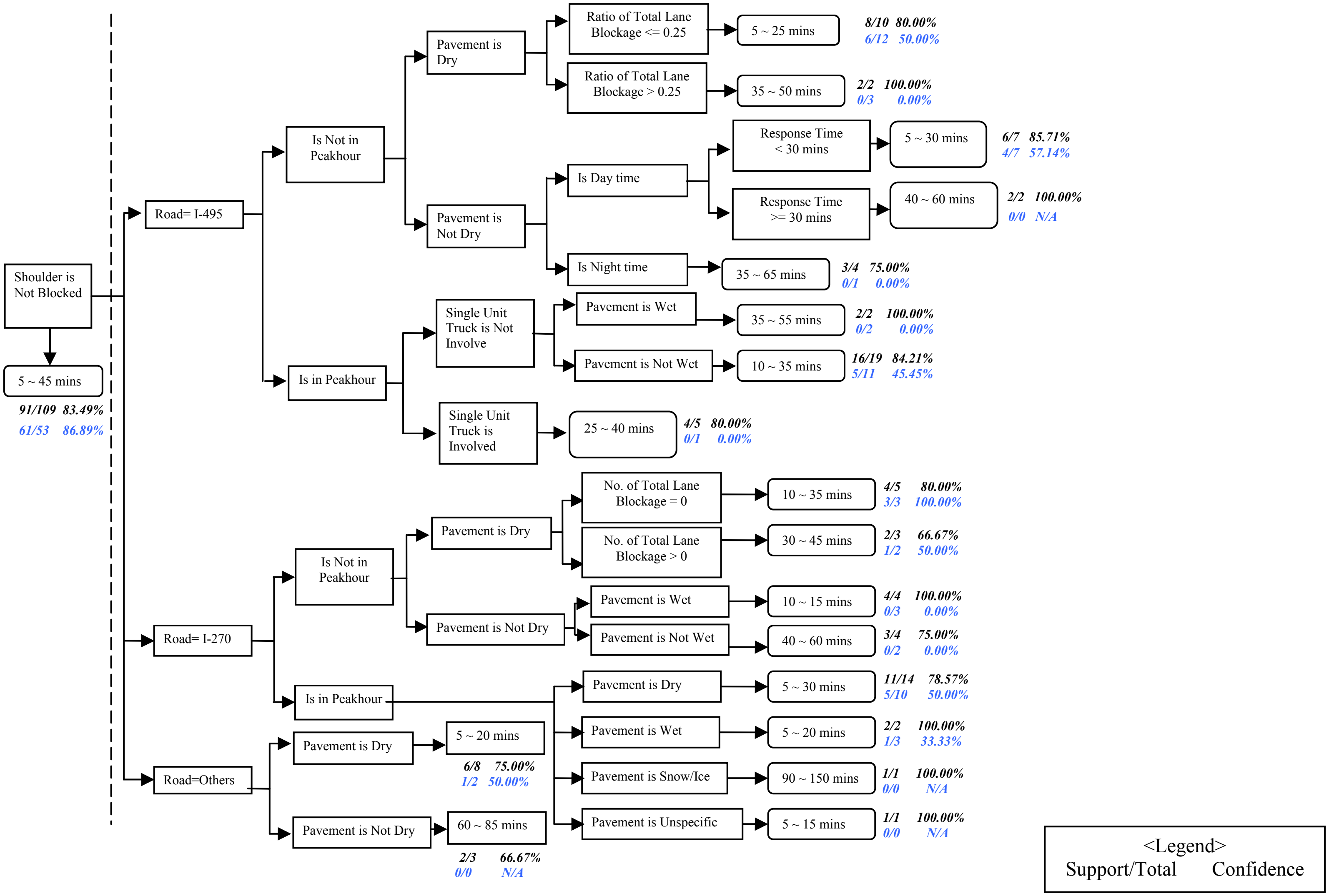


Figure A2.3(b) Rule Based Tree Model for Subsets for CPD-Sub-Model II

