AN ABSTRACT OF THE THESIS OF

Lijuan Wang for the degree of Master of Science in Industrial Engineering presented on September 19, 2008.
Title: An Application of Regression Tree Methodology in Freeway Travel Time Estimation Using Speed as a Proxy.

Abstract approved:

Rasaratnam Logendran

Accurate freeway travel time estimation is of increasing importance for the travelers' information and route guidance system. A non-parametric statistical methodology known as regression trees is deployed in this research for dynamically and accurately estimating freeway travel times for the I5-I205 loop in the Portland Metro area of Oregon using speed as a proxy. In the absence of historical travel time data on PORTAL (Portland Oregon Regional Transportation Archive Listing), which is the source of data collection in this research, regression tree models are built to predict speeds first and the predicted speeds are in turn used to estimate travel times by mid-point algorithm.

The regression tree models in this research are built based on historical data sets, including not only the traffic flow data but also the incident related data, weather data and time of day. This ensures the models will maintain stable prediction ability under both free flow conditions and non-free flow conditions on freeways. Model construction and validation are implemented in the statistical software package S-PLUS. A full regression tree model is constructed on one test data set including 227 daily test data sets randomly selected from the total of 342 daily test data sets collected in the entire year of 2005.

To determine what kind of regression tree model should be selected to predict speed or estimate travel time for a certain day under dynamic conditions, a characterization approach is deployed and four characterization standards are setup to track the characteristics of both test data sets and validation data sets.

Two experimental designs are constructed to evaluate and compare the performances of eleven regression tree models ─ the full regression tree model and the ten characterization regression tree models. The results show that these eleven tree models possess the ability to accurately predict speeds or estimate travel times. In addition, meaningful results are obtained showing which of these eleven tree models are best to choose for dynamically estimating travel times for a future day.

An Application of Regression Tree Methodology in Freeway Travel Time Estimation Using Speed as a Proxy

by

Lijuan Wang

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented September 19, 2008
Commencement June 2009

Master of Science thesis of Lijuan Wang presented on September 19, 2008

APPROVED:

Major Professor, representing Industrial Engineering

Head of the School of Mechanical, Industrial & Manufacturing Engineering

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Lijuan Wang, Author

ACKNOWLEDGEMENTS

First of all, my sincere gratitude goes to Dr. Rasaratnam Logendran, my Major Professor, from whom the idea of the research originates. I would like to thank him for his skillful and patient guidance throughout my Master's study and the pursuit of this research. I also appreciate his work on review and revision of this thesis. His scientific and earnest attitude in research will definitely be beneficial to my future work and life.

I would like to express my gratitude to the faculty and the staff of the School of Mechanical, Industrial & Manufacturing Engineering (MIME) for their guidance and support throughout my Master's study.

I would like to thank my committee members: Dr. Lan Xue, my Minor Professor, Dr. David Kim, my committee member, and Dr. John Sessions, my graduate council representative, for taking the time to evaluate my thesis.

I am also very grateful of the financial support from the Oregon Department of Transportation.

Finally, I would like to express my deepest gratefulness to my parents for their support and encouragement. Without them, I would not have had the opportunity to come to OSU to pursue my Master's degree. I also would like to thank my boyfriend, Kevin Claborn, for his comfort and support whenever I ran into difficulties in the research.

TABLE OF CONTENTS

Page

1. INTRODUCTION 1

2. LITERATRUE REVIEW 4

3. PROBLEM STATEMENT 9

4. REGRESSION TREE METHODOLOGY 11

5. MODEL DEVELOPMENT 14

5.1. Data Collection 14

5.2. Raw Data Reorganizations 18

6. ALGORITHM IMPLEMENTATION IN S-PLUS 22

6.1. Model Construction 22

6.1.1. Model Construction Using the Example Test Data 22

6.1.2. Full Regression Tree Model Construction 25

6.2. Model Validation 27

7. TRAVEL TIME ESTIMATION 30

8. CHARACTERIZATION APPROACH 33

9. EXPERIMENTAL DESIGN AND RESULT DISCUSSIONS 37

9.1. One Sample T-test 38

9.2. RCBD Experimental Design 40

9.3. Analysis of Results 43

10. CONCLUSIONS AND FUTURE RESEARCH 46

BIBLIOGRAPHY 51

APPENDICES 54

LIST OF FIGURES

Figure Page

4.1 Example of splitting the test data starting with volume....................................12

6.1 Results summary of the example regression tree model...................................23

6.2 Regression tree plot of the example test data set 24

6.3 Results summary of the full regression tree model...26

6.4 Tree plot of the full regression tree model 26

7.1 Influence area around the detector station 30

9.1 G*Power window of sample size determination for t-tests 39

LIST OF TABLES

Table Page

4.1 Example test data for regression tree model construction 12

5.1 Raw volume data at station I-205 NB Gladstone on 03/23/05 (9:10 – 10:10 am). 15

5.2 Incident data at the station I-205 NB Gladstone on March 23, 2005................16

5.3 Partial hourly weather data (3:00 – 11:00 am)..17

5.4 Test data with traffic flow data and incident data (9:10 – 10:10 am)................20

5.5 Test data with traffic flow data, incident data and weather data.......................21

6.1 Example validation data set 28

6.2 Example validation data set with fitted speeds and MSE 29

7.1 MSE result of estimated travel times by full regression tree model 31

8.1 Characterization results for test data sets and validation data sets 35

9.1 T-test results of the performances of the eleven regression tree models...........40

9.2 Number of validation data sets used for RCBD...42

9.3 Results of ANOVA and multiple comparisons for ten RCBDs.........................44

LIST OF APPENDICES

Appendix Page

A. DATA COLLECTION FOR TRAFFIC FLOW VARIABLES IN PORTAL ... 55

B. DATA COLLECTION FOR INCIDENT RELATED VARIABLES IN PORTAL ……………………………………………………………………………58

C. DATA COLLECTION OF WEATHER DATA VARIABLES IN PORTAL 61

D. EXCEL VBA PROGRAMS FOR RAW DATA REORGANIZATIONS......... 63

E. MODEL CONSTRUCTION IN S-PLUS USING EXAMPLE TEST DATA .. 76

F. RANDOM NUMBER GENERATION PROGRAM....................................... 81

G. VALIDATION OF REGRESSION TREE MODEL IN S-PLUS 83

H. MACRO IN EXCEL VBA FOR CHARACTERIZATION 87

I. REGRESSION TREE MODELS OF TEN CHARACTERIZATIONS 91

J. THE USE OF G*POWER TO DETERMINE SAMPLE SIZES 114

K. CALCULATION OF MSE VALUES USED IN RCBD............................... 116

L. RCBD FOR VALIDATION DATA SETS IN TEN CHARACTERIZATIONS……………………………………………...……….118

M. ANOVA AND MULTIPLE COMPARISONS FOR RCBDS IN S-PLUS 124

LIST OF APPENDIX FIGURES

Figure Page

A.1 Screen shot of PORTAL system for traffic flow data collection......................56

B.1 Screen shot of the PORTAL system for incident data collection 58

B.2 Incident data graph on I-205 NB on March 23, 2005 59

B.3 Incident data table on I-205 NB on March 23, 2005.......................................59

C.1 Screen shot of PORTAL system for weather data collection 62

D.1 Raw daily data set at the station I-205 NB Gladstone on Jan 10th, 2006 (0:00 – 1:55 am) 64

D.2 Record new Macro 65

D.3 Personal macro workbook..65

D.4 Unhide PERSONAL.XLS..65

D.5 Exit Excel..65

D.6 Open Visual Basic Editor...66

D.7 Insert a new module 66

D.8 Missing data 68

D.9 Erroneous data..68

D.10 The organized test data – after the first Macro is run....................................69

D.11 The organized test data – after the second Macro is run 70

D.12 The organized test data – after the third Macro is run...................................71

D.13 The organized test data – after the fourth Macro is run 72

LIST OF APPENDIX FIGURES (Continued)

Figure Page

E.1 Import test data set into S-PLUS 76

E.2 Imported test data set 77

E.3 Open tree models window 77

E.4 “Model” tab 77

E.5 “Results” tab 77

E.6 “Plot” tab 77

E.7 Results summary of the example regression tree model 79

E.8 Regression tree plot of the example test data set 80

G.1 Importation of validation data set into S-PLUS 83

G.2 Validation data set at I-205 NB Gladstone on Aug 2, 2006 (7:05 – 9:05 pm). 84

G.3 Open tree models window 85

G.4 “Tree Models” window – "Predict" tab 85

H.1 Record new Macro 89

H.2 Personal Macro workbook 89

H.3 Unhide PERSONAL.XLS 89

H.4 Exit Excel 89

H.5 Open Visual Basic Editor 89

LIST OF APPENDIX FIGURES (Continued)

Figure Page

H.6 Insert a new Module 89

H.7 Open Macro window 90

H.8 Macro window 90

I.1 Tree plot of regression tree model of characterization 3 92

I.2 Tree plot of regression tree model of characterization 4 93

I.3 Tree plot of regression tree model of characterization 5 95

I.4 Tree plot of regression tree model of characterization 6 98

I.5 Tree plot of regression tree model of characterization 7 100

I.6 Tree plot of regression tree model of characterization 8 102

I.7 Tree plot of regression tree model of characterization 11 103

I.8 Tree plot of regression tree model of characterization 12 104

I.9 Tree plot of regression tree model of characterization 14 109

I.10 Tree plot of regression tree model of characterization 16 113

J.1 G*Power software 114

K.1 MSE calculation 117

L.1 RCBD for validation data sets in characterization 3 118

L.2 RCBD for validation data sets in characterization 4 119

LIST OF APPENDIX FIGURES (Continued)

Figure Page

L.3 RCBD for validation data sets in characterization 5 119

L.4 RCBD for validation data sets in characterization 6 120

L.5 RCBD for validation data sets in characterization 7 120

L.6 RCBD for validation data sets in characterization 8 121

L.7 RCBD for validation data sets in characterization 11..................................... 121

L.8 RCBD for validation data sets in characterization 12 122

L.9 RCBD for validation data sets in characterization 14 122

L.10 RCBD for validation data sets in characterization 16 123

M.1 Open ANOVA window 126

M.2 “Model” tab in ANOVA 126

M.3 “Compare” tab in ANOVA 126

LIST OF APPENDIX TABLES

Table Page

A.1 Partial traffic flow data (9:10 – 10:10 am) 55

A.2 Raw volume data at station I-205 NB Gladstone on March 23, 2005 57

B.1 Incident data at the station I-205 NB Gladstone on March 23, 2005 58

C.1 Partial hourly weather data (0:00 – 11:00 am) 61

G.1 MSE result of validation data on August 2, 2006 86

M.1 Adjusted data set of RCBD for characterization 11 125

This thesis is dedicated to

My parents

An Application of Regression Tree Methodology in Freeway Travel Time Estimation Using Speed as a Proxy

1. INTRODUCTION

With the occurrence of frequent congestion on urban freeways, travelers experience frequent delays in their journeys. The extent of delays in a given route of the journey is unknown information for the travelers before they start driving in the route. Therefore, accurate freeway travel time information is of increasing importance for the travelers' information and the route guidance system. Travel times provide valuable information not only for traveler routing but also for transportation scheduling and management. Since travel times give a real-time measurement of congestion, they are also used to detect incidents. However, a traditional method for collecting travel time data, the floating car surveys technique, is usually costly to perform for accurate travel time measurements (Rickman *et al.,* 1990). Turner (1996) compared several advanced techniques for travel time data collection, including electronic distance-measuring instruments (DMIs), computerized and video license plate matching, cellular phone tracking, automatic vehicle identification (AVI), automatic vehicle location (AVL), and video imaging. He pointed out that these techniques either require a significant investment to be employed or are still in testing stages, with uncertainty about their cost and accuracy. As a result, accurately estimating travel time has become a high priority.

In the last decade advanced traveler information systems (ATISs) have been developed and installed in a number of U.S. cities to inform travelers about current travel conditions and travel time estimates and, therefore, improve freeway efficiency. If the travel time estimates provided by ATISs are accurate and timely even under dynamic conditions, travelers can make wise decisions about route selection to avoid congestion and incidents. A number of different models and algorithms have been developed for estimating travel time, such as linear regression models (Wong and Sussman, 1973; Kwon *et al.*, 2000), Kalman filtering models (Chu *et al.* 2005), artificial neural network (ANN) models (Park and Rilett, 1998; Park *et al.*, 1999; Rilett and Park, 2001), etc. In most travel time estimation studies, the input data typically consists of traffic volume and occupancy from loop detectors (Dailey, 1997; Coifman, 2002), travel time data from probe vehicles (Chen and Chien, 2001) and AVI (Park and Rilett, 1998). The above studies will be reviewed briefly in the next chapter.

This research aims at employing a non-parametric statistical methodology known as regression trees in freeway travel time estimations by using speed as a proxy. Regression trees cannot be used to directly estimate travel times because historical travel time data are not available on PORTAL (Portland Oregon Regional Transportation Archive Listing), which is the source of data collection in this research. Regression trees, also known as hierarchical tree-based regression (HTBR), are introduced by Brieman *et al.* (1984) in their classic text titled "*Classification and Regression Trees*." A regression tree is constructed by recursively partitioning the data into homogeneous regions within which constant or linear estimates are generally fitted (Lee et al., 2006). Within the last 20 years, there has been an increasing interest in the use of regression tree analysis. Regression tree methodology has been applied in several studies related to traffic security and accident analysis (Golias and Karlaftis, 2001; Karlaftis and Golias, 2002; Chong *et al*., 2004; Chang and Chen, 2005; Tesema *et al*., 2005; Chang and Wang, 2006). These studies will be reviewed briefly in the next chapter.

The regression tree models in this research are built based on the daily historical data sets, including not only the traffic flow data but also the incident related data, weather data and time of day. This ensures the models will maintain stable prediction ability among different flow conditions on the freeway. This research focuses on examining the use of regression trees in estimating travel time in the I5-I205 loop in the Portland Metro area in Oregon. All of the data used in the research are collected from the PORTAL system managed by Portland State University. As historical travel time data are not available in the PORTAL system, the regression tree models are built to predict speed first, and then the mid-point algorithm is used to estimate travel time.

To determine what kind of regression tree model should be selected to predict speed or estimate travel time for a certain day, a characterization approach is deployed and four characterization standards are setup to track the characteristics of both test data sets and validation data sets. The objective of this research is to dynamically and accurately estimate freeway travel times using regression trees. Dynamic estimation of travel time is important because both the travelers' information system and the route guidance system need to provide travel time estimates under dynamic conditions. The characterization approach provides the regression tree models the capability to estimate travel times dynamically. At the same time, a comprehensive full regression tree model is constructed based on the historical data collected in one entire year.

The performance of the characterization regression tree models and the full regression tree model are then compared through a randomized complete block design (RCBD) and multiple comparisons are also performed using Tukey's method and Fisher Least Significant Difference (LSD) method. Regression tree analysis, RCBD and multiple comparisons are performed by use of the statistical software package S-PLUS version 7.0 (The Insightful Corp., 2005). Finally, each of the characterization regression tree models and the full regression tree model are evaluated individually by using the one-sample t-test.

2. LITERATRUE REVIEW

Travel time information is of increasing importance for the real time travelers' information and route guidance system. It provides valuable information for traveler routing and transportation scheduling. Most of the advanced techniques for travel time data collection, including electronic DMIs, computerized and video license plate matching, cellular phone tracking, AVI, AVL and video imaging either require a significant investment for application or are still in testing stages and have uncertainty about their accuracy (Turner, 1996). Therefore, the development of algorithms and models for travel time estimation is more preferable for the travelers' information and route guidance system.

The source of data for most of the studies in travel time estimation is the traffic surveillance system, i.e. loop detector system. The output from loop detectors contains information on traffic volumes, occupancy levels and arrival patterns. These data may be applied directly or may be used in functions relating them to other important parameters defining the performance of a freeway network, such as travel time. As discussed below, there have been a number of studies that have attempted to estimate freeway travel times directly using flow measurements from the loop detectors.

The basis for the following three works was constructing stochastic models of traffic flow to estimate travel times. Dailey (1993) used cross-correlation of the freeway flow at the upstream and downstream detectors to estimate travel time between two single loop detectors. This statistical method, however, does not work well under congested traffic conditions, because the correlation disappears in such situations. The modeling approach adopted by Nam and Drew (1996) for estimating freeway travel times was based on stochastic queuing theory and the principle of conservation of vehicles. By assuming that during a given interval of time the travel times may be regarded as draws from the same probability distribution, another stochastic model (Petty *et al.*, 1998) was developed to estimate travel times directly using the flow and occupancy data collected by the single loop detectors. Probe vehicle data and travel time estimates from dual loop detectors were used to corroborate the accuracy of their methods. Most of these studies have focused on overcoming the problems in speed estimation from single loop data, assuming that the travel times estimated from dual loop data are accurate. However, the travel times estimated from dual loop speeds may also be inaccurate under congested traffic conditions.

Among the studies attempted to develop relationships between travel time and loop detector data (flow, occupancy, or both), the vast majority of existing work had focused on the use of linear regression analysis to estimate travel times. Wong and Sussman (1973) proposed a regression model that predicted route travel time on the basis of historical mean route travel time and real-time route travel time information by updating the parameter values of the regression model. A linear regression with stepwise variable selection method was applied by Kwon *et al.* (2000) to estimate future travel times on a freeway using flow and occupancy data from single loop detectors and historical travel time information from probe vehicles. They found that using the linear regression method on the current flow and occupancy measurements are favorable for short-term travel time forecasts (up to 20 minutes), while historical data are better predictors for longer-range travel time predictions. Zhang and Rice (2003) presented a time-varying coefficient linear model in which the coefficients vary as smooth functions of the departure time. The relationship between the anticipated travel time and a travel time estimate was modeled as being transiently approximately linear, using the observed travel times from probe vehicles and the speeds measured by dual loop detectors.

