Final report of ITS Center project: Incident duration forecasting

A Research Project Report

For the National ITS Implementation Research Center

A U.S. DOT University Transportation Center

Forecasting the Clearance Time of Freeway Accidents

Kevin Smith
Department of Civil Engineering

Dr. Brian L. Smith
Department of Civil Engineering
Email: briansmith@Virginia.EDU

Center for Transportation Studies
University of Virginia
CTS Website http://cts.virginia.edu
351 McCormick Road, P.O. Box 400742
Charlottesville, VA 22904-4742
434.924.6362

Smart Travel Lab Report No. STL-2001-01

Center for Transportation Studies at the University of Virginia produces outstanding transportation professionals, innovative research results and provides important public service. The Center for Transportation Studies is committed to academic excellence, multi-disciplinary research and to developing state-of-the-art facilities. Through a partnership with the Virginia Department of Transportation's (VDOT) Research Council (VTRC), CTS faculty hold joint appointments, VTRC research scientists teach specialized courses, and graduate student work is supported through a Graduate Research Assistantship Program. CTS receives substantial financial support from two federal University Transportation Center Grants: the Mid-Atlantic Universities Transportation Center (MAUTC), and through the National ITS Implementation Research Center (ITS Center). Other related research activities of the faculty include funding through FHWA, NSF, US Department of Transportation, VDOT, other governmental agencies and private companies.

Abstract

Freeway congestion is a major and costly problem in many U.S. metropolitan areas. From a traveler's perspective, congestion has costs in terms of longer travel times and lost productivity. From the traffic manager's perspective, congestion causes a freeway to operate inefficiently and below capacity. There are also environmental costs associated with congestion such as increased pollution and noise. Researchers have estimated that "non-recurring" congestion due to freeway incidents such as accidents, disabled vehicles, and weather events accounts for one-half to three-fourths of the total congestion on metropolitan freeways in this country.

The objective of this study is to develop a forecasting model that can predict the clearance time of a freeway accident. This can aid traffic managers in making decisions regarding the appropriate response to freeway incidents. Three models were investigated in this paper; a stochastic model, nonparametric regression model, and classification tree model. The stochastic model was not applied to forecasting future accidents due to the lack of a probabilistic distribution to fit the clearance time data. The Weibull and lognormal distributions have been applied to incident duration in the past, but were not applicable to the accident clearance time data used in this study. The other two models were developed but suffered from poor performance in predicting the clearance time of future accidents. However, the classification tree model appears to be well suited for forecasting the phases of incident duration given a database of incidents with reliable and informative characteristics.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Project Definition

Freeway congestion is a major and costly problem in many U.S. metropolitan areas. From a traveler's perspective, congestion has costs in terms of longer travel times and lost productivity. From the traffic manager's perspective, congestion causes a freeway to operate inefficiently and below capacity. There are also environmental costs associated with congestion such as increased pollution and noise. The type of congestion most people are familiar with is the "recurring" congestion patterns of rush hour. Traffic managers and politicians have been fighting this congestion for many years through the use of High-Occupancy Vehicle (HOV) lanes, ride-sharing programs, transit incentives, and ramp metering. However, the "non-recurring" congestion due to unpredictable incidents and events warrants immediate response. The actions taken by traffic managers require a full understanding of the nature and tendencies of freeway incidents.

1.2 Problem Rationale

Past researchers have estimated that "non-recurring" congestion due to freeway incidents such as accidents, disabled vehicles, and weather events accounts for one-half to three-fourths of the total congestion on metropolitan freeways in this country (Giuliano, 1989). Thus, the specific field of Incident Management has become an important component of traffic management. Incident Management involves the steps of clearing traffic incidents quickly and then minimizing the congestion effects on the traffic flow. Clearing an incident quickly involves managerial support among agencies, clear guidelines for action, and immediate identification of the incident. Minimizing the incident congestion involves the use of traveler information systems such as dynamic message signs and advisory radio, reversible direction lanes, and vehicle re-routing. Ideally, these techniques should be employed as soon as possible instead of waiting for the congestion to begin. The difficult task, from a traffic manager's perspective, is estimating the duration of the incident to help decide on the appropriate course of action.

This situation can be illustrated with an example. A traffic manager observes a freeway accident near a major interchange that has resulted in the closure of one lane. The manager knows that if the queue of stopped vehicles reaches the major interchange, they will need to activate the variable message signs to alert motorists on both roadways about the stopped traffic at the interchange. Queuing models are in place currently that predict queue characteristics based on demand flows, speeds, and available capacity. However, the queue length depends on the length of time that the incident is active and the capacity is reduced by the one lane closure. In this case, if the manager can anticipate the clearance time of the accident, they can determine the length of the queue and make a decision on the activation of variable message signs.

At the start of a freeway incident, traffic managers have an impression on the nature of the incident. These people have constant interaction with traffic incidents and may be able to use past experiences to predict the duration of the current incident. Only a few studies have been undertaken to provide a quantitative model to support the managers' projection based on personal experience.

1.3 Project Purpose and Scope

The goal of this project is to develop methods to forecast the clearance time of a freeway accident based on its characteristics. The accident data to support model development will come from the Smart Travel Lab at the University of Virginia. The Smart Travel Lab receives traffic data from VDOT's Smart Traffic Center in Virginia Beach, VA. It is anticipated that the forecasting models will be applicable to any freeway system.

It should be stressed that this project will attempt to forecast accident clearance time, which is the length of time that emergency and other personnel are present on the freeway. It is assumed that the clearance time is a good indication of the total duration of an incident. The importance of understanding incident duration is that is a major factor in determining queues, delay, and other non-recurring congestion effects.

1.4 Report Overview

The remainder of this report is composed of the following chapters:

- Chapter 2 – a review of past research on forecasting phases of incident duration
- Chapter 2 – an overview of the different forecasting techniques used in this project
- Chapter 3 – the experimental setup of the project including data collection
- Chapters 4, 5, 6 – an evaluation of the three forecasting models
- Chapter 7 – the project conclusions and contributions to transportation engineering
- Chapter 7 - recommendations and suggested future research

Chapter 2: Review of Relevant Literature

2.1 Freeway Traffic Incidents

A freeway incident is defined as any planned or unplanned event that effects the traffic flow on the roadway (Sethi, et al., 1994). Some examples of freeway incidents include accidents and crashes, disabled or abandoned vehicles, vehicle fires, weather events, road debris, construction, etc. The Highway Capacity Manual (TRB, 1994) states that incidents

- Disrupt the level of service being provided,
- Reduce capacity radically, and
- Present hazards to motorists, particularly those directly involved.

Incidents reduce the level of service by lowering speeds (Giuliano, 1989). Other motorists slow down to allow emergency vehicles to respond, avoid debris and vehicles, and also to "rubberneck" or look at the incident. Capacity is also reduced during incidents due to lane closures or impediments. One study claims that a single lane blockage on a three lane roadway reduces capacity by fifty percent (TRB, 1994). Additionally, it should not be overlooked that freeway incidents do result in fatalities, personal injuries, and property damage.

The duration of an incident is composed of four important and distinct components; detection, response, clearance, and recovery (TRB, 1994) as shown in Figure 2-1.

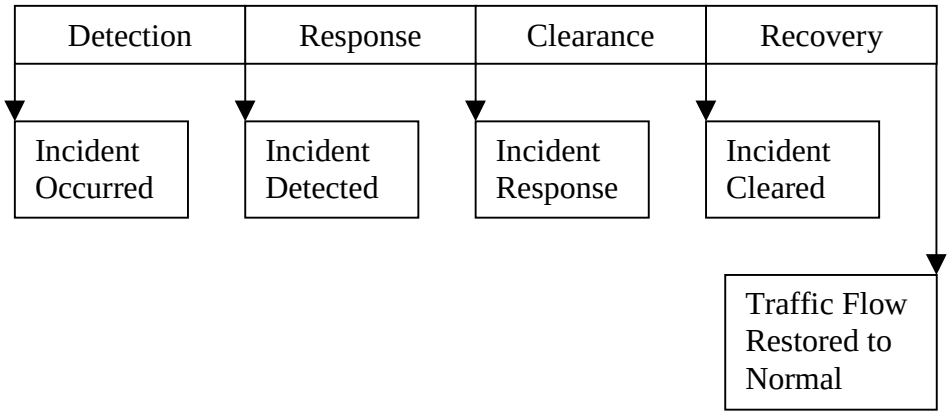

Figure 2-1: The 4 phases of a freeway incident over time.

- Detection Phase - the period of time between the occurrence of the incident and detection by the traffic managers, police, or freeway response team. Included in this phase is the verification of the incident as severe enough to warrant a response.
- Response Phase - the period of time between the detection of an incident and the arrival of emergency or response vehicles.
- Clearance Phase - the period of time when responding agencies treat victims, close lanes, and eventually remove vehicles and debris.

- Recovery Phase - the period of time after the clearance of an incident for the traffic flow to return to normal conditions.

Together the four phases represent the total duration of the incident or the period of time from the occurrence of an incident to the return of normal traffic flow conditions.

Even though research has dissected incident duration into these four phases, it is possible for an incident to not exhibit all of the phases. For example, an incident may not have a response phase if police or response teams discover the incident while patrolling the area. Likewise, if an incident is observed occurring on a surveillance camera, there will not be a detection phase. Also, for minor incidents that have short detection, response, and clearance phases there might not be an effect on traffic flow and no noticeable recovery phase. Finally, it would appear that the clearance phase would always be present, but some minor incident may not necessitate emergency vehicles or police and can treated by the people involved without assistance.

2.2 Past Research on Incident Duration Prediction

There has been research in the past looking into predictive techniques applied to incident duration. The results from these studies have been mixed and comparisons between different methods are difficult due to data issues. Almost each study uses a different source of incident data with different descriptive variables and reporting techniques. Some studies suffer from a small sample size, and others from inaccurate data or data with missing values.

2.2.1 Probabilistic Distributions

A simple method to predict incident duration is to model the duration value as a random variable and attempt to fit a probability density function to the data. From this distribution, the traffic manager has an idea on the mean and variance of incident duration. Another useful piece of information is the ability to say there is an x probability of the incident lasting over y minutes.

In 1987, Golob et al. analyzed freeway accidents that involved trucks. The data used in the analysis was 332 freeway and 193 ramp accidents around Los Angeles, California over a two-year period. The authors theorized that an incident is comprised of the sequential phases listed in the previous section, but added that the length of each phase is influenced by the length of the preceding phases (Golob, et al., 1987). From this hypothesis, they were able to theorize that the total duration of an incident is modeled according to a lognormal distribution. Kolmogorov-Smirnov tests of the truck data supported the lognormal distribution of all incidents and each specific incident type (Golob, et al., 1987). Lacking in the analysis was a test of the assumption that each incident phase is time-dependent on the previous phases. This assumption could not be tested since the incident data only contained the total duration of the incident.

Other research studies by Giuliano in 1989, Garib et al. in 1997, and Sullivan in 1997 have supported the use of a lognormal distribution to describe freeway incident duration. Jones et al. used a similar distribution, the log-logistic distribution, in 1991 to a specific data set from the Seattle area. Nam and Mannering in 2000 found that the Weibull distribution could also be used to describe some incident data. The common

theme among all three of these distributions is a shift to the left that shows a larger proportion of short-duration incidents (see Figure 2-2).

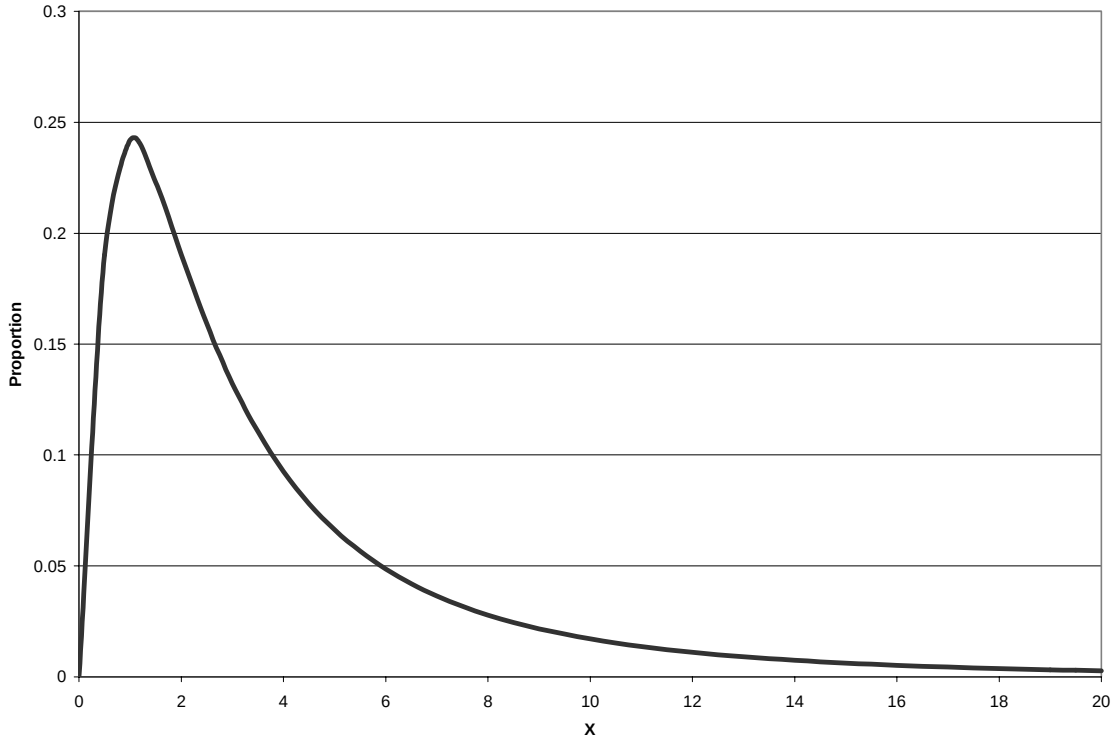

Figure 2-2: General shape of a lognormal probability distribution.

Recent research by Ozbay and Kachroo in 1999 went one step further in terms of the probabilistic distribution of freeway incidents. Using 650 incidents from Northern Virginia over a one-year period, it was found that the incident duration had a shape similar to a lognormal distribution, but was rejected by several statistical significance tests. However, they found that if the set of incidents is divided into subsets of incidents that have the same type and similar severity a normal distribution of duration is found. This conclusion supports the theory that the duration of similar incidents are random variables.

2.2.2 Linear Regression Models

Another simple prediction model is a linear regression function. In terms of incident duration, the regression usually has a number of binary variables that represent certain incident characteristics. A 1991 unpublished paper from Northwestern University (Ozbay and Kachroo, 1999) studies incident clearance data of 121 incidents from the Chicago area and found 9 statistically significant variables: heavy wrecker (WRECKER), assistance from other response agencies (OTHER), sand/salt pavement operations (SAND), number of heavy vehicles involved (NTRUCK), heavy loading (HEAVY), liquid or uncovered broken loadings in heavy vehicles (NONCON), severe injuries in

vehicles (SEVINJ), freeway facility damage caused by incident (RDSIDE), and extreme weather conditions (WX). Two other variables were deemed useful but not statistically significant: response time (RESP) and incident report (HAR). The regression model developed has the form

Clearance Time = 14.03 + 35.57(HEAVY) + 16.47(WX) + 18.84(SAND) – 2.31(HAR) + 0.69(RESP) + 27.97(OTHER) + 35.81(RDSIDE) + 18.44(NTRUCK) + 32.76(NONCON) + 22.90(SEVINJ) + 8.34(WRECKER)

Ozbay and Kachroo do not report on the validity of this regression model, such as r-squared values, or any testing techniques.

In 1997, Garib et al. also developed a linear regression model to predict incident duration. The analysis consisted on 205 incidents over a two-month period from Oakland, California, and found six significant variables: number of lanes affected (X_1), number of vehicles involved (X_2), binary variable for truck involvement (X_5), binary variable for time of day (X_6), natural logarithm of the police response time (X_7), and a binary variable for weather conditions (X_8). The log-based regression model is given by

$$\text{Log(Duration)} = 0.87 + 0.027\ X_1\ X_2 + 0.2\ X_5 - 0.17\ X_6 + 0.68\ X_7 - 0.24\ X_8$$

The adjusted R-square value of this regression model is 0.81 (Garib, et al., 1997). The authors make the conclusion that the model is thus 81% accurate at predicting incident duration without performing any tests on incidents not used to develop the model.

2.2.3 Conditional Probabilities

Another use of probability in incident duration is to develop conditional probabilities. Traffic managers may be interested in the probability of an incident lasting 30 minutes given that it has already been active for 15 minutes, or similar cases. Most research has focused on unconditional probabilities such as the probability of an incident lasting exactly 30 minutes. Jones et al. reported on conditional probabilities in 1991. Nam and Mannering followed up on the concept by applying hazard-based models developed in the biometrics and industrial engineering fields to incident duration. Hazard-based models also use conditional probabilities to find the likelihood that an incident will end in the next short time period given its continuing duration (Nam and Mannering, 2000). The use of conditional probabilities is based on the theory developed by Golob et al. in 1987 that each incident phase is influenced by the length of previous phases of the incident. To date, these types of models have been used to find the accident characteristics that have the greatest influence on incident duration instead of explicitly forecasting the duration for empirical testing purposes.

2.2.4 Time Sequential Models

A 1995 paper by Khattak et al. makes the statement that most incident duration prediction models have no operational value since they require knowledge about all

incident variables. In the field, accident information is acquired sequentially and this progression should be reflected in the model.

To develop the time sequential model, the authors identified ten distinct stages of the incident duration based on the availability of information (Khattak, et al., 1995). The length of time for each stage differs for each incident, but it is truncated after a maximum of 10 minutes. Each stage has a separate truncated regression model, and the models progressively add more variables. The time sequential model was not tested or validated in the study due to a small sample size of 109 freeway accidents. This study intends to demonstrate the methodology of time sequential models rather than show its performance in traffic operations (Khattak, et al., 1995). It does not appear that this model approach was ever applied to a large sample size or used in any future study on forecasting incident duration.

2.2.5 Decision Trees

All of the methods of incident duration prediction discussed so far have had a probabilistic basis. This may be preferable in many cases because you can add confidence intervals or other probabilities to the forecasted output. For example, a model could tell a traffic operator that the current incident will last 20 minutes with a 95% confidence level. However, if these models do not produce accurate results, it is of no use to the operator to know that an incident will last 20 minutes with a 60% confidence level. In this case new models and methods for duration prediction are needed that can identify patterns in data without an underlying probabilistic distribution.

One such pattern-recognition model that has recently been applied to incident duration prediction is a decision tree. A specific type of decision tree, a classification tree, will be discussed thoroughly in the next section. In 1999, Ozbay and Kachroo used decision trees to predict incident clearance times in the Northern Virginia region. Their published work describes a comprehensive study in a step-by-step manner to show all of the data collection and analysis processes. After collecting a large sample of incident data, the authors followed a series of trial prediction methods with poor results. They first tried linear regression techniques with a low R-square value (about 0.35), and then found the duration values did not follow either a lognormal or log-logistic distribution (Ozbay and Kachroo, 1999). The next step was to develop a decision tree similar to the construction of the classification and regression trees (CART) developed by Breiman et al. and will be defined later in this chapter

Before constructing the decision tree, Ozbay and Kachroo first determined the significant independent variables using ANOVA tests of the data. Some types of incidents, such as HAZMAT and weather related incidents, the number of samples in the database was too small to make any conclusion on the variables' importance. Thus, these variables were also excluded from the construction of the decision tree. It should also be noted that the intended output of the model is an average duration of past incidents that are similar to the current. Other outputs could be a range of possible durations or a minimum and maximum duration value.

A portion of the final decision tree is included in Figure 2-3 (Ozbay and Kachroo, 1999).

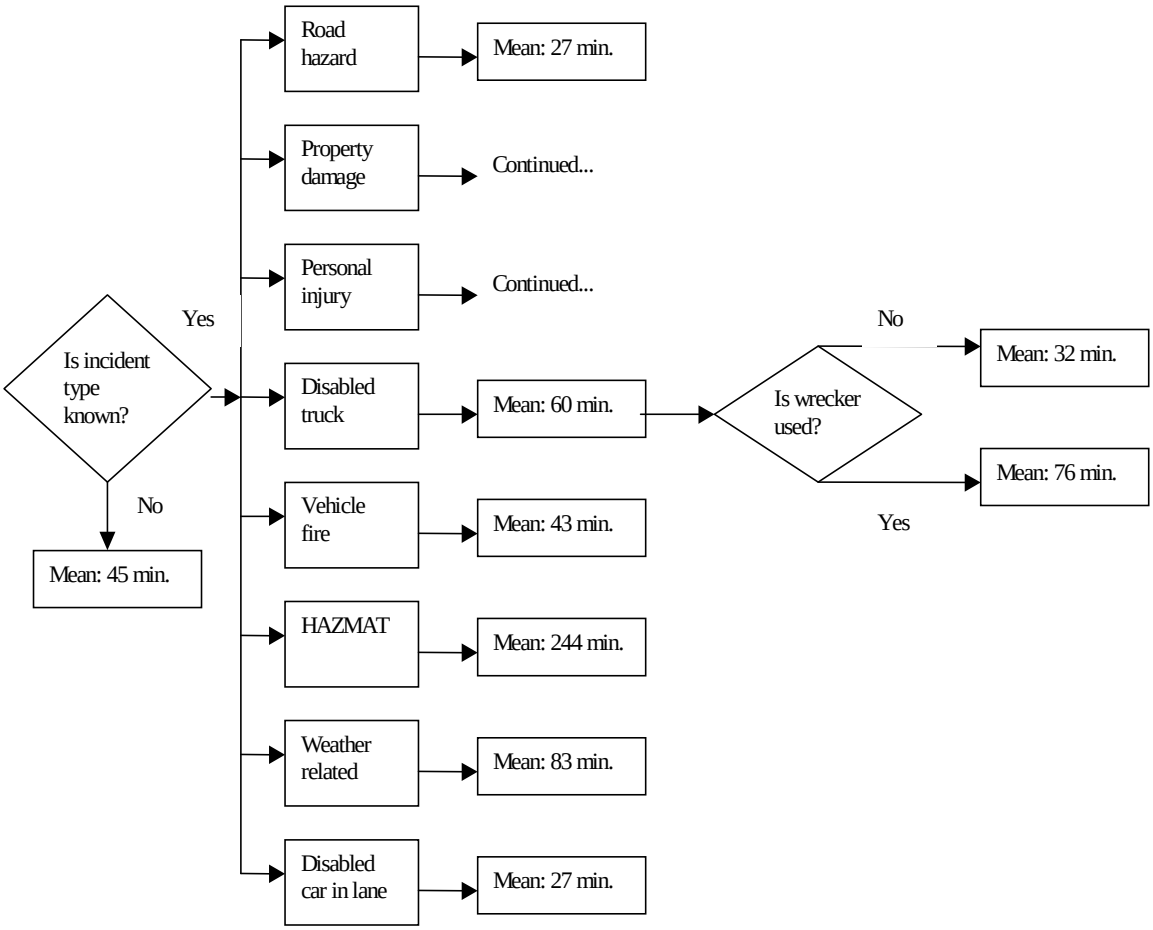

Figure 2-3: Decision tree for incident clearance time prediction (Ozbay and Kachroo, 1999).