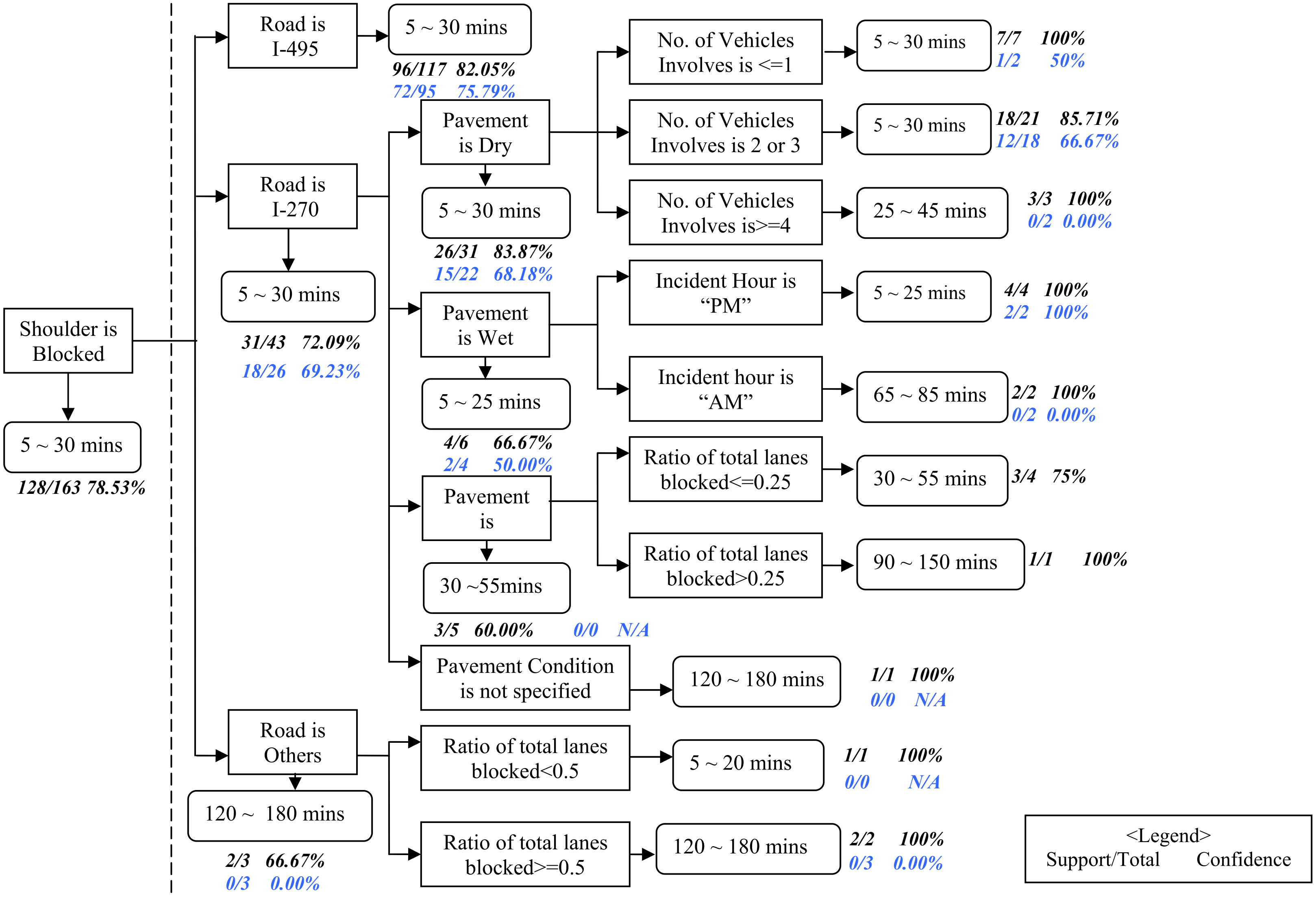


Figure A2.3(c) Rule Based Tree Model for Collision-Property Damage in Montgomery County