A thorough review of linear regression methods to determine travel times was given by Sisiopiku and Rouphail (1994). They pointed out that most of the existing linear models for travel time estimation are site-specific, which means that these models are sensitive to location. Site dependency limits the applicability and transferability of the models under different demand, control, and geometric configurations. Furthermore, Carey and McCartney (2003) proved that most of the linear models for travel time estimation exhibited a form of pseudo-periodicity due to the linearity of the travel time functions. More specifically, they showed that if there is a sharp change (increase or decrease) in the inflow rate this generates a sharp change in the outflow rate. This sharp change in outflows in turn generates an infinite series of sharp changes in the outflow rate, which damp out over time. On the other hand, if traffic flows are not changing very sharply over time, these linear models could be very acceptable approximations.

The Kalman filtering method has been applied to many traffic studies, such as the prediction of travel times. Dailey (1997) presented an algorithm to estimate travel time using speed estimates, which were obtained by the Kalman filter approach using volume and occupancy data from a series of single loop detectors. The Kalman filtering method was also used to perform travel time predictions based on real-time travel time data collected by probe vehicles on a path and its consisting links (Chen and Chien, 2001). Although path travel time at a given time period can be obtained by adding the travel times on all consisting links (i.e., links that make up the path), Chen and Chien showed that directly measuring path travel time could generate a more accurate prediction than adding consisting-link travel times. However, their models were based on the assumption of recurrent traffic (incident-free) conditions and are sensitive to incidents, meaning that the performance of their models is not stable in the presence of incidents. Chu *et al.* (2005) proposed a method for section travel time estimation by applying the Adaptive Kalman filter technique that incorporated two data sources, i.e. point detector data and area-wide probe data. Compared to the methods of using either probe data or dual loop detector data only, the proposed algorithm outperformed under both recurrent and non-recurrent traffic conditions despite the errors in loop detectors. However, in practice, the probe data are not always available.

A few studies deployed ANNs for real-time travel time forecasting. Park and Rilett (1998) used a modular neural network (MNN) that combined artificial intelligence clustering techniques with a conventional ANN for predicting travel times. It was found that the MNN approach required a more labor intensive effort on the part of the modeler compared with that required by a conventional ANN. Park *et al.* (1999) and Rilett and Park (2001) proposed spectral neural networks (SNN) that combine a pre-transformation of the input features and conventional ANN for forecasting link travel times and corridor travel times, respectively. All of these three studies used link travel times collected by AVI as data input, which required a large investment in new detector infrastructure.

This research focuses on applying a non-parametric statistical methodology known as regression trees in freeway travel time estimations using speed as a proxy. There has been an increasing interest in the use of regression tree analysis. Regression tree methodology has been applied in several studies related to traffic security and accident analysis. Golias and Karlaftis (2001) applied HTBR, also known as regression trees, to identify which external factors affect the related aspects of self-reported driver behavior, and found that regression trees are extremely robust to the effects of outliers and the multicollinearity between the independent variables. Karlaftis and Golias (2002) applied HTBR to analyze the effects of road geometry and traffic characteristics on accident rates for rural two-lane and multilane roads. Their study concluded that HTBR (a non-parametric model) without any assumption of the underlying distribution of the model, has both theoretical and applied advantages over multiple linear and negative binomial regression models (parametric models) in analyzing highway accident rates. Chong *et al.* (2004) modeled the severity of injury resulting from traffic accidents using ANNs and regression trees. Their experimental results revealed that in all the cases regression trees outperformed ANNs. Tesema *et al.* (2005) built a decision support system to handle road traffic accident analysis for the Addis Ababa city traffic office using adaptive regression trees. The tree model they developed helped decision makers to understand all the issues causing accidents resulting in fatalities with an accuracy of 87.47% to formulate better traffic safety control policies. Chang and Chen (2005) used classification and regression tree (CART) models to establish a relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. By comparing the analysis and prediction results of negative binomial regression models, this study demonstrated that CART is a good alternative for analyzing freeway accident frequency. Chang and Wang (2006) applied CART models to analyze risk factors that can influence injury severity in traffic accidents. They demonstrated that CART models effectively deal with large data sets containing a large number of explanatory variables and can produce useful results by using only a few important variables.

Besides the applications of regression trees in traffic security and accident analysis, Lederer *et al.* (2005) performed a study using regression trees to examine the effects of geometrical and operational characteristics on changes in vehicle activity levels on entrance ramps. Lee *et al.* (2006) adopted a regression tree algorithm to analyze the winter maintenance on highways and found that it was very effective to analyze the large amount of data without bias. The tree models developed can explain various relationships that exist between variables without sacrificing prediction accuracy, which is really needed in building models for travel time estimation. Moreover, the regression tree method can overcome the limitation of the large amount of outliers and complex relationships among all of the variables considered for travel time estimation, which existing models of travel time estimation could not deal with.

One thing worth noting is that Kwon *et al.* (2000) compared their methodology based on linear regression with regression trees in their research on the performances of both methods in travel time estimation. They pointed out that the errors of the tree method are consistently larger than those of linear regression; however, the differences are slight and not statistically significant. Furthermore, their regression trees were constructed only upon two types of data, i.e. flow and occupancy data from single loop detectors and travel time data from probe vehicles.

3. PROBLEM STATEMENT

Most of the existing studies on travel time estimation focus on developing models using only traffic flow data, such as speed, occupancy and flow rate. However, there are more factors that have significant influence on travel time than simply traffic flow factors. These factors may include incidents, weather, construction, special events, and time of day, all of which may have a relationship with recurring congestion or non-recurring congestion on the freeway (Ishimaru *et al.*, 2003). Therefore, the congestion that occurs as a result of such factors would affect travel time estimation in a negative way, if they are not considered in travel time estimation. Kwon *et al.* (2000) also mentioned in the conclusions of their study that other relevant parameters, such as major events and weather conditions, should be included in their model. The higher the number of independent variables considered in a parametric model, the higher the likelihood of multicollinearity among the independent variables. The parametric models, such as linear regression models, tend to be the most common models used in travel time estimation. However, multicollinearity causes problems in using linear regression models to draw conclusions about the relationships between predictors and outcome. That is the reason why most of the previous studies on travel time estimation were limited to traffic flow variables, rather than to include more relevant factors in the model development. Therefore, in order to consider these related factors in this research and avoid the negative effect of possible multicollinearity, a non-parametric statistical methodology known as regression trees is employed to estimate freeway travel time. Regression trees are robust to the multicollinearity among the independent variables and the effects of outliers (Brieman *et al.*, 1984).

This research focuses on examining the use of regression trees in estimating travel time in the I5-I205 loop in the Portland Metro area in Oregon. All of the data used in the research are collected from the PORTAL system managed by Portland State University. Ideally, the regression tree model would be constructed to estimate travel time directly. However, because the actual historical travel time data are not yet available in the PORTAL system, regression tree models are built to predict speeds first, and speeds in turn is used in the mid-point algorithm to estimate travel times. The traffic flow data, incident data and weather data collected from PORTAL are all detector-station wise, instead of segment-of-freeway wise. Therefore, the regression tree models constructed based on the data in this research are station-specific. To estimate travel time at one detector station, only the regression tree models developed based upon the data from the same station can be used. To estimate travel times in a segment of highway, the travel times at every station in that segment of highway need to be estimated first by using the predicted speeds obtained from using the regression tree model constructed for each station. Based on the assumption that all the detector stations included in the segment are independent, the travel time for the segment of highway may be estimated by adding the estimated travel time at each station in that segment of highway.

The objectives of this research are as follows:

(i) To develop a full regression tree model capable of estimating freeway travel time by using speed as a proxy in the mid-point algorithm. The model is to be based on all pertinent data, including traffic flow data, incident data, weather data, and time of day.

(ii) To characterize the test data and validation data, and to construct characterization regression tree models capable of dynamically estimating travel time.

(iii) To construct an experimental design to compare the performance of characterization regression tree models with that of the full regression tree model.

(iv) To evaluate the performances of the characterization regression tree models and the full regression tree model individually and show how capable each is in accurately predicting speed or estimating travel time.

4. REGRESSION TREE METHODOLOGY

Regression trees were first introduced by Breiman *et al.* (1984) in their classic text titled “Classification and Regression Trees.” The tree model is constructed through binary recursive partitioning by which the data are consecutively split along the explanatory variables. Each explanatory variable is evaluated sequentially, and the variable which results in the largest decrease of the deviance in the response variable is selected. Deviance is calculated based on a threshold value in the explanatory variable and this threshold value generates two mean values for the response variable: one mean above the threshold and the other below the threshold. Splitting continues until no further reduction in deviance can be obtained or the data points are too sparse. If the response variable is a qualitative factor, the tree is called a classification tree. If the response variable is continuous, the tree is called a regression tree. The data set which is split to construct the tree model is called test data, while the data which is fed into the tree model for prediction purposes is called validation data. The regression tree algorithm can be further explained by using the following example.

In the test data shown in Table 4.1, speed is the response variable, while volume and occupancy are two explanatory variables. Since speed is a continuous variable, the tree model based on this test data would be a regression tree model. To construct a regression tree model using the above described procedure, it starts with assessing any explanatory variable, i.e. volume or occupancy in this case. For example, starting with volume, the assessment steps are as follows:

- Select a threshold value, say 306, of the explanatory variable volume (the vertical dotted line in Figure 4.1).
- Calculate the mean value of the response variable speed, above and below this threshold, which are 62.19 and 59.00, respectively (the two horizontal solid lines in Figure 4.1).
- Use the two means to calculate the deviance. The deviance is defined as

$$D = \sum_i \sum_j (y_{ij} - \mu_i)^2$$

where μ_i is the mean value of the response variable speed, above or below the

Table 4.1 Example test data for regression tree model construction

Speed	Volume	Occupancy
58.00	252.00	1.00
61.00	192.00	0.67
62.33	324.00	0.67
58.00	288.00	0.67
63.00	432.00	1.00
64.00	492.00	2.00
62.33	360.00	1.33
61.67	408.00	1.00
68.33	480.00	1.33
66.33	372.00	0.67
61.67	384.00	1.33
61.50	324.00	0.67
60.00	564.00	1.67
62.33	432.00	1.67
60.33	516.00	1.67
59.00	396.00	1.00
61.00	588.00	1.33
61.33	708.00	2.00
61.00	984.00	3.00
61.00	876.00	2.33

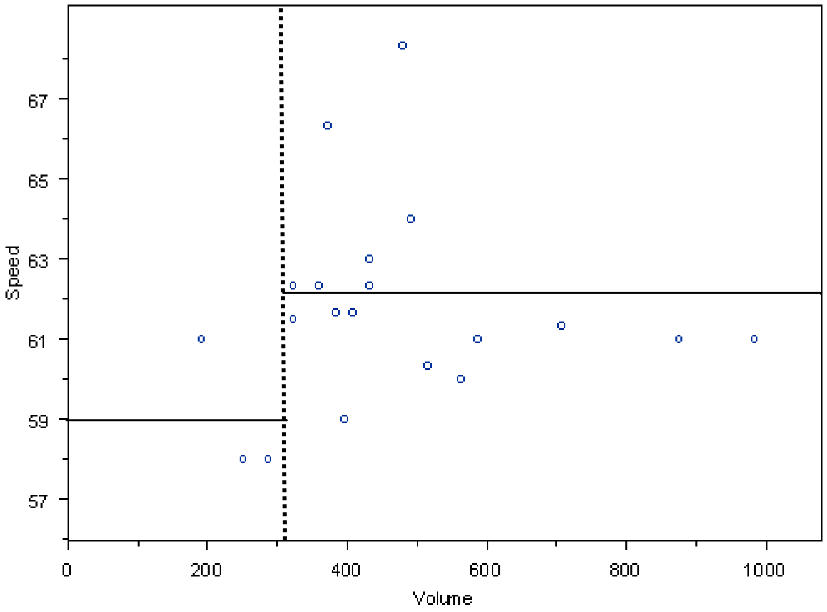


Figure 4.1 Example of splitting the test data starting with volume

threshold selected in the first step (say i = 1 is above the threshold and i = 2 is below the threshold); y_{ij} is the value of the response variable speed, above or below the threshold; i = the total number of all the subsets separated by all the selected threshold values on the explanatory variables (i = 2 in this case); j = the number of all the values of the response variable in a certain subset separated by the threshold on the explanatory variables.

- Look to see which value of the threshold gives the lowest deviance.
- Split the data into high and low subsets on the basis of the threshold for the variable volume.
- Repeat the whole procedure on each subset of the data.
- Continue until no further reduction in deviance is obtained, or there are too few data points to merit further subdivision.

After the regression tree is constructed, a validation data set, which contains the same response variable and explanatory variables as the test data set does, can be used to run through the constructed tree to obtain the predicted values for the response variable. Each row of validation data set is used to run through the constructed tree by following the splits of the tree. The predicted value of the response variable for that row of validation data can be obtained after a leaf node of the tree is reached.

5. MODEL DEVELOPMENT

Instead of the model being fixed as in traditional mathematical programming models, the regression tree model is dependent on test data. Thus, to develop the regression tree model to predict speed and then estimate travel time, the explanatory variables over which the test data are split along need to be illustrated first. In developing the regression tree-based model to predict speed, not only the traffic flow variables, but also the incident presence related variables, weather data variables, and time of day, are considered as explanatory variables. Traffic flow variables are considered for free-flow conditions, and the incident related variables, weather data variables and time of day are considered for non-free flow conditions. These ensure that the model has the same prediction ability among different flow conditions on freeways.

Although the data for the explanatory variables considered can be collected on PORTAL based on different time periods, such as one day, three days, a week, etc., in this research the data are collected on a daily basis, the shortest time period, to better track the traffic pattern. In the test data for developing the regression tree model, the response variable is speed, and all the explanatory variables considered are classified into four types: traffic flow variables, incident related variables, weather data variables and time of day variable. By using the detector station I-205 Northbound (NB) Gladstone as an example, the data collected at this station on a certain day, say March 23, 2005, for the four types of explanatory variables to form the test data is demonstrated below.

5.1. Data Collection

- Traffic flow variables

The traffic flow variables include two explanatory variables – (i) volume (number of vehicles per lane per hour); (ii) occupancy (percentage of time when the detector stayed occupied while vehicles passed) – and the response variable speed (miles per hour). Speed, volume and occupancy were collected in the “Grouped Data” archive on PORTAL on a daily basis in 5-minute increments, which is the smallest time increment available on PORTAL, in order to more accurately track traffic patterns in regression tree model construction. The traffic flow data at the station I-205 NB Gladstone on March 23, 2005 is collected as shown in Appendix A. Only a part of the collected raw volume data from 9:10 to 10:10 am is shown in Table 5.1 in the interest of space. The unit “vplph” of the raw volume data means “number of vehicles per lane per hour”. The data of “Avg. Percentage Good Data” shown in the third column in Table 5.1 is not always accurate according to PORTAL staff and therefore the data in that column is not used.

Table 5.1 Raw volume data at station I-205 NB Gladstone on March 23, 2005 (9:10 – 10:10 am)

Time	Avg Volume (vplph)	Avg Percentage Good Data
9:10	1008	1
9:15	1080	0.93333
9:20	928	1
9:25	1032	1
9:30	1232	1
9:35	1264	1
9:40	1196	1
9:45	1248	1
9:50	1188	1
9:55	1208	1
10:00	1144	1
10:05	1004	1
10:10	1300	1

- Incident related variables

Seven incident related variables – start time of incident, duration of incident (the time period from the occurrence of an incident until it is cleared), incident type (e.g. stall, crash, etc.), affected lanes by incident (such as right lanes, left lanes), number of affected lanes, hazard materials (hazmat) and number of fatalities – are considered to track the impact of incidents on speed/travel time in the regression tree model. The data for these seven variables are collected on a daily basis in the “Timeseries” archive on PORTAL. The process used to collect the incident data for these seven explanatory variables is shown in Appendix B. The raw incident data at the station I-205 NB Gladstone on March 23, 2005 collected from PORTAL is shown in Table 5.2.

Table 5.2 Incident data at the station I-205 NB Gladstone on March 23, 2005

ID	Primary Route	Location	Number of Lanes Affected	Start Time (hh:mm:ss)	Duration (min)	Incident Type	Affected Lanes	Hazmat	Number of Fatalities
421624	"I-205"	"I-205 NB GLADSTONE"	0	9:32:55	14	Debris	All Lanes	no	0

- Weather data variables

Adverse weather, such as heavy rainfall, snowfall, low visibility, etc, is a considerable cause of an increased risk of traffic accidents and compromised traffic flow on highway. Therefore, the test data would be preferable if the constructed regression tree model is capable of predicting speed even in non-free flow conditions related to weather. Three weather data variables, namely wind speed (miles per hour), rainfall (millimeters of rainfall) and visibility (miles), are considered because strong wind, heavy rainfall and low visibility could affect speed significantly. Another weather data variable “temperature” is not considered because the temperature data in the PORTAL system was found to be incomplete and also because extreme temperature conditions do not occur often in the I5-I205 loop in the Portland Metro area. The method used to collect the weather data on the PORTAL system is shown in Appendix C, with Table 5.3 showing the partial hourly weather data (3:00 – 11:00 am) for the station I-205 NB Gladstone on

March 23, 2005.

Table 5.3 Partial hourly weather data (3:00 – 11:00 am)

Time	Temp f	Wind speed ms	Visibility mi	Rainfall
3/23/2005 3:00	44.06	3	10	0
3/23/2005 4:00	44.06	0	10	0
3/23/2005 5:00	46.04	0	10	0
3/23/2005 6:00	46.04	9	10	1
3/23/2005 7:00	46.04	10	10	0
3/23/2005 8:00	46.04	0	10	1
3/23/2005 9:00	46.04	4	10	0
3/23/2005 10:00	46.94	4	10	1
3/23/2005 11:00	46.04	5	7	2

- Time of day variable

The time of day variable is important because of the existence of recurring congestion. During recurring congestion, speed usually gets lowered notably. To better track the traffic patterns, the smallest time increment available in the PORTAL system, 5 minutes, is used in the final test data set.