The decision tree above is the main tree where the first decision is based on incident type. For example, in the above tree an incident where the type is unknown is immediately given a mean duration of 45 minutes. A disabled truck is assumed to have a clearance time with a mean of 60 minutes. But, if more information is available about the use of a wrecker the prediction is refined to 32 minutes for no wrecker or 76 minutes for a wrecker. The decision tree can handle differing levels of information knowledge about the current incident.

Ozbay and Kachroo tested the decision tree and found a satisfactory performance where 44 out of 77 test incidents were predicted with less than 10 minutes of prediction error. One important finding was that there were a number of outliers that had a large difference between actual and predicted durations. These outliers have the potential to skew some performances of measure like the Mean Absolute Error (MSE) that average the difference in actual and predicted durations for all test incidents.

2.3 New Forecasting Techniques

The following sections will describe two fairly new forecasting techniques that will be used for this project. The techniques are classification tress and nonparametric regression.

2.4 Classification Trees

A classification tree is a type of decision tree that represents of a number of yes/no questions that sort object into distinct classes. The difference between a classification tree and a decision tree similar to the one described above is that the classification tree assigns a class instead of a deterministic value. When a mechanic is inspecting a car to find why it doesn't run they progress through a checklist of different parts to inspect. The result of each question leads the mechanic down a different path. If the checklist were plotted as a number of nodes and links, a classification tree would be formed (see Figure 2-4).

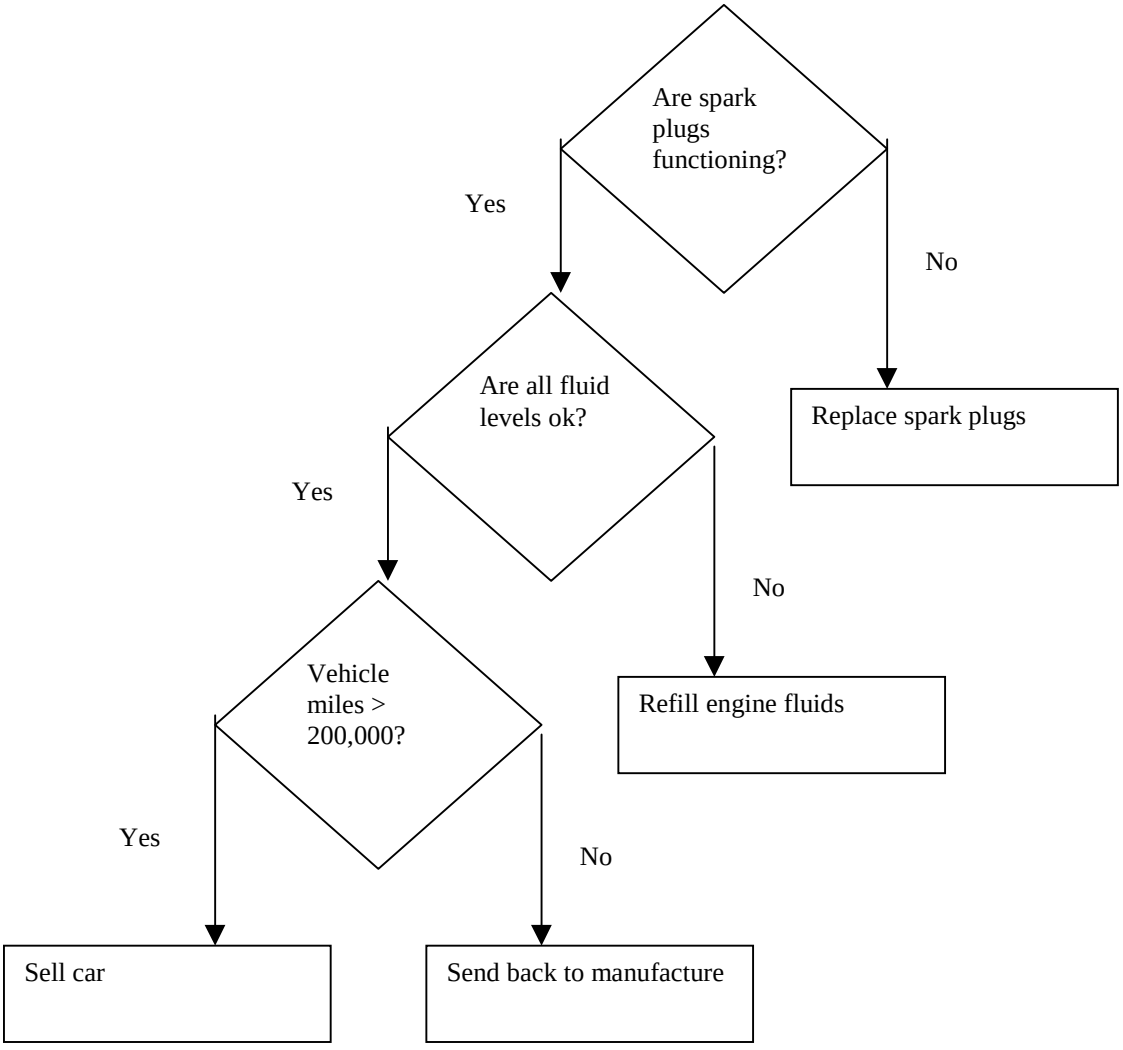

Figure 2-4: Example of a mechanic's classification tree.

The goal of a classification tree is to take a set of objects with characteristics or measurements and find a systematic method of assigning the objects to a number of distinct classes. This systematic method of splitting a sample into two sub-samples is referred to as a classifier (or classification rule), and the combination of numerous classifiers forms a classification tree (Breiman, et al., 1984).

Data is the most important key to constructing a classifier (Breiman, et al., 1984). As seen in the example above, the mechanic's checklist is based on past experience with similar vehicles and problems. The data used in a classifier can have either numerical or categorical variables. An example of a numerical variable would be the number of miles on the vehicle, while a categorical variable would be the type of engine in the vehicle. The ability to find patterns in both numerical and categorical variables makes decision trees very useful for modeling real-world processes like diagnostic problems (Park, 1995).

To date, there has been no published study on the use of classification trees to forecast incident duration. Ozbay and Kachroo have shown that decision trees are a promising technique that needs to be developed further. Classification trees have the advantage of sorting classes of incident duration that can be defined by the user. In addition, the categorical output may be better suited for the categorical inputs that are used in describing the nature of a freeway incident.

2.4.1 Tree Construction

This section deals specifically with binary tree structured classifiers, which consist of simple yes or no splitting decisions. The classifier splits a sample into two descendant subsets. A classification tree is formed by repeated splits of descendant samples (Breiman, et al., 1984). For a binary tree structure, the splitting rule is of the nature, "Is x = y?" with a path for objects where this statement is true and another path for a false statement. At some point a tree must stop splitting and terminal subsets (or nodes) are declared. Each terminal subset is the assigned to a distinct class and it is possible for two or more subsets to belong to the same class (Breiman, et al., 1984). See Figure 2-5 for the structure of a classification tree.

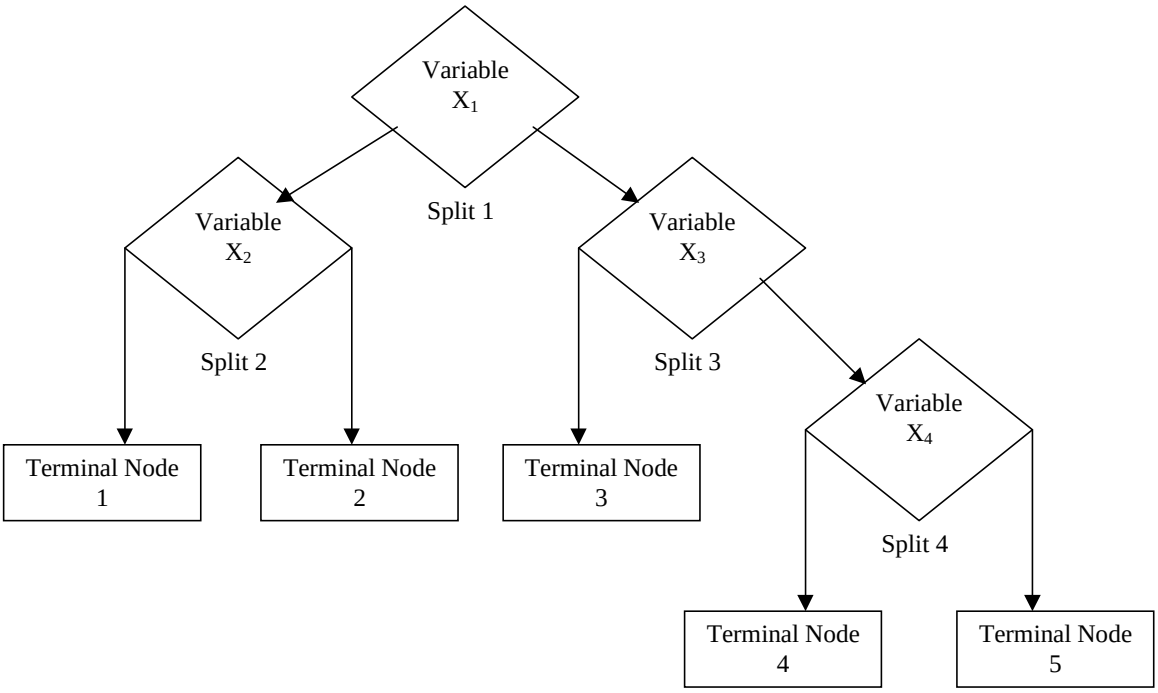

Figure 2-5: Classification tree structure.

The construction of a classification tree is dependent on three processes (Breiman, et al., 1984):

- The selection of the splitting criteria,
- The decision to declare a terminal node or continue splitting, and
- The assignment of each terminal node to a class.

2.4.2 Classification and Regression Tree (CART) Software

The CART software program is based on the decision tree methodology developed by Breiman, Friedman, Olshen, and Stone in 1984. The software program incorporates a binary-recursive portioning algorithm, where parent nodes are always split into two child nodes and each child node is considered a future parent node if it is determined that a split is needed (Salford Systems, 2000). One key feature of the CART program is that the classifiers are nonparametric and thus do not require a prior knowledge about the probabilistic distribution of the underlying data (Salford Systems, 2000).

The first problem in constructing a classification tree is the method to determine the splits that will divide the parent node data into two smaller samples. CART is based on the fundamental idea that each split should be selected so that the data in each descendant subset are "purer" than the data in the parent node (Breiman, et al., 1984). This measure of impurity is based on the proportion of cases in the node belonging to each class. Node impurity is largest when all classes are equally mixed together and smallest when the node contains only one class (Breiman, et al., 1984). Consider a parent node t, which contains data belonging to J number of classes. The proportions for each class in the node are given by

$p(j \mid t)$ where $j = 1, 2, \ldots, J$ and $\Sigma\, p(j \mid t) = 1$ for all j

Based on the Gini diversity index, the impurity function i(t) for node t is given by

$$i(t) = 1 - \sum_j p^2(j \mid t)$$

Thus, consider a parent node t that uses splitting rule δ to split into two nodes t_L and t_R, where p_R and p_L are the respective proportions of cases from the parent node in each sub-sample. The original impurity of the parent node is given by i(t), and the impurity for the two new nodes by $i(t_L)$ and $i(t_R)$. The decrease in impurity of this split δ is given by

$$\Delta i(\delta, t) = i(t) - p_L\, i(t_L) - p_R\, i(t_R)$$

When selecting a node split it is important to note that there are a definite number of possible splits of the data (Breiman, et al., 1984). Thus, CART uses a brute force method that examines each possible split for each possible variable. Using the impurity function discussed above, the split with the largest decrease in impurity is chosen for that node. The assumption is that the impurity will never increase from a split. If the impurity can not be decreased, then a terminal node is declared and that portion of the tree stops growing. The class assigned to the terminal node is based on which class became "purer" from the final split. It should be noted that using the Gini rule for class

assignment is different from the plurality rule. Thus, it is possible that a terminal node will be assigned one class, when there is a higher proportion of another class in the node.

The conclusion of Breiman et al. was that the properties of the final classification tree are insensitive to the specific splitting rule used in development. A much more criterion is the pruning method, used to determine the best size tree to use. CART constructs the largest possible tree such that the impurity of all terminal nodes can not be decreased. Pruning involves taking the large tree and recombining splits into parent nodes (Cios, et al., 1998). The pruning moves upward and produces a decreasing sequence of sub-trees. The sub-trees are then tested for their predictive accuracy using a separate test data set or cross-validation techniques (testing techniques will be discussed later). The sub-tree that has the lowest misclassification rate of the test data is selected as the optimal classification tree for that data (Breiman, et al., 1984). Pruning is widely accepted in the construction of classification trees as one method to avoid overfitting a test data set (Cios, et al., 1998). Overfitting is the phenomena where a process of fitting a model or method to a data set goes too far and attempts to define or explain every instance in the data set.

Since classification trees are data-dependent, there must be adequate testing of the tree with data not used to develop the classifiers. CART uses two testing procedures: learning samples and cross validation (Salford Systems, 2000). When there is a large data set to develop a tree, the sample can be divided into learning and testing sub-samples. CART uses the learning sample to develop potential trees during the pruning process, while the testing sample is used to compare the tree performance and select the optimal tree. When there is a small data sample available, CART uses cross validation for testing. In this process the data set is divided into ten equal samples. Nine of the samples are then used as a learning sample, with the remaining one used as a testing sample. This process of growing a tree and testing continues until each sample has been used for testing. The results from these 10 trees are then combined to form error rates for trees of each possible size (Salford Systems, 2000). The optimal testing situation would be to use a learning and testing sample, and CART has historically performed about 10 to 15 percent better using testing samples than cross validation (Salford Systems, 2000).

2.5 Non-parametric Regression

Nonparametric regression is a forecasting technique that has been used in the past for predicting traffic flows in the short-term. The technique has provided positive results and is considered a viable choice for traffic condition forecasting for freeway management systems, and is especially important when there are difficulties developing parametric models (Smith, et al., 2001). The basis of nonparametric regression is to make current decisions based on past, similar experience. Thus, it relies heavily on data describing the relationship between dependent and independent variables. The basic approach is to locate the state of the current system (as defined by the independent variables) in a neighborhood of past, similar states. Once a neighborhood is defined, the past cases in the neighborhood are used to estimate the value of the current dependent variable (Smith, et al., 2001).

To date, there has been no published study of the use of nonparametric regression for predicting incident duration. For such an application, the system state of an incident

can be described using a number of independent variables such as time of day and the number of vehicles involved. The dependent variable and forecasting output would be the duration of the incident. One attractive feature of nonparametric regression as a forecasting tool is that the knowledge of the relationship is in the data instead of the model (Smith, et al., 2001)

The key to the effective use of nonparametric regression is defining an appropriate neighborhood and then generating a forecast based on the cases within the given neighborhood.

2.5.1 Neighborhood Definition

The accuracy of nonparametric regression is dependent directly on the quality of the neighborhood and its ability to include similar cases (Smith, et al., 2001). The two basic approaches to defining neighborhoods are kernel and nearest neighbor (Altman, 1992). Kernel neighborhoods have a constant bandwidth, and thus occupy a specific range on the independent variable space (Smith, et al., 2001). Nearest neighbor neighborhoods are defined as containing a constant number of past cases. This is commonly referred to as k-nearest neighbor (KNN) nonparametric regression, where k is the number of past cases used to define the neighborhood (Smith, et al., 2001). The two methods of neighborhood definition are best seen on a graph. Figures 2-6 and 2-7 show a sample of data points with corresponding independent and dependent variable values. In this example the problem is how to define a neighborhood for an independent variable value of 35 (dashed line). Figure 2-6 uses a kernel size of 6 to return a neighborhood of 16 data points. Figure 2-7 uses a nearest neighbor approach with k=8 to return a neighborhood of 8 data points.

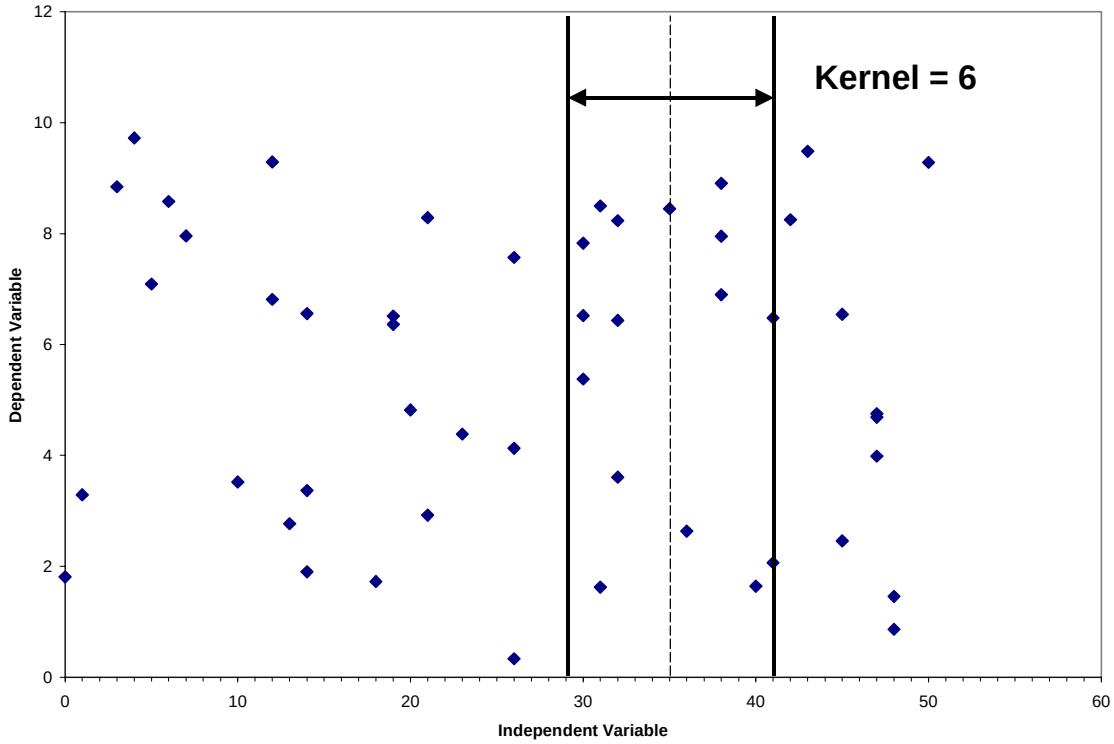

Figure 2-6: Example of a kernel neighborhood of size 6.

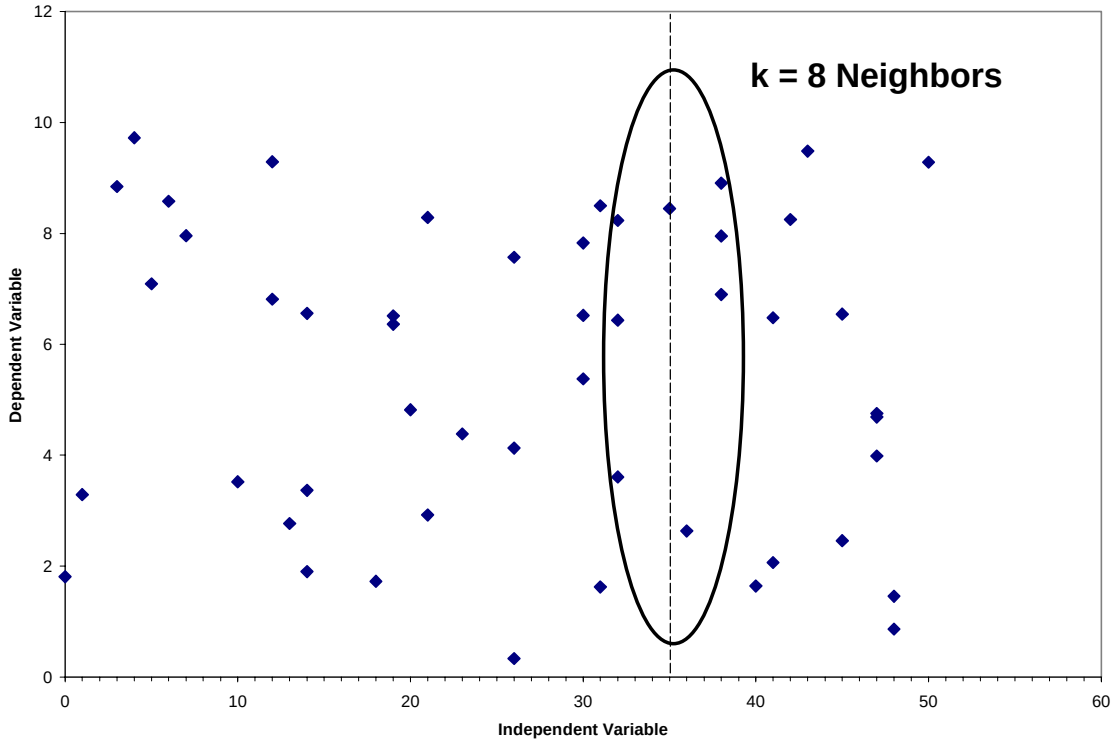

Figure 2-7: Example of a nearest neighbor neighborhood (k=8).

An issue with neighbor definition is how to measure the distance between cases in the database. The basic method is to use the Euclidean distance to measure proximity. This method is applicable for a single numerical independent variable. When more than one independent variable is present, it may be beneficial to add weight factors to each variable to determine the total distance. Assigning weight factors to rank variables is heuristic in nature and requires careful consideration by the model developer (Smith, et al., 2001). A similar situation is present when independent variables are categorical as opposed to numerical in value.

2.5.2 Forecast Generation

Once a neighborhood of past cases has been defined, it is necessary to generate a prediction based on those neighbors. A straightforward method is to compute a simple average of the dependent variable values of the cases in the neighborhood. The weakness of this approach is that it ignores the distance metric information developed in the neighborhood creation (Smith, et al., 2001). A more appropriate approach may be to weigh the average so that past cases nearer to the current case should have more importance in generating the forecast. Other approaches involve linear regression of the dependent values of the cases in the neighborhood, and other weighting techniques. As with neighborhood definition, forecast generation has many possible approaches and the final choice of a technique should be thoroughly tested and evaluated.

2.6 Summary

The purpose of this chapter was to investigate past attempts at forecasting incident duration and present new methods that will be developed in this project. The next chapter will present the structure of the project and information on the data to be used for model development and testing.

Chapter 3: Research Framework

3.1 Methodology

The main goal of this study is to investigate models to forecast the clearance time of freeway accidents. The two main methods that will be discussed in depth are nonparametric regression and classification trees. Below is the methodology to construct and evaluate the forecasting methods.

- Collect accident data from a specific freeway or freeway system for an extended time period.
- Clean data by removing entries with missing or unrealistic values.
- Identify potential independent variables and determine their significance to accident clearance time.
- Select appropriate forecasting models for evaluation.
- Apply forecasting models to the accident data to predict the clearance time.
- Evaluate the performance of each forecasting.
- Make recommendations on the use of the developed models for forecasting accident clearance time.

3.2 Data Source

The majority of the forecasting models studied in the previous sections have had an empirical rather than theoretical basis for model development. Having a large sample of past accidents with reliable information is the most important key to producing accurate predictions.