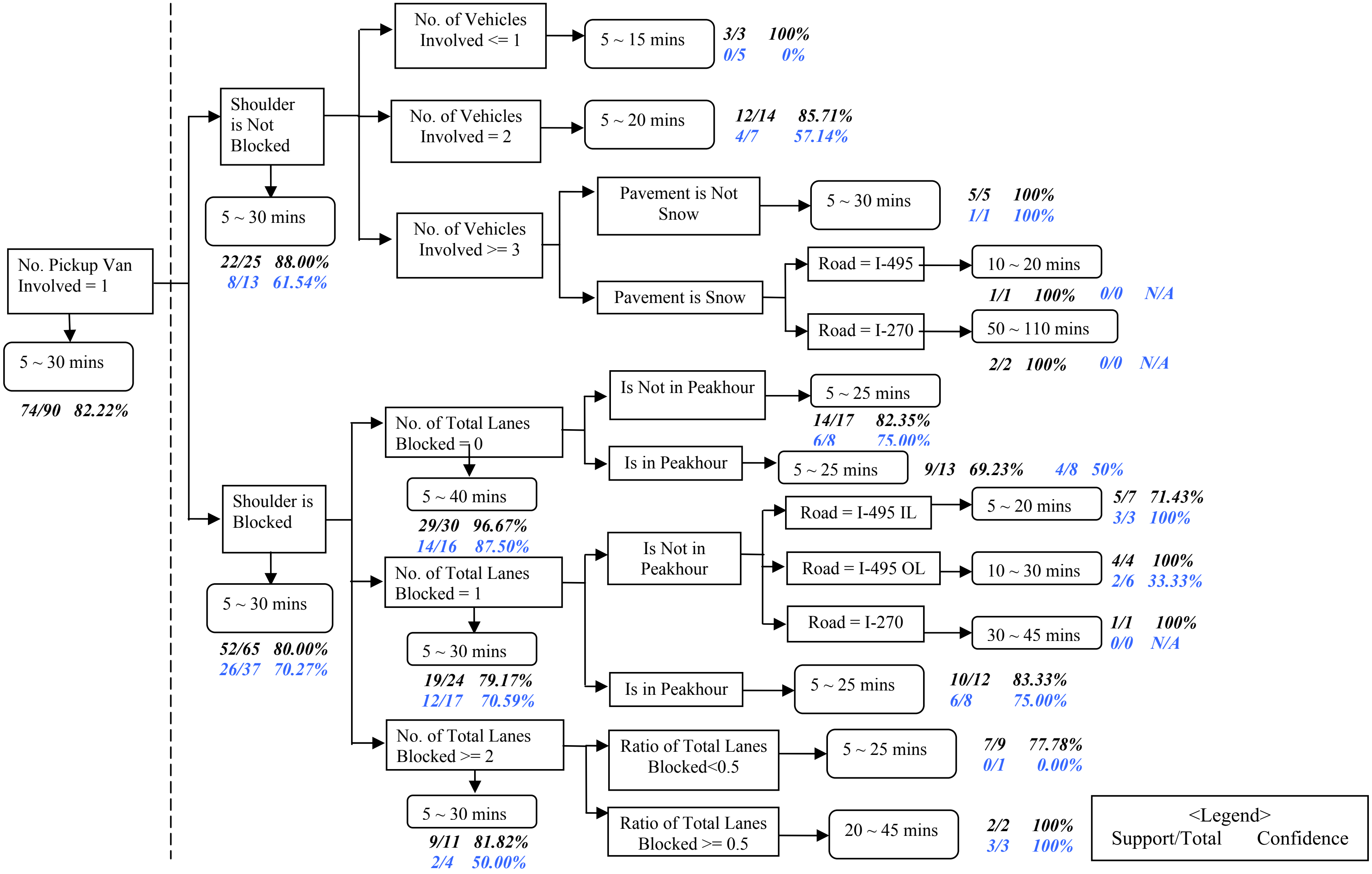


Figure A2.3(d) Rule Based Tree Model for Collision-Property Damage in Montgomery County