Besides considering these four types of explanatory variables, the preliminary investigations in this research focused on including another explanatory variable — vehicle classification — because the fraction of vehicles in each vehicle class on freeways can affect the estimated freeway travel times. However, the vehicle classification was later ignored, upon finding out that the data on vehicle classification was not available in the PORTAL system. With the above thirteen explanatory variables considered, daily test data and validation data need to be collected for regression tree model construction and evaluation. As mentioned earlier in Chapter 3, although the data for the explanatory variables can be collected on a daily basis for a station or a segment of highway on PORTAL, the traffic flow data for a segment of highway on PORTAL is shown as traffic flow data at every station in that segment of highway instead of single values for the entire segment of highway. At the same time, incident data are also station-specific on PORTAL. Thus, the research reported below would focus on speed prediction and travel time estimation at detector stations. The detector station I-205 NB Gladstone with milepost 11.05 is randomly selected to collect the daily test data sets and validation data sets.

To capture all the characteristics over a whole year in the regression tree model, test data sets are collected by collecting the data for the explanatory variables in the whole year of 2005 at the selected station. Thus 342 daily test data sets are collected (23 days of data are incomplete for unknown reasons). Validation data sets were used to validate the regression tree models by analyzing MSEs obtained from the validation results in the later experimental design. Since MSE is the response variable in the experimental design, a large amount of validation data sets means that a large amount of MSEs can be obtained in the later design, i.e., a large sample size for the experimental design. A large enough sample size can lead to a smaller effect size and high power of test in the experimental design. Therefore, 532 daily validation data sets are collected at the same station in the whole year of 2006 and the first half year of 2007 (i.e., a duration of 1.5 years and 14 days of data are incomplete for unknown reasons). All of the data sets are collected manually by copying from the PORTAL system into Excel files. The raw data for the four types of explanatory variables in the same day are kept in the same Excel file.

5.2. Raw Data Reorganizations

After the raw data are collected for the response variable and the four types of explanatory variables described above, there arises a question as to how the data collected for the different variables, which are in different formats, can be organized in one test data to construct a regression tree model or in one validation data to use the model to predict speed. Therefore, the raw data reorganizations as described below need to be performed. At the same time, since the regression tree algorithm is implemented in S-PLUS, the reorganized test data and the reorganized validation data must be compatible in S-PLUS. Because raw data reorganizations are needed for every raw daily data set collected, it means 874 daily data sets in total need to be reorganized, including 342 collected test data sets and 532 validation data sets. To save time and increase accuracy, four macros written in Excel Visual Basic Application (VBA) are developed to reorganize daily raw data saved in Excel files as described in Appendix D.

- Traffic flow variables

Speed, volume and occupancy were collected on a daily basis grouped by 5 minutes, which is consistent with the time setup in the final daily test data set. The only change needed is to delete the unnecessary column "Avg. Percentage Good Data" collected with traffic flow data.

- Incident related variables

Two of the seven incident related variables, the start time and duration, need to be shown in the final test data set indirectly. That is, the other five data items—incident type, affected lanes, number of affected lanes, hazard materials and number of fatalities—are inserted into the final daily data set according to the start time and duration. Since the time frame in the final daily data sets is in 5-minute increments, which is decided by the time frame of the traffic flow variables, the time point for insertion of those five incident related data items can be found by rounding the start time of the incident to the closest 5-minute increment of time. For example, if the start time of one incident is 8:01:36 am, then the inserting time point will be 8:00 am, instead of 8:05 am. The incident data collected at the station I-205 NB Gladstone on March 23, 2005, shown in Table 5.2, will be used as an example to illustrate how to insert the raw data for the seven incident variables into the final test data set.

It is easy to see in Table 5.2 that the incident debris occurred at 9:32:55, which can be rounded to 9:35 in a 5-minute increment of time. Thus, the incident data (incident type, affected lanes and number of affected lanes) are inserted into the test data to start at 9:35 and end at 9:50, as shown in Table 5.4, because the duration of this incident is 14 minutes and the cleared time of 9:49 can be rounded to 9:50.

Table 5.4 Test data with traffic flow data and incident data (9:10 – 10:10 am)

Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities
9:10	3024	60	8.67	None	None	0	No	0
9:15	3240	59.67	10.67	None	None	0	No	0
9:20	2784	58.33	9.33	None	None	0	No	0
9:25	3096	59	9.33	None	None	0	No	0
9:30	3696	56	12.33	None	None	0	No	0
9:35	3792	57.67	12	Debris	All lanes	0	No	0
9:40	3588	58.33	11.33	Debris	All lanes	0	No	0
9:45	3744	55.67	12.33	Debris	All lanes	0	No	0
9:50	3564	58	11.33	Debris	All lanes	0	No	0
9:55	3624	58	12	None	None	0	No	0
10:00	3432	61	11	None	None	0	No	0
10:05	3012	57.33	9	None	None	0	No	0
10:10	3900	56.33	11.67	None	None	0	No	0

- Weather data variables

The smallest time frame of the data for these three weather variables on PORTAL is on a daily basis grouped by hour. Thus, to insert the data for weather variables into the final test data in 5-minute increments, the weather data for one hour needs to be inserted into all the time points in that hour in the test data as shown in Table 5.5, which is the partial test data at the station I-205 NB Gladstone on March 23, 2005 (9:10 – 10:10 am).

As shown in Table 5.5, all the time points from 9:10 to 9:55 share the same hourly weather data at 9 o'clock taken from Table 5.3, while the time points from 10:00 to 10:10 share the same hourly weather data at 10 o'clock.

Table 5.5 Test data with traffic flow data, incident data and weather data

Time	Volume	Speed	OC	Incident Type	Affected Lanes	NAL	Hazmat	NF	Temp	WS	Rainfall	VS
9:10	3024	60	8.67	None	None	0	No	0	46.04	4	0	10
9:15	3240	59.67	10.67	None	None	0	No	0	46.04	4	0	10
9:20	2784	58.33	9.33	None	None	0	No	0	46.04	4	0	10
9:25	3096	59	9.33	None	None	0	No	0	46.04	4	0	10
9:30	3696	56	12.33	None	None	0	No	0	46.04	4	0	10
9:35	3792	57.67	12	Debris	All lanes	0	No	0	46.04	4	0	10
9:40	3588	58.33	11.33	Debris	All lanes	0	No	0	46.04	4	0	10
9:45	3744	55.67	12.33	Debris	All lanes	0	No	0	46.04	4	0	10
9:50	3564	58	11.33	Debris	All lanes	0	No	0	46.04	4	0	10
9:55	3624	58	12	None	None	0	No	0	46.04	4	0	10
10:00	3432	61	11	None	None	0	No	0	46.94	4	1	10
10:05	3012	57.33	9	None	None	0	No	0	46.94	4	1	10
10:10	3900	56.33	11.67	None	None	0	No	0	46.94	4	1	10

OC = Occupancy, NAL = Number of affected lanes, NF = Number of fatalities, WS = Wind speed, VS = Visibility

- Time of day variable

Time of day, which is in five-minute increments, needs to be adjusted into sequential integer numbers starting from 1, because test data containing data in time format cannot be processed by S-PLUS.

6. ALGORITHM IMPLEMENTATION IN S-PLUS

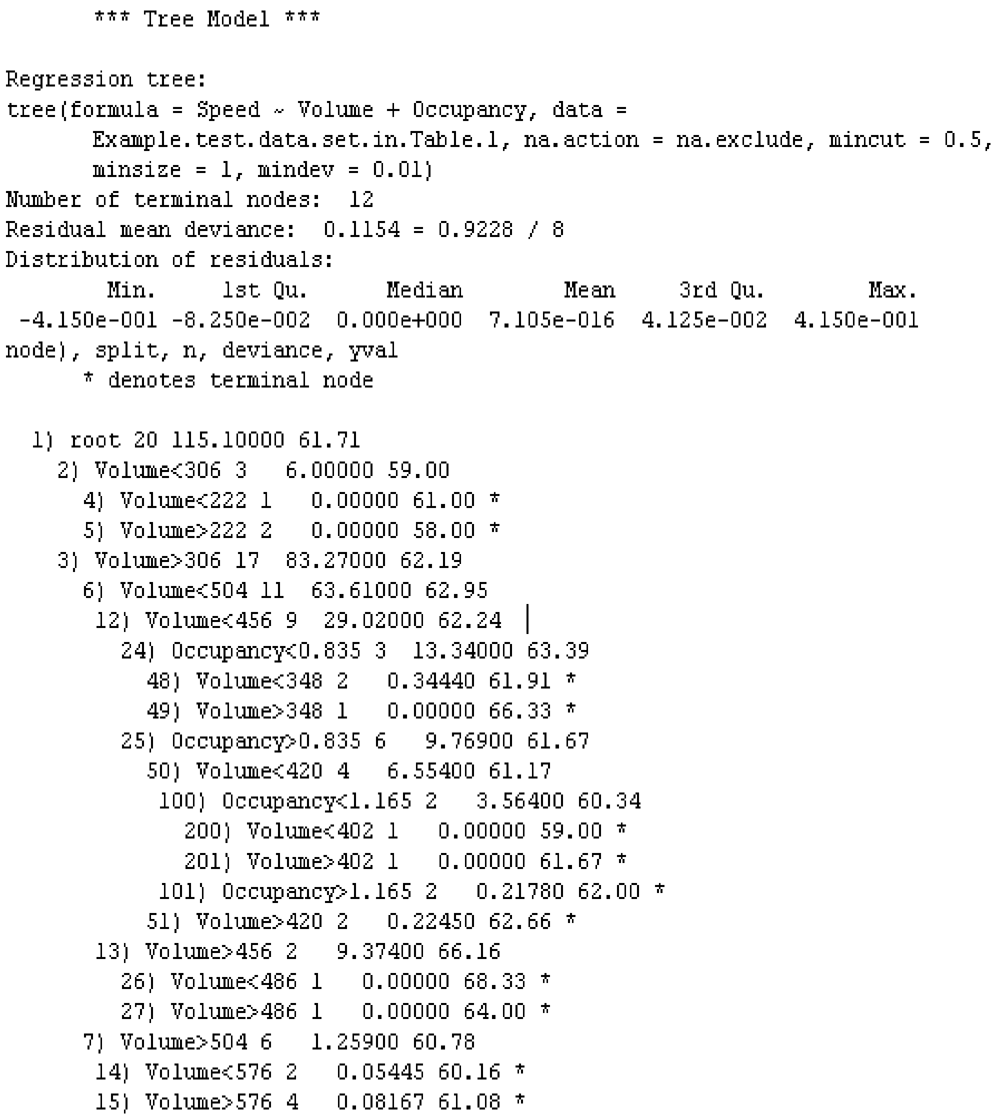
S-PLUS is a statistical software package designed for data analysis and statistical modeling by Insightful Corp. (2005). The regression tree algorithm is implemented in S-PLUS. Thus, in this research, regression tree model construction and validation are both performed in S-PLUS by importing test data and validation data into the software. The S-PLUS software used in this research is version 7.0.

6.1. Model Construction

The response variable and four types of explanatory variables considered in the regression tree model in this research have been illustrated in section 5.1. The data collection and the raw data reorganization for these variables have also been demonstrated. By using S-PLUS, the regression tree models can be constructed upon the test data sets obtained after data collection and reorganization. Before introducing the construction of the regression tree model used in this research, the implementation of the regression tree algorithm in S-PLUS is first described using the example test data set shown in Table 4.1.

6.1.1. Model Construction Using the Example Test Data

The procedure to construct a regression tree model using S-PLUS upon the example test data, shown in Table 4.1, is demonstrated in Appendix E. The results summary and the regression tree plot for the regression tree model constructed upon the example test data are displayed as shown in Figures 6.1 and 6.2.



```
           *** Tree Model ***

1  Regression tree:
2  tree(formula = Speed ~ Volume + Occupancy, data =
3         Example.test.data.set.in.Table.1, na.action = na.exclude, mincut = 0.5,
4         minsize = 1, mindev = 0.01)
5  Number of terminal nodes:  12
6  Residual mean deviance:  0.1154 = 0.9228 / 8
7  Distribution of residuals:
8          Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
9   -4.150e-001 -8.250e-002  0.000e+000  7.105e-016  4.125e-002  4.150e-001
10 node), split, n, deviance, yval
11       * denotes terminal node
12
13   1) root 20 115.10000 61.71
14     2) Volume<306 3   6.00000 59.00
15       4) Volume<222 1   0.00000 61.00 *
16       5) Volume>222 2   0.00000 58.00 *
17     3) Volume>306 17  83.27000 62.19
18       6) Volume<504 11  63.61000 62.95
19        12) Volume<456 9  29.02000 62.24  |
20          24) Occupancy<0.835 3  13.34000 63.39
21            48) Volume<348 2   0.34440 61.91 *
22            49) Volume>348 1   0.00000 66.33 *
23          25) Occupancy>0.835 6   9.76900 61.67
24            50) Volume<420 4   6.55400 61.17
25             100) Occupancy<1.165 2   3.56400 60.34
26               200) Volume<402 1   0.00000 59.00 *
27               201) Volume>402 1   0.00000 61.67 *
28             101) Occupancy>1.165 2   0.21780 62.00 *
29            51) Volume>420 2   0.22450 62.66 *
30        13) Volume>456 2   9.37400 66.16
31          26) Volume<486 1   0.00000 68.33 *
32          27) Volume>486 1   0.00000 64.00 *
33       7) Volume>504 6   1.25900 60.78
34        14) Volume<576 2   0.05445 60.16 *
35        15) Volume>576 4   0.08167 61.08 *
```

Figure 6.1 Results summary of the example regression tree model

In the results summary shown in Figure 6.1, lines 1 to 9 are summary descriptions of the regression tree model and lines 10 to 35 show the complete tree model, with lines 10 and 11 showing the interpretations of the complete tree. In every line of the lines 13 to 35, the first number with right bracket is the node number. The node number for the root of the regression tree is 1. The node numbers for the two splits below one branch is 2n and 2n+1, respectively, if the node number of that branch is n. The second term in the line is the splitting standard of that branch, including the explanatory variable and the threshold value of that explanatory variable that the test data set split along. For example, in line

14, the splitting standard for the branch with node number 2, is “Volume<306.” The third term in the line is the number of observations in the branch. For example, still in line 14, the number of observations in the branch with node number 2 is 3, which is equal to the sum of the number of observations of the two splits (nodes 4 and 5) under this branch. The fourth term in the line is the node deviance and the last term is the mean value of the response variable in the branch. As stated in line 11, the node with * is a terminal node. For example, in lines 14 and 15, nodes 4 and 5 with * are terminal nodes also as shown in the tree plot in Figure 6.2.

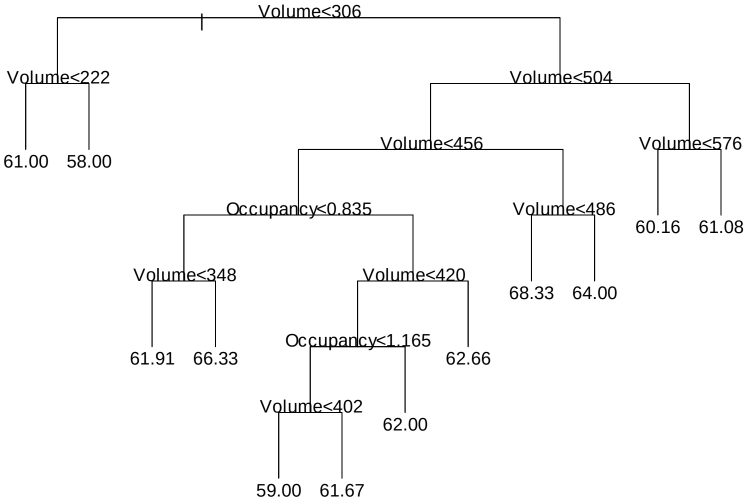


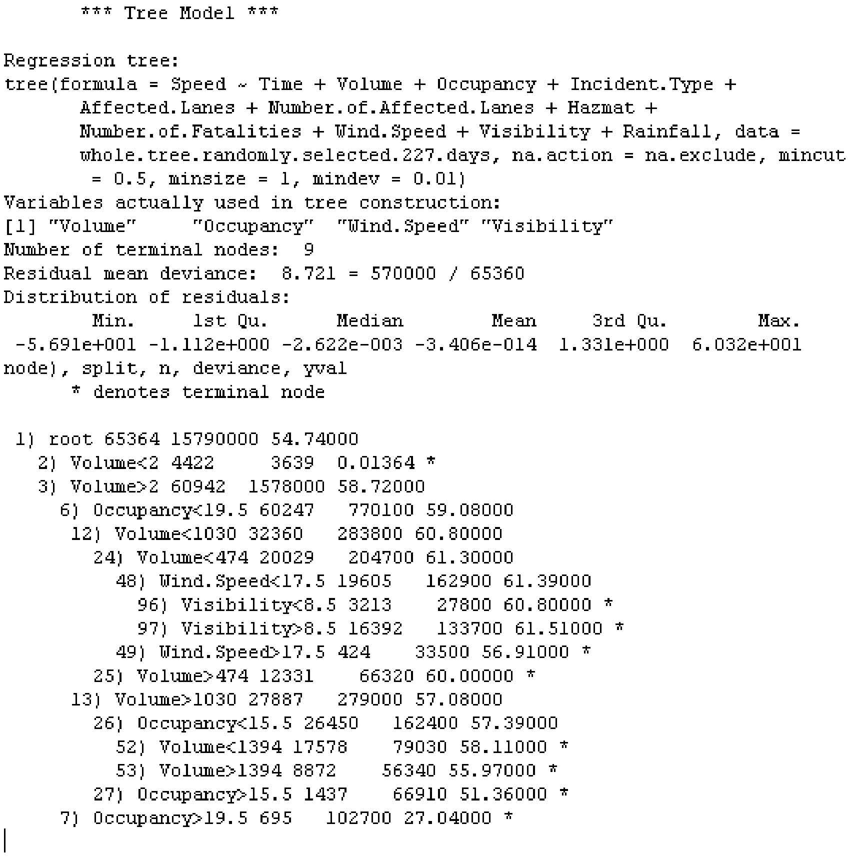
Figure 6.2 Regression tree plot of the example test data set

The way that the regression tree plot is displayed in S-PLUS is a little confusing because of the splitting conditions marked on the tree plot. For example, in Figure 6.2, the splitting condition marked above the first two splits is “Volume<306,” which is actually the splitting condition for the split on the left, and “Volume≥306” is the splitting condition for the split on the right. The splitting conditions of these two splits can be also found in lines 14 and 17 as shown in Figure 6.1.