The accident data used in this project was obtained from the Smart Travel Lab located at the University of Virginia in Charlottesville, Virginia. The Smart Travel Lab was created through a partnership of the Virginia Department of Transportation (VDOT) and the Department of Civil Engineering at U.Va. The Lab is a state-of-the-art facility for research and education in the field of Intelligent Transportation Systems (ITS). The Lab maintains a number of direct data connections with VDOT facilities around the state. One such VDOT facility that shares data with the Lab is the Hampton Roads Smart Traffic Center (HRSTC) in Virginia Beach, Virginia. The HRSTC is a freeway management system that monitors traffic in the Norfolk and Virginia Beach region using 203 detector stations and 38 surveillance cameras. This area encompasses the I-64 and I-264 corridors (see Figure 3-1).

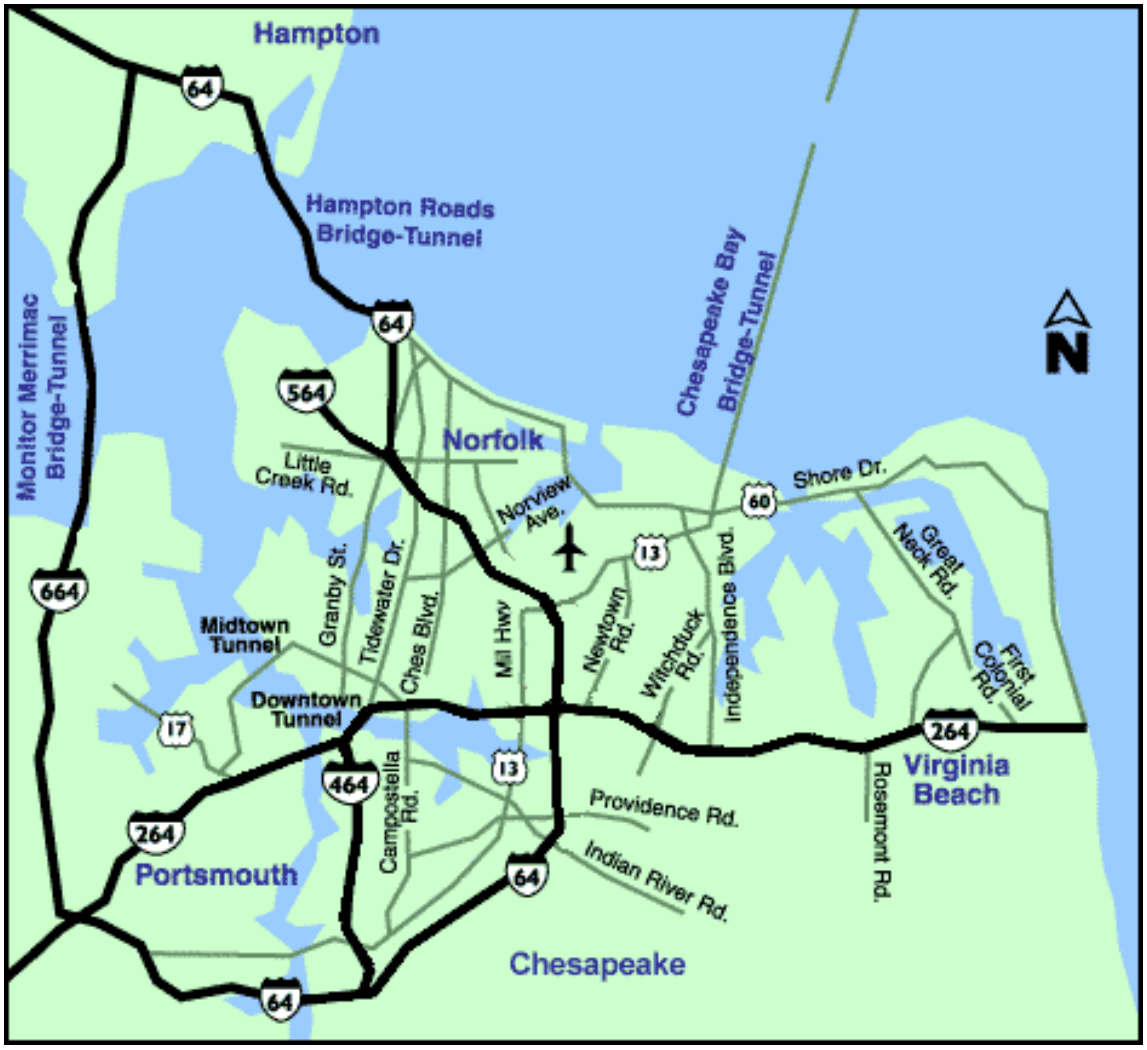

Figure 3-1: Map of the Hampton Roads region of Virginia.

The HRSTC also serves as the headquarters for the Freeway Incident Response Team (FIRT) that patrols the freeways and assists motorists and emergency vehicles. The Smart Travel Lab receives video, station data, and incident data directly from the HRSTC.

The Hampton Roads Incident Database is maintained by the operators and personnel at the HRSTC. This includes the persons monitoring the freeway cameras and other devices and the supervisors. Freeway incidents are identified by the operators watching the cameras, the incident response team on the freeways, state police radio, phone calls from motorists, and other sources. The incident is then manually entered into a database using a graphical user interface program at the HRSTC. All of the information is entered by hand into the database. The database began collecting incidents in January of 1997 and is still in use today with all new entries being sent to the Smart Travel Lab.

3.3 Database Structure

The actual incident database in the Lab is built on 5 different tables. Each unique incident recorded at the HRSTC is given a unique ID number (named the TMS call number) that is used to join the tables in the database. The structure of the database is shown in Figure 3-2.

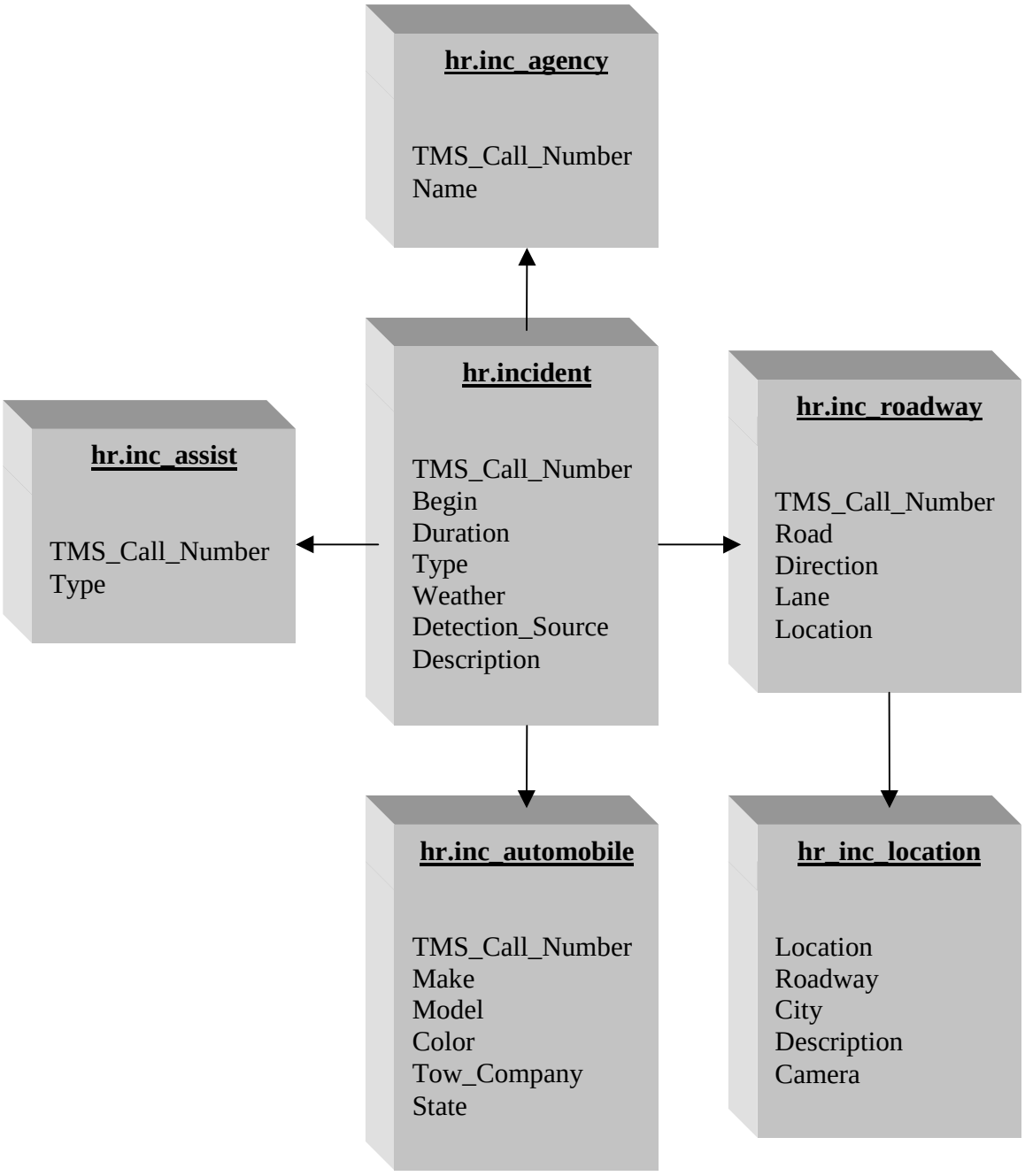

Figure 3-2: Structure of incident database tables.

3.3.1 Incident Table

The main table in the database is the incident table (hr.incident). This table contains important information on the beginning time and date of the incident along with the ending time and date. A single entry in the start and end fields contains both the date and time together, such as MM/DD/YYYY HH24:MI. The duration of the incident is defined as the distance between the start and end times.

The type of incident is also recorded in this table. The options for the incident type are

- Abandoned – an unoccupied vehicle left on the freeway shoulder
- Accident – a crash involving one or more vehicles
- Bridge – any type of incident that occurs on a bridge
- Condition Change – action taken by traffic manager
- Debris – material on the freeway affecting travel
- Disabled – an occupied vehicle that has broken down in a travel lane or shoulder lane
- Other
- TEOC – a severe weather advisory
- Tunnel – any type of incident that occurs in a tunnel
- VMS Change – action taken by traffic manager to provide traveler information by variable message signs (VMS)

TEOC refers to VDOT's Transportation Emergency Operations Center, a statewide coordinating unit that informs VDOT agencies and the public on significant weather conditions that may affect traffic conditions (VDOT, 2000). These TEOC entries only represent about 2 percent of the total entries in the database.

The next field in the table is the weather. The person entering the incident into the database has the choices of

- Clear,
- Cloudy,
- Cold / Ice,
- Cool,
- Fog,
- Fog / Rain,
- Hail,
- Hot / Humid,
- Natural Disaster,
- Rain,
- Sleet,
- Snow, and
- Warm.

This field is subjective from the operators' perspective and is often left blank during data entry.

The next field is the detection source. This is the manner in which the traffic manager was informed of the incident. Some of the possible detection methods are

- FIRT (Freeway Incident Response Team),
- HRSTC Cameras,

- Virginia State Police (VSP) Radio,
- Bridge Personnel,
- Phone Calls,
- TEOC, and
- VDOT Personnel

In the database there are a total of over 1000 unique entries for this field. Sometimes a specific police officer or camera operator is called by name in the field. Also, sometimes the camera number is recorded. Again, sometimes the entry person leaves this field blank.

The final field in the incident table is a text description of the incident. This field is considered optional and is only used by the operator to provide additional useful information that was not included in the other entries. Often, this description field is used to provide police codes on the specific type of incident, or the presence of personal injuries. Also, the database only stores a certain number of characters for this field, but the operator can enter as long a description as needed. Thus, some description entries are cut off in mid sentence and some information is not stored in the database. Overall, it is not possible to search this field for information due to the inconsistent nature of the entries.

3.3.2 Agency Table

This table (hr.inc_agency) lists the specific agencies that responded to the scene during the course of the incident. The entries are listed in alphabetical order in the database, so it is not clear which agency was the first on-scene. The amount of time each agency was on-scene is also not reported. This table is connected to the main incident table through the unique TMS call number. Some of the most frequently recorded responding agencies are

- Emergency Medical Services (EMS),
- Fire Department,
- FIRT (Freeway Incident Response Team),
- Hazardous Material Team,
- Local Police,
- Virginia State Police (VSP), and
- VDOT.

Overall, there are over 100 unique entries for the responding agency. In some cases the actual city is listed instead of local police. Other entries are unidentifiable or include specific names of police officers or FIRT personnel.

3.3.3 Assist Table

The next table in the incident database is the assist table (hr.inc_assist). This table lists the specific assistance given by the FIRT team on–scene or the traffic operators back at the HRSTC. Some of the most frequently recorded entries for the assistance are

- Call AAA for vehicle assistance,
- Call local tow truck,

- Call taxi,
- Change flat tire,
- Fill vehicle with gasoline,
- Extinguish vehicle fire,
- Traffic control around scene,
- Remove road debris,
- Push vehicle to shoulder,
- Activate Variable Message Signs (VMS), and
- Activate Highway Advisory Radio (HAR).

This field also suffers from many different entries, over 1000 unique entries are in the database. Many of these entries are similar, but each operator has their own personal phrases or spellings that they use to enter the assistance. It should be noted that the assistance listed applies to only FIRT and the HRSTC and not the other responding agencies such as EMS or police.

3.3.4 Automobile Table

This table (hr.inc_automobile) lists the specific automobiles that were involved in the incident. Two fields record the make and model of each vehicle. In some cases, a tractor-trailer is listed as the vehicle make, but in other cases the truck is listed by the specific make and model such as a Volvo 5100. This situation is also present for other types of vehicles such as motorcycles, buses, and emergency equipment. The license plate number and originating state are recorded at the HRSTC, but the plate number is stripped when the data is passed along to the Smart Travel Lab. Also included in this table is the towing company that was used for each vehicle. Again, the entries in the automobile table are joined to the main incident table through the unique TMS call number.

3.3.5 Roadway Table

This table (hr.inc_roadway) is concerned with the location of the incident as opposed to the incident characteristics. However, the location entries in this table are joined to the incident characteristics in the main incident table through the unique TMS call number. The first field in the table is the specific road or interstate where the incident occurred. The choices for this entry are

- I-64,
- I-264 (recently renamed from Route 44),
- I-464,
- I-564,
- I-664,
- Off-highway, and
- Bridge/Tunnel.

The interstates are straightforward, but the others are not. In the Hampton Roads area, there are two river interstate river crossings; the Hampton Roads Bridge Tunnel (HRBT), and the Monitor Merrimac Memorial Bridge Tunnel (MMMBT). The HRBT is used by

I-64, while the MMMBT is serviced by I-664. However, if an incident occurs on either of these systems the road entry is only given as bridge/tunnel. Thus, it is impossible to differentiate between incidents on the HRBT and MMMBT. The off-highway entry is unclear, but probably refers to incidents that occurred on major arterial roads in the region that interchange with the interstates, and thus may cause back-ups on some of the interstate exit ramps. It should also be noted that the road entry does not say whether the incident occurred on the main lanes or HOV lanes on the road.

The road direction is also given in this table to differentiate between the two opposing travel lanes. The lane field states which travel lanes are affected by the incident. In this field, some entries state the lanes by name (left and center lanes) or by lane number (1 and 2) depending on the method preferred by the operator. In addition, the lane field includes shoulder lanes, ramps, and reversible lanes. The lane field is the place where main lines are differentiated from HOV lanes.

Perhaps the most important field in this table is the specific location of the incident. The HRSTC has defined specific sections of the interstates as different zones, using names such as W64-01 (see Appendix D for a complete map). The east and west part of the zone name does not refer to the direction of travel, but rather if the zone is located east or west of the large I-64/I-264 interchange in Norfolk. The zone boundaries are the interchanges along the roadway, so a location zone may be 1 to 2 miles in length and refer to both directions of travel. This is the most specific location of an incident that is available in the incident database.

3.3.6 Location Table

The location table is the only table in the incident database that is not joined to the main table through the unique TMS call number. This table instead is joined to the roadway table through the unique location zone name. The table lists some important information about each location zone. The road and corresponding city of the location is given in two fields. The text description field gives the name of the two interchanges that bound the location. The final field is for the HRSTC and tells which traffic cameras are located within the zone.

3.4 Data Collection

The Incident Database from the Smart Travel Lab includes all types of incidents from January of 1997 and is updated daily. This project uses accident data up to the end of December 2000, which gives a total of 7,396 unique freeway accidents.

3.4.1 Data Reduction

As with any project that includes a large amount of data, the first step in analysis is to determine which data is of use to the project. This involves reducing the data by eliminating useless data. A number of accidents had missing values in the database, especially in the automobile and weather fields, and were thus removed from the analysis. The focus of this project is on freeway accidents, so any other accidents were removed. Some entries in the incident database listed an ‘off-highway’ entry in the

roadway field, and were excluded from the analysis. Other accidents were removed due to errors in the duration. Accidents that have a zero or negative value of duration are assumed to be entry errors by the HRSTC. Also, accidents with a duration greater than 12 hours were assumed to be operation error and removed from the analysis data. The 12 hour cutoff was used because it includes the case where the operator incorrectly enters the AM or PM part of the time.

After data reduction, there are 6,828 accidents that are assumed to be valid in terms of the clearance time and characteristics. This population of accidents is divided into learning and testing samples. The accidents in the learning sample will be used for model development and calibration, while the testing sample accidents will be used to evaluate the performance of the forecasting model. The testing sample is comprised of one-quarter of the accident population or 1707 accidents. Thus, the learning sample consists of 5121 accidents. It should be noted that the accidents for the testing sample were chosen chronologically rather than randomly. The testing sample represents the most recent accidents of the total population. Normally, a random sample of the total population is used for the testing sample. However, the goal of the forecasting models used in this study is to predict the clearance time of future accidents using knowledge from past accidents. For this reason, a chronological division was used for the learning and testing samples.

3.5 Potential Independent Variables

The goal of a forecasting model is to emulate a relationship between the dependent and independent variables. For this example, the dependent variable is the duration of the accident. Numerous independent variables are possible from the large amount of data recorded in the incident database for each accident. Table 3-1 gives a summary of the independent variables considered for the forecasting models. All of the independent variables are categorical with 2 or 3 possible values.

Table 3-1: Potential model independent variables.

Variable		Name	Value
Physical	Time of Day	PEAK	1 = Peak (6-8am, 4-6pm) 0 = Off-peak
	Day of the Week	WEEKDAY	1 = Weekday 0 = Weekend
	Weather	WEATHER	1 = Normal (Clear, Cloudy, Cool, Hot/Humid, Warm) 0 = Adverse (Cold/Ice, Fog, Natural Disaster, Rain, Sleet, Snow)
Response	EMS Response	EMS	1 = Yes 0 = No
	Fire Response	FIRE	1 = Yes 0 = No
	FIRT Response	FIRT	1 = Yes 0 = No
	Hazardous Material Agency	HAZMAT	1 = Yes 0 = No
	Police Response	POLICE	1 = Yes 0 = No
	VDOT Response	VDOT	1 = Yes 0 = No
	Tow Truck Response	TOW	1 = Yes 0 = No
Vehicle	Number of Vehicles	NUMVEH	1 = Single Vehicle 2 = Two Vehicles 3 = Three or More Vehicles
	Truck Involvement	TRUCK	1 = Yes 0 = No
	Passenger Bus Involvement	BUS	1 = Yes 0 = No

3.5.1 Physical Independent Variables

The physical independent variables describe the nature of the accident in terms of time and place. The first independent variable is the accident time of day. This variable has possible values of peak and off-peak. In this case, the peak hours are defined as 6am to 8am inclusive and 4pm to 6pm inclusive. Thus, off-peak hours are 8:01am to 3:59pm and 6:01pm to 5:59am. These peak hours were chosen because they correspond to the hours of operation for the High Occupancy Vehicle (HOV) reversible lanes that run along the median of I-64 in the region. The next variable is the day of the week. This variable has possible values of weekday (Monday to Friday) or weekend (Saturday and Sunday). For both the time of day and day of week, the variable value is based on the start time of

the accident regardless of the duration. A final physical variable is the weather, which takes on values of normal or adverse. Adverse weather is defined as fog, rain, ice, snow, sleet, or natural disaster. Normal weather includes all other conditions.

3.5.2 Vehicle Independent Variables

The vehicle independent variables attempt to provide information about the number and types of vehicles involved in the accident. The number of vehicles variable has three different values; single vehicle, two vehicles, and three or more vehicles. A number of other variables were used to reflect the types of vehicles involved in the accident. The involvement of a truck or tractor-trailer will give the truck variable a yes value. Similarly, the involvement of a passenger bus will give the bus variable a yes value. The assumption is that the accident only involved passenger automobiles unless otherwise noted by the truck and bus variables.

3.5.3 Accident Response Independent Variables

Another important independent variable related to accident clearance time is which emergency agencies responded to the scene. These variables give some sense of the severity of the accident. Binary variables were used to note the response of EMS, Fire Department, FIRT, Hazardous Material Agency, Police (local and state), Virginia Department of Transportation (VDOT) personnel, and tow-trucks. It should be noted that there are no variables to distinguish the response order or time of the above agencies.

3.6 ANOVA Significance Test

The above independent variables were identified from the available accident data. It is possible that some of the independent variables are not significant with regards to affecting accident clearance time. For example, some of the variables may have an influence on other factors than clearance time such as accident frequency, accident severity, and accident detection time. Thus, it was necessary to perform statistical significance tests on the independent variables using ANOVA for the proposed dependent variable of accident clearance time.

Analysis of variance (ANOVA) refers to a collection of experimental situations and statistical procedures for the analysis of quantitative responses from experimental units (Devore, 1995). A single-factor ANOVA table analyzes data from two or more population samples where one factor is used to differentiate between the populations. The null hypothesis being tested is given below (Devore, 1995).

$$H_o: \mu_1 = \mu_2 = \ldots = \mu_I$$

Versus the alternative hypothesis

H_a: at least two of the μ_I's are different

Where

I = the number of samples being compared
μ_1 = the mean of sample 1 when a single factor is applied to population
μ_I = the mean of sample I when a single factor is applied to population

This ANOVA is easily able to handle samples with different sample sizes. The major assumption with single-factor ANOVA is that each sample of the population is normally distributed with the same variance (Devore, 1995). However, Ozbay and Kachroo used ANOVA tables successfully to test variable significance and even found a normal distribution for incident duration. As data sets of incidents were divided into smaller data sets so that the incidents were all of the same severity and nature, a normal distribution trend was found and confirmed by statistical tests (Ozbay and Kachroo, 1999). Thus, the ANOVA assumption appears to be valid.

For this project, an ANOVA table was applied to each independent variable. For example, for the time of day independent variable, the single factor used was peak versus off-peak. The total population of accidents was divided into two samples, peak accidents and off-peak accidents. Each sample has a sample mean and sample variance. The ANOVA test compared the two samples to determine if the underlying mean of each sample was the same (null hypothesis) or significantly different (alternative hypothesis). The output of the ANOVA table is a p-value, which is the smallest level of significance (α) at which the null hypothesis (that the two sample means are the same) can be rejected (Devore, 1995). Common levels of significance are 0.05 and 0.01. If the p-value is less than or equal to the level of significance, the null hypothesis is rejected and we can say that the two samples have different means and the independent variable is significant in terms of clearance time. Thus, the clearance time of an accident is assumed to be dependent on all significant variables.

The ANOVA table was applied to each independent variable and the corresponding p-values are given below. The full ANOVA results for each independent variable are given in Appendix A.

Table 3-2: Independent variable significance test results.