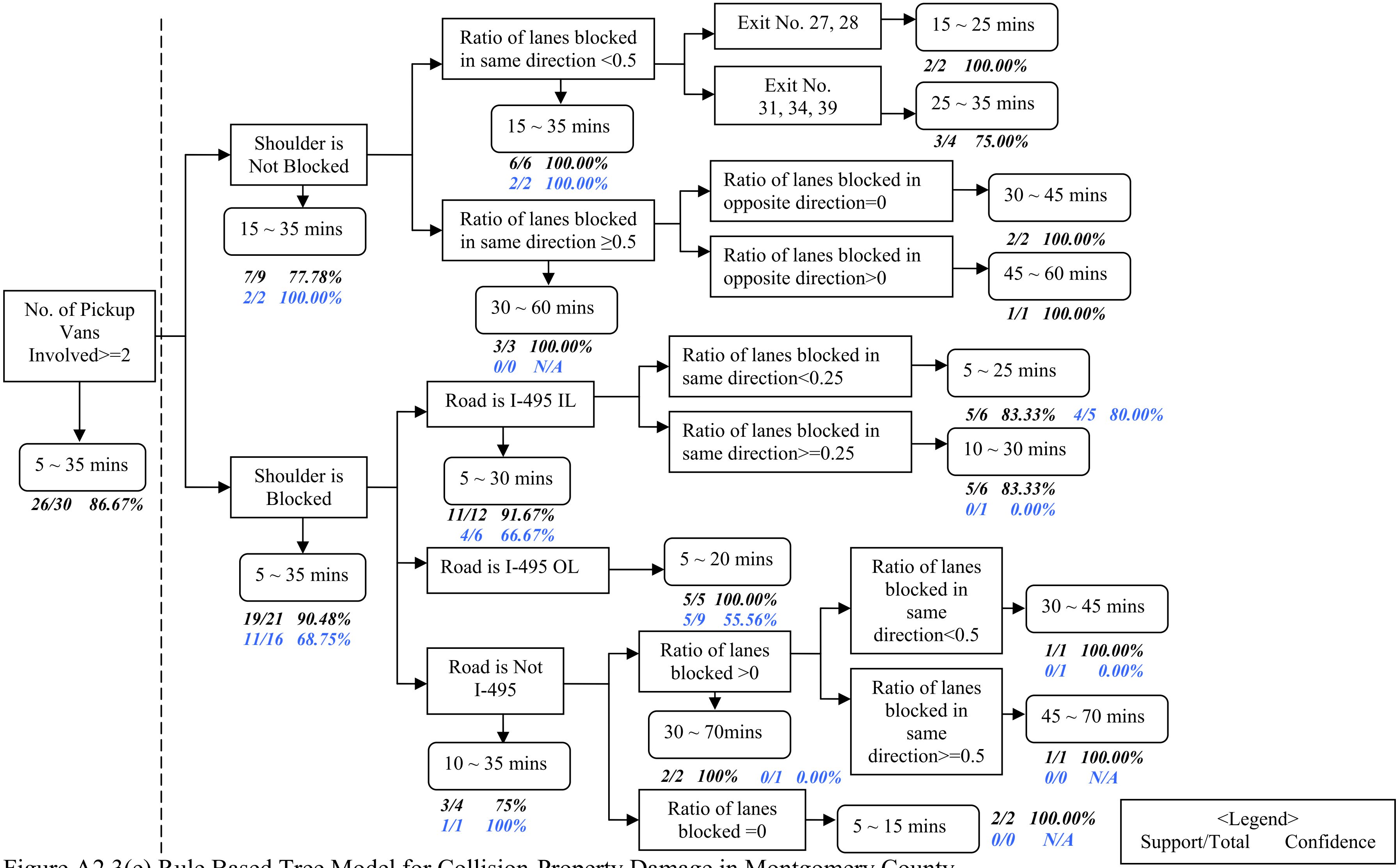


Figure A2.3(e) Rule Based Tree Model for Collision-Property Damage in Montgomery County