6.1.2. Full Regression Tree Model Construction

In section 5.1 it has been documented that to construct a comprehensive regression tree model containing all the characteristics over a whole year, test data sets are collected in the whole year of 2005 at the selected station I-205 NB Gladstone. The comprehensive regression tree model in this research is called a full regression tree model, which represents all of the collected daily test data sets, i.e., 342 daily test data sets in the entire year of 2005 (23 days of data are incomplete for unknown reasons).

Thus, to construct the full regression tree model in S-PLUS, these 342 test data sets need to be put consecutively into one Excel file to form one test data set. However, since every daily test data set has 288 rows of data (24 hours of data in 5-minute increments, which means 24*12=288 rows of data) and one Excel file only holds 65536 rows of data, the test data set for the full model can only include 227 daily test data sets (227*288=65376). These have to be randomly selected from the total of 342 daily test data sets. The random number generation is performed by using a Macro written in Excel VBA, as shown in Appendix F. After these randomly selected 227 daily test data sets are put consecutively into one Excel file to form one test data set, this test data set can be imported into S-PLUS and be used to construct the full regression tree model using the same procedure described in section 6.1.1. The only difference is when selecting independent variables in the “Model” tab in the “Tree Models” window, all of the explanatory variables described in section 5.1 need to be selected, including Time, Volume, Occupancy, Incident Type, Affected Lanes, Number of Affected Lanes, Hazmat, Number of Fatalities, Wind Speed, Visibility and Rainfall. The results summary and the tree plot of the full regression tree model constructed are shown in Figures 6.3 and 6.4, respectively.



```
        *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
        Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
        Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
        whole.tree.randomly.selected.227.days, na.action = na.exclude, mincut
         = 0.5, minsize = 1, mindev = 0.01)
Variables actually used in tree construction:
[1] "Volume"     "Occupancy"  "Wind.Speed" "Visibility"
Number of terminal nodes:  9
Residual mean deviance:  8.721 = 570000 / 65360
Distribution of residuals:
        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
 -5.691e+001 -1.112e+000 -2.622e-003 -3.406e-014  1.331e+000  6.032e+001
node), split, n, deviance, yval
      * denotes terminal node

 1) root 65364 15790000 54.74000
   2) Volume<2 4422     3639  0.01364 *
   3) Volume>2 60942  1578000 58.72000
     6) Occupancy<19.5 60247   770100 59.08000
      12) Volume<1030 32360   283800 60.80000
        24) Volume<474 20029   204700 61.30000
          48) Wind.Speed<17.5 19605   162900 61.39000
            96) Visibility<8.5 3213    27800 60.80000 *
            97) Visibility>8.5 16392   133700 61.51000 *
          49) Wind.Speed>17.5 424    33500 56.91000 *
        25) Volume>474 12331    66320 60.00000 *
      13) Volume>1030 27887   279000 57.08000
        26) Occupancy<15.5 26450   162400 57.39000
          52) Volume<1394 17578    79030 58.11000 *
          53) Volume>1394 8872    56340 55.97000 *
        27) Occupancy>15.5 1437    66910 51.36000 *
     7) Occupancy>19.5 695   102700 27.04000 *
```

Figure 6.3 Results summary of the full regression tree model

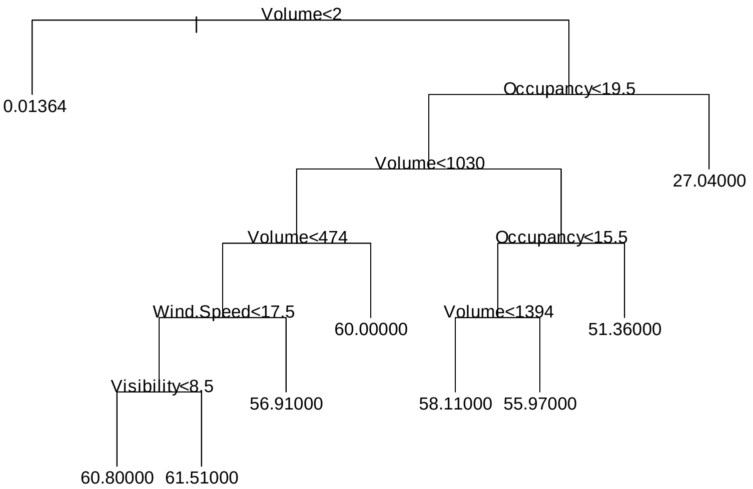


Figure 6.4 Tree plot of the full regression tree model

6.2. Model Validation

After the regression tree model is constructed in S-PLUS on the test data set containing the four types of explanatory variables, model validation can also be accomplished in S-PLUS to evaluate the prediction accuracy of the regression tree model. As mentioned earlier, actual travel time data are not available on PORTAL and speed is used as a proxy to estimate travel time. Therefore, to validate the prediction ability of the regression tree model, the model will be used to predict speeds of daily validation data sets first by using S-PLUS and the Mean Squared Errors (MSE) will be used to estimate the accuracy of the predicted speeds compared to the actual speeds of the validation data sets.

The regression tree model validation implementation in S-PLUS is shown in Appendix G. Here the regression tree model shown in Figure 6.2, which was built based on the example test data in Table 4.1, and a small validation data, as shown in Table 6.1, are used to demonstrate the algorithm of the regression tree model validation.

To validate the regression tree model in Figure 6.2 using the validation data in Table 6.1, every row of validation data, including only the data for the two explanatory variables “Volume” and “Occupancy,” is used to run through the regression tree model to obtain the fitted speed value for that row of validation data. For example, the first row of validation data is 232 for Volume and 0.667 for Occupancy. Since the first split in the regression tree model in Figure 6.2 is “Volume<306,” the first row of validation data needs to go to the left branch after the first split. (The data goes to the left branch if it satisfies the splitting condition above the split, or goes to the right branch if it does not.) After the first row of validation data goes to the left branch of the first split, it comes to the second split “Volume<222” and this time the first row of validation data goes to the right branch because its Volume data is 232, which is larger than 222. Then the first row of data reaches a leaf node with the speed value 58.00. So the fitted speed for the first row of validation data is 58.00.

Table 6.1 Example validation data set

Speed	Volume	Occupancy
65.67	232	0.667
64.00	328	1.000
61.33	228	1.000
58.67	260	1.000
62.00	332	1.333
61.67	240	1.000
59.00	304	1.333
60.33	364	1.333
63.00	376	1.333
66.33	416	1.667
66.00	424	1.667
64.67	412	2.000
62.67	384	1.333
64.00	400	1.667
62.33	516	1.667
61.33	380	1.333
65.00	420	1.333
62.00	512	1.667
64.33	520	1.667
62.33	568	2.333

Similarly, the second row of validation data (Volume is 328 and Occupancy is 1.00) reaches the leaf node 59.00 by going through “Volume<306,” the right branch, “Volume<504,” the left branch, “Volume<456,” the left branch, “Occupancy<0.835,” the right branch, “Volume<420,” the left branch, “Occupancy<1.165,” the left branch, “Volume<402” and the left branch. After every row of validation data goes through the regression tree model, the fitted speed values can be obtained for the whole validation data set as shown in Table 6.2, which are same as the results given by S-PLUS. After the fitted speed values are obtained, MSE is used to evaluate the validation results as also shown in Table 6.2.

Table 6.2 Example validation data set with fitted speeds and MSE

Speed	Volume	Occupancy	Fitted Speed	Squared Error
65.67	232	0.667	58.00	58.78
64.00	328	1.000	59.00	25.00
61.33	228	1.000	58.00	11.11
58.67	260	1.000	58.00	0.44
62.00	332	1.333	62.00	0.00
61.67	240	1.000	58.00	13.44
59.00	304	1.333	58.00	1.00
60.33	364	1.333	62.00	2.78
63.00	376	1.333	62.00	1.00
66.33	416	1.667	62.00	18.78
66.00	424	1.667	62.66	11.12
64.67	412	2.000	62.00	7.11
62.67	384	1.333	62.00	0.44
64.00	400	1.667	62.00	4.00
62.33	516	1.667	60.16	4.70
61.33	380	1.333	62.00	0.44
65.00	420	1.333	62.66	5.45
62.00	512	1.667	60.16	3.37
64.33	520	1.667	60.16	17.37
62.33	568	2.333	60.16	4.70
			MSE	9.55

The MSE value of 9.55 shown in Table 6.2 appears to be high because the regression tree model based on the example test data, shown in Table 4.1, includes only two explanatory variables: volume and occupancy. Instead the full regression tree model used to predict speed in this research includes four types of explanatory variables. With four types of explanatory variables, the MSE value of predicted speed evaluation is fairly low as shown in Figure G.7 in Appendix G.

7. TRAVEL TIME ESTIMATION

As mentioned earlier, after speed is predicted by the regression tree model, the predicted speeds will be used to estimate travel time using mid-point algorithm, which is also used in the PORTAL system to generate the estimated travel time data. MSEs are used to evaluate the estimated travel time by using the predicted speed obtained by the regression tree model compared with the estimated travel time data stored in PORTAL.

The standard mid-point algorithm used in PORTAL is based on ODOT's travel time algorithm which is used to generate travel time estimates for display via dynamic message signs. The key feature of this algorithm is the use of influence areas around each detector station as shown in Figure 7.1 (Kothuri *et al.*, 2006). It is assumed that the detector station is at the midpoint of each influence area. Travel time for each influence area of a station is estimated by calculating the ratio of the length of influence area of a station to the measured speed at the station, which in this research is the predicted speed by use of the regression tree model.

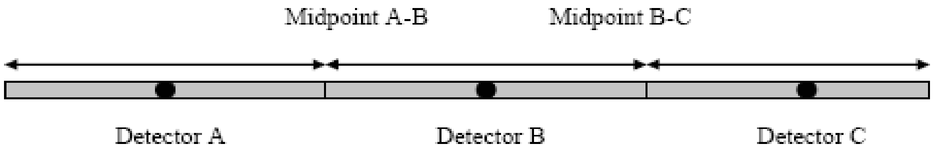


Figure 7.1 Influence area around the detector station

For example, if the predicted speeds between 7:05 and 9:05 pm (between 230 and 254 in time order) on August 2, 2006 are obtained by using the full regression tree model, the travel time in this time period can be estimated and compared with the estimated travel time data stored in PORTAL. The station length of the station I-205 NB Gladstone is found to be 1.75 miles in the PORTAL system. Therefore, to use mid-point algorithm to estimate travel time at I-205 NB Gladstone, the station length of 1.75 miles needs to be divided by the predicted speeds. The estimated travel times at I-205 NB Gladstone between 7:05 and 9:05 pm on August 2, 2006 and the MSE are shown in Table 7.1. The MSE between the estimated travel time by using the predicted speeds of the full regression tree model and the estimated travel time in PORTAL is 0.0011, which is also fairly low for the travel time errors.

Table 7.1 MSE result of estimated travel times by full regression tree model

Time Order No.	Actual Speed	Predicted Speed	Estimated Travel Time (min) (Station length/Predicted speed)	Estimated Travel Time in PORTAL (min)	Squared Errors
230	59.67	60	1.75	1.77	0.0003
231	61.33	60	1.75	1.73	0.0006
232	58.33	58.11	1.81	1.82	0.0001
233	58.67	58.11	1.81	1.81	0.0000
234	59.33	60	1.75	1.77	0.0004
235	62.67	60	1.75	1.68	0.0050
236	62.33	60	1.75	1.70	0.0022
237	60.33	60	1.75	1.75	0.0000
238	61	60	1.75	1.74	0.0001
239	59.67	60	1.75	1.78	0.0010
240	60	60	1.75	1.76	0.0000
241	60	60	1.75	1.75	0.0000
242	61.33	60	1.75	1.73	0.0006
243	62.67	60	1.75	1.67	0.0062
244	61	60	1.75	1.76	0.0001
245	59.67	60	1.75	1.77	0.0005
246	62	60	1.75	1.71	0.0019
247	61.67	60	1.75	1.71	0.0014
248	61.67	60	1.75	1.73	0.0003
249	61	60	1.75	1.72	0.0009
250	58.67	60	1.75	1.80	0.0021
251	61.33	60	1.75	1.73	0.0003
252	59.33	60	1.75	1.77	0.0006
253	59.33	60	1.75	1.78	0.0009
254	59	60	1.75	1.79	0.0019
				MSE	0.0011

It has been demonstrated how predicted speeds can be used to estimate travel time at a station. Now the travel time estimation in a segment of highway can be briefly illustrated. Since the traffic flow data in a segment of highway can only be collected station-wisely, the regression tree model can only be developed station-wisely too. Therefore, to estimate travel times in a segment of highway, the travel times at every station in this segment of highway need to be estimated first by using the predicted speeds obtained through the regression tree model constructed for each station. Based on the assumption that all the detector stations included in this segment are independent, the estimated travel time for the segment of highway can be obtained by adding the travel time estimates of all the stations in this segment of highway.

8. CHARACTERIZATION APPROACH

After the daily data sets are collected and used as test data sets to build regression trees, there arises a question: regression tree models based on what kind of test data sets should be selected to predict speed for a certain day, for example Monday, with good weather (normal temperature and wind speed, no rainfall and clear visibility) and no incidents. A characterization approach is deployed to answer this question. All of the collected daily data sets, including test data sets and validation data sets, are characterized first. Then daily test data sets in the same characterization are put together to construct one regression tree model representing that characterization. To predict speed/estimate travel time in a certain day in the future, the daily validation data sets in the same characterization as that day can be randomly selected to run through the regression tree model to get the fitted speeds. Therefore, by bringing in the characterization approach, the regression tree models may be used to dynamically estimate travel time in a certain day in the future.

Four standards are setup to track the characteristics of both test data sets and validation data sets, including “Outliers”, “Good weather”, “Incidents” and “Weekday or Weekend.” “Outliers” is to check if there are missing data or erroneous data that exist in traffic flow data due to detector errors. The preliminary research showed that regression tree models are robust to outliers in the test data sets. Furthermore, compared to the regression tree models built on the test data containing no outliers, those built on the test data containing outliers may have more stable prediction ability, i.e., ability of predicting speeds for both the validation data sets with outliers and those without. For “Good Weather”, based on published sources (Pisano and Goodwin, 2004; Maze *et al.*, 2006), a daily data set is regarded as having good weather if wind speed is lower than 15 mph, visibility is higher than 8 miles and rainfall is less than 3 mm per hour and no good weather if any of the three conditions is not satisfied. The main reason why temperature is not considered here is that the temperature data in weather data in the PORTAL system is incomplete. It is also because that in Portland Metro area (I5-I205 loop) extreme temperatures is not common. “Incidents” is to check whether any incidents existed in the daily data sets or not. “Weekday or Weekend” is used to track the characteristic of day of week in the data sets, since the traffic flow patterns between weekdays and weekends are surely different. Since there are two levels for each of four standards, there are 16 combinations or characterizations, into which all the test data sets and validation data sets have to be distributed.

Referring back to section 5.1, the daily test data sets and validation data sets were collected at the selected station I-205 NB Gladstone with milepost 11.05. 342 daily test data sets in the entire year of 2005 are collected (23 days of data are incomplete for unknown reasons) as test data sets. 532 daily data sets at the same station in the whole year of 2006 and the first half year of 2007 (i.e. a duration of 1.5 years and 14 days of data are incomplete for unknown reasons) are collected as validation data sets, because the weather data and incidents data in the second half year of 2007 are not yet available in the PORTAL system.

As mentioned in section 5.2, to save time and increase accuracy, four macros written in Excel VBA are developed to reorganize the collected daily data sets, as shown in Appendix D. For the same reason, another macro, shown in Appendix H, is developed to characterize all the daily data sets *automatically* with characterization results shown in Table 8.1.

There are no test data sets and validation data sets in characterization 9. And there are less than or equal to 3 validation data sets in characterizations 1, 2, 10, 13 and 15. Since both the test data sets and validation data sets cover a considerably long time period of 1 and 1.5 years, respectively, the characterization results show that characterizations 1, 2, 9, 10, 13 and 15 are not worthwhile to be further considered in this research. Thus, only ten characterizations in total, to include characterization 3, 4, 5, 6, 7, 8, 11, 12, 14 and 16, are considered.

Table 8.1 Characterization results for test data sets and validation data sets

Characterization No.	Outliers	Good Weather	Incidents	Weekday or Weekend	Number of Test Data Sets	Number of Validation Data Sets
1	Yes	Yes	Yes	Weekday	5	2
2	Yes	No	Yes	Weekday	6	3
3	Yes	Yes	No	Weekday	51	42
4	Yes	No	No	Weekday	64	57
5	No	Yes	Yes	Weekday	6	13
6	No	Yes	No	Weekday	44	106
7	No	No	Yes	Weekday	8	14
8	No	No	No	Weekday	63	145
9	Yes	Yes	Yes	Weekend	0	0
10	Yes	No	Yes	Weekend	3	1
11	Yes	Yes	No	Weekend	8	11
12	Yes	No	No	Weekend	7	21
13	No	Yes	Yes	Weekend	3	2
14	No	Yes	No	Weekend	32	52
15	No	No	Yes	Weekend	4	3
16	No	No	No	Weekend	38	60
				Total	342	532

The ultimate challenge in this research is to dynamically estimate travel time. That is, for a certain time period in a certain day in the future, what kind of regression tree model should be selected to predict speed or estimate travel time. Thus, by including the characterization approach, it can be determined if there is a regression tree model of a specific characterization, capable of better predicting speeds/travel times for existing conditions on the road such as weather, incident, time of day, etc., than the regression tree models of other characterizations. Or if the full regression tree model, which is constructed on the 227 daily test data sets randomly selected from the total of 342 collected daily test data sets to represent a general regression tree model, outperforms the regression tree models representing the specific characterizations. Therefore, the prediction abilities for speeds/travel times of the regression tree models of ten characterizations and the full regression tree model need to be compared. Before that, the regression tree models representing the ten characterizations need to be constructed first.