Independent Variable	ANOVA p-value
PEAK	2.30×10^{-5}
WEEKDAY	7.83×10^{-6}
WEATHER	0.235
EMS	1.49×10^{-66}
FIRE	1.31×10^{-60}
FIRT	0.958
HAZMAT	6.21×10^{-28}
POLICE	2.70×10^{-31}
VDOT	2.18×10^{-11}

NUMVEH	1.41×10^{-44}
TRUCK	1.95×10^{-19}
BUS	0.0440
TOW	1.80×10^{-181}

The ANOVA analysis shows that all of the independent variables are significant except for weather and FIRT response. The bus involvement variable is significant at a 0.05 level, but not a 0.01 level. This borderline independent variable was included in the forecasting models none the less, because it intuitively appears to have a significant impact on accident clearance time. With a passenger bus, there is the probability of numerous victims and a relatively large vehicle to evacuate the accident scene.

3.7 Model Selection

The next step in the experimental framework is to select potential models for forecasting accident clearance time. There is a wide range of forecasting techniques that may be applicable to accident clearance time. This study will focus on three different forecasting models.

The first model to be evaluated is a stochastic model using probability density functions to describe clearance time. Bast research on incident duration has shown that the duration of an incident can be modeled as a random variable using a lognormal or Weibull distribution. The second forecasting model is a nonparametric regression model. Nonparametric regression techniques have been used successfully to forecast other traffic conditions such as flow. The final forecasting model is a classification tree model. This model was chosen based on promising research performed recently using decision trees to predict incident clearance times. The next three chapter will outline the development of the three forecasting models and investigate the performance of the each model.

Chapter 4: Stochastic Model

4.1 Model Background

Many events in nature are assumed to behave in some random manner. Even though the events are random, there may be tendencies and trends in the behavior of the events that can be used to describe the system as a whole. A stochastic model attempts to describe the randomness of the events (Higgins and Keller-McNulty, 1995). For example, flipping a coin produces two possible results in showing heads or tails. A single event of flipping the coin is not dependent on any factors, so there is a random outcome of heads or tails. However, over a long period of time and many trials, it is expected that the proportion of heads outcomes will be 50 percent. This is a simple example of a stochastic model of the random event of flipping a coin.

4.2 Probability Density Functions

One method to describe the behavior of a random event is through a probability density function. The probability distribution shows how probability density is distributed across the possible values of a random variable (Higgins and Keller-McNulty, 1995). The equation to describe continuous random variables for a specific distribution are referred to as a probability density functions

Past research on incident duration has shown that the duration tends to show a Weibull or lognormal probabilistic distribution. One major deficiency with a number of these results is the relatively small sample size used to test different probability density functions. If such distributions are applicable, it may be stated that accident duration can be modeled as a random variable with a known distribution.

A random variable is said to have a Weibull distribution if the probability density function of the random variable is given by (Devore, 1995).

$$f(x;\alpha,\beta) = \left\{ \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} \right\} \quad for\ x \geq 0$$

where x is the value of the random variable, α is a shape parameter, and β is a scale parameter.

Likewise, a random variable is said to have a lognormal distribution if the log of the variable has a normal distribution. The probability density function of a lognormal distribution is given by (Devore, 1995).

$$f(x;\mu,\sigma) = \left\{ \frac{1}{\sqrt{2\pi}\,\sigma\, x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} \right\} \quad for\ x \geq 0$$

where x is the value of the random variable, μ is a scale parameter, and σ is a shape parameter.

4.2.1 Goodness-of-fit Test

The assumption from past theoretical and quantitative research is that the Weibull or lognormal distributions can describe incident duration. In this project there is a large sample to verify or disprove this assumption using statistical goodness-of-fit tests (Ang and Tang, 1975). One common goodness-of-fit test that will be applied to the above probability density functions is the chi-square test.

The chi-square test compares the observed interval frequencies with the theoretical frequencies for the distribution to be tested (Ang and Tang, 1975). The test statistic is given as

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

where k is the number of intervals, n_i is the observed frequency for the i^{th} interval, and e_i is the theoretical frequency of the i^{th} interval. The test statistic, χ^2, will approach the chi-square distribution χ_f^2 with f=k-1 degrees of freedom (Ang and Tang, 1975). The critical value of the χ_f^2 distribution at the cumulative probability of $1-\alpha$ is given by $c_{1-\alpha,f}$ where α is referred to as the level of significance (Ang and Tang, 1975). Thus, the assumed distribution is an acceptable fit at the α significance level if χ^2 is less than $c_{1-\alpha, f.}$ Otherwise, the assumed theoretical distribution is not supported by the observed data (Ang and Tang, 1975).

4.3 Model Development

This chapter discusses a collection of different stochastic models for a certain accident characteristic. Two factors were tested to develop the models, accident severity and time of day. In addition a stochastic model was developed for all accidents regardless of the severity and time of day. The distributions that are emphasized in the analysis are the Weibull and lognormal distributions. The ExpertFit program was used to select the optimal probability density function parameters for 30 possible distributions (Law, 2001). It is worth noting that either the lognormal or Weibull distribution was the best fitting distribution for each case. Once the stochastic models were developed, the chi-square goodness-of-fit test was used to evaluate the fit of the probabilistic distribution.

4.3.1 Model for All Accidents

The simplest stochastic model for this project is one that models any accident, regardless of the accident characteristics. Figure 4-1 shows a histogram of the clearance time of all accidents.

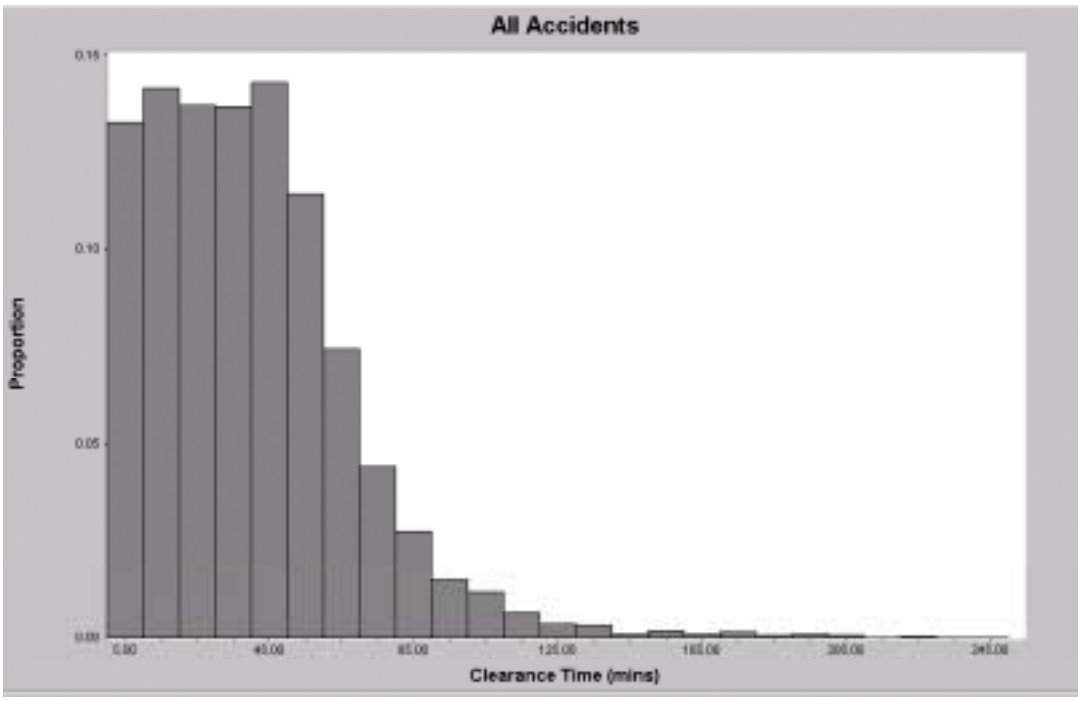

Figure 4-1: Clearance time histogram for all accidents.

This graph shows a definite left-shifted tendency towards accidents with smaller clearance times. The ExpertFit program identified the Weibull distribution as the best candidate distribution. The parameters for the Weibull and lognormal probability density functions are given in Table 4-1.

Table 4-1: Distribution parameters for all accidents.

Weibull Distribution		***Lognormal Distribution***	
β	α	μ	σ
43.3	1.33	3.34	0.968

Overlaying the two probability density functions on the original histogram gives a comparison between the two distributions.

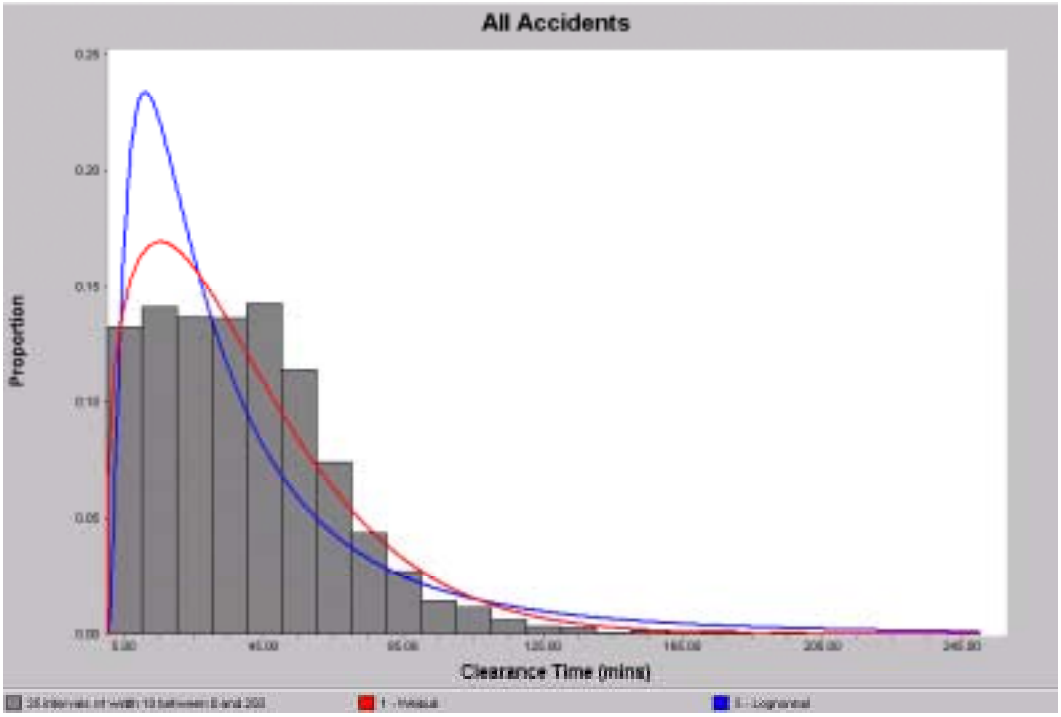

Figure 4-2: Histogram and distribution overlay for all accidents.

It appears from the graph that the Weibull distribution is a better fit for the clearance time data. The chi-square test was used to test the assumption that the accident data follows the Weibull and lognormal distributions.

Table 4-2: Chi-square test for all accidents.

Number of samples, N	6,828
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	612.996
Lognormal Test Statistic, χ^2	2,005.369

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

The chi-square results show that both the Weibull and lognormal stochastic models do not adequately describe the clearance time values for all accidents. The next step was to introduce accident severity into the stochastic model.

4.3.2 Model for Accident Severity

This stochastic model attempts to fit a probabilistic density function to three different subcategories of accidents; single vehicle, two vehicle, and three or more vehicle accidents. The same procedure outlined for all accidents was used for these models. First, histograms were prepared for each category of accident severity.

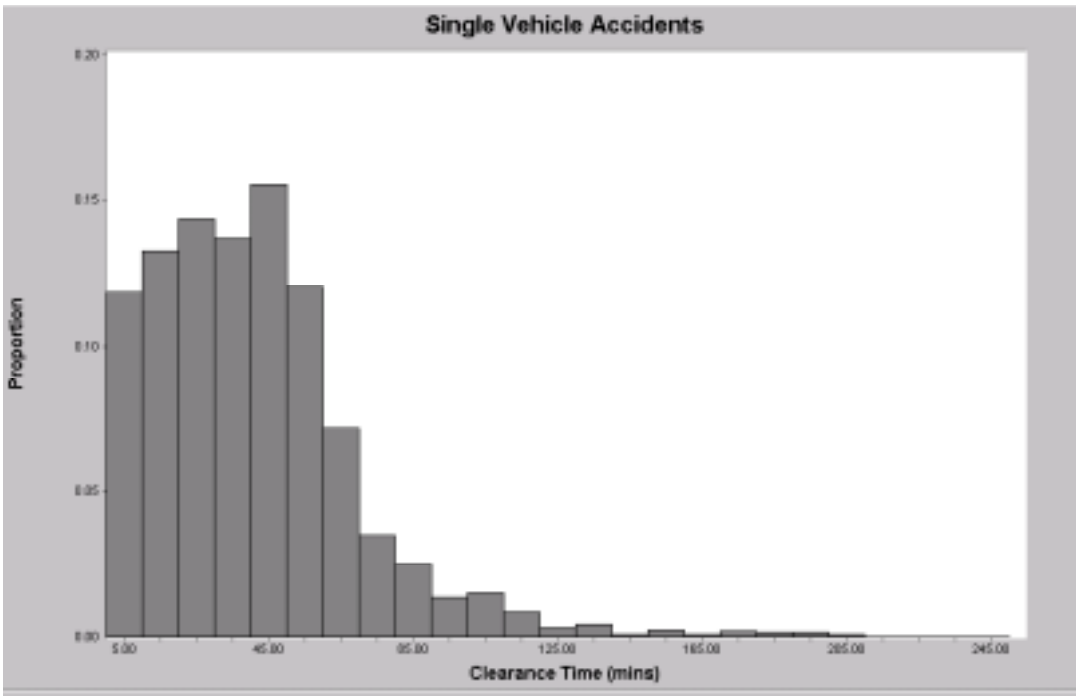

Figure 4-3: Clearance time histogram for single vehicle accidents.

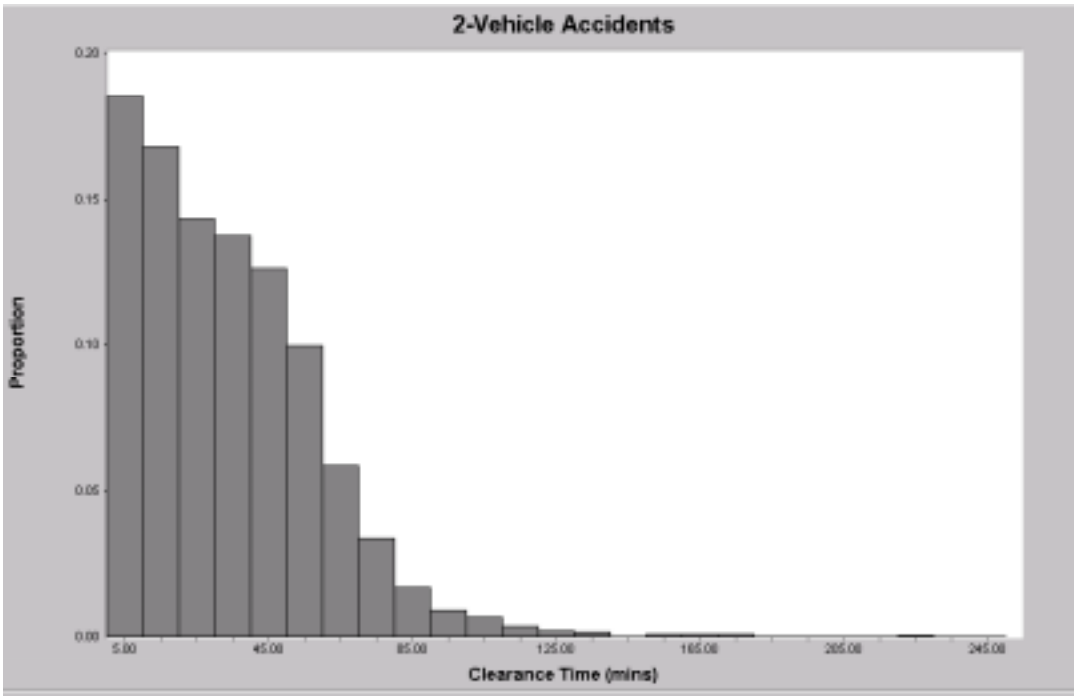

Figure 4-4: Clearance time histogram for two vehicle accidents.

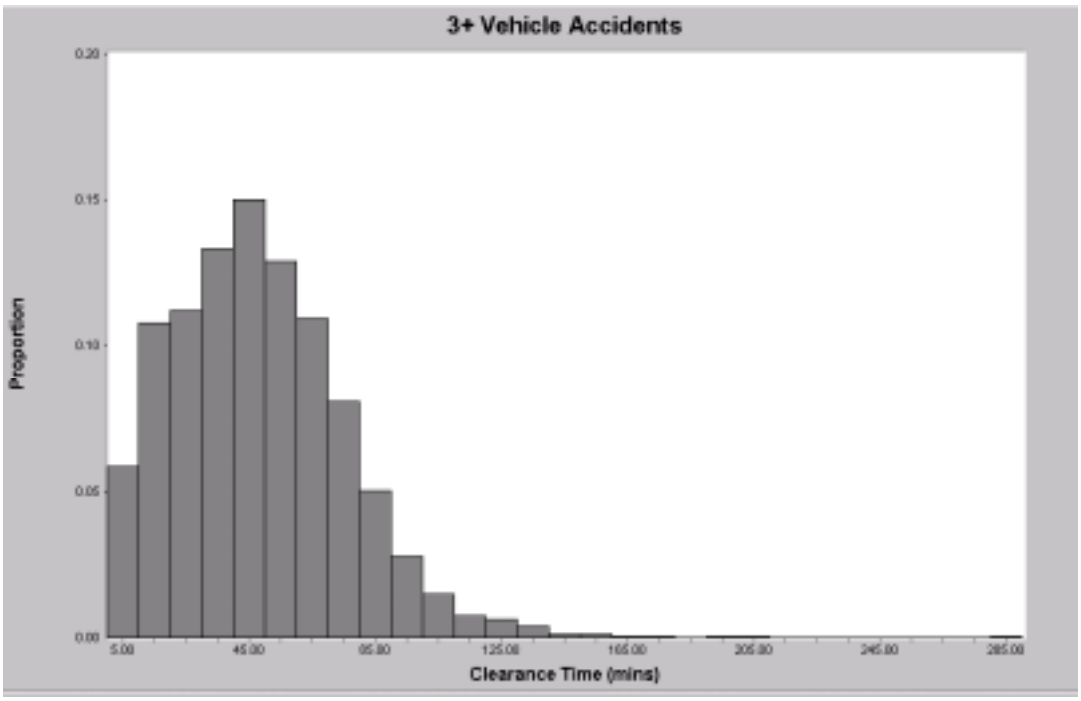

Figure 4-5: Clearance time histogram for three or more vehicle accidents.

ExpertFit evaluated a large number of probabilistic distributions and found that the Weibull distribution was the best fit for all three of the histograms above. The parameters for the Weibull and lognormal distributions are given below.

Table 4-3: Distribution parameters for single vehicle accidents.

Weibull Distribution		***Lognormal Distribution***	
β	α	μ	σ
44.8	1.30	3.37	0.980

Table 4-4: Distribution parameters for two vehicle accidents.

Weibull Distribution		*Lognormal Distribution*	
β	α	μ	σ
36.6	1.27	3.15	1.00

Table 4-5: Distribution parameters for three or more vehicle accidents.

Weibull Distribution		*Lognormal Distribution*	
β	α	μ	σ
53.5	1.72	3.64	0.779

For a visual comparison, the two distributions are overlaid on the histograms created above.

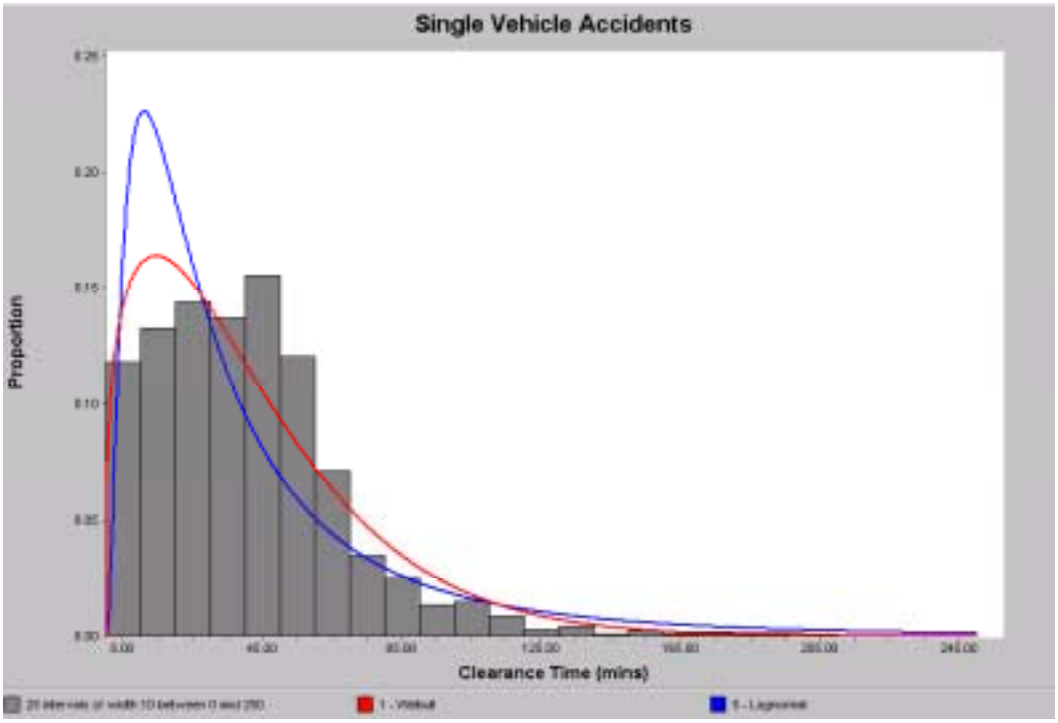

Figure 4-6: Histogram and distribution overlay for single vehicle accidents.

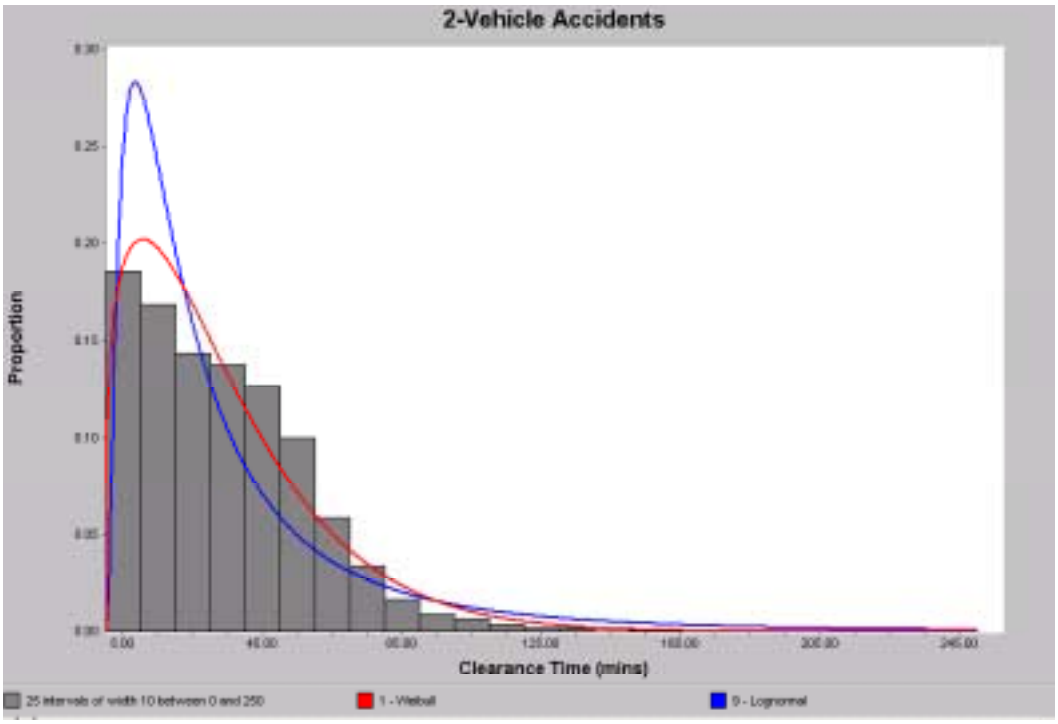

Figure 4-7: Histogram and distribution overlay for two vehicle accidents.