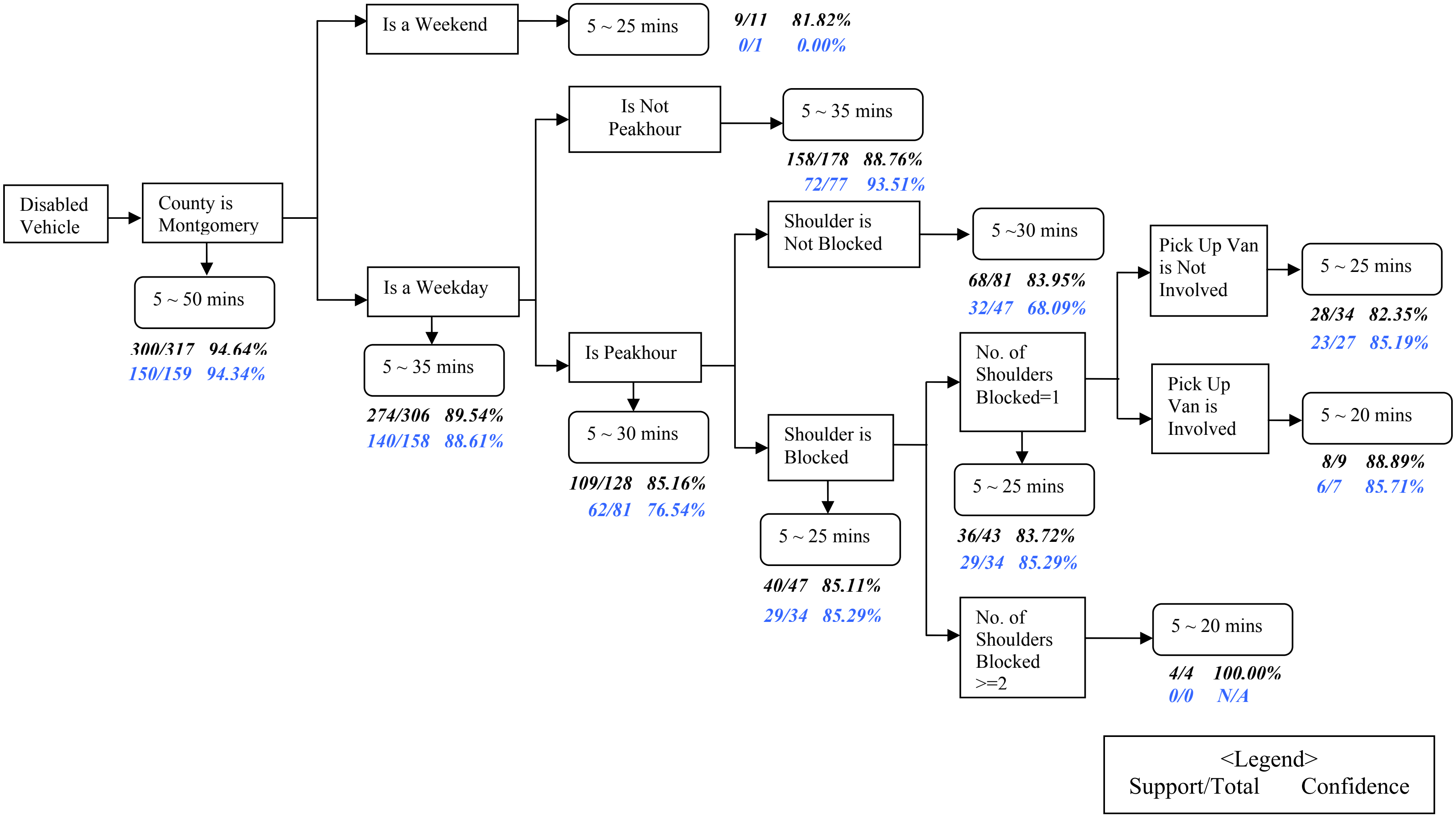


Figure A2.4 Rule Based Tree Model for Disabled Vehicles in Montgomery County