Similar to the construction of the full regression tree model presented in section 6.1.2, to build a regression tree model representing a specific characterization, all of the daily test data sets in the same characterization need to be put consecutively into one Excel file to form one test data set to construct the regression tree model. For example, to form a test data set to construct the regression tree model representing characterization 3, the 51 daily test data sets in characterization 3, as shown in Table 8.1, need to form one test data set. The regression tree models representing ten characterizations are presented in Appendix I.

9. EXPERIMENTAL DESIGN AND RESULT DISCUSSIONS

The development of the regression tree models in this research and the construction of the full regression tree model have been documented in Chapters 5 and 6. Ten characterization regression tree models are constructed, with the characterization approach introduced and illustrated in Chapter 8. The ultimate challenge in this research is to determine what kind of regression tree model or even models should be selected to predict speed or estimate travel time for a certain day under dynamic conditions. To address this question, with the full regression tree model and the ten characterization regression tree models constructed, it is necessary to first evaluate the performances of these eleven regression tree models. Then two types of comparisons of the performances of the regression tree models need to be performed: one between one characterization tree model and other characterization tree models, and the other between the full regression tree model and the characterization tree models. That is the focus of this chapter. To evaluate a constructed regression tree model, the MSE is used to estimate the accuracy of the predicted speeds by the regression tree model compared with the actual speeds of the validation data set. As shown in Figure G.7 in Appendix G, the MSE value between the actual speeds for the validation data of August 2, 2006 and the predicted speeds using the full regression tree model is as low as 1.70, which means that the errors of the predicted speeds are approximately less than ±1.3 mph. However, this result is not persuasive because it is only based on one validation data set. To establish more compelling evidence, the experimental designs need to be developed to evaluate the performances of these eleven regression tree models.

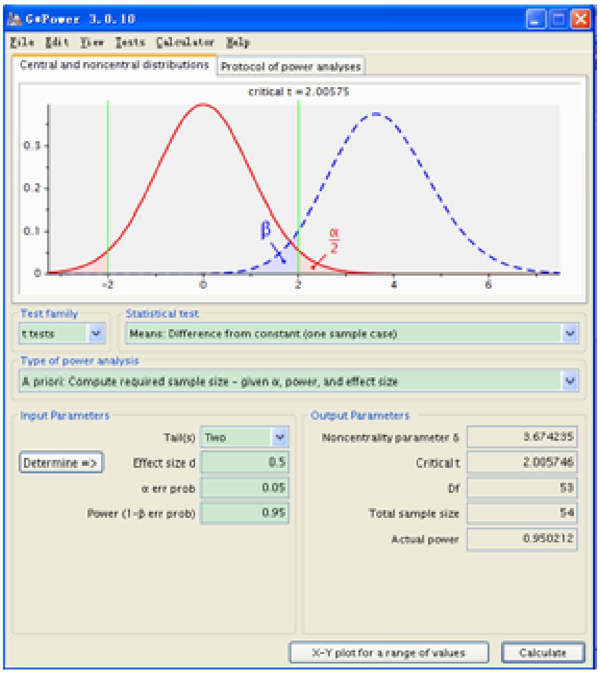
First, a one-sample t-test is used to evaluate the individual performance of these eleven regression tree models, with a significance level of α=0.05 and a power of statistical test of 1-β=0.95. Then to compare the performances of the ten characterization regression tree models and the full regression tree model, a randomized complete block design (RCBD) is constructed with a significance level of α=0.05. Multiple comparisons are also performed using Tukey's method and the Fisher LSD method (Montgomery, 2005).

9.1. One Sample T-test

After the ten characterization regression tree models and the full regression tree model are constructed, each of them is used to predict speeds for all of the collected validation data sets, i.e., 532 daily validation data sets collected in the whole year of 2006 and the first half year of 2007. The MSEs are calculated for the predicted speeds of each validation data set, i.e., 532 MSEs are obtained for each of these eleven models. T-tests with a significance level of α=0.05 and a power of statistical test of 1-β=0.95 need to be constructed to evaluate the 532 MSEs for each of these models. To construct the t-tests, the sample size for these t-tests needs to be determined first.

Sample sizes are strongly related to the effect size and the power of test of an experimental design. Effect size is a measure of the strength of the relationship between two variables. In an experimental design, the effect size helps to determine whether a statistically significant difference is a difference of practical concern. In other words, given a sufficiently large sample size, it is always possible to show that there is a difference between two means being compared out to some decimal position. The effect size helps to know whether the difference observed is a difference that matters. To determine a reasonable sample size, a software program G*Power is used to determine the sample size for the one-sample t-tests. G*Power is a general power analysis program developed by Erdfelder *et al.* (1996), which can be downloaded from the following website: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/.

By using G*Power, it shows that, to maintain a medium effect size of 0.5 (according to the effect size conventions) and a relatively high power of test of 95%, a sample size of at least 54 out of the whole population of MSEs (532 MSEs) for the t-test, as shown in Figure 9.1, is required. Therefore, a sample size of 54 is used and 54 MSEs are randomly selected from the 532 MSEs for each of the eleven regression tree models. The use of G*Power to determine the sample size is introduced in Appendix J.


G*Power 3.0.10
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 2.00575
0.3
0.2
0.1
0
-2
0
2
4
6
β
α/2
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Tail(s): Two
Effect size d: 0.5
α err prob: 0.05
Power (1-β err prob): 0.95

Output Parameters

Noncentrality parameter δ: 3.674235
Critical t: 2.005746
Df: 53
Total sample size: 54
Actual power: 0.950212

Figure 9.1 G*Power window of sample size determination for t-tests

The results of the t-tests performed on the 54 MSEs obtained by the eleven regression tree models are shown in Table 9.1, with normality assumption checked and holding true for the eleven randomly-selected MSEs samples. In Table 9.1, the 95% confidence interval (C.I.) is calculated for the MSEs of each of the eleven models. The square roots of these C.I.s represent the 95% C.I.s of the mean speed errors by using these eleven models. Therefore, the 95% C.I.s for the errors of the predicted speeds by the ten characterization regression tree models and the full regression tree model are ± [2.65, 3.03] mph, ± [2.86, 3.19] mph, ± [3.66, 4.58] mph, ± [3.56, 4.41] mph, ± [3.89, 4.68] mph, ± [3.63, 4.43] mph, ± [2.99, 3.46] mph, ± [3.07, 3.56] mph, ± [3.22, 4.20] mph, ± [3.10, 3.74] mph and ± [2.62, 2.97] mph, respectively. For the errors of the predicted speeds, these results are all fairly low and all significantly less than ± 5 mph. Of these eleven models, the full tree model has the lowest 95% C.I., i.e., ± [2.62, 2.97] mph.

Table 9.1 T-test results of the performances of the eleven regression tree models

Model	Sample Mean	Sample Var. (S^2)	Sample S.D. (S)	$t_{\alpha/2,n-1}$	d.f.	95% C.I. of MSE		95% C.I. of Mean Speed	
						Lower Bound	Upper Bound	Lower Bound	Upper Bound
C. 3	8.09	14.99	3.87	2.006	53	7.04	9.15	2.65	3.03
C. 4	9.17	13.84	3.72	2.006	53	8.16	10.19	2.86	3.19
C. 5	17.19	194.50	13.95	2.006	53	13.38	20.99	3.66	4.58
C. 6	16.06	150.68	12.28	2.006	53	12.71	19.41	3.56	4.41
C. 7	18.51	154.42	12.43	2.006	53	15.12	21.90	3.89	4.68
C. 8	16.41	138.79	11.78	2.006	53	13.19	19.62	3.63	4.43
C.11	10.45	29.68	5.45	2.006	53	8.97	11.94	2.99	3.46
C.12	11.05	34.42	5.87	2.006	53	9.45	12.65	3.07	3.56
C.14	13.99	174.40	13.21	2.006	53	10.39	17.60	3.22	4.20
C.16	11.77	64.56	8.03	2.006	53	9.58	13.97	3.10	3.74
F.R.T.	7.83	13.09	3.62	2.006	53	6.84	8.82	2.62	2.97

Var. = Variance, S.D. = Standard Deviance, C.I. = Confidence interval, C.3 = Characterization 3, F.R.T. = Full regression tree

9.2. RCBD Experimental Design

The randomized complete block design (RCBD) (Montgomery, 2005) is probably the most frequently used design. The experimental units are divided into homogeneous groups of material (called blocks), each of which constitutes a single replication of the experiment. The word “complete” indicates that each block contains all the treatments. In this research, blocks are the daily validation data sets. Each daily validation data set constitutes a “day” block, in which the validation result (MSE of the predicted speeds) of this validation data set is calculated for each of the eleven regression tree models, including the ten characterization regression tree models and the full regression tree model. Compared with a completely randomized design (CRD), RCBD effectively improves the accuracy of the comparisons among the eleven regression tree models by eliminating the variability among different daily validation data sets.

With the daily validation data sets of ten characterizations, ten randomized complete block designs can be constructed. In each RCBD, each daily validation data set constitutes one block. The regression tree model is the only factor, which has eleven treatments—the ten characterization regression tree models and the full regression tree model. The response variable is MSE, which is used to estimate the accuracy of the predicted speeds by the regression tree model compared with the actual speeds of the validation data set. The total sample size of each RCBD can be calculated by multiplying the number of treatment levels (11) by the number of blocks (the number of daily validation data sets in each of the ten characterizations). Since sample sizes are strongly related to the effect size and the power of test of the experimental designs, to determine a reasonable sample size, operating characteristic (OC) curves are usually used. However, for this situation, it is difficult to apply OC curves to determine the sample sizes because the number of treatment levels (11) cannot be found on the OC curves. Thus, the software program G*Power is also used to determine the sample sizes for each RCBD.

G*Power shows that, to maintain a reasonable effect size of 0.25, which is the medium effect size for F-test according to Cohen's (1988) conventions of effect size measures and a relatively high power of test of 95%, a sample size of 407 for each RCBD is required, as shown in Figure J.1 in Appendix J. It means that at least 37 blocks are needed for each RCBD (37*11 = 407). Therefore, in each of the ten characterizations, at least 37 daily validation data sets need to be randomly selected for the validation of each of the eleven models and to further construct a RCBD. Forty daily validation data sets are decided to be used for characterizations 3, 4, 6, 8, 14 and 16, which contain more than 37 daily validation data sets, as shown in Table 8.1. For characterizations 5, 7, 11 and 12, there are less than 37 daily validation data sets available. Characterization 11 has the least number of daily validation data sets of 11. Thus, to obtain meaningful conclusions, there are only two choices: either to sacrifice effect size (use higher effect size) to get higher power of test or to sacrifice power of test to get lower effect size. By testing in G*Power, it is found out that even for characterization 11 with only 11 daily validation data sets, the large effect size for F-test of 0.40 according to Cohen's conventions of effect size measures and a power of test of 90% are still guaranteed. Consequently, for characterization 5, 7, 11 and 12, all the daily validation data sets available in these characterizations are used in constructing RCBDs. Table 9.2

shows the number of daily validation data sets needed for each of the ten RCBDs. The macro presented in Appendix F can also be used here to randomly select 40 daily validation data sets for characterizations 3, 4, 6, 8, 14 and 16. Then all the validation data sets selected for RCBDs are imported in S-PLUS to validate the eleven constructed regression tree models, the process of which has been demonstrated in section 6.2.

Table 9.2 Number of validation data sets used for RCBD

Characterization No.	Outliers	Good Weather	Incidents	Weekday or Weekend	Number of VDS	Number of VDS for RCBD
3	Yes	Yes	No	Weekday	42	40
4	Yes	No	No	Weekday	57	40
5	No	Yes	Yes	Weekday	13	13
6	No	Yes	No	Weekday	106	40
7	No	No	Yes	Weekday	14	14
8	No	No	No	Weekday	145	40
11	Yes	Yes	No	Weekend	11	11
12	Yes	No	No	Weekend	21	21
14	No	Yes	No	Weekend	52	40
16	No	No	No	Weekend	60	40

VDS = Validation data sets

RCBD is used to test if there is significant difference among the prediction abilities of the eleven regression tree models. Multiple comparisons using Tukey’s method and Fisher LSD method are performed to compare the prediction abilities of speed/travel time of one characterization regression tree model vs. each of the other characterization regression tree models and each of the ten characterization regression tree models vs. full regression tree model. The response variable in each RCBD is MSE values from validation of regression tree models by using validation data sets. Each of the validation data sets, serving as one block in RCBD, is used to validate 11 different regression tree models. Thus, 11 MSE values will be calculated for each block (each of the validation data sets). The computation of

MSE values used in RCBDs is shown in Appendix K. The ten RCBDs constructed using MSEs are shown in Appendix L.

9.3. Analysis of Results

In section 9.2, RCBD has been used to compare the prediction abilities of speed/travel time of one characterization regression tree model vs. each of the other characterization regression tree models and each of the ten characterization regression tree models vs. full regression tree model. After ten RCBDs for validation data sets in ten characterizations are constructed as shown in Appendix L, the analysis of variance (ANOVA) and multiple comparisons are performed for each RCBD in S-PLUS, as shown in Appendix M, to compare the prediction abilities of characterization regression tree models and full regression tree model. The normality assumption and equal variance have been checked and are holding true. The results of ANOVA and multiple comparisons are shown in Table 9.3.

The above table shows the results of ANOVA and multiple comparisons for ten RCBDs in two sections, characterization model vs. characterization model and characterization model vs. full model. The first section compares the regression tree model in the same characterization as the validation data sets with the other nine characterization regression tree models. For example, for validation data sets in characterization 3 (the first row), the prediction ability of regression tree model representing characterization 3 is compared with that of regression tree models representing characterization 4, 5, 6, 7, 8, 11, 12, 14 and 16. The first column in this section shows if there is significant difference among these ten characterization models with a significance level of α=0.05. The second column to the fourth column in this section show the multiple comparison results between the regression tree model in the same characterization as the validation data sets and the other nine characterization regression tree models using Tukey's method except characterization 14 using Fisher LSD method. Still using validation data sets in characterization 3 as an example, the second column "No Difference" shows that there is no significant difference between characterization 3 model and

Table 9.3 Results of ANOVA and multiple comparisons for ten RCBDs

C. no.	Characterization model vs. characterization model				Characterization model vs. full model		
	Significant difference?	No difference	Positive difference	Negative difference	Better than full model	Worse than full model	No difference than full model
3	Yes	11, 12, 16, 4	None	5, 6, 7, 8, 14	None	14, 5, 6, 7, 8	11, 12, 16, 3, 4
4	Yes	11, 12, 16, 3	None	5, 6, 7, 8, 14	None	14, 5, 6, 7, 8	11, 12, 16, 3, 4
5	Yes	14, 16, 3, 4, 6, 7, 8	None	11, 12	None	11, 12	14, 16, 3, 4, 5, 6, 7, 8
6	Yes	16, 3, 4, 5, 7, 8	None	11, 12, 14	None	11, 12	14, 16, 3, 4, 5, 6, 7, 8
7	Yes	14, 16, 3, 4, 5, 6, 8	None	11, 12	None	12	11, 14, 16, 3, 4, 5, 6, 7, 8
8	Yes	16, 3, 4, 5, 6, 7	None	11, 12, 14	None	11, 12, 14	16, 3, 4, 5, 6, 7, 8
11	Yes	12, 16, 3, 4	5, 6, 7, 8, 14	None	None	14, 5, 6, 7, 8	11, 12, 16, 3, 4
12	Yes	11, 16, 3, 4	5, 6, 7, 8, 14	None	None	14, 5, 6, 7, 8	11, 12, 16, 3, 4
14*	Yes	3, 4, 5, 6, 7, 8, 11, 12, 16	None	None	None	None	3, 4, 5, 6, 7, 8, 11, 12, 14, 16
16	Yes	11, 12, 14, 3, 5	4, 7, 8	6	6	4, 7, 8	11, 12, 14, 16, 3, 5

* : Fisher LSD method is used for multiple comparison.
C. no. = Characterization no. of validation data sets

characterization 11, 12 16 and 4 models, which means these five models are equally good to predict validation data sets in characterization 3. The third column "Positive Difference" shows that no characterization models outperform characterization 3 model significantly. The fourth column "Negative Difference" shows that characterization 3 model significantly outperforms the regression tree models representing characterizations 5, 6, 7, 8 and 14.

For validation data sets in characterization 14, although there are significant differences that exist among the regression tree models representing ten characterizations, no significant differences exist either between the regression tree model of characterization 14 and each of the other characterization regression tree models or between the full regression tree model and each of the ten characterization regression tree models. Significant differences, however, exist just among the other nine regression tree models except the regression tree model of characterization 14 and the full regression tree model.

The second section in Table 9.3 shows the multiple comparison results between the full regression tree model and the ten characterization regression tree models. Still using validation data sets in characterization 3 as an example, the first column in this section shows that no characterization models significantly outperform the full model to predict validation data sets in characterization 3. The second column shows that the full model significantly outperforms the regression tree models representing characterizations 5, 6, 7, 8 and 14. The third column shows that there is no significant difference between the prediction ability of the full model and that of characterization 11, 12 16, 3 & 4 models, which means these six models are equally good to predict validation data sets in characterization 3.

10. CONCLUSIONS AND FUTURE RESEARCH

The objective of this research is to accurately and dynamically estimate freeway travel times in the I5-I205 loop in the Portland Metro area of Oregon using regression tree methodology. Because historical travel time data are not yet available on PORTAL (Portland Oregon Regional Transportation Archive Listing), which is the source of data collection in this research, speeds are predicted first by using the regression tree models and speeds are in turn used as a proxy in travel time estimation. The objective of this research is important in industry practice, because accurate freeway travel time information is of increasing importance for the travelers' information and route guidance system. Travel times provide valuable information not only for traveler routing but also for transportation scheduling and management.