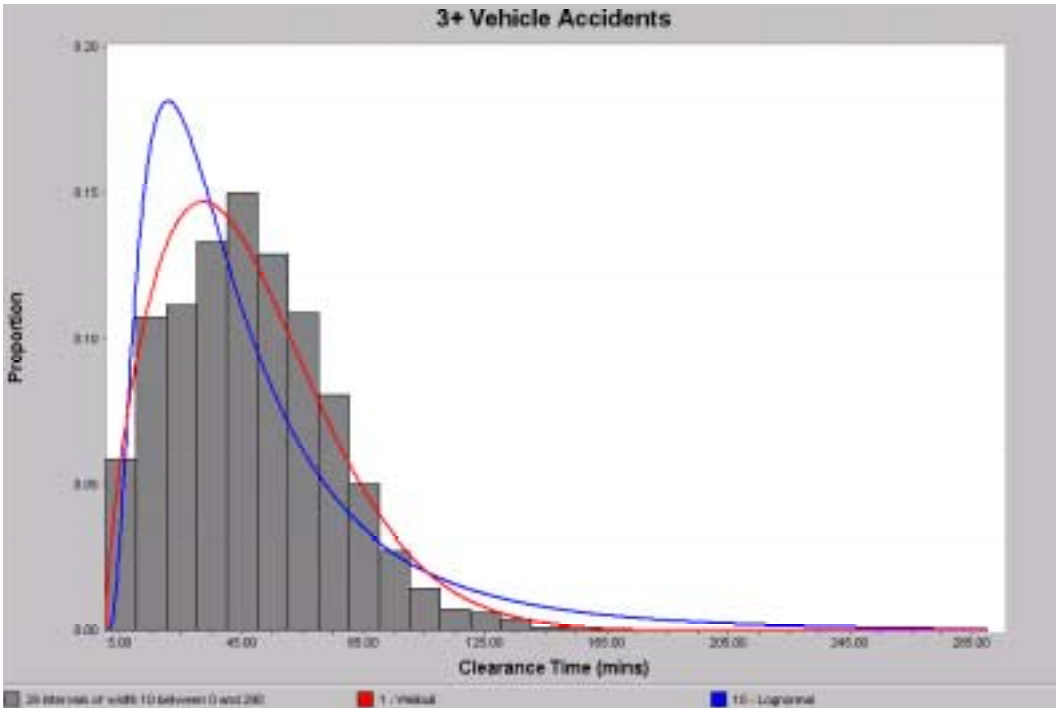

Figure 4-8: Histogram and distribution overlay for three or more vehicle accidents.

Using the distribution parameters the models were tested using the chi-square test.

Table 4-6: Chi-square test for single vehicle accidents.

Number of samples, N	2,716
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	381.290
Lognormal Test Statistic, χ^2	992.365

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

Table 4-7: Chi-square test for two vehicle accidents.

Number of samples, N	2,687
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	336.044
Lognormal Test Statistic, χ^2	780.317

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

Table 4-8: Chi-square test for three or more vehicle accidents.

Number of samples, N	1,425
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	116.305
Lognormal Test Statistic, χ^2	353.947

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

Again, these chi-square results show that accident data does not support the Weibull or lognormal distributions for the three models. A full range of significance levels were tested and for each case, the data overwhelmingly rejected the assumption of the probabilistic distribution.

4.3.3 Model for Accident Time of Day

This stochastic model attempts to fit probabilistic distributions to accident clearance time based on the time of day. Three different categories of accidents were

used; peak period weekday, off-peak period weekday, and weekend accidents. The histograms for the clearance times of each category are given below.

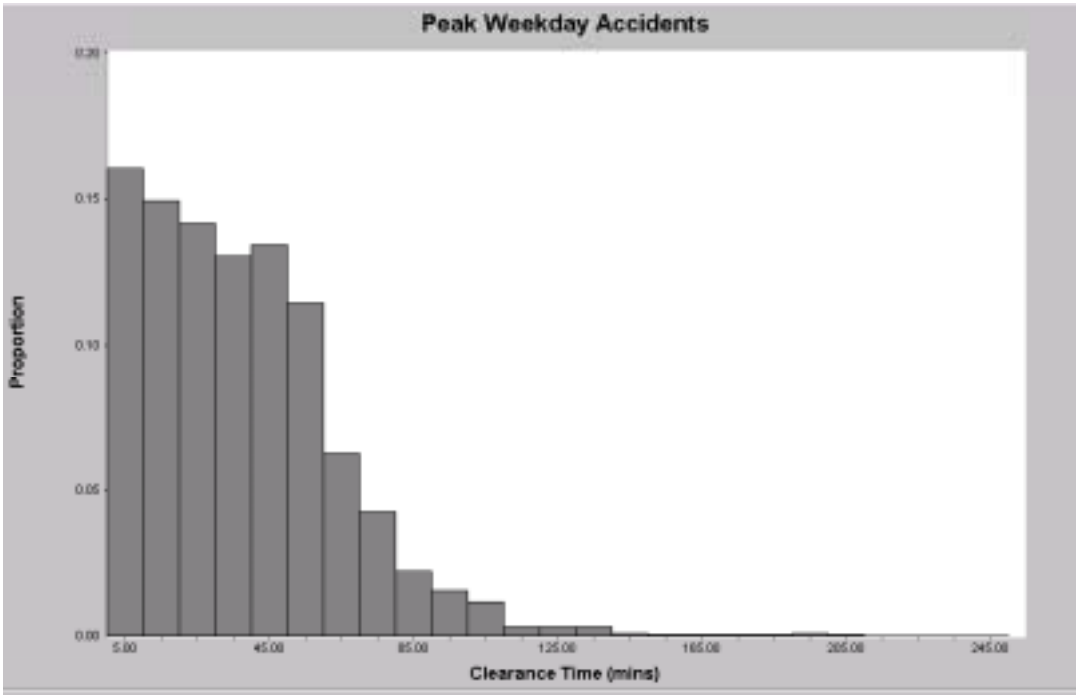

Figure 4-9: Clearance time histogram of peak weekday accidents.

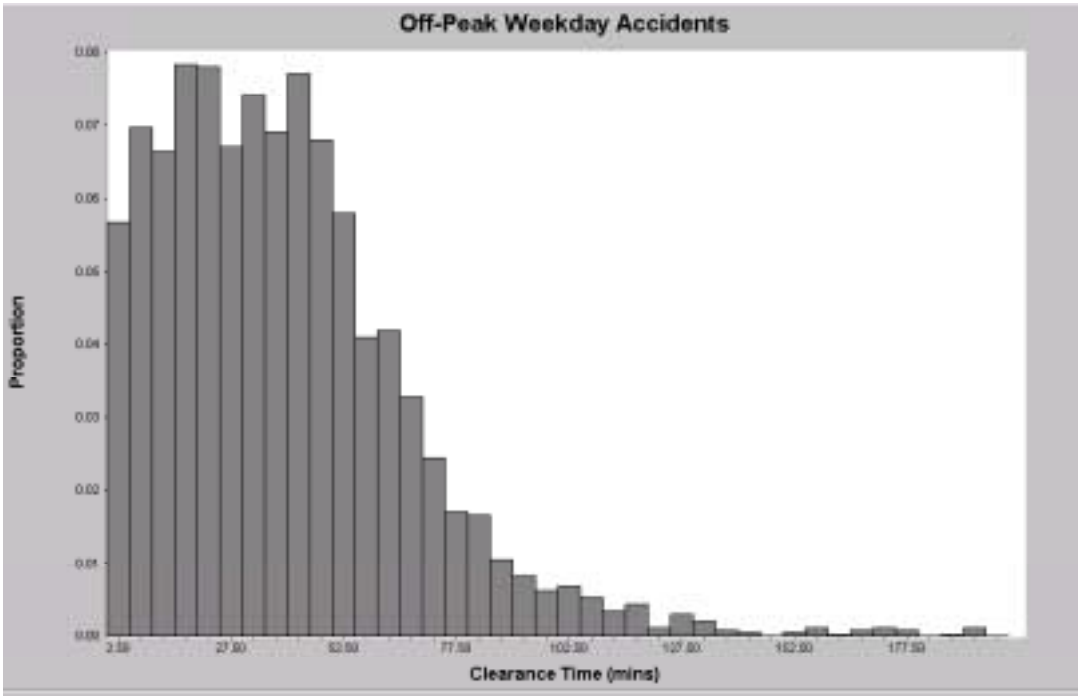

Figure 4-10: Clearance time histogram of off-peak weekday accidents.

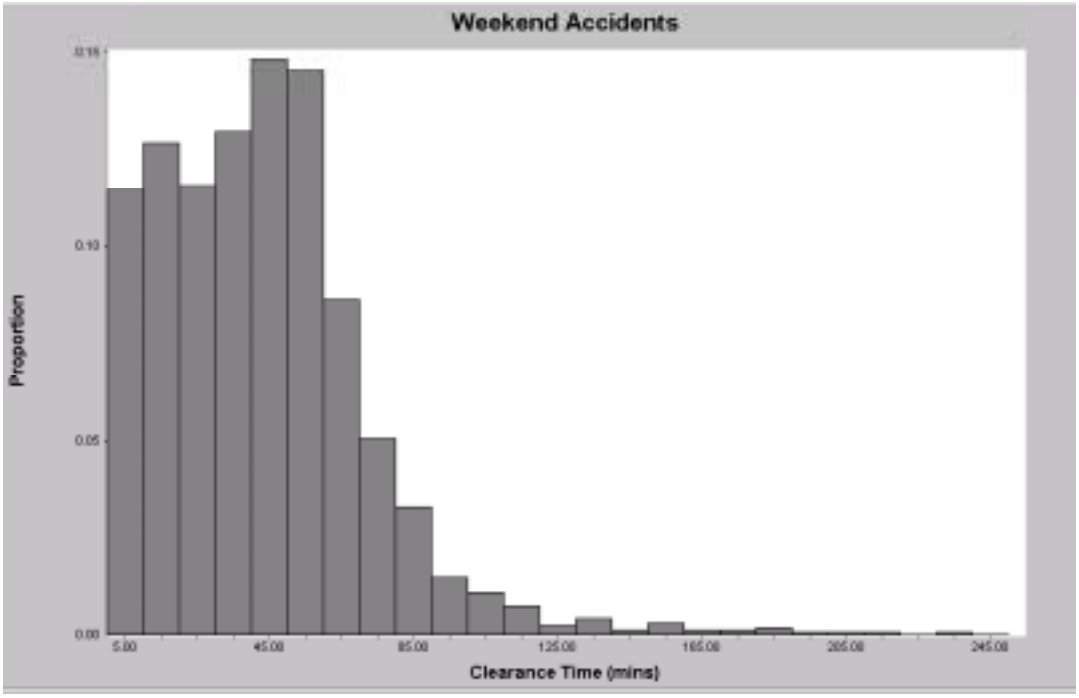

Figure 4-11: Clearance time histogram of weekend accidents.

Again, ExpertFit evaluated a number of different distributions and selected the Weibull distribution as the best candidate distribution for the three clearance time samples. The distribution parameters for the Weibull and lognormal distributions are given below.

Table 4-9: Distribution parameters for peak weekday accidents.

Weibull Distribution		***Lognormal Distribution***	
β	α	μ	σ
40.4	1.34	3.26	0.973

Table 4-10: Distribution parameters for off-peak weekday accidents.

Weibull Distribution		***Lognormal Distribution***	
β	α	μ	σ
43.3	1.32	3.34	0.958

Table 4-11: Distribution parameters for weekend accidents.

Weibull Distribution		***Lognormal Distribution***	
β	α	μ	σ
46.7	1.37	3.42	0.976

For a visual comparison, the two distributions are overlaid on the histograms created above.

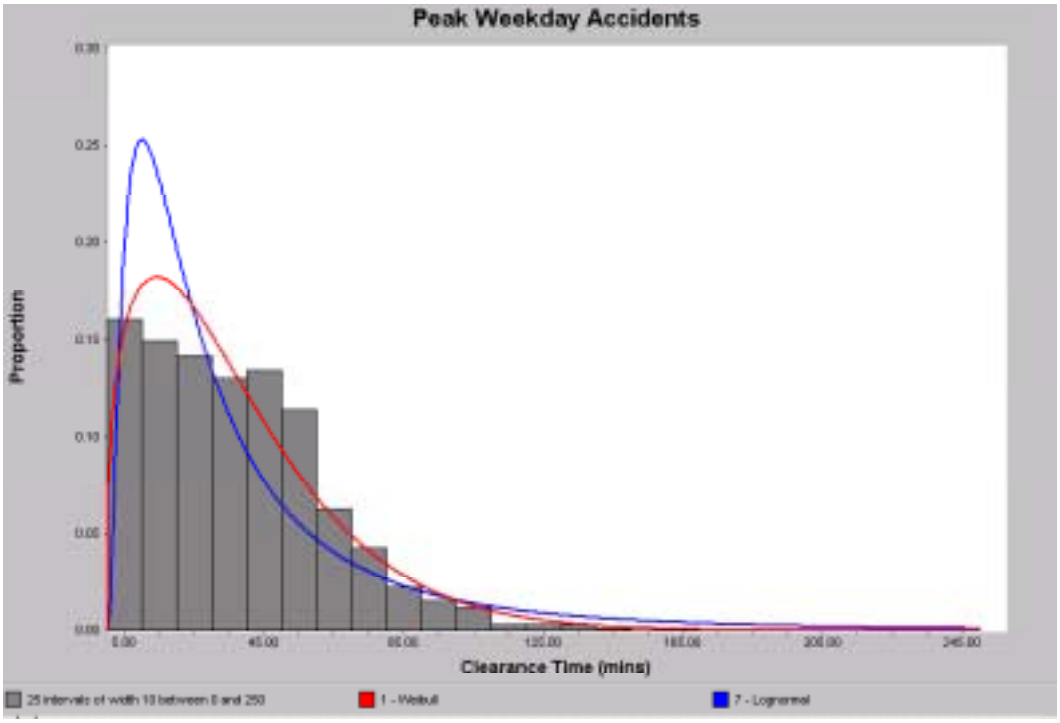

Figure 4-12: Histogram and distribution overlay for peak weekday accidents.

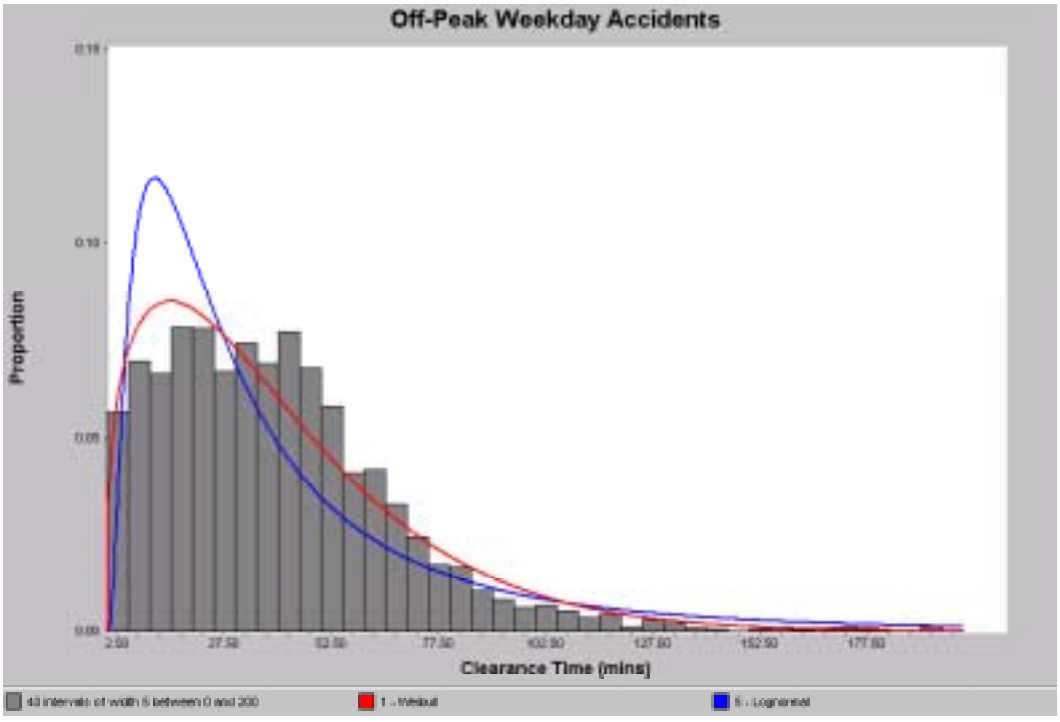

Figure 4-13: Histogram and distribution overlay for off-peak weekday accidents.

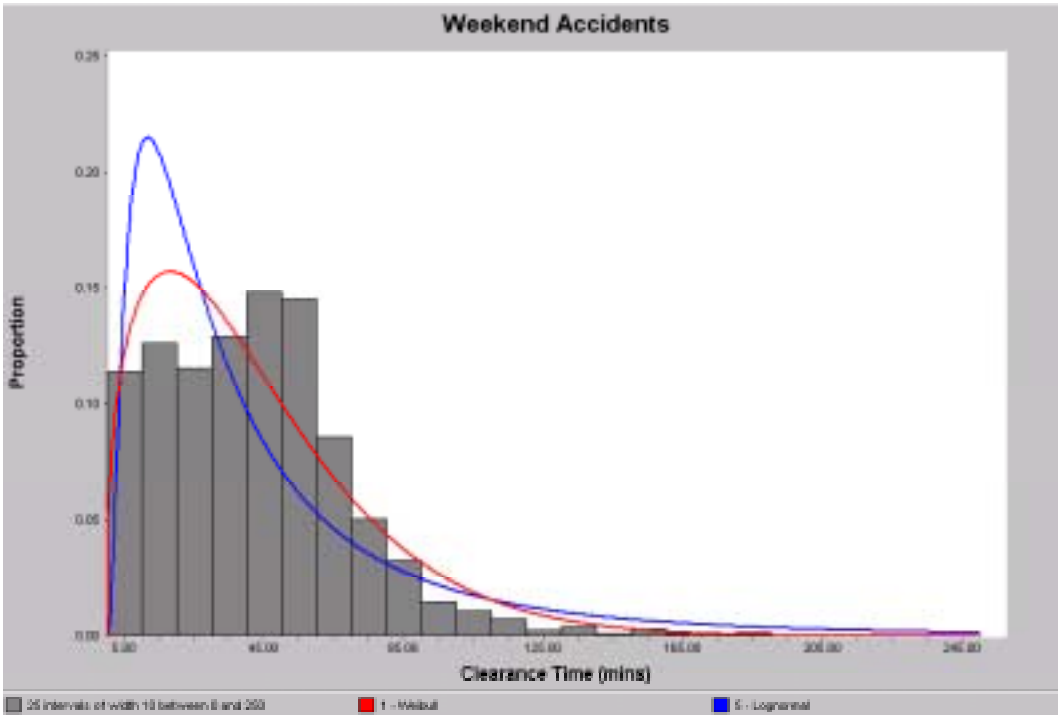

Figure 4-14: Histogram and distribution overlay for weekend accidents.

Using these distribution parameters the models were tested using the chi-square test.

Table 4-12: Chi-square test for peak weekday accidents.

Number of samples, N	1,797
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	195.810
Lognormal Test Statistic, χ^2	494.130

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

Table 4-13: Chi-square test for off-peak weekday accidents.

Number of samples, N	3,384
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	337.631
Lognormal Test Statistic, χ^2	874.770

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

Table 4-14: Chi-square test for weekend accidents.

Number of samples, N	1,647
Number of Intervals	40
Degrees of Freedom	39
Weibull Test Statistic, χ^2	227.706
Lognormal Test Statistic, χ^2	676.910

Significance Level α	**Critical Value $c_{1-\alpha,f}$**	**Accept Weibull distribution?**	**Accept Lognormal distribution?**
0.25	44.539	No	No
0.15	48.126	No	No
0.10	50.660	No	No
0.05	54.572	No	No
0.01	62.428	No	No

As with the previous stochastic models, the Weibull and lognormal distributions are rejected based on the available clearance time data.

4.4 Summary

In this chapter accident clearance time data was used to produce a number of stochastic models. Unfortunately, no stochastic models were applied to future accident scenarios due to the inability to accurately fit any probabilistic distribution to the accident data. It is possible that some of the variance in accident clearance time may be explained by more specific accident characteristics. The next chapter investigates two deterministic models that incorporate independent variables gathered from accident characteristics in the incident database.

Chapter 5: Nonparametric Regression Model

5.1 Model Development

The second forecasting model developed for this study was a nonparametric regression model. This model attempts to emulate a deterministic relationship between the accident characteristics and the clearance time. The nonparametric regression model was presented in general terms in Section 2.4. This chapter will give the specific model characteristics that were developed for this project.

5.1.1 Neighborhood Definition

This nonparametric regression model defines neighborhoods based on a constant sample size. This is referred to a k-nearest-neighbor (KNN) nonparametric regression technique, where k is the number of samples in the neighborhood. The optimal value of k will be found through the empirical testing of numerous values, based on the model measures of effectiveness that will be discussed in a later section.

5.1.2 Distance Metric

The neighborhood is determined by selecting the k number of past accidents that are closest to the current accident. This requires some measure of "closeness' to find the distance between two accidents. The state of the accident depends on the model independent variables, so these variables were used in the distance metric. A summary of these independent variables is given below in Table 5-1.

Table 5-1: Nonparametric regression independent variables.

Variable	Name	W_x	Value
PEAK	A	3.43	1 = Peak (6-8am, 4-6pm) 0 = Off-peak
WEEKDAY	B	3.90	1 = Weekday 0 = Weekend
EMS	C	16.07	1 = Yes 0 = No
FIRE	D	15.28	1 = Yes 0 = No
HAZMAT	E	97.27	1 = Yes 0 = No
POLICE	F	9.17	1 = Yes 0 = No
VDOT	G	24.78	1 = Yes 0 = No
TOW	H	20.83	1 = Yes 0 = No

NUMVEH	I	$W_I^{12} = 7.44$ $W_I^{13} = 6.39$ $W_I^{23} = 13.83$	1 = Single Vehicle 2 = Two Vehicles 3 = Three or More Vehicles
TRUCK	J	16.10	1 = Yes 0 = No
BUS	K	11.01	1 = Yes 0 = No

Since the independent variables are categorical, it was not possible to use the euclidean distance to measure distance. A new distance metric was developed that is based on the number of matching independent variables between the two accidents. The distance is increased for each non-matching variable. The unique feature of the distance metric for this model is that each mismatch is given a different weight factor for each independent variable. The weight factors for each variable are given in the W_x field in the table above. The values for the weight factors were calculated from the absolute difference in means of the two samples used in the ANOVA table for that particular independent variable. For example, for the ANOVA table the accident population was divided into a peak accident sample and a non-peak accident sample. The absolute difference between the mean of these two samples was 3.43 minutes, so this was used as the value of W_A.