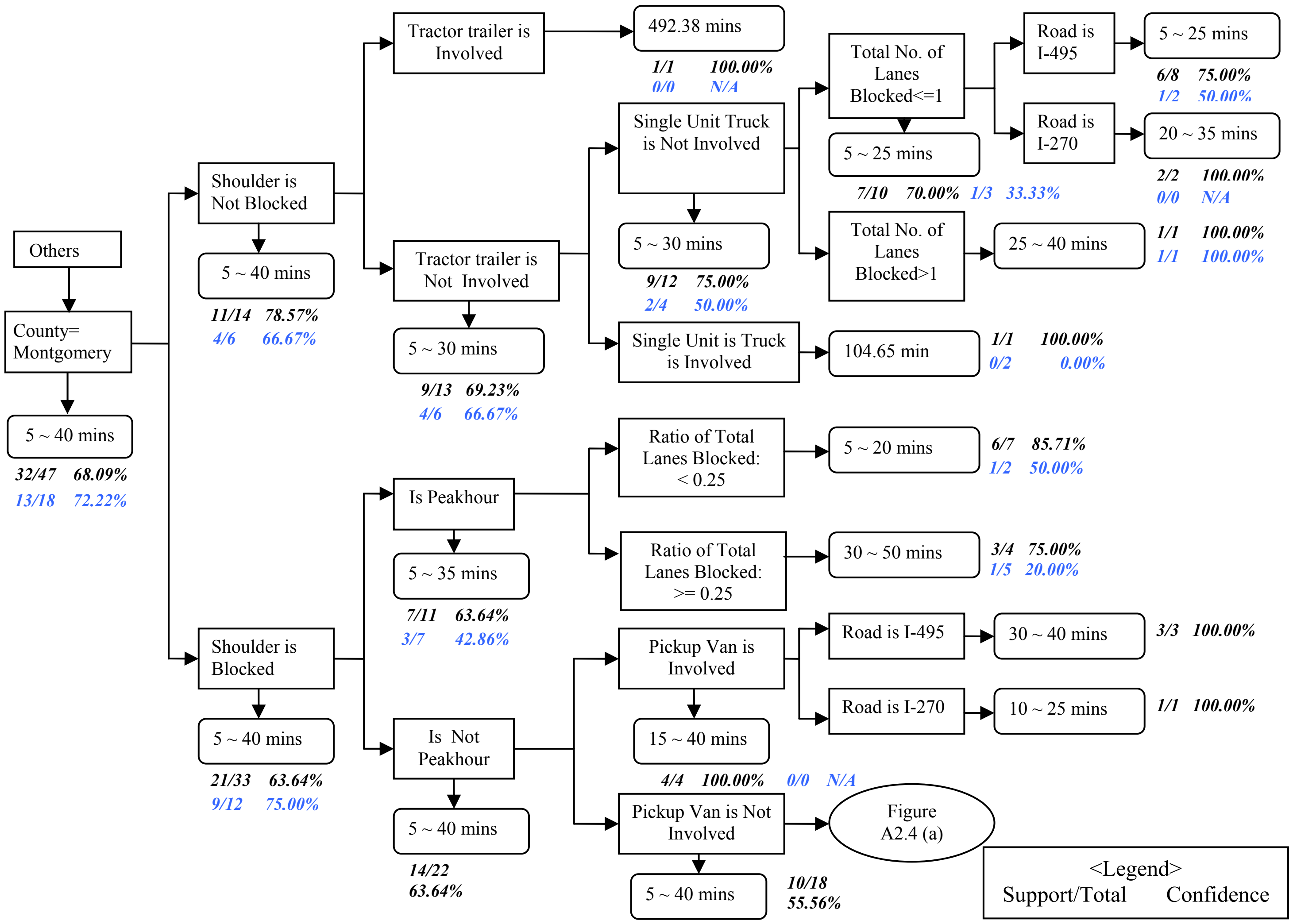


Figure A2.4 Rule Based Tree Model for Other Incident Natures in Montgomery County