Following the introduction of the regression tree methodology, the development of the regression tree model has been demonstrated. Thirteen explanatory variables in four types, traffic flow variables, incident related variables, weather data variables and time of day variable are considered in the data sets for regression tree model construction. This ensures that the regression tree models in this research have the same prediction ability among different flow conditions on freeways. Since the data for the traffic flow variables and incident related variables are both station-specific on PORTAL, data collection in this research is accomplished only at the detector station, instead of a segment of highway. The detector station I-205 NB Gladstone with milepost 11.05 is randomly selected to collect the daily test data sets and validation data sets. To capture all the characteristics in the regression tree models, 342 daily test data sets in the entire year of 2005 are collected. 532 daily validation data sets are collected at the same station in the whole year of 2006 and the first half year of 2007 (i.e., a duration of 1.5 years and 14 days of data are incomplete for unknown reasons). For the total of 874 collected daily data sets, the accuracy and the efficiency of the raw data reorganization became a challenge. Thus, four macros written in Excel VBA have been developed to *automatically* reorganize the collected daily data sets.

After the reorganizations, the daily data sets are ready to be imported into the statistical software package S-PLUS for regression tree model constructions and validations. The implementation of the regression tree algorithm in S-PLUS has then been illustrated in both model construction and validation. By importing the test data sets into S-PLUS, the regression tree model based on the test data sets can be constructed using the built-in functions in S-PLUS and the regression tree plot can also be obtained. The full regression tree model is constructed on one test data including 227 daily test data sets randomly selected from the total of 342 daily test data sets. To compare the predicted speeds of the validation data sets by the regression tree models with their actual speeds, mean squared errors (MSEs) are used. As pointed out in Figure G.7 in Appendix G, the MSE value of predicted speeds using the full regression tree model is fairly low, which shows the promising potential to accurately estimate travel times using the regression tree model in this research.

The estimation of travel time using predicted speeds as a proxy through the mid-point algorithm has been illustrated. Since the travel time data on PORTAL were also estimated by the mid-point algorithm using the actual speeds, if the predicted speeds by the regression tree models are accurate and close to the actual speeds, the travel time estimates in this research would also be close to the travel time estimates on PORTAL. As mentioned earlier, since the data collection can only be accomplished at the detector station, the regression tree model can only be developed station-wise too. Therefore, to estimate travel times in a segment of highway, the travel times at every station in this segment of highway need to be estimated first by using the predicted speeds obtained through the regression tree model constructed for each station. Based on the assumption that all the detector stations included in the segment are independent, the estimated travel time for the segment of highway can be obtained by adding the travel time estimates of all the stations in that segment of highway.

To dynamically estimate travel time for a certain day in the future using regression tree models, the characterization approach and how it is applied in the regression tree analysis for travel time estimation have been addressed. The regression tree models representing ten characterizations are then constructed. With the full regression tree model and the ten characterization regression tree models constructed, it is necessary to evaluate the performances of these eleven regression tree models. One sample t-test has been used to evaluate the individual performance of these eleven regression tree models, with a significance level of α=0.05 and a power of statistical test of 1-β=0.95. Then to compare the performances of the ten characterization regression tree models and the full regression tree model, a randomized complete block design (RCBD) has been constructed with a significance level of α=0.05. Multiple comparisons have also been performed using Tukey's method and Fisher LSD method. The analysis of the results of these two experimental designs reveals four main promising findings:

- For the errors of the predicted speeds by the ten characterization regression tree models and the full regression tree model, the 95% confidence intervals (C.I.s) are all fairly low and significantly less than ± 5 mph, as illustrated in section 9.1. It shows that all of these eleven models possess the ability to accurately predict speeds/estimate travel times. Of these eleven models, the full tree model has the lowest 95% C.I. of ± [2.62, 2.97] mph.
- To predict speed/travel time for a certain day (within a certain characterization), several regression tree models have been shown to be equally effective, and the use of the regression tree models do not need to be limited to one of the same characterization as that day. For example, to predict speed/travel time for a day in characterization 3, the full regression tree model and the regression tree models representing characterization 4, 11, 12 and 16 are equally good as that of characterization 3.
- To predict speed/travel time for a day in characterization 11, 12 or 16, there are several characterization regression tree models that outperform the regression tree model of the same characterization as that day (i.e. characterization 11, 12 or 16).
- The full regression tree model is expected to have better or at least as good a prediction ability as the characterization regression tree models. The full regression tree model covers the test data sets of all the characterizations and, therefore, should have more stable prediction ability. However, this research has revealed that, to predict speed/travel time for a day in characterization 16,

the regression tree model of characterization 6 is significantly better than the full regression tree model (α=0.05).

In spite of the highlighted findings above, the characterization approach increases the power of the full regression tree model in its applicability to predict speed/travel time in the future. For example, without using the characterization approach, to predict speed/travel time on a future Monday with expected good weather and no incidents, a group of validation data sets need to be randomly selected to be run through the full regression tree model to get the predicted values. The average value of the predicted speeds/travel times of all the randomly selected validation data sets would be used as the estimated value for the desired day. This approach may lead to an inaccurate estimated value, because of the possibility of having features different in the randomly selected validation data sets than in the desired day. However, using the characterization approach, the validation data sets in the same characterization as that of the desired day can be selected to be run through the full regression tree model, which may increase the accuracy of the prediction ability of the full regression tree model. Moreover, the characterization approach helps to construct the regression tree models of specific characterizations, one of which (the regression tree model of characterization 6) is proven to outperform the full regression tree model in the prediction of validation data sets in characterization 16.

In this research, the regression tree models are employed to predict speed first and then predicted speeds are used as a proxy to estimate travel time. Thus the regression tree models are not directly applied to estimate travel time. This limitation is due to the fact that the historical travel time data are not available on PORTAL. In the future, if the actual travel time measurements are made available by ODOT, the current regression tree models demonstrated in Chapter 5 can be adjusted to estimate travel time directly. To make the adjustments, the daily travel time data in five-minute increments need to be collected first and then be incorporated into the current daily test data set for regression tree model construction. The travel time would then serve as the response variable in the regression tree model, while speed would serve as one of the explanatory variables in the group of traffic flow variables. By making these adjustments, the regression tree methodology described in this research can be applied to estimate travel time directly.

BIBLIOGRAPHY

Brieman, L., Friedman, J.H., Olshen, R. A. and Stone, C. J. (1984) Classification and Regression Trees. Belmont, California:Wadsworth.

Carey, M. and McCartney, M. (2003) Pseudo-periodicity in a travel-time model used in dynamic traffic assignment. Transportation Research Part B, 37(9), 769-792.

Chang, L. and Chen, W. (2005) Data Mining of Tree-based Models to Analyze Freeway Accident Frequency. Journal of Safety Research, 36(4), 365-375.

Chang, L. and Wang, H. (2006) Analysis of Traffic Injury Severity: An Application of Non-parametric Classification Tree Techniques. Accident Analysis and Prevention, 38(5), 1019-1027.

Chen, M. and Chien, S. (2001) Dynamic Freeway Travel Time Prediction with Probe Vehicle Data: Link-based vs. Path-based. Transportation Research Record, 1768, 157-161.

Chong, M. M., Abraham, A. and Paprzycki, M. (2004) Traffic accident analysis using decision trees and neural networks. IADIS International Conference on Applied Computing, Portugal, 2, 39-42.

Chu, L., Oh, J. S. and Recker, W. (2005) Adaptive Kalman filter based freeway travel time estimation. The 84th Transportation Research Board Annual Meeting (CD-ROM), Washington D.C.

Cohen, J. (1988) Statistical power for the behavioral sciences (2nd edition). Hillsdale, NJ: Erlbaum, 286-287.

Coifman, B. (2002) Estimating travel times and vehicle trajectories on freeways using dual loop detectors. Transportation Research Part A, 36(4), 351-364.

Dailey, D. J. (1993) Travel time estimation using cross-correlation techniques. Transportation Research Part B, 27(2), 97-107.

Dailey, D. J. (1997) Travel time estimation using a series of single loop volume and occupancy measurements. The 76th Transportation Research Board Annual Meeting, Washington D.C.

Erdfelder, E., Faul, F., & Buchner, A., 1996, “GPOWER: A general power analysis program,” Behavior Research Methods, Instruments, & Computers, 28, 1-11.

Golias, I. and Karlaftis, M. G. (2002) An International Comparative Study of Self-reported Driver Behavior. Transportation Research Part F: Psychology and Behavior, 4(4), 243-256.

Ishimaru, J. M., Nee, J. and Hallenbeck, M. E. (2003) Measurement of recurring versus non-recurring congestion: Technical report. Washington State Transportation Center (TRAC), October.

Karlaftis, M. G. and Golias, I. (2002) Effects of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates. Accident Analysis and Prevention, 34(3), 357-365.

Kwon, J., Coifman, B. and Bickel, P. (2000) Day-to-day travel time trends and travel time prediction from loop detector data. Transportation Research Record, 1717, 120-129.

Lederer, P. R., Cohn, L. F., Guensler, R. and Harris, R. A. (2005) Effect of on-ramp geometric and operational factors on vehicle activity. Journal of Transportation Engineering, 131(1), 18-26.

Lee, C., Ran, B. and Qin, X. (2006) The Analysis of Winter Maintenance Logs using Regression Tree Algorithm. Proceedings of the 85th Transportation Research Board Annual Meeting (CD-ROM), Washington DC, January 21-26.

Maze, T. H., Agarwal, M. and Burchett, G. (2006) Whether weather matters to traffic demand, traffic safety and traffic operations and flow. Transportation Research Record, 1948, 170-176.

Montgomery, D. C. (2005) Design and Analysis of Experiments (6th edition). Wiley, New Jersey.

Nam, D. H. and Drew, D. R. (1996) Traffic dynamics: method for estimating freeway travel time in real time from flow measurements. Journal of Transportation Engineering, 122(3), 185-191.

Park, D. and Rilett, L. R. (1998) Forecasting multiple-period freeway link travel times using modular neural networks. Transportation Research Record, 1617, 163-170.

Park, D., Rilett, L. R. and Han, G. (1999) Spectral basis neural networks for real-time travel time forecasting. Journal of Transportation Engineering, 125(6), 515-523.

Petty, K. F., Bickel, P., Jiang, J., Ostland, M., Rice, J., Ritov, Y. and Schoenberg, F. (1998) Accurate estimation of travel times from single-loop detectors. Transportation Research Part A, 32(1), 1-17.

Pisano, P. A. and Goodwin, L. C. (2004) Research needs for weather-responsive traffic management. Transportation Research Record, 1867, 127-131.

Rickman, T. D., Hallenbeck, M. E. and Schroeder, M. (1990) Improved method for collecting travel time information. Transportation Research Record, 1271, 79-88.

Rilett, L. R. and Park, D. (2001) Direct forecasting of freeway corridor travel times using spectral basis neural networks. Transportation Research Record, 1752, 140-147.

Sisiopiku, V. P. and Rouphail, N. M. (1994) Toward the use of detector output for arterial link travel time estimation: A literature review. Transportation Research Record, 1457, 158-165.

S-PLUS version 7.0 (2005) The Insightful Corp., Seattle, WA.

Tesema, T. B., Abraham, A. and Grosan, C. (2005) Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation, 6(10 & 11), 80-94.

Turner, S. M. (1996) Advanced techniques for travel time data collection. Transportation Research Record, 1551, 51-58.

Wong, H. K. and Sussman, J. M. (1973) Dynamic travel time estimation on highway networks. Transportation Research, 7(4), 355-370.

Zhang, X. and Rice, J. A. (2003) Short-term travel time prediction using a time-varying coefficient linear model. Transportation Research Part C, 11(3 & 4), 187-210.

APPENDICES

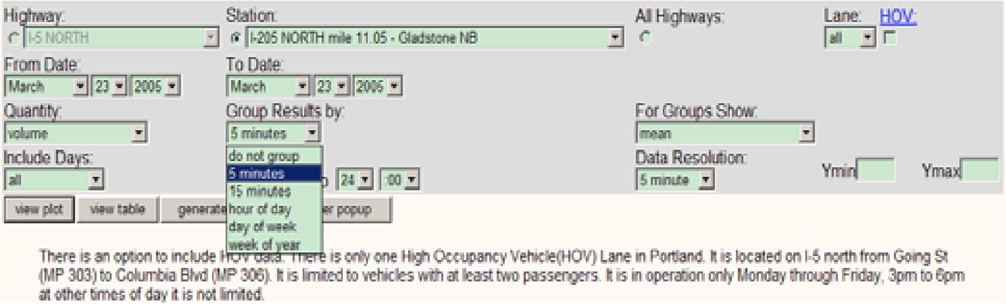
APPENDIX A. DATA COLLECTION FOR TRAFFIC FLOW VARIABLES IN PORTAL

In free flow conditions, only the traffic flow data needs to be considered in the test data for travel time estimation on freeways. The following traffic flow data shown in Table A.1 was collected at the station I-205 NB Gladstone on March 23, 2005 from 9:10 to 10:10 am. In the interest of space, the complete traffic flow data collected at this station on this day is not shown here.

Table A.1 Partial traffic flow data (9:10 – 10:10 am)

Time	Volume	Speed	Occupancy
9:10	3024.00	60.00	8.67
9:15	3240.00	59.67	10.67
9:20	2784.00	58.33	9.33
9:25	3096.00	59.00	9.33
9:30	3696.00	56.00	12.33
9:35	3792.00	57.67	12.00
9:40	3588.00	58.33	11.33
9:45	3744.00	55.67	12.33
9:50	3564.00	58.00	11.33
9:55	3624.00	58.00	12.00
10:00	3432.00	61.00	11.00
10:05	3012.00	57.33	9.00
10:10	3900.00	56.33	11.67

The traffic flow data can be collected as shown in Figure A.1, which is the screen shot taken from the PORTAL system for traffic flow data.



Highway: ○ I-5 NORTH
Station: ● I-205 NORTH mile 11.05 - Gladstone NB
All Highways: ○
Lane: all HOV: □
From Date: March 23 2005
To Date: March 23 2005
Quantity: volume
Group Results by: 5 minutes
For Groups Show: mean
Include Days: all
24 :00
Data Resolution: 5 minute
Ymin Ymax

There is an option to include HOV data. There is only one High Occupancy Vehicle(HOV) Lane in Portland. It is located on I-5 north from Going St (MP 303) to Columbia Blvd (MP 306). It is limited to vehicles with at least two passengers. It is in operation only Monday through Friday, 3pm to 6pm at other times of day it is not limited.

Figure A.1 Screen shot of the PORTAL system for traffic flow data collection

After clicking the archive “Grouped Data” on the homepage of the PORTAL system, the screen like that shown in Figure A.1 can be seen. Different stations or segments of highway can be selected in Station or Highway as shown in Figure A.1. Single day or a time period can be selected by appropriately choosing “From Date” and “To Date.” Different data items can be selected to show by choosing in “Quantity,” such as volume, speed, etc. To collect the traffic flow data, volume, speed or occupancy needs to be selected in “Quantity.” Five minutes is chosen in “Group Results by” because it is the smallest time increment we can choose to better track the data pattern.

After all the items on the webpage as described above are selected appropriately, a table of results for the selected data item in “Quantity” can be obtained by clicking “view table.” For example, by selecting all the items shown in Figure A.1, the volume data in Table A.2 is obtained.

Similarly, speed and occupancy data at the station I-205 NB Gladstone on March 23, 2005 can be collected. Then the raw traffic flow data, including time, volume, speed and occupancy can be reorganized in one data table as shown in Table A.1.

Table A.2 Raw volume data at station I-205 NB Gladstone on March 23, 2005 (9:10 – 10:10 am)

Time	Avg Volume (vplph)	Avg Percentage Good Data
9:10	1008	1
9:15	1080	0.93333
9:20	928	1
9:25	1032	1
9:30	1232	1
9:35	1264	1
9:40	1196	1
9:45	1248	1
9:50	1188	1
9:55	1208	1
10:00	1144	1
10:05	1004	1
10:10	1300	1

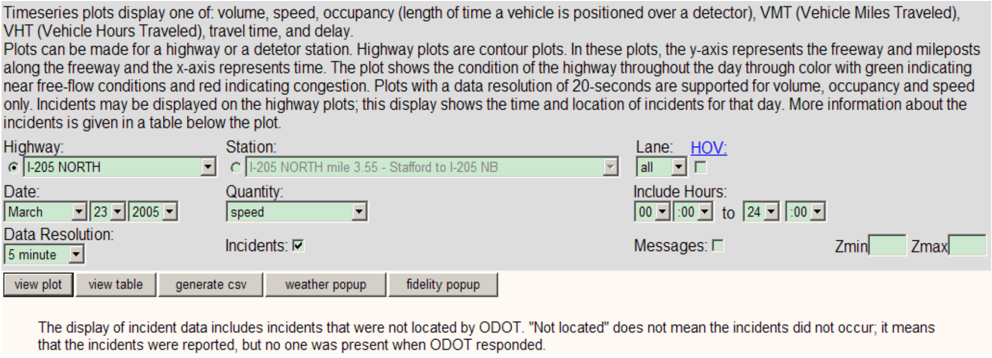
APPENDIX B. DATA COLLECTION FOR INCIDENT RELATED VARIABLES IN PORTAL

An incident would typically result in a reduced speed between detector stations on the I5-I205 loop, which could lead to a non-recurring congestion. The incident data related variables, such as the start time of an incident, the time the incident got cleared, incident type, etc., are very useful to comprehensively analyze the impact of incident data on the travel time in this research. The raw incident data at the station I-205 NB Gladstone on March 23, 2005 collected from the PORTAL system is shown in Table B.1.