The distance metric between two accidents is given by

$$\textbf{Distance} = W_A |A_1 - A_2| + W_B |B_1 - B_2| + W_C |C_1 - C_2| + W_D |D_1 - D_2| + W_E |E_1 - E_2| + W_F |F_1 - F_2| + W_G |G_1 - G_2| + W_H |H_1 - H_2| + W_J |J_1 - J_2| + W_K |K_1 - K_2| + W_I^*$$

Given A_1 = the value of the PEAK variable for the current accident
A_2 = the value of the PEAK variable for the past accident
B_1 = the value of the WEEKDAY variable for the current accident
B_2 = the value of the WEEKDAY variable for the past accident
etc.
W_A = the weight factor for the PEAK variable
W_B = the weight factor for the WEEKDAY variable
etc.

Where $W_I^* = W_I^{12}$ if ($I_1 = 1$ and $I_2 = 2$) or ($I_1 = 2$ and $I_2 = 1$)
$W_I^* = W_I^{13}$ if ($I_1 = 1$ and $I_2 = 3$) or ($I_1 = 3$ and $I_2 = 1$)
$W_I^* = W_I^{23}$ if ($I_1 = 2$ and $I_2 = 3$) or ($I_1 = 3$ and $I_2 = 2$)

In the distance equation, the terms $|A_1 - A_2|$, $|B_1 - B_2|$, and such determine if the two accidents have equal values for a single independent variable. The terms take on a value of 0 if the two values match and 1 if they do not match.

The distance metric is best understood through an example to find the distance between two accidents. The accident characteristics are given below.

Table 5-2: Example of distance metric.

Characteristic	Accident 1 (Current Accident)		Accident 2 (Past Accident)	
Time of day	7:30 am	($A_1 = 1$)	4:15 pm	($A_2 = 1$)
Day of week	Wednesday	($B_1 = 1$)	Sunday	($B_2 = 0$)
EMS response?	Yes	($C_1 = 1$)	No	($C_2 = 0$)
Fire Department response?	No	($D_1 = 0$)	No	($D_2 = 0$)
HAZMAT Agency response?	No	($E_1 = 0$)	No	($E_2 = 0$)
Police Department response?	Yes	($F_1 = 1$)	Yes	($F_2 = 1$)
VDOT response?	No	($G_1 = 0$)	No	($G_2 = 0$)
Tow-truck response?	Yes	($H_1 = 1$)	Yes	($H_2 = 1$)
Number of vehicles	1	($I_1 = 1$)	2	($I_2 = 2$)
Truck involvement?	No	($J_1 = 0$)	No	($J_2 = 0$)
Bus involvement?	No	($K_1 = 0$)	No	($K_2 = 0$)

This example involved finding the distance between a single vehicle accident with injuries and a 2-vehicle accident without injuries. Using the values from the above table and the weight factors in Table 6-1 the distance metric becomes

Distance = 3.43|1-1| + 3.90|1-0| + 16.07|1-0| + 15.28|0-0| + 97.27|0-0| + 9.17|1-1| + 9.17|1-1| + 24.78|0-0| + 20.83|1-1| + 16.10|0-0| + 11.01|0-0| + 7.44

Distance = 27.41

This value is a numerical representation of the distance between the categorical values of the two accidents.

This approach to the distance metric is new. The use of categorical independent variables necessitates a creative solution for a model that requires a numerical distance metric. An alternate approach would be to disregard the weight factors and the distance metric would be the number of mismatch variables between the two accidents. It was assumed that a more realistic approach would be to add weight factors to show that some variables tend to have a greater effect on accident clearance time.

5.1.3 Forecast Generation

This nonparametric model uses a straight average of the clearance time of each accident in the neighborhood. However, this average is stripped of the decimal places so the model outputs an integer value. This approach does not give special consideration to past accidents that are "closer" to the current accident. The basic assumption is that each past accident in the neighborhood is considered of equal importance in predicting the clearance time of the current accident.

5.2 Model Algorithm

The basic function of the nonparametric regression model is to compare a current accident with each accident in a historical database to define a neighborhood. This is outline in the pseudo-code shown below

```
procedure Nonparametric Regression
begin
        select neighborhood size k
        read current accident
        evaluate all independent variables of current accident
        repeat
                read past accident from historical database
                evaluate all independent variables of past accident
                calculate distance between current and past accidents
        until all accidents in historical database are compared
        sort list of past accidents by increasing distance from current accident
        calculate prediction from average duration of first k accidents in list
end
```

Figure 5-1: Pseudo-code for nonparametric regression procedure.

This procedure will output the expected clearance time for a current accident. For this project, we need to test the nonparametric regression model to find the optimal neighborhood size. This procedure is outlined below and makes use of the learning and testing samples that were developed from the accident population.

```
procedure Nonparametric Regression Testing
begin
    repeat
        select neighborhood size k
        repeat
            read the current accident from the testing sample
            evaluate all independent variables of current accident
            repeat
                read past accident from learning sample
                evaluate all independent variables of past accident
                calculate distance between current and past accidents
            until all accidents in learning sample are compared
            sort list of past accidents by increasing distance from current
                accident
            calculate prediction from average duration of first k accidents in
                list
        until all accidents from the testing sample are used
        increment k
    until all neighborhood sizes have been tested
end
```

Figure 5-2: Pseudo-code for nonparametric regression testing procedure

The actual testing model was developed in Microsoft's EXCEL spreadsheet program. The testing and learning samples were recorded on separate worksheets with all of the corresponding independent variable values. The Visual Basic programming language was used to perform the analysis and output the results. The complete programming code for the nonparametric regression model is included in Appendix B.

5.3 Measures of Effectiveness

There are numerous methods to determine the accuracy or effectiveness of the nonparametric regression model. First and foremost, every accident from the testing sample will have a predicted clearance time in addition to the actual clearance time that was recorded in the incident database. The difference between the predicted and actual duration is defined as the prediction error and is given in minutes. The value of the prediction error will positive for an overestimated duration, zero for an exact prediction, and negative for underestimated duration. Since each accident in the testing sample will have a prediction error, the main measure of effectiveness is the mean absolute prediction error (MAPE), or the average of all the prediction errors in the testing sample.

$$MAPE = \frac{\sum_{i}^{N} \left| TP_i - TA_i \right|}{N}$$

Where N = the total number of test accidents
TP_i = the predicted clearance time for the i^{th} test accident
TA_i = the actual clearance time for the i^{th} test accident

Some secondary measures of effectiveness are also related to the prediction error of the test accidents. It is useful to know the percentage of predictions that were within a certain tolerance of their actual clearance times. These tolerances used include 5, 10, 15, 30 and 60 minutes. For example, the percentage of clearance times that were predicted within 10 minutes of the actual time is found by counting the number of absolute predictions errors for the test sample that were less than or equal to 10 minutes. This number is then divided by the total number of test accidents to find the percentage.

Due to the nature of the nonparametric regression model, each of these measures of effectiveness must be found for a wide range of neighborhood sizes. The measures will be used to select an appropriate or best performing neighborhood size.

5.4 Selection of Neighborhood Size

Nonparametric regression is a unique forecasting technique in that the model parameter values are determined through empirical testing of the model. Fortunately, the model algorithm used in this project is simple and fast enough to test a large range of neighborhood size values.

A number of different measures of effectiveness were defined in the above section. Because this analysis uses numerous measures, the final neighborhood size selected may not be the optimal value for each measure. Thus, the selection of a neighborhood size is subjective and based on the recommendation of the model developer.

The nonparametric regression model was run for the 1707 accidents in the test sample for a wide range of neighborhood sizes and the measures of effectiveness were evaluated. It should be noted that a naïve forecast was used for a neighborhood size of zero. The naïve forecast gave every test accident the same predicted clearance time from the mean clearance time of the learning sample accidents. Figure 5-3 below shows the mean absolute error of the 1707 test accidents for a given neighborhood size. The complete results for all neighborhood sizes and all measures of effectiveness are given in Appendix C.

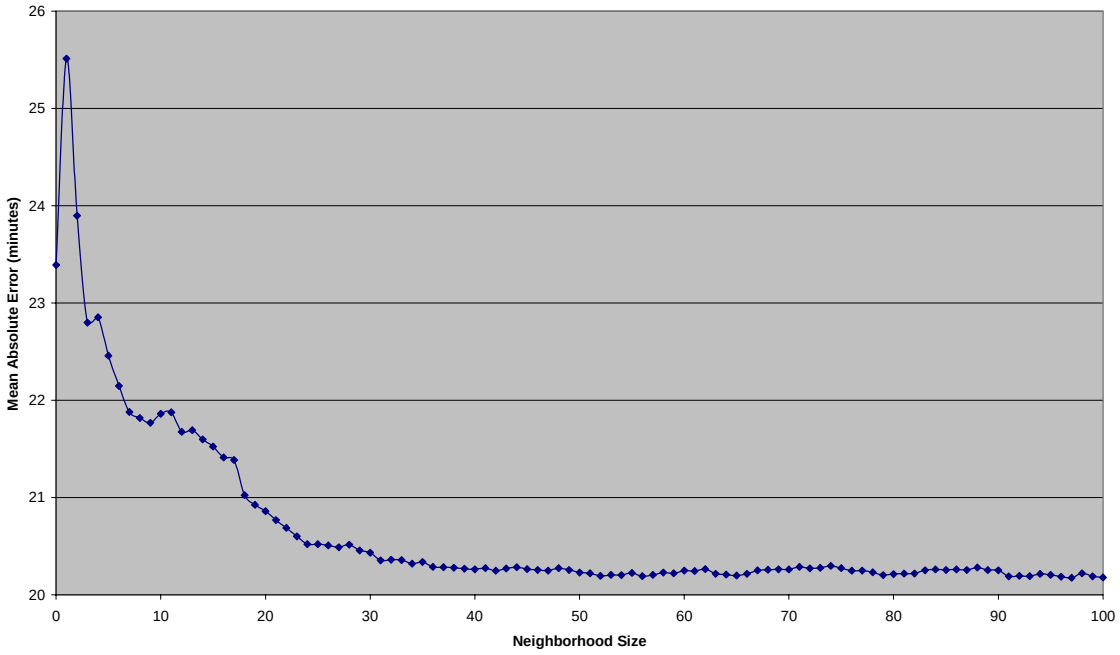

Figure 5-3: Mean absolute prediction error for range of neighborhood sizes.

This chart shows that the mean prediction error decreases for increasing neighborhood sizes. The prediction error also appears to reach a minimum around a neighborhood size of 30 and continues asymptotically around 20.2 minutes for the range of neighborhood sizes. It should be noted that the prediction error encompasses a small range from 20.3 minutes to 25.5 minutes. Selecting the best neighborhood size from this measure is difficult due to the small difference between neighborhood sizes of 30 and greater.

Figures 5-4 shows the results of the secondary measures of effectiveness, the number of test accidents predictions that were within a specific threshold of the actual clearance time. From this perspective it appears that the performance is almost insensitive to neighborhood size, and show an asymptote that begins around a size of 30 neighbors. Figures 5-5 through 5-9 are zoomed in for a specific tolerance value.

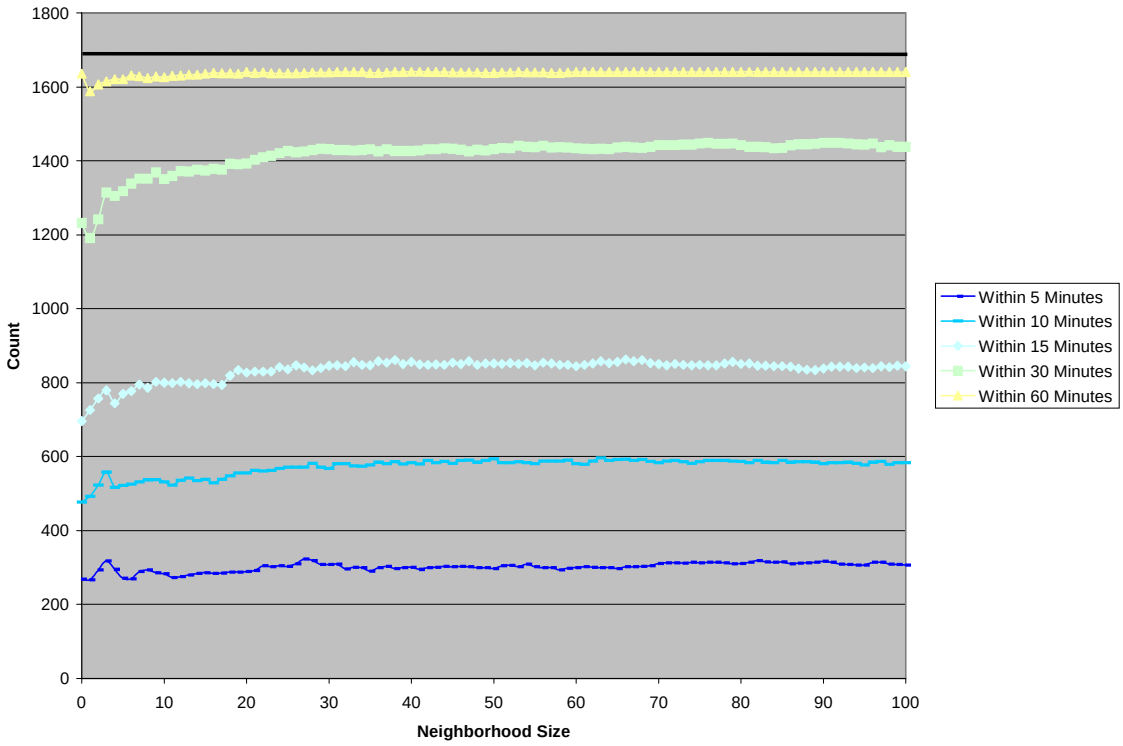

Figure 5-4: Number of test accidents predictions within X minutes of actual.

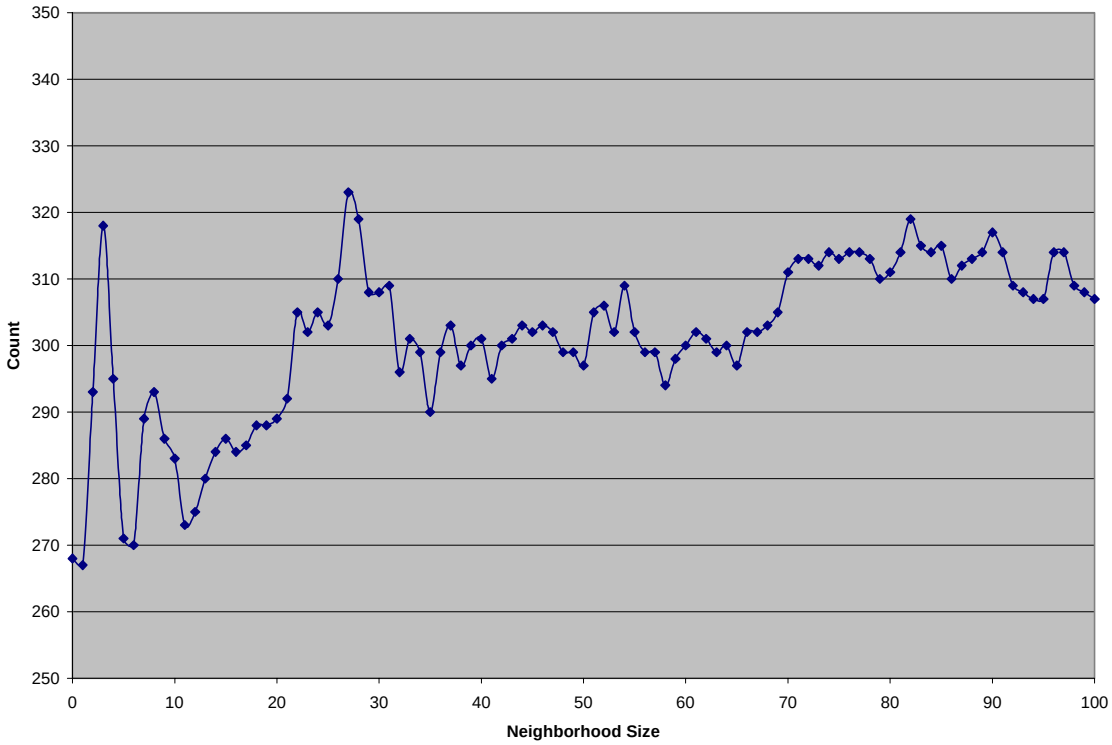

Figure 5-5: Number of prediction errors less than or equal to 5 minutes.

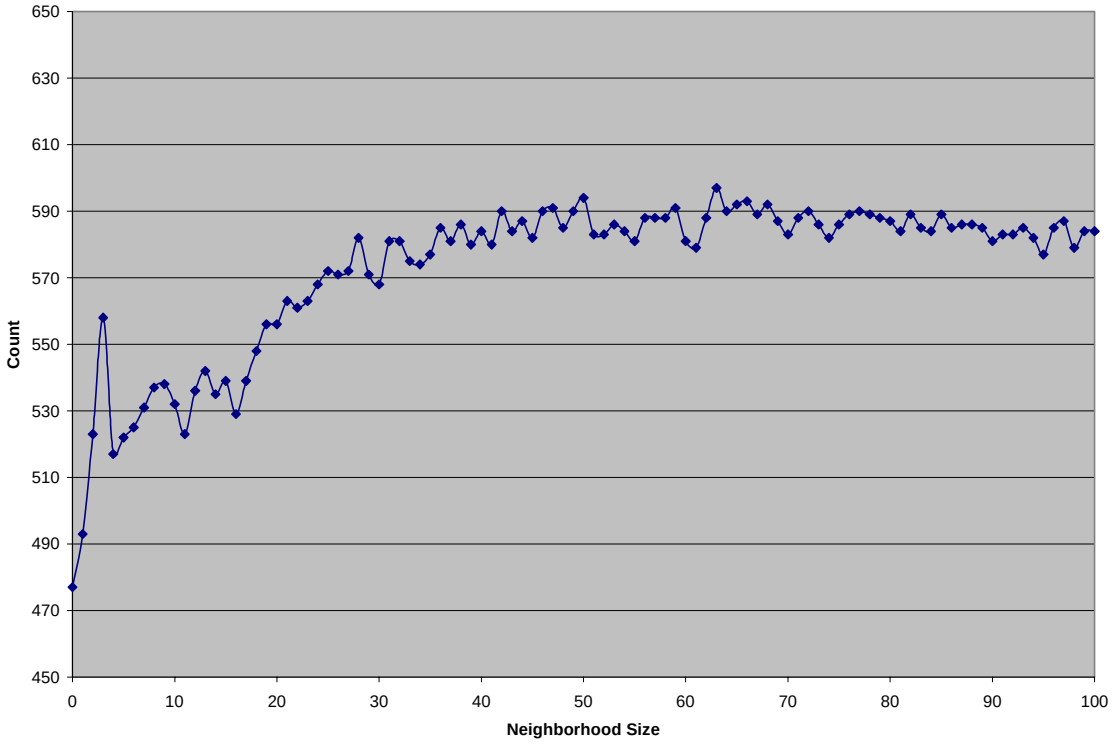

Figure 5-6: Number of prediction errors less than or equal to 10 minutes.

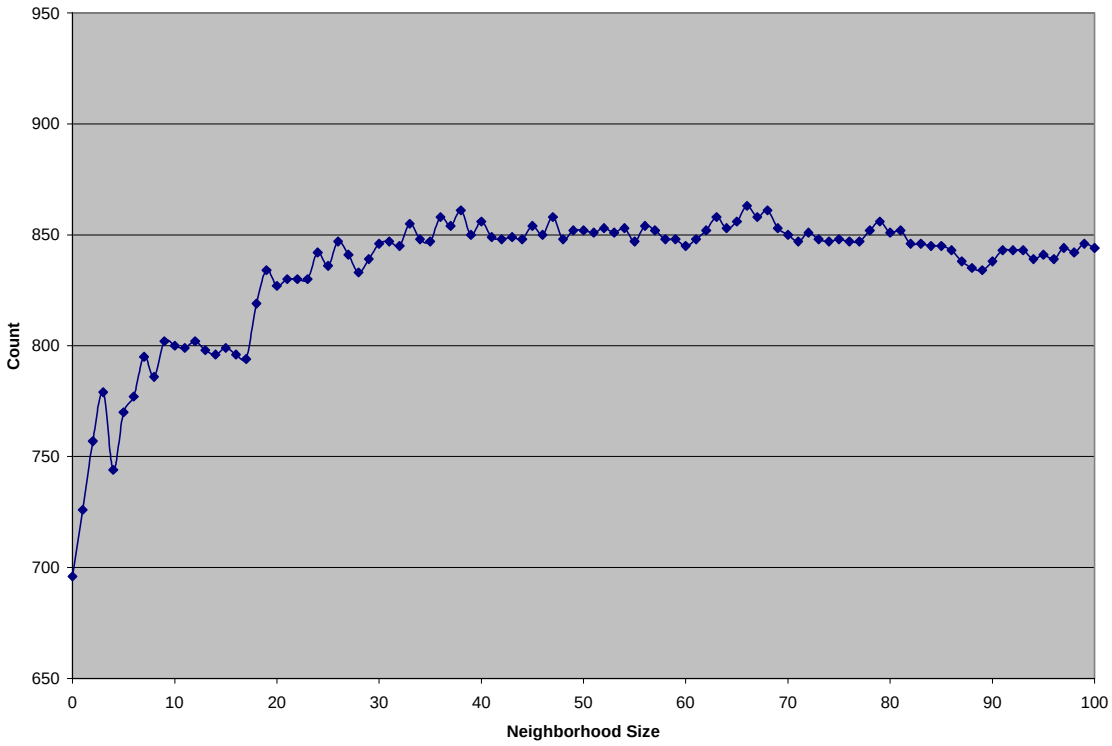

Figure 5-7: Number of prediction errors less than or equal to 15 minutes.