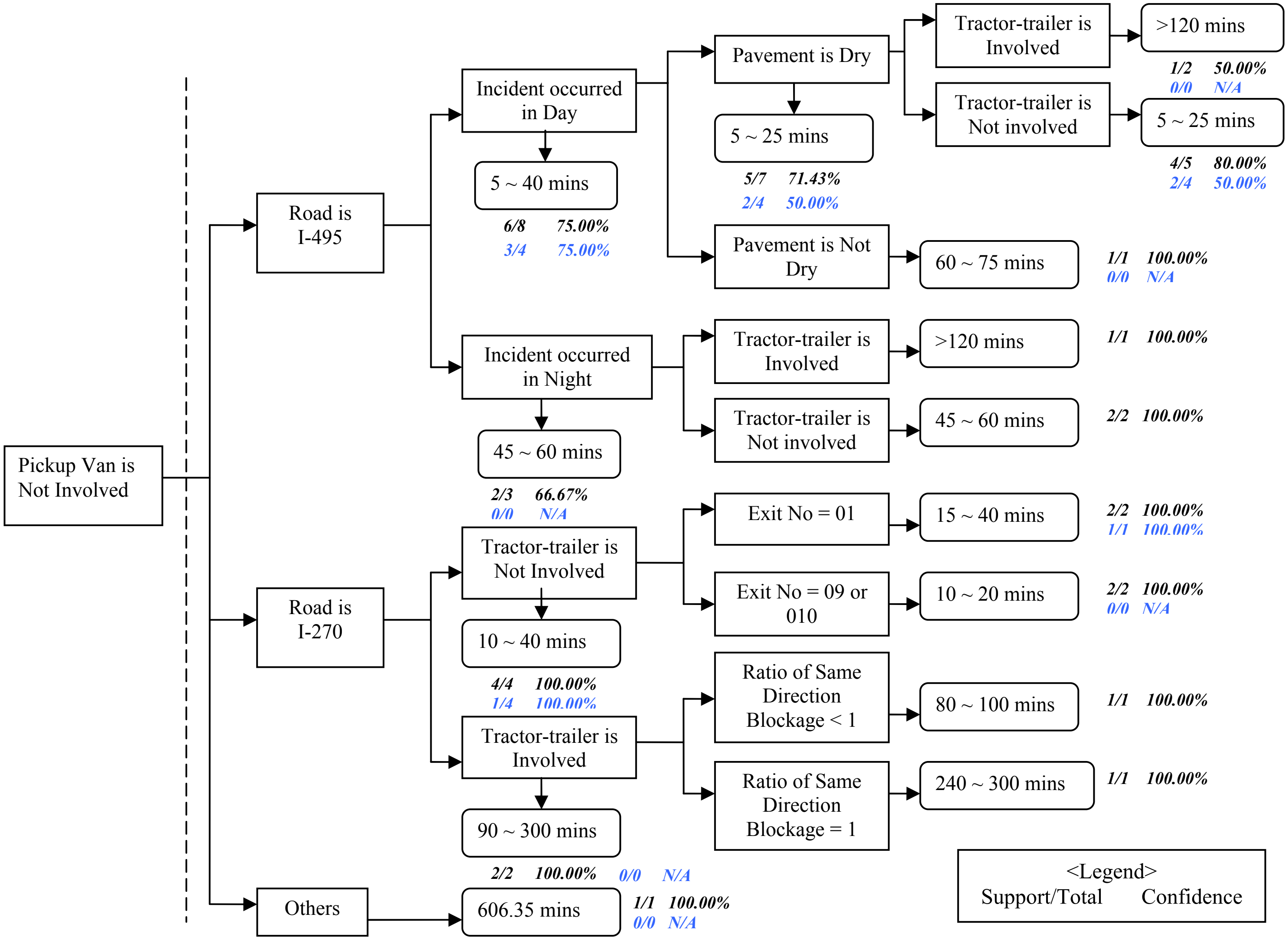


Figure A2.4(a) Rule Based Tree Model for Other Incident Natures in Montgomery County (Cont'd)

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.

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.

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.

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.

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). (1994). "Special Report 209: Highway Capacity Manual." 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.