Table B.1 Incident data at the station I-205 NB Gladstone on March 23, 2005

ID	Primary Route	Location	Number of Lanes Affected	Start Time (hh:mm:ss)	Duration (min)	Incident Type	Affected Lanes	Hazmat	Number of Fatalities
421624	"I-205"	"I-205 NB GLADSTONE"	0	9:32:55	14	Debris	All Lanes	no	0

Figure B.1 is a screen shot of the PORTAL system for incident data from which the incident data above in Table B.1 can be collected.



Timeseries plots display one of: volume, speed, occupancy (length of time a vehicle is positioned over a detector), VMT (Vehicle Miles Traveled), VHT (Vehicle Hours Traveled), travel time, and delay.
Plots can be made for a highway or a detetor station. Highway plots are contour plots. In these plots, the y-axis represents the freeway and mileposts along the freeway and the x-axis represents time. The plot shows the condition of the highway throughout the day through color with green indicating near free-flow conditions and red indicating congestion. Plots with a data resolution of 20-seconds are supported for volume, occupancy and speed only. Incidents may be displayed on the highway plots; this display shows the time and location of incidents for that day. More information about the incidents is given in a table below the plot.

Highway: ● I-205 NORTH
Station: ○ I-205 NORTH mile 3.55 - Stafford to I-205 NB
Lane: all HOV: □
Date: March 23 2005
Quantity: speed
Include Hours: 00 :00 to 24 :00
Data Resolution: 5 minute
Incidents: ☑
Messages: □
Zmin Zmax

The display of incident data includes incidents that were not located by ODOT. "Not located" does not mean the incidents did not occur; it means that the incidents were reported, but no one was present when ODOT responded.

Figure B.1 Screen shot of the PORTAL system for incident data collection

After clicking the archive "Timeseries" on the homepage of the PORTAL system, the screen as in Figure B.1 can be seen. To check the incident data at certain station, the segment of highway to which this station belongs to must be selected in "Highway," instead of the station itself in "Station." Then select the date of the incident data needed to be viewed and any item in "Quantity" (speed or volume, doesn't really matter which). Check "Incidents" and then click "view plot." The graph as shown in Figure B.2 and the incident data table for the whole segment of highway as shown in Figure B.3 will be seen.

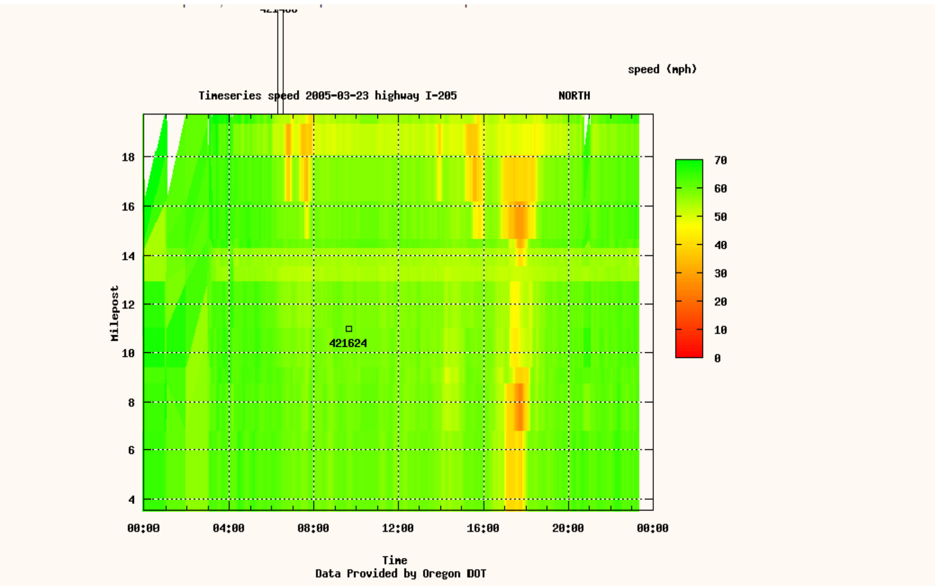
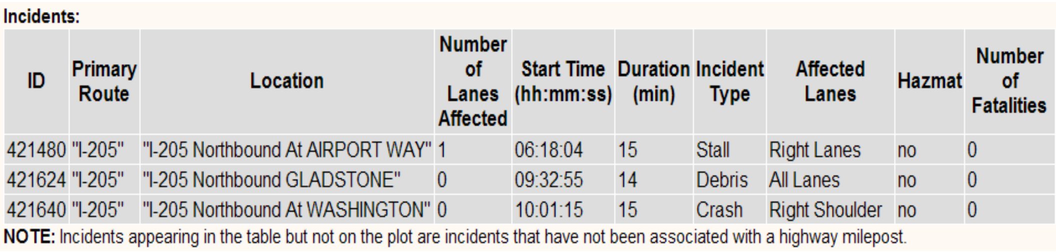


Figure B.2 Incident data graph on I-205 NB on March 23, 2005



Incidents:

ID	Primary Route	Location	Number of Lanes Affected	Start Time (hh:mm:ss)	Duration (min)	Incident Type	Affected Lanes	Hazmat	Number of Fatalities
421480	"I-205"	"I-205 Northbound At AIRPORT WAY"	1	06:18:04	15	Stall	Right Lanes	no	0
421624	"I-205"	"I-205 Northbound GLADSTONE"	0	09:32:55	14	Debris	All Lanes	no	0
421640	"I-205"	"I-205 Northbound At WASHINGTON"	0	10:01:15	15	Crash	Right Shoulder	no	0

NOTE: Incidents appearing in the table but not on the plot are incidents that have not been associated with a highway milepost.

Figure B.3 Incident data table on I-205 NB on March 23, 2005

Since the station I-205 NB Gladstone (milepost 11.04) is the selected station, as in Figure B.2, it can be seen that the incident with ID 421624 occurred around milepost 11.04 and its detailed information can be found in Figure B.3 with its ID.

APPENDIX C. DATA COLLECTION OF WEATHER DATA VARIABLES IN PORTAL

Adverse weather, such as heavy rainfall, snowfall, low visibility, etc, is a considerable cause of an increased risk of traffic accidents and compromised traffic flow on highway. Thus, considering weather data variables in the formation of test data would make the test data capable of predicting speed even in a non-free flow condition related to severe weather conditions. The partial hourly weather data (from 0:00 to 11:00) at the station I-205 NB Gladstone on March 23, 2005 is shown in Table C.1.

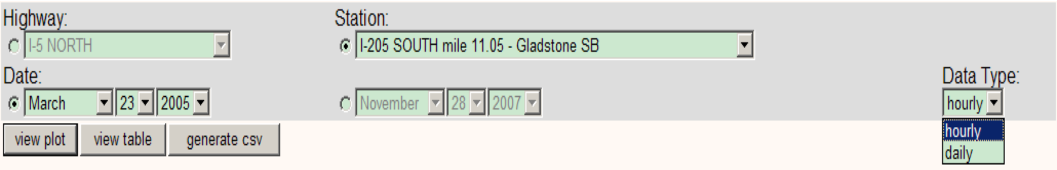
Table C.1 Partial hourly weather data (0:00 – 11:00 am)

Time	Temp f	Wind speed ms	Visibility mi	Rainfall
3/23/2005 0:00	46.04	0	10	0
3/23/2005 1:00	46.04	0	10	0
3/23/2005 2:00	44.96	6	10	0
3/23/2005 3:00	44.06	3	10	0
3/23/2005 4:00	44.06	0	10	0
3/23/2005 5:00	46.04	0	10	0
3/23/2005 6:00	46.04	9	10	1
3/23/2005 7:00	46.04	10	10	0
3/23/2005 8:00	46.04	0	10	1
3/23/2005 9:00	46.04	4	10	0
3/23/2005 10:00	46.94	4	10	1
3/23/2005 11:00	46.04	5	7	2

The above weather data can be collected as shown in Figure C.1, which is a screen shot of the PORTAL system for weather data collection.

After clicking the archive “Weather” on the homepage of the PORTAL system, the screen as shown in Figure C.1 can be seen. To access the weather data at certain station on certain day, the station and the day need to be selected in “Station” and

"Date," respectively. And "Data Type" should be set as hourly to track the weather data pattern more accurately. By clicking "view table," the weather data table as shown in Table C.1 can be obtained.



Highway: ○ I-5 NORTH

Station: ● I-205 SOUTH mile 11.05 - Gladstone SB

Date: ● March 23 2005

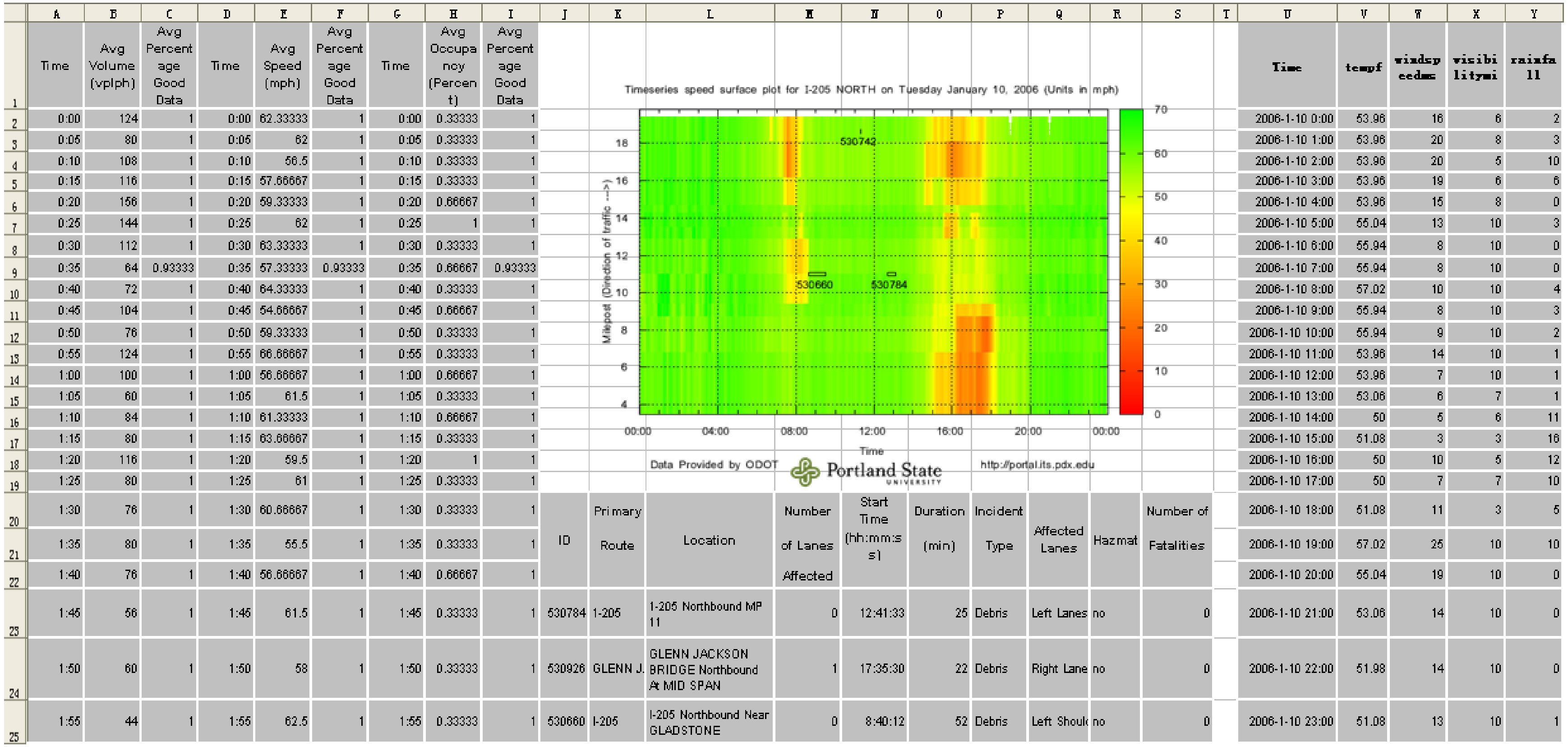
○ November 28 2007

Data Type: hourly

Figure C.1 Screen shot of the PORTAL system for weather data collection

APPENDIX D. EXCEL VBA PROGRAMS FOR RAW DATA REORGANIZATIONS

As described in section 5.2, raw data reorganizations are needed for the raw data collected for the four types of explanatory variables considered in the test data set for regression tree model construction. Because raw data reorganizations are needed for every raw daily data set collected, it means 874 daily data sets in total need to be reorganized, including 342 collected test data sets and 532 validation data sets. To save time and increase accuracy, four macros written in Excel Visual Basic Application (VBA) are developed to reorganize daily raw data saved in Excel files. Before describing the VBA programs, the raw daily data set collected at the station I-205 NB Gladstone on Jan 10th, 2006 is shown in Figure D.1 as an example of the raw daily data sets. In the interest of space, the example raw daily data set is only shown from 0:00 to 1:55 am for traffic flow data in Figure D.1. The daily raw data collected for the four types of explanatory variables need to be copied into one Excel file, with traffic flow data (including time of day) in Columns A to I, incident related data in Columns J to S (plot copied in Rows 1 to 19 and table copied, starting from Row 20) and weather data in Columns U to Y, as shown in Figure D.1.


Timeseries speed surface plot for I-205 NORTH on Tuesday January 10, 2006 (Units in mph)
Milepost (Direction of traffic --->)
18
16
14
12
10
8
6
4
530742
530660
530784
00:00
04:00
08:00
12:00
16:00
20:00
00:00
Time
70
60
50
40
30
20
10
0
Data Provided by ODOT
Portland State University
http://portal.its.pdx.edu

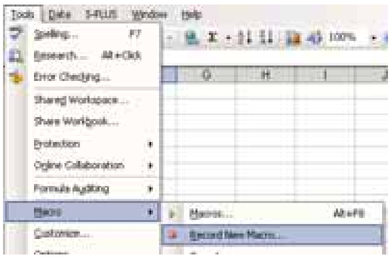
Time	Avg Volume (vplph)	Avg Percentage Good Data	Time	Avg Speed (mph)	Avg Percentage Good Data	Time	Avg Occupancy (Percent)	Avg Percentage Good Data
0:00	124	1	0:00	62.33333	1	0:00	0.33333	1
0:05	80	1	0:05	62	1	0:05	0.33333	1
0:10	108	1	0:10	56.5	1	0:10	0.33333	1
0:15	116	1	0:15	57.66667	1	0:15	0.33333	1
0:20	156	1	0:20	59.33333	1	0:20	0.66667	1
0:25	144	1	0:25	62	1	0:25	1	1
0:30	112	1	0:30	63.33333	1	0:30	0.33333	1
0:35	64	0.93333	0:35	57.33333	0.93333	0:35	0.66667	0.93333
0:40	72	1	0:40	64.33333	1	0:40	0.33333	1
0:45	104	1	0:45	54.66667	1	0:45	0.66667	1
0:50	76	1	0:50	59.33333	1	0:50	0.33333	1
0:55	124	1	0:55	66.66667	1	0:55	0.33333	1
1:00	100	1	1:00	56.66667	1	1:00	0.66667	1
1:05	60	1	1:05	61.5	1	1:05	0.33333	1
1:10	84	1	1:10	61.33333	1	1:10	0.66667	1
1:15	80	1	1:15	63.66667	1	1:15	0.33333	1
1:20	116	1	1:20	59.5	1	1:20	1	1
1:25	80	1	1:25	61	1	1:25	0.33333	1
1:30	76	1	1:30	60.66667	1	1:30	0.33333	1
1:35	80	1	1:35	55.5	1	1:35	0.33333	1
1:40	76	1	1:40	56.66667	1	1:40	0.66667	1
1:45	56	1	1:45	61.5	1	1:45	0.33333	1
1:50	60	1	1:50	58	1	1:50	0.33333	1
1:55	44	1	1:55	62.5	1	1:55	0.33333	1

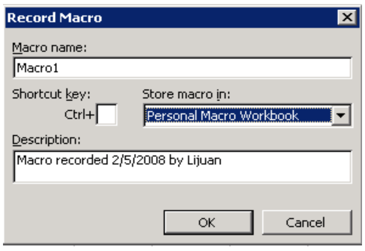
ID	Primary Route	Location	Number of Lanes Affected	Start Time (hh:mm:ss)	Duration (min)	Incident Type	Affected Lanes	Hazmat	Number of Fatalities
530784	1-205	1-205 Northbound MP 11	0	12:41:33	25	Debris	Left Lanes	no	0
530926	GLENN J.	GLENN JACKSON BRIDGE Northbound At MID SPAN	1	17:35:30	22	Debris	Right Lane	no	0
530660	I-205	I-205 Northbound Near GLADSTONE	0	8:40:12	52	Debris	Left Shoul	no	0

Time	tempf	windspeedms	visibilitymi	rainfall
2006-1-10 0:00	53.96	16	6	2
2006-1-10 1:00	53.96	20	8	3
2006-1-10 2:00	53.96	20	5	10
2006-1-10 3:00	53.96	19	6	6
2006-1-10 4:00	53.96	15	8	0
2006-1-10 5:00	55.04	13	10	3
2006-1-10 6:00	55.94	8	10	0
2006-1-10 7:00	55.94	8	10	0
2006-1-10 8:00	57.02	10	10	4
2006-1-10 9:00	55.94	8	10	3
2006-1-10 10:00	55.94	9	10	2
2006-1-10 11:00	53.96	14	10	1
2006-1-10 12:00	53.96	7	10	1
2006-1-10 13:00	53.06	6	7	1
2006-1-10 14:00	50	5	6	11
2006-1-10 15:00	51.08	3	3	16
2006-1-10 16:00	50	10	5	12
2006-1-10 17:00	50	7	7	10
2006-1-10 18:00	51.08	11	3	5
2006-1-10 19:00	57.02	25	10	10
2006-1-10 20:00	55.04	19	10	0
2006-1-10 21:00	53.06	14	10	0
2006-1-10 22:00	51.98	14	10	0
2006-1-10 23:00	51.08	13	10	1

Figure D.1 Raw daily data set at the station I-205 NB Gladstone on Jan 10th, 2006 (0:00 – 1:55 am)

Programs written in Excel VBA language for raw data reorganizations are saved as Macros in a special Excel file PERSONAL.XLS, which can make Macros applicable for any opened Excel files. To access PERSONAL.XLS, the software Excel needs to be opened first and then followed by clicking Tools>Macro>Record New Macro as shown in Figure D.2. A dialog window will show up and Personal Macro Workbook needs to be selected in "Store macro in:" as shown in Figure D.3. After clicking OK, a new file called PERSONAL.XLS will be created automatically in Excel. Then by clicking Window>Unhide as shown in Figure D.4, "PERSONAL" needs to be selected in a popped out window "Unhide workbook." Now we can close all the opened Excel files by clicking Yes in a popped-out confirmation window as shown in Figure D.5.