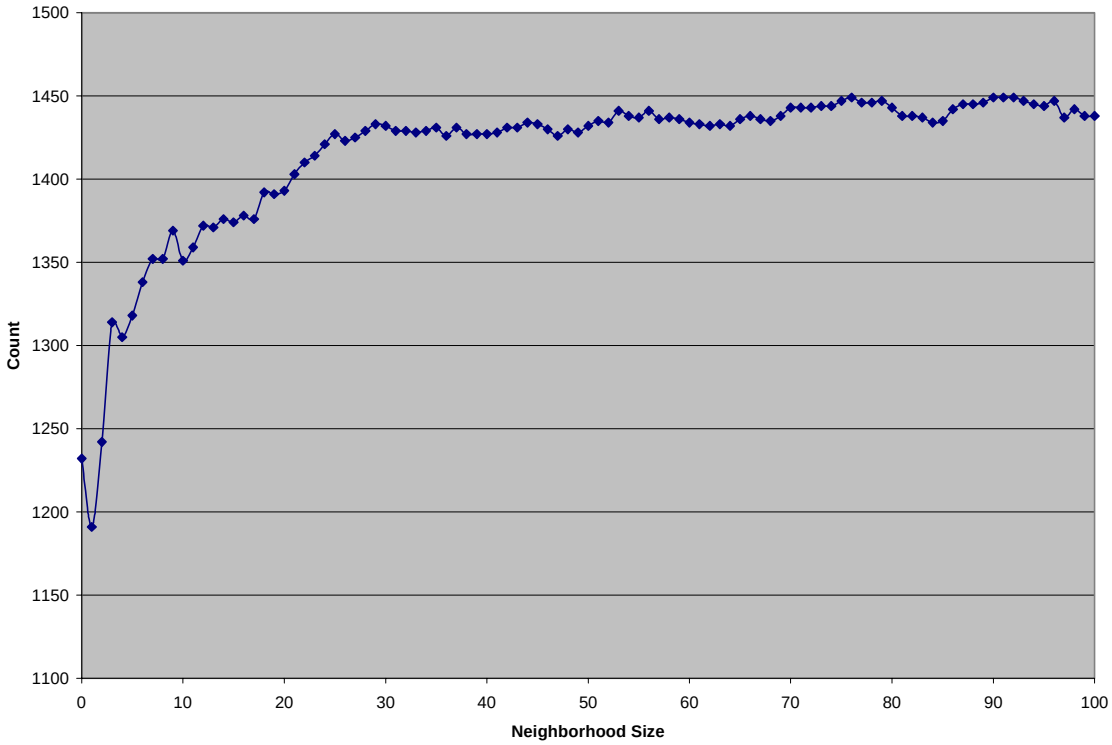

Figure 5-8: Number of prediction errors less than or equal to 30 minutes.

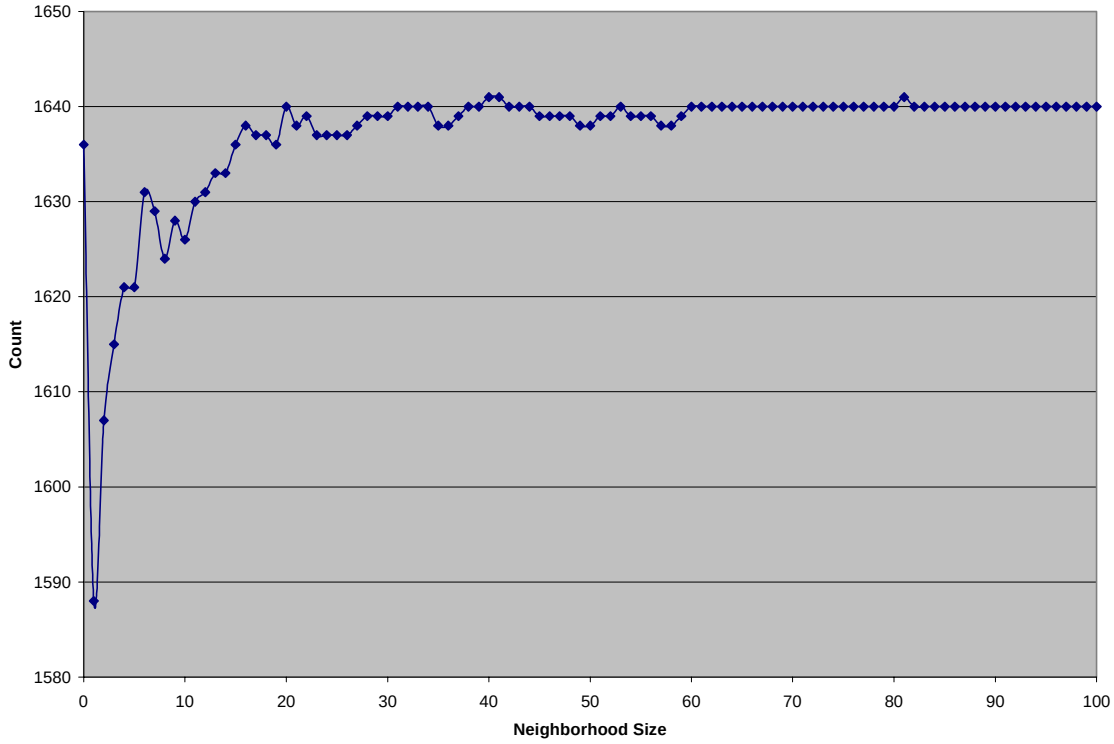

Figure 5-9: Number of prediction errors less than or equal to 60 minutes.

The charts shown above reveal a number of different features about the neighborhood size. First, the maximum value for each prediction error tolerance is reached at a different neighborhood size. Thus, the question arises as to the relative importance of having prediction errors with each tolerance. Is it more important to have the predicted accident clearance time be within 5 or 60 minutes of the actual clearance time? Obviously, the smaller error tolerances are more important to traffic managers and the charts of smaller prediction error tolerances should have more weight.

Also, with the exception of the 5 minute tolerance the number of prediction errors less than the tolerance increases with increasing neighborhood size and levels off at some point. This gives the impression that larger neighborhood sizes give better results in terms of the measures of effectiveness used in this project. This may be because the effect of past accidents with extremely large or small clearance times are absorbed by the average of a large number of samples.

There is some support in the above charts for a small neighborhood size. The neighborhood size of 3 represents a local optimum value on a number of measures of effectiveness. For prediction tolerances of 5, 10, 15, and 30 minutes, the number of test accidents increases initially up to a neighborhood size of 3 and then drops dramatically before eventually surpassing the value for the neighborhood size of 3. This local optimum is more pronounced for the smaller prediction tolerances. This suggests that a neighborhood size of 3 performs significantly better than the other smaller neighborhood

sizes. The only problem with such a small neighborhood size is that some important past accidents may be excluded from consideration.

The final recommendation for the neighborhood size to be used in the nonparametric regression model is 30 neighbors. This size is approximately the point at which the mean prediction error starts to level off, and also shows good performance with the number of predictions within the 5, 10, and 15-minute tolerances. Smaller neighborhood sizes are preferable because it forces the model to find past accidents that are more similar to the current accident. With larger neighborhood sizes, each prediction is the average of a large number of past clearance times and thus each prediction is progressing towards a common value. Thirty neighbors include enough pertinent past accidents to hopefully cancel out the effect of any outliers during the straight average forecast generation.

5.4 Model Results

From the previous section, the neighborhood size of 30 neighbors was selected as the most appropriate value for this nonparametric regression model. The model results for the 1707 accidents in the test sample for the measures of effectiveness are given below in Table 5-2.

Table 5-3: Nonparametric Regression Model Results.

Performance Measure	Value	Percent of Test Accidents
Mean Prediction Error	20.4 minutes	
Prediction Error ≤ 5 minutes	308	18.0%
Prediction Error ≤ 10 minutes	568	33.3%
Prediction Error ≤ 15 minutes	846	49.6%
Prediction Error ≤ 30 minutes	1432	83.9%
Prediction Error ≤ 60 minutes	1639	96.0%
Number of Overestimated Predictions	914	53.6%
Number of Underestimated Predictions	765	44.8%

Overall, this model averages over 20 minutes of error between the predicted and actual clearance times. Slightly less than half of the predicted clearance times were within 15 minutes of the actual time. The model also tends to overestimate the clearance time the majority of the time. This indicates that there may be some outliers with large clearance times that are influencing the predicted time.

5.5 Result Summary

Overall, the results from the nonparametric regression model are not encouraging. The model has a very large average error that in most cases is larger than the model prediction value. Considering the small percentage of test accidents with prediction errors less than 5 and 10 minutes, it does not appear that this model is applicable for use

in an existing incident management system. The next step in the methodology is to try another forecasting model. The classification tree model described in the next section has a different approach to clearance time prediction that may provide for better model performance.

Chapter 6: Classification Tree Model

6.1 Model Development

The second forecasting model that was developed for this project was a classification tree model. Classification trees were introduced in Section 2.3 in general terms. This model differs dramatically from the nonparametric regression is many ways. First, the nonparametric regression model is mathematical, in that clearance time forecasts are calculated from the clearance time average of selected past accidents. The classification tree model is more of a sorting tool based on accident characteristics. The clearance time forecasts are assigned instead of mathematically calculated. Also, unlike the nonparametric regression model the classification tree model does not require special modifications for categorical variables. This model does not explicitly compare two accidents and thus a distance metric is not required.

This model uses the same categorical independent variables that were used for the nonparametric regression model and shown in Table 5-1. The dependent variable is again the accident clearance time, but for this model the clearance time is a categorical value. The different values of the clearance time are short, medium, and long. A short clearance time is defined as 1-15 minutes, medium is 16-30 minutes, and long is 31 or greater minutes. These class divisions are based on practical experience from the HRSTC. The traffic managers prefer to categorize freeway incidents based on these classifications. The assumption from the traffic managers' standpoint is that any accident with a clearance time greater than 30 minutes is considered seriously detrimental to traffic operations.

The CART software program was used to develop the classification tree model. CART constructs classification using a brute force method. For each level of the tree where a decision node is present, CART considers each independent variable as the splitting criteria. The one split that provides the best results is selected for that decision node. This process continues until the largest possible tree has been created. Each new split creates a new classification tree that is a candidate for the optimal tree. Next, CART uses a pruning technique to determine this optimal tree. Starting from the largest tree, the testing sample is run through the classification tree to find the prediction accuracy. This continues for each smaller tree until the tree with the best prediction accuracy is found. This tree growing and pruning technique assures that the best possible splits and size are found.

6.2 Measures of Effectiveness

The main measure of effectiveness with the nonparametric regression model was the prediction error. For the classification tree, the predicted and actual clearance times are classes, so the prediction error is either a correct or incorrect prediction. Thus, the primary measure of effectiveness will be the percentage of test accidents where the clearance time was predicted correctly, termed the prediction accuracy. This prediction performance is used by CART to evaluate the potential classification trees, so the optimal tree is guaranteed to have the greatest prediction accuracy.

There are also some secondary measures of effectiveness related to the prediction accuracy. Accidents with long clearance times are most important to traffic managers, so another measure of effectiveness is the percentage of long clearance times that were predicted correctly. Likewise, it is important not to overestimate a short clearance time accident, so the prediction accuracy of short clearance time accidents will also be investigated.

6.3 Model Results

The optimal classification tree model from the CART methodology is shown below in Figure 6-1.

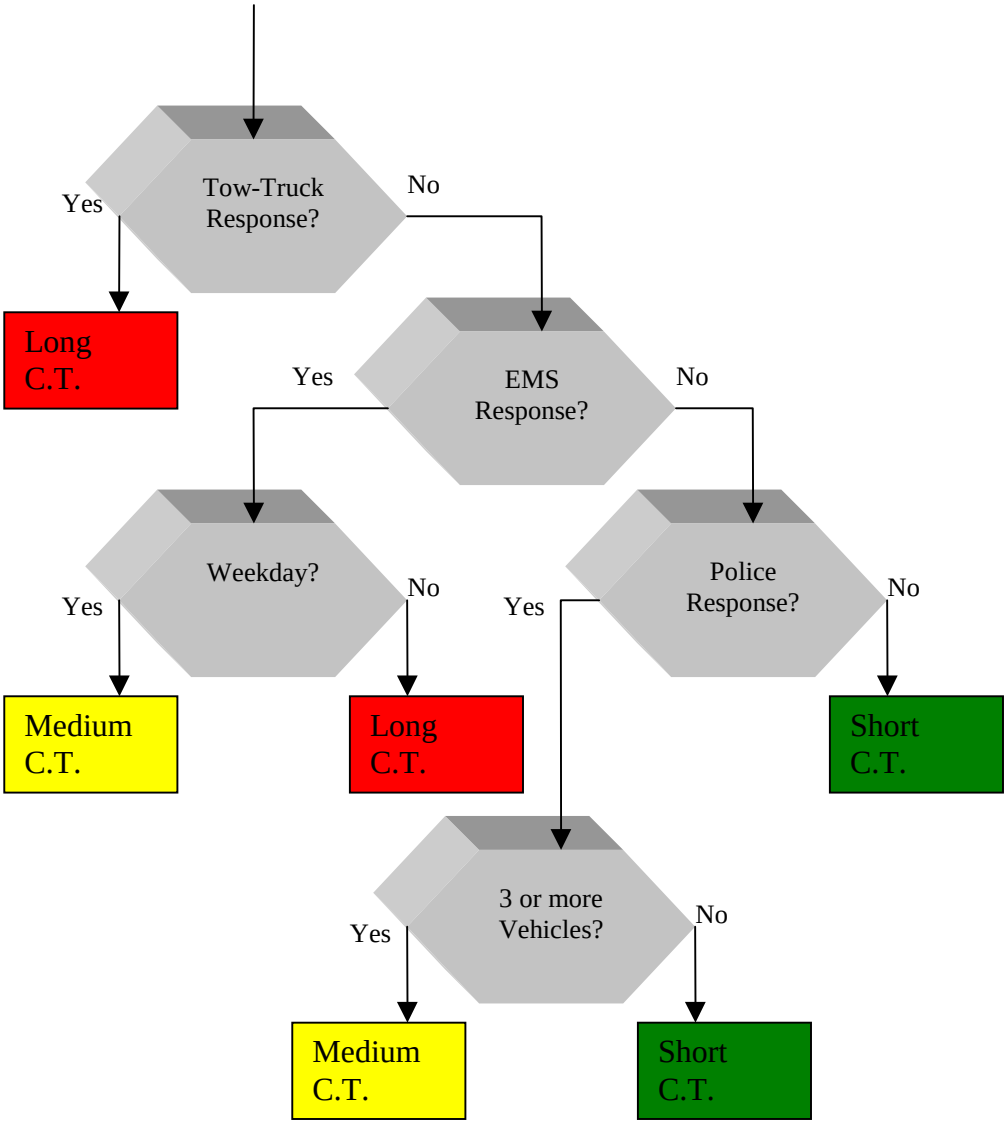

Figure 6-1: Classification tree model diagram.

The classification tree follows a path from top to bottom. Thus, the first step in predicting the clearance time of a current accident is to determine if a tow-truck was called or not. The appropriate path is followed to lead to the next decision. Eventually, the path will reach a termination node, at which time the current accident is assigned to a clearance time class or short, medium, or long.

One feature of the classification tree model is that not all of the independent variables are used. This may seem counter-intuitive since the independent variables were previously tested for their significance, and now the classification tree seems to decide that additional independent variables are not significant. The reason is that the ANOVA table tested the significance of the variable in terms of the means (real numbers) of the two samples. However, the classification tree is concerned with differentiating the class of the clearance time. Thus, the independent variables used in the classification tree model are significant in their affect on the class of the accident clearance time.

Another important note is that the classification tree model does not follow a chronological progression. The tree does not represent a series of events, but rather a series of decisions on past knowledge. The presence of a two-truck is the first decision in the tree but at the accident scene is one of the last decisions to be made. However, a tow-truck response has the greatest effect on the clearance time and thus deserves to be the most important decision in the classification tree. Since the tree does not follow a chronological progress, it is best to have all of the accident information before making a complete prediction on the clearance time.

6.3.1 Prediction Accuracy

The performance of this classification tree model is shown in Table 6-1.

Table 6-1: Classification tree model prediction accuracy.

			Predicted Class		
Actual Class	**Cases**	**% Correct**	**Short**	**Medium**	**Long**
Short	361	76.73	277	41	43
Medium	324	19.14	170	62	92
Long	1022	64.48	235	128	659
Total	1707	58.47	682	231	794

The model was tested using the test sample of accidents. The rows in Table 6-1 represent the different classes of clearance time that make up the test sample. The last three columns of the table show how the model predicted the classes. For example, the test sample contained 361 accidents with short clearance times. Of these accidents, 277 were predicted to be short, 41 to be medium, and 43 to be long. The model was only able to correctly predict the class of 277 of the 361 short accidents, or 76.73%. The main measure of prediction accuracy is the total accuracy for all of the test accidents. The table shows that only 58.47% of the test accidents inputted into the model resulted in correct clearance time predictions.

It is interesting to note that the model would perform better if all accidents were assigned long clearance times. Since accidents with long clearance times make up 60% (1022 of 1707) of the total accidents in the test sample, the classification tree model would improve to 60% accuracy. Since the model attempts to find a relationship between the accident characteristics and clearance time, this fact insinuates that the relationship may be weak and accident clearance time is an independent event.

These results show that the classification tree model performed best at accurately predicting short clearance times, and to a lesser extent the long clearance times. However, the model was ineffective at predicting medium clearance times (19% accuracy). The model appears to favor the two extremes of the clearance time classes in predictions. Table 6-1 shows that of the 1707 clearance time predictions made, 682 (40%) of these predictions were for the short class and 794 (46%) were for the long class. This means that the model only made medium clearance time predictions for 231 (14%) of the accidents. It is also possible that the characteristics that make an accident have a medium clearance time are not reflected in the independent variables used in the model.

It is also important to note that a misclassification makes no statement about the prediction error in terms of minutes. The classification model only predicts classes of clearance time. It is possible for accidents that have clearance times near the class boundaries (15 and 30 minutes) to have similar characteristics as accidents in both classes. In essence, an accident could have a short clearance time of 15 minutes, but the model would predict it have a medium clearance time (16-30 minutes) because it has similar features as most of the medium clearance time accidents in the database. This misclassification would not be severe if the prediction accuracy was measured in terms of minutes instead of classes.

6.4 Result Summary

The results from the classification model are not promising. They also do not show a marked performance improvement from the nonparametric regression models, despite the radically different approach to predicting accident clearance time. Overall, the classification tree model is only correct in predicting accident clearance time 58% of the time. This accuracy is not enough for a traffic manager to recommend this model for implementation into an operational incident management system.

Chapter 7: Conclusion

7.1 Project Conclusions

Predicting the duration an incident is one of the most important steps of the freeway incident management process (Ozbay and Kachroo, 1999). This project has attempted to predict accident clearance time, since clearance time is a major factor in the total incident duration. Using a large sample size of past freeway accidents, different models were evaluated and tested. None of the forecasting models produced results that were accurate enough to warrant implementation in an operational incident management system. The reason for the poor performance can be attributed to the choice of forecasting models and/or the quality of the accident data.

It is the conclusion of the author that the shortcomings of the accident data had the greatest contribution to the poor performance of the forecasting models. Both nonparametric regression and classification trees are proven forecasting techniques that have been applied successfully to traffic management systems. However, with unreliable input data even the best forecasting models will still output unreliable information. This is the case with the accident data that was collected from the Hampton Roads Smart Traffic Center for this project.

Some information on the data quality is seen upon examining the results of the classification tree model. For example, consider the accidents with short clearance times as recorded in the database that were predicted to have long clearance times. One of these accidents is a three-car accident with personal injuries as reported in the text description in the database along with responding agencies of state police, EMS, FIRT, and fire department. The reported clearance time of this accident was 9 minutes in the database, which seems to be inaccurate and shorter than possible. There are other examples of specific accidents with clearance times that do not appear to match the accident characteristics. The recommendations for improving the accident data are included in the following section.

7.2 Recommendations

7.2.1 Forecasting Models

This paper presented two different forecasting models with different approaches to predicting accident clearance time. Although, both models performed unsatisfactorily with a test database, the classification tree model stands out as being a better choice for an incident management system.

First of all the classification tree model is easy to understand and comprehend how a prediction is made. The series of yes/no decisions are simple and can avoid any ambiguities. The flow of the decision may not follow a chronological path, but it does mimic the thought pattern that traffic managers have used to predict accident clearance time based on past experience. The tree also reveals which accident characteristics are most important to understanding the expected clearance time. This gives traffic managers information on which areas of clearing an accident can be improved to decrease the total clearance time.

Another advantage of the classification tree model is the speed of outputting a prediction. The nonparametric regression model takes a current accident and compares it to every other accident in the historical database of past accidents. For large growing databases this has the potential to be a time consuming process. On the other hand, the classification tree model was constructed using past accidents, but it does not require access to the past accidents to make a prediction. Of course, from time to time the model should be reconstructed to reflect the addition of more accidents, but this can be done off-line without interfering with the model process. Overall, the classification tree model adopted for a computer prediction program is just a series of if/then statements that can be processed very quickly for even the most complicated tree structures.

The final feature of classification tree models that makes them advantageous to traffic managers is the output type. This model predicts a class or range of values instead of a single numerical value. In most cases, ranges of expected clearance time have more meaning than a single time.

7.2.2 Incident Databases

The accident data used in this study was suspected to have some inconsistencies in terms of how the data was collected and recorded. Most of these issues only come to light when examining the incident database as a historical collection of past accidents to be used for analysis purposes.

The incident data used in this study had an intended audience of traffic managers currently tracking the progress of the accident. A large amount of information was included in the text description (and not in other data fields) for other operators to read. Also, some of the fields did not have consistent entries, but the information is understandable when read in person. In general, a historical database is not read by a person, but rather searched for specific keywords by a computer. Thus, a historical database should conform to a non-human audience.

7.2.3 Data Entry Procedure

The first recommendation for the incident database maintained by the Hampton Roads Smart Traffic Center (HRSTC) is the elimination of the text description field. This field often contained the only information about personal injuries, but a computer can not search for all instances of personal injuries in the database for a few reasons. First, computers can not read sentences and have to search for patterns of letters. Second, there are numerous means to note a personal injuries in a sentence using abbreviations, incorrect spellings and other methods (P.I., PI, p.i., pi, Pers. Inj., 1 PI, at least 2 personal injur., personel injurys, etc.). The easiest way to note a personal injury is to have a separate data field where the value in that field is the number of injuries. This eliminates any confusion on the part of the computer that is searching the historical database.

The second recommendation is that each data field contains only one piece of information. This may result in a larger database memory due to an increased number of fields, but it makes computer queries more effective. The field that notes the lanes affected by the accident contains entries in the form of a list with each entry separated by

a comma. Another method of designing the database would be to include a separate field for each lane where the operator enters a 1 if that lane is included in the accident.

The key to an effective historical database is consistent and easy to discern data entries. This may include adding more fields, drop down lists to select entries, and clear instructions for the traffic operators who record accident information. For example, the fields in the database that record the detection source and responding agencies have over 1000 and 100 unique entries respectively. However, these unique entries only represent about 15 unique agencies. This makes the field have a much more limited use in any forecasting model.

7.2.4 Needed Accident Information

There were numerous potential independent variables that were lacking from the model development because the information was not available in the database. Some of these may have significant impacts on the clearance time of an accident.

- The number of lanes blocked due to the accident,
- The number of personal injuries, and
- The number of vehicles responding from each agency.

The more information recorded the better the possibility to create important independent variables for the forecasting models.

7.2.5 Incident Duration

The HRSTC incident database recorded only the clearance time for an accident. A more important value is the total duration of the incident. This is the length of time from the first occurrence of the incident to the time when traffic conditions return to normal. The total incident duration is an important input for predicting traffic delay on queues on the freeway.

In some cases it is not possible to record the time of the total incident duration. For example, the HRSTC incident database is a record of activity by the incident response team and thus only measures clearance time. No matter which incident phase is recorded, there must be a clear definition and guidelines to insure consistency among each incident.

Along these same lines, the length of time recorded needs to be accurate. In some cases an incident entry in the database in started, but not completed until some time after the incident has expired. When examining past events from a historical database, there needs to be confidence in the data entries that they are correct and without error. This is applied to all entries in addition to the phase duration.

7.3 Future Research

It is convenient to fit accident clearance time to a known probabilistic distribution. If accident clearance time is a random variable with known distributions, it is easy to model the expected value of future accidents. However, this paper has not proven that accident clearance time behaves like a random variable and conforms to convenient probabilistic distribution. One reason for the inability to fit a probabilistic

distribution to the accident data is the goodness-of-fit statistical fit used for the stochastic model. All goodness-of-fit tests such as the chi-square test may not be accurate for extremely large sample sizes (Law, 2001). Further research is needed on goodness-of-fit tests to determine the validity of this statement and other possible statistical tests for large sample data.

Nonparametric regression models should not be considered irrelevant for use in forecasting phases of incident duration, based on my recommendation for classification trees models in incident management systems. More research is needed on methods to define the distance metric based on categorical variables. It is possible that other weighting techniques may be more appropriate. Also, for some extremely large database, the best method may be to only search for exact matches in the database. Research is also needed to determine the best forecast generation approaches, and identify other possibilities than the simple straight average approach used in this project.

The application of classification tree models would also benefit from future research. The model results may be sensitive to the number and size of the classes used to classify accident clearance time. However, it is important to remember the audience of the forecasting models and avoid extremely small or large classes that have no meaning for transportation managers.

7.5 Summary

The output of a forecasting model is only as good as the data used to develop the model. For that reason, if the HRSTC wishes to develop an accurate incident duration prediction program, they should consider changes to their method of recording incident information. Overall, the database needs to be structured to allow for easy computer queries without the need for human interpretation. This will not impede their current use for the incident database, which is to inform other operators about a current accident. What it will require is more attention to details of the incident and consistent data entry.

References

Altman, N. S. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *The American Statistician*. Vol. 46, 1992.

Ang, Alfredo H-S and Wilson H. Tang. *Probability Concepts in Engineering Planning and Design*. New York: John Wiley & Sons, 1975.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen and Charles J. Stone. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth, 1984.

Cios, Krzysztof J., Witold Pedrycz and Roman W. Swiniarski. *Data Mining Methods for Knowledge Discovery*. Boston, MA: Kluwer Academic, 1998.

Devore, Jay L. *Probability and Statistics for Engineering and the Sciences*. Pacific Grove, CA: Brooks/Cole, 1995.

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

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

Golob, Thomas F., Wilfred W. Recker and John D. Leonard. "An Analysis of the Severity and Incident Duration of Truck-Involved Freeway Accidents." *Accident Analysis & Prevention*. Vol. 19, No. 4, August 1987. pp 375-395.

Higgins, James J. and Sallie Keller-McNulty. *Concepts in Probability and Stochastic Modeling*. Belmont, California: Duxbury Press, 1995.

Jones, Bryan, Lester Janssen and Fred Mannering. "Analysis of the Frequency and Duration of Freeway Accidents in Seattle." *Accident Analysis & Prevention*. Vol. 23, No. 4, August 1991. pp 239-255.

Khattak, Asad J., Joseph L. Schofer and Mu-Ham Wang. "A Simple Time Sequential Procedure for Predicting Freeway Incident Duration." *IVHS Journal*. Vol. 2, No. 2, 1995. pp 113-138.

Law, Averill M. *ExpertFit User's Guide*. January 2001.

Nam, Doohee and Fred Mannering. "An Exploratory Hazard-Based Analysis of Highway Incident Duration." *Transportation Research – A*. Vol. 34A, No. 2, 2000. pp 85-102.

Ozbay, Kaan and Pushkin Kachroo. *Incident Management in Intelligent Transportation Systems*. Boston, MA: Artech House, 1999.

Park, Han Chung. *An Empirical Study of Methods for Producing Multi-Variate Decision Trees*. University of Virginia Masters Thesis, August 1995.

Salford Systems. *Salford Systems White Paper Series*. 2000. [*available at* http://www.salford-systems.com/whitepaper.html].

Sethi, Vaneet, Frank S. Koppelman, Clayton P. Flannery, Nikhil Bhandari and Joseph L. Schofer. *Duration and Travel Time Impacts of Incidents – ADVANCE Project Technical Report TRF-ID-202*. Evanston, IL: Northwestern University, November 1994.

Smith, Brian L., Billy M. Williams and R. Keith Oswald. "Comparison of Parametric and Nonparametric Models for Traffic Condition Forecasting."

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

Transportation Research Board (TRB). *Special Report 209: Highway Capacity Manual*. Washington, DC: National Research Council, 1994.

Virginia Department of Transportation (VDOT). *Organization Guide*. October 2000.

Appendix A: ANOVA Significance Test Results

Time of Day (Peak)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Peak=1	2079	77823	37.4329	786.8568
Peak=0	4749	194074	40.86629	1021.251
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	17045.46	1	17045.46	17.94456	2.3E-05	3.842814
Within Groups	6483988	6826	949.8958			
Total	6501034	6827				

Day of the Week (Weekday)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Weekday=1	5181	201439	38.88033	928.537
Weekday=0	1647	70458	42.7796	1015.924
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	19001.14	1	19001.14	20.00943	7.83E-06	3.842814
Within Groups	6482033	6826	949.6093			
Total	6501034	6827				

Weather

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Weather=1	5430	215005	39.59576	995.1114
Weather=0	1398	56892	40.69528	785.4189
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1344.05	1	1344.05	1.411527	0.234844	3.842814
Within Groups	6499690	6826	952.196			
Total	6501034	6827				

EMS Response (EMS)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
EMS=1	1331	70224	52.76033	1285.827
EMS=0	5497	201673	36.68783	821.3385
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	276807.1	1	276807.1	303.5695	1.49E-66	3.842814
Within Groups	6224227	6826	911.841			
Total	6501034	6827				

FIRE Response (FIRE)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
FIRE=1	1343	69964	52.09531	1246.241
FIRE=0	5485	201933	36.8155	834.5546
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	251880.9	1	251880.9	275.1315	1.31E-60	3.842814
Within Groups	6249153	6826	915.4927			
Total	6501034	6827				

FIRT Response (FIRT)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
FIRT=1	6735	268178	39.81856	955.9196
FIRT=0	93	3719	39.98925	694.2281
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.67258	1	2.67258	0.002806	0.957755	3.842814
Within Groups	6501031	6826	952.3925			
Total	6501034	6827				

HAZMAT Response (HAZMAT)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
HAZMAT=1	12	1643	136.9167	15323.17
HAZMAT=0	6816	270254	39.64994	912.5677
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	113330.3	1	113330.3	121.1065	6.21E-28	3.842814
Within Groups	6387704	6826	935.7902			
Total	6501034	6827				

Police Response (POLICE)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
POLICE=1	4551	195146	42.87981	1015.957
POLICE=0	2277	76751	33.70707	769.216
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	127695.1	1	127695.1	136.7645	2.7E-31	3.842814
Within Groups	6373339	6826	933.6857			
Total	6501034	6827				

VDOT Response (VDOT)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
VDOT=1	70	4504	64.34286	5682.576
VDOT=0	6758	267393	39.56688	897.796
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	42528.9	1	42528.9	44.94884	2.18E-11	3.842814
Within Groups	6458505	6826	946.1625			
Total	6501034	6827				

Number of Vehicles (NUMVEH)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
NUMVEH=1	2716	112480	41.41384	1148.235
NUMVEH=2	2687	91294	33.97618	761.2258
NUMVEH=3	1425	68123	47.80561	807.1553
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	189533.5	2	94766.74	102.4769	1.41E-44	2.997048
Within Groups	6311500	6825	924.762			
Total	6501034	6827				

Truck Involvement (TRUCK)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
TRUCK=1	311	17164	55.18971	2727.999
TRUCK=0	6517	254733	39.08746	856.1065
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	76963.99	1	76963.99	81.77934	1.95E-19	3.842814
Within Groups	6424070	6826	941.1178			
Total	6501034	6827				

Passenger Bus Involvement (BUS)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
BUS=1	32	1625	50.78125	1235.854
BUS=0	6796	270272	39.76928	950.5313
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3862.248	1	3862.248	4.05772	0.044009	3.842814
Within Groups	6497172	6826	951.8271			
Total	6501034	6827				

Tow Truck Called (TOW)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
TOW=1	3386	170380	50.31896	937.9337
TOW=0	3442	101517	29.49361	751.4851
	6828			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	740268.1	1	740268.1	877.1524	1.8E-181	3.842814
Within Groups	5760766	6826	843.9446			
Total	6501034	6827				

Appendix B: Visual Basic Code for Nonparametric Regression Model

```
'Name:         Nonparametric Regression Model
'Author:       Kevin W. Smith
'              Department of Civil Engineering
'              University of Virginia
'Last Update:  March 15, 2001

'Language:     Visual Basic for Applications
'Application:  Microsoft EXCEL

'This is the code for the nonparametric regression model outlined
'in the paper. The learning sample accidents are loaded in Sheet2,
'while the testing sample accidents are loaded in Sheet1. Sheet3
'is used to perform the sorting procedures and for temporary
'storage of the output. This program predicts the duration of a
'set of test accidents, given a specific neighborhood size.

'Variable Declaration
Option Explicit
'The learning and testing samples will be copied into array
'variables for faster access during run time.
Public LearnArray(5200, 12) As Integer
Public TestArray(1800, 12) As Integer
'Dummy and place holder variables
Public i, j, k, q As Integer
Public distance, nsum As Long
Public nsize, npredict As Integer

'This subroutine is the main subroutine that calls the functions
'and other subroutines. It houses the basic structure of the model.
Public Sub NPR_Main()
   'call the subroutine to load the learn and test sample data
   NPR_LoadData
   'cycle through each test accidents
   For i = 1 To 1707
      'cycle through each learn accident
      For j = 1 To 5121
         'call the subroutine to calculate the distance between
         'the current test accident and current learn accident
         NPR_Distance
         'record the distance and duration of the current
         'learn accident to the temp sheet
         Sheet3.Cells(j + 1, 1) = distance
```

```
            Sheet3.Cells(j + 1, 2) = LearnArray(j, 1)
        'move to the next learn accident
        Next j
        'sort all learn accidents in order of increasing distance
        'to the current test accident
        Sheet3.Range("A2:B5122").Sort Key1:=Rows(1), _
              Order1:=xlAscending, Orientation:=xlSortColumns
        'call the subroutine to find the neighborhood of learn
        'accidents and make a prediction
        NPR_Neighborhood
        'clear the sorted list on the temp sheet
        Sheet3.Range("A2:B5122").Clear
        'record the prediction on the test sheet
        Sheet1.Cells(i + 1, 25) = npredict
        'display results to track run time progression
        Sheet3.Cells(1, 4) = i
        Sheet3.Cells(1, 5) = npredict
    'move to the next test accident
    Next i
End Sub

'This subroutine calculates the distance between two accidents based
'on their independent variable values. If the two accidents have
'different values for an independent variable, a penalty is added to
'the distance. "Distance" variable is the sum of all penalties.
Public Sub NPR_Distance()
    'reset the dummy distance variable each time
    distance = 0
    'calculation for PEAK variable
    If TestArray(i, 2) <> LearnArray(j, 2) Then
        'if unequal add penalty
        distance = distance + 3.43
    End If
    'calculation for the WEEKDAY variable
    If TestArray(i, 3) <> LearnArray(j, 3) Then
        'if unequal add penalty
        distance = distance + 3.9
    End If
    'calculation for the EMS variable
    If TestArray(i, 4) <> LearnArray(j, 4) Then
        'if unequal add penalty
        distance = distance + 16.07
    End If
    'calculation for the FIRE variable
    If TestArray(i, 5) <> LearnArray(j, 5) Then
        'if unequal add penalty
```

```
    distance = distance + 15.28
End If
'calculation for the HAZMAT variable
If TestArray(i, 6) <> LearnArray(j, 6) Then
    'if unequal add penalty
    distance = distance + 97.27
End If
'calculation for the POLICE variable
If TestArray(i, 7) <> LearnArray(j, 7) Then
    'if unequal add penalty
    distance = distance + 9.17
End If
'calculation for the VDOT variable
If TestArray(i, 8) <> LearnArray(j, 8) Then
    'if unequal add penalty
    distance = distance + 24.78
End If
'calculation for the NUMVEH variable with
'three possible values
If TestArray(i, 9) <> LearnArray(j, 9) Then
    'if unequal add penalty
    If TestArray(i, 9) = 1 Then
        If LearnArray(j, 9) = 2 Then
            distance = distance + 7.44
        Else
            distance = distance + 6.39
        End If
    ElseIf TestArray(i, 9) = 2 Then
        If LearnArray(j, 9) = 1 Then
            distance = distance + 7.44
        Else
            distance = distance + 13.83
        End If
    ElseIf TestArray(i, 9) = 3 Then
        If LearnArray(j, 9) = 1 Then
            distance = distance + 6.39
        Else
            distance = distance + 13.83
        End If
    End If
End If
'calculation for the TRUCK variable
If TestArray(i, 10) <> LearnArray(j, 10) Then
    'if unequal add penalty
    distance = distance + 16.1
End If
```

```
    'calculation for the BUS variable
    If TestArray(i, 11) <> LearnArray(j, 11) Then
        'if unequal add penalty
        distance = distance + 11.01
    End If
    'calculation for the TOW variable
    If TestArray(i, 12) <> LearnArray(j, 12) Then
        'if unequal add penalty
        distance = distance + 20.83
    End If
End Sub

'This subroutine defines the neighborhood of learn accidents for the
'current test accident. The neighborhood size is defined in this
'subroutine and needs to be updated manually for each trial run. This
'subroutine also generates a forecast using a simple average of the
'duration of each learn accident in the neighborhood.
Public Sub NPR_Neighborhood()
    'define the neighborhood size for all test accidents
    nsize = 65
    'reset the dummy variable
    nsum = 0
    'from the list of learn accidents sorted by distance, select the
    'first "nsize" accidents and sum the duration values
    For k = 1 To nsize
        nsum = nsum + Sheet3.Cells(k + 1, 2)
    Next k
    'generate forecast using simple average of learn accidents in the
    'neighborhood from above
    npredict = Int(nsum / nsize)
End Sub

'This subroutine loads the dependent and independent variables of
'the learn and test accidents into array variables. This allows
'faster access to the values during run time.
Public Sub NPR_LoadData()
    For q = 1 To 5121                'load each learning incident
        LearnArray(q, 1) = Sheet2.Cells(q + 1, 2)    'load duration
        LearnArray(q, 2) = Sheet2.Cells(q + 1, 6)    'load peak
        LearnArray(q, 3) = Sheet2.Cells(q + 1, 8)    'load weekday
        LearnArray(q, 4) = Sheet2.Cells(q + 1, 11)   'load ems
        LearnArray(q, 5) = Sheet2.Cells(q + 1, 12)   'load fire
        LearnArray(q, 6) = Sheet2.Cells(q + 1, 14)   'load hazmat
        LearnArray(q, 7) = Sheet2.Cells(q + 1, 15)   'load police
        LearnArray(q, 8) = Sheet2.Cells(q + 1, 16)   'load vdot
        LearnArray(q, 9) = Sheet2.Cells(q + 1, 18)   'load numveh
```

```
      LearnArray(q, 10) = Sheet2.Cells(q + 1, 19)  'load truck
      LearnArray(q, 11) = Sheet2.Cells(q + 1, 20)  'load bus
      LearnArray(q, 12) = Sheet2.Cells(q + 1, 21)  'load tow
   Next q
   For q = 1 To 1707                'load each testing incident
      TestArray(q, 1) = Sheet1.Cells(q + 1, 2)    'load duration
      TestArray(q, 2) = Sheet1.Cells(q + 1, 6)    'load peak
      TestArray(q, 3) = Sheet1.Cells(q + 1, 8)    'load weekday
      TestArray(q, 4) = Sheet1.Cells(q + 1, 11)   'load ems
      TestArray(q, 5) = Sheet1.Cells(q + 1, 12)   'load fire
      TestArray(q, 6) = Sheet1.Cells(q + 1, 14)   'load hazmat
      TestArray(q, 7) = Sheet1.Cells(q + 1, 15)   'load police
      TestArray(q, 8) = Sheet1.Cells(q + 1, 16)   'load vdot
      TestArray(q, 9) = Sheet1.Cells(q + 1, 18)   'load numveh
      TestArray(q, 10) = Sheet1.Cells(q + 1, 19)  'load truck
      TestArray(q, 11) = Sheet1.Cells(q + 1, 20)  'load bus
      TestArray(q, 12) = Sheet1.Cells(q + 1, 21)  'load tow
   Next q
End Sub
```

Appendix C: Nonparametric Regression Model Results for all Neighborhood Sizes

Neighborhood Size	Mean Prediction Error (mins)	Number of Underestimates	Number of Overestimates	Number of Exact Predictions	Prediction Error <= 5 Minutes	Prediction Error <= 10 Minutes	Prediction Error <= 15 Minutes	Prediction Error <= 30 Minutes	Prediction Error <= 60 Minutes
0	23.390744	758	924	25	268	477	696	1232	1636
1	25.51318102	934	749	24	267	493	726	1191	1588
2	23.89806678	851	830	26	293	523	757	1242	1607
3	22.79906268	844	834	29	318	558	779	1314	1615
4	22.85354423	831	849	27	295	517	744	1305	1621
5	22.45869947	854	829	24	271	522	770	1318	1621
6	22.14762742	853	826	28	270	525	777	1338	1631
7	21.87932045	836	845	26	289	531	795	1352	1629
8	21.81898067	830	861	16	293	537	786	1352	1624
9	21.76918571	825	861	21	286	538	802	1369	1628
10	21.86233158	818	867	22	283	532	800	1351	1626
11	21.87639133	803	884	20	273	523	799	1359	1630
12	21.67603984	806	877	24	275	536	802	1372	1631
13	21.69244288	810	880	17	280	542	798	1371	1633
14	21.59929701	812	878	17	284	535	796	1376	1633
15	21.52489748	804	885	18	286	539	799	1374	1636
16	21.41300527	801	880	26	284	529	796	1378	1638
17	21.38664323	787	892	28	285	539	794	1376	1637
18	21.02460457	776	903	28	288	548	819	1392	1637
19	20.92560047	779	900	28	288	556	834	1391	1636
20	20.85940246	769	904	34	289	556	827	1393	1640
21	20.76918571	768	902	37	292	563	830	1403	1638
22	20.68892794	774	902	31	305	561	830	1410	1639
23	20.6016403	771	904	32	302	563	830	1414	1637
24	20.52138254	773	905	29	305	568	842	1421	1637
25	20.52196837	769	910	28	303	572	836	1427	1637
26	20.50790861	771	908	28	310	571	847	1423	1637
27	20.4897481	780	896	31	323	572	841	1425	1638
28	20.51611013	775	901	31	319	582	833	1429	1639
29	20.456942	765	912	30	308	571	839	1433	1639
30	20.4340949	765	914	28	308	568	846	1432	1639
31	20.35559461	760	913	34	309	581	847	1429	1640
32	20.36203866	771	908	28	296	581	845	1429	1640
33	20.35852373	772	908	27	301	575	855	1428	1640

34	20.32044523		765	904	38		299	574	848	1429	1640
35	20.33919156		770	912	25		290	577	847	1431	1638
36	20.28705331		770	910	27		299	585	858	1426	1638
37	20.28412419		762	912	33		303	581	854	1431	1639
38	20.27885179		769	910	28		297	586	861	1427	1640
39	20.27006444		768	911	28		300	580	850	1427	1640
40	20.26303456		767	906	34		301	584	856	1427	1641
41	20.27475103		765	908	34		295	580	849	1428	1641
42	20.24780316		775	901	31		300	590	848	1431	1640
43	20.27123609		783	896	28		301	584	849	1431	1640
44	20.28471002		777	900	30		303	587	848	1434	1640
45	20.26537786		775	907	25		302	582	854	1433	1639
46	20.25717633		771	905	31		303	590	850	1430	1639
47	20.25073228		776	906	25		302	591	858	1426	1639
48	20.27475103		773	907	27		299	585	848	1430	1639
49	20.25659051		775	906	26		299	590	852	1428	1638
50	20.23022847		774	896	37		297	594	852	1432	1638
51	20.22202695		775	907	25		305	583	851	1435	1639
52	20.19566491		773	901	33		306	583	853	1434	1639
53	20.20503808		774	903	30		302	586	851	1441	1640
54	20.20445226		772	902	33		309	584	853	1438	1639
55	20.22554189		774	901	32		302	581	847	1437	1639
56	20.19214997		776	900	31		299	588	854	1441	1639
57	20.2056239		780	890	37		299	588	852	1436	1638
58	20.22964265		776	893	38		294	588	848	1437	1638
59	20.22144112		780	889	38		298	591	848	1436	1639
60	20.25014646		774	900	33		300	581	845	1434	1640
61	20.24545987		779	900	28		302	579	848	1433	1640
62	20.26420621		779	899	29		301	588	852	1432	1640
63	20.21792619		778	896	33		299	597	858	1433	1640
64	20.20913884		776	890	41		300	590	853	1432	1640
65	20.19859402		784	888	35		297	592	856	1436	1640
66	20.21792619		784	892	31		302	593	863	1438	1640
67	20.2513181		782	897	28		302	589	858	1436	1640
68	20.25834798		782	894	31		303	592	861	1435	1640
69	20.26244874		785	890	32		305	587	853	1438	1640
70	20.26069127		790	893	24		311	583	850	1443	1640
71	20.28705331		778	890	39		313	588	847	1443	1640
72	20.27357938		783	891	33		313	590	851	1443	1640
73	20.27768014		780	894	33		312	586	848	1444	1640
74	20.29759813		780	892	35		314	582	847	1444	1640
75	20.2741652		781	894	32		313	586	848	1447	1640
76	20.24838899		782	893	32		314	589	847	1449	1640
77	20.24956063		787	895	25		314	590	847	1446	1640
78	20.23140012		784	896	27		313	589	852	1446	1640
79	20.20445226		788	898	21		310	588	856	1447	1640
80	20.21441125		784	901	22		311	587	851	1443	1640
81	20.21909783		778	904	25		314	584	852	1438	1641
82	20.22026948		780	898	29		319	589	846	1438	1640
83	20.25307557		777	901	29		315	585	846	1437	1640
84	20.26186292		778	901	28		314	584	845	1434	1640

85	20.25659051		770	904	33		315	589	845	1435	1640
86	20.26069127		769	903	35		310	585	843	1442	1640
87	20.25717633		771	900	36		312	586	838	1445	1640
88	20.28060926		768	905	34		313	586	835	1445	1640
89	20.25483304		770	900	37		314	585	834	1446	1640
90	20.25248975		768	911	28		317	581	838	1449	1640
91	20.19156415		773	903	31		314	583	843	1449	1640
92	20.19507909		767	902	38		309	583	843	1449	1640
93	20.19214997		770	901	36		308	585	843	1447	1640
94	20.21558289		764	904	39		307	582	839	1445	1640
95	20.20503808		764	907	36		307	577	841	1444	1640
96	20.18746339		773	899	35		314	585	839	1447	1640
97	20.17574692		775	898	34		314	587	844	1437	1640
98	20.22144112		766	908	33		309	579	842	1442	1640
99	20.19097832		770	905	32		308	584	846	1438	1640
100	20.17984769		769	909	29		307	584	844	1438	1640

Appendix D: Map of HRSTC Location Zones