Record Macro

Macro name: Macro1

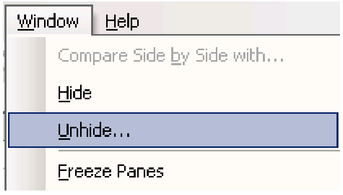
Shortcut key: Ctrl+

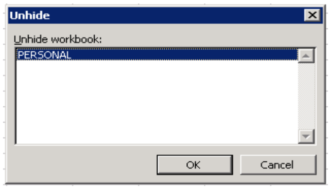
Store macro in: Personal Macro Workbook

Description: Macro recorded 2/5/2008 by Lijuan

Figure D.2 Record new macro

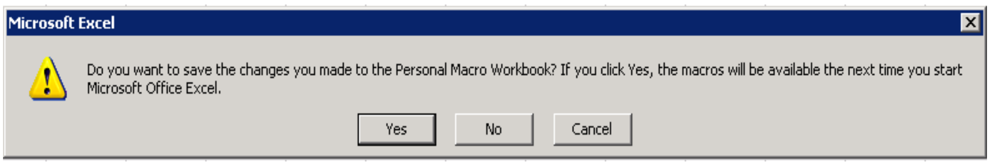
Figure D.3 Personal macro workbook




Unhide

Unhide workbook: PERSONAL

Figure D.4 Unhide PERSONAL.XLS


Microsoft Excel

Do you want to save the changes you made to the Personal Macro Workbook? If you click Yes, the macros will be available the next time you start Microsoft Office Excel.

Figure D.5 Exit Excel

Next time no matter which Excel file is opened, the file PERSONAL.XLS will open automatically. Now we can start writing programs in EXCEL VBA as Macros in the opened file PERSONAL.XLS by clicking Tools>Macro>Visual Basic Editor as shown in Figure D.6. After a window named “Microsoft Visual Basic – PERSONAL.XLS” pops out, we can right click “Sheet 1” under “VBAProject (PERSONAL.XLS)” and then click Insert>Module as shown in Figure D.7. A blank window will then pop out for programs writing (or code imputing) to create Macros in the file PERSONAL.XLS.

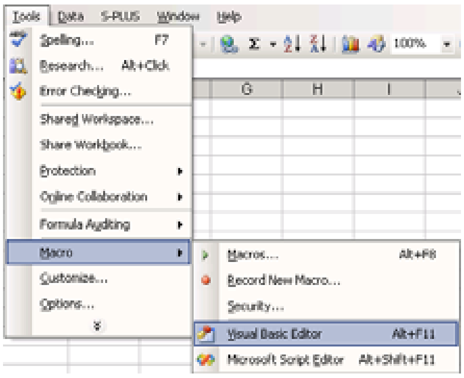


Figure D.6 Open Visual Basic Editor

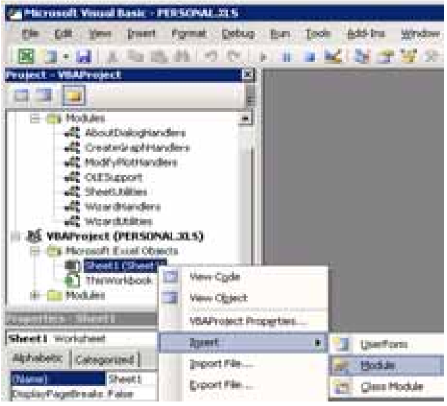


Figure D.7 Insert a new module

The Excel VBA programs are written as four Macros to reorganize the raw data of four types, which are described in the following.

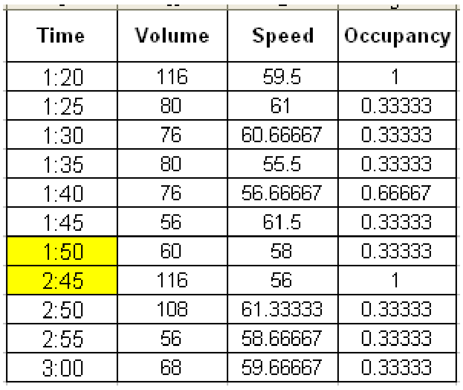
- Traffic flow data: the following program can be copied to the new module created as shown in Figure D.7 as a Macro with the name “Traffic_flow_data.” The raw data shown in Figure D.1 is kept in the same worksheet “Sheet 1” in one Excel file and the reorganized data will be kept in the worksheet “Sheet 2” in the same Excel file.

```
Sub Traffic_flow_data()
    Dim i, j As Integer
    Workbooks(1).Activate           ///Activate the workbook “PERSONAL.XLS”
    Range("a1").EntireColumn.Copy     ///Copy the complete data of time of day
                                        ///stored in the first column
```

```
        Workbooks(2).Activate      ///Activate the workbook where the raw data are stored
        Worksheets(2).Select                         ///Activate the “Sheet 2” of this workbook
        Range("a1").Select
        ActiveSheet.Paste                          ///Paste the copied data into the first column
        Worksheets(1).Select                         ///Activate the “Sheet 1” of this workbook
        Range("a1:i289").Copy      ///Copy the traffic flow data stored from column A to I
        Worksheets(2).Select
        Range("b1").Select
        ActiveSheet.Paste                            ///Paste into “Sheet 2” from column B
        Range("d1").EntireColumn.Delete
        Range("d1").EntireColumn.Delete
        Range("e1:f289").Delete
        Range("f1").EntireColumn.Delete    ///Delete the columns of good data percentage
        Range("a1").Copy
        Range("b1:m289").PasteSpecial xlPasteFormats
        Range("a2").Copy
        Range("b2").EntireColumn.PasteSpecial xlPasteFormats
        Range("b2:m289").Font.Bold = False          ///Setup the format for the area where
                                                    ///the reorganized
        With Range("b1")                              ///Data will be stored in “Sheet 2”
            .Offset(0, 1).Value = "Volume"
            .Offset(0, 2).Value = "Speed"
            .Offset(0, 3).Value = "Occupancy"
            .Offset(0, 4).Value = "Incident Type"
            .Offset(0, 5).Value = "Affected Lanes"
            .Offset(0, 6).Value = "Number of Affected Lanes"
            .Offset(0, 7).Value = "Hazmat"
            .Offset(0, 8).Value = "Number of Fatalities"
            .Offset(0, 9).Value = "Wind Speed"
            .Offset(0, 10).Value = "Visibility"
            .Offset(0, 11).Value = "Rainfall"       ///Name the twelve columns as shown
                                                      ///in Figure D.10
        End With
        For i = 2 To 289
            If Not Cells(i, 1).Value = Cells(i, 2).Value Then
                For j = i + 1 To 289
                     If Cells(j, 1).Value = Cells(i, 2).Value Then
                        Range(Cells(i, 2), Cells(289, 5)).Cut
           Destination:=Range(Cells(j, 2), Cells(289 + j - i, 5))
                        Range(Cells(i, 3), Cells(j - 1, 5)).Value = 0
                     End If
                Next j
            End If
        Next I       ///Detect the missing flow data and fill zero values for the missing data
        Columns(2).Delete    ///Delete the incomplete data of time of day due to the
                                  ///missing data
    End Sub
```

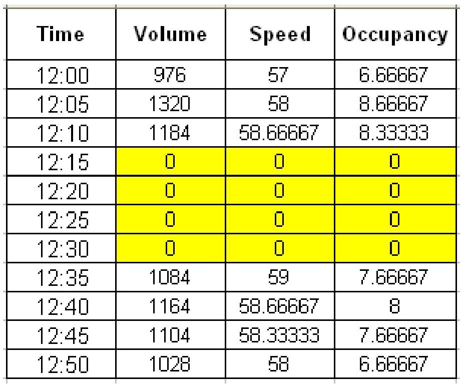
The above program cannot only copy the raw traffic flow data into “Sheet 2” and delete the unnecessary columns, but also can detect and clear the outliers. The outliers in the collected data are mainly the missing data and the erroneous data existing in traffic flow data due to detector errors, as shown in Figure D.8 and

Figure D.9, respectively. In Figure D.8, the traffic flow data between 1:50 and 2:45 are missing and needs to be filled in with zero values for the missing part. At the same time, the incomplete Time column needs to be substituted with a complete time column. In Figure D.9, from 12:15 to 12:30, the traffic flow data all have zero values, which are impossible in real life and show that these data are erroneous data. Since the erroneous data already have zero values filled in and the time column is complete, no reorganizations are needed for the erroneous data.



Time	Volume	Speed	Occupancy
1:20	116	59.5	1
1:25	80	61	0.33333
1:30	76	60.66667	0.33333
1:35	80	55.5	0.33333
1:40	76	56.66667	0.66667
1:45	56	61.5	0.33333
1:50	60	58	0.33333
2:45	116	56	1
2:50	108	61.33333	0.33333
2:55	56	58.66667	0.33333
3:00	68	59.66667	0.33333

Figure D.8 Missing data

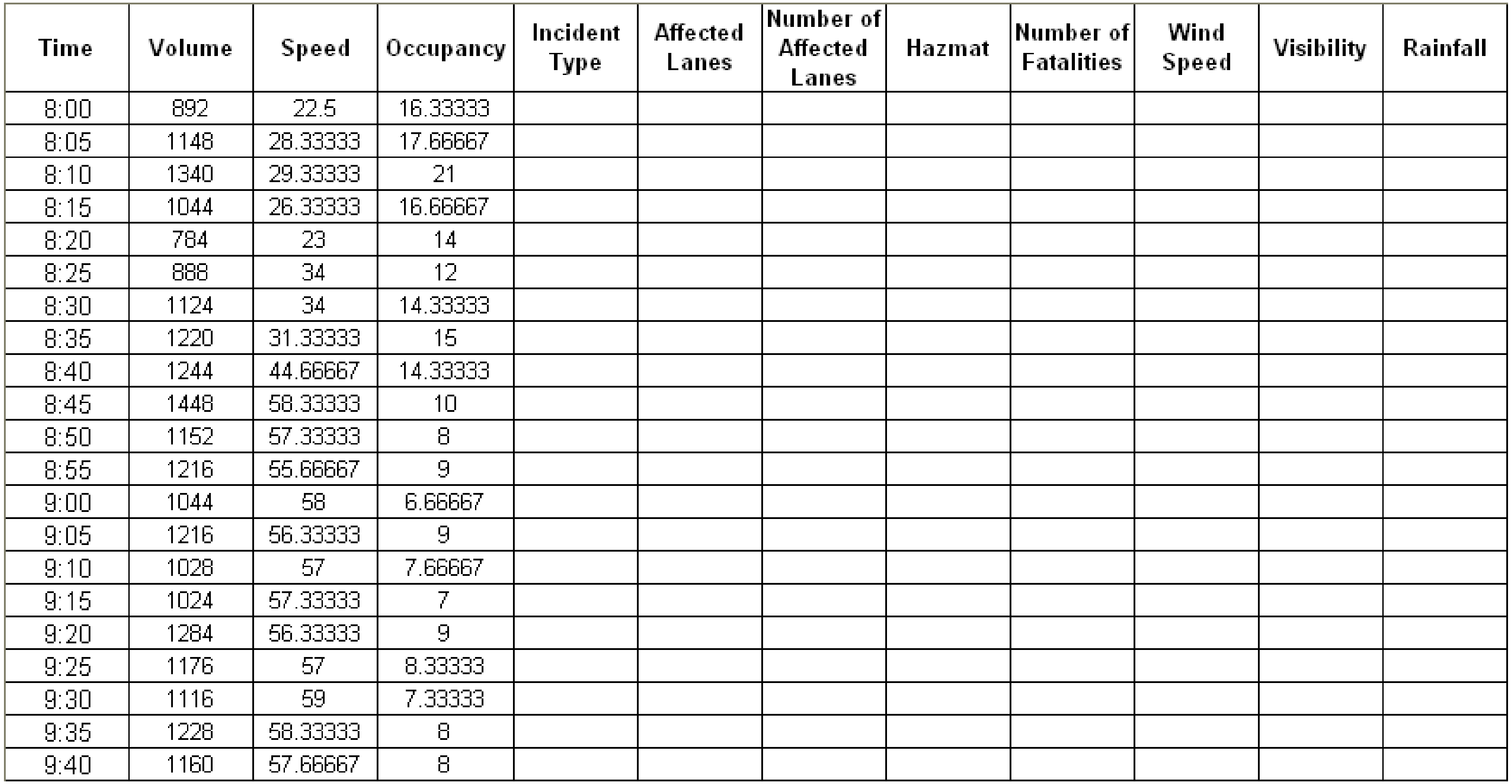


Time	Volume	Speed	Occupancy
12:00	976	57	6.66667
12:05	1320	58	8.66667
12:10	1184	58.66667	8.33333
12:15	0	0	0
12:20	0	0	0
12:25	0	0	0
12:30	0	0	0
12:35	1084	59	7.66667
12:40	1164	58.66667	8
12:45	1104	58.33333	7.66667
12:50	1028	58	6.66667

Figure D.9 Erroneous data

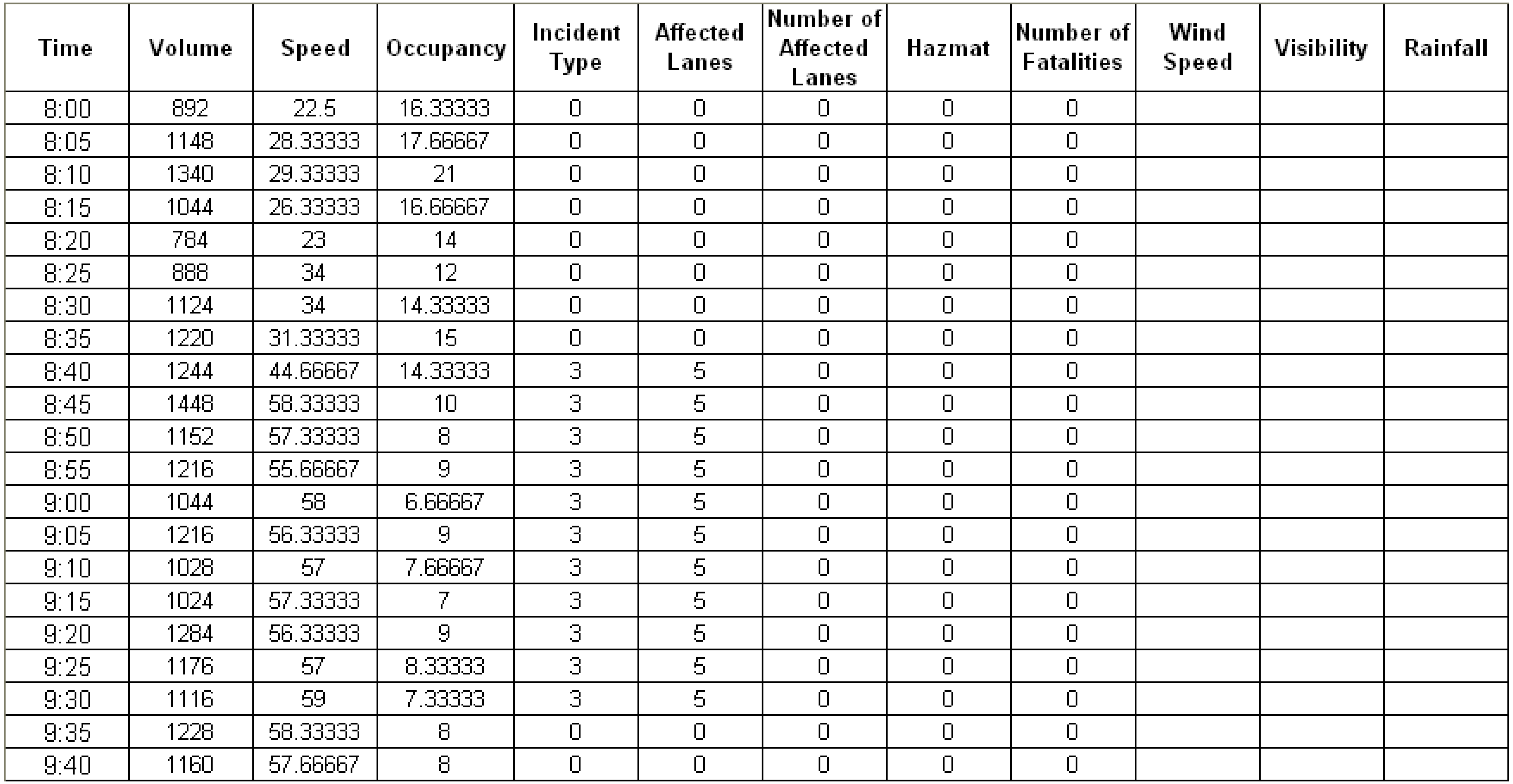
Since the missing traffic flow data may lead to incomplete data for time of day in 5-minute increments, the complete data for time of day in 5-minute increments should be setup in the first column in PERSONAL.xls for later use by the program.

To show how these four Macros work in reorganizing the raw data, four screen shots of the organized data are taken after running each one of the four Macros as shown in Figures D.10 to D.13. In order to show the reorganization changes made to the raw data of four types (especially to the raw incident data), the organized data in Figures D.10 to D.13 are only shown in the time period 8:00 – 9:40 am, including the time period in which the incident occurred at 8:40 am and lasted for 52 minutes at the station I-205 NB Gladstone on Jan 10th, 2006, as shown in the raw data in Figure D.1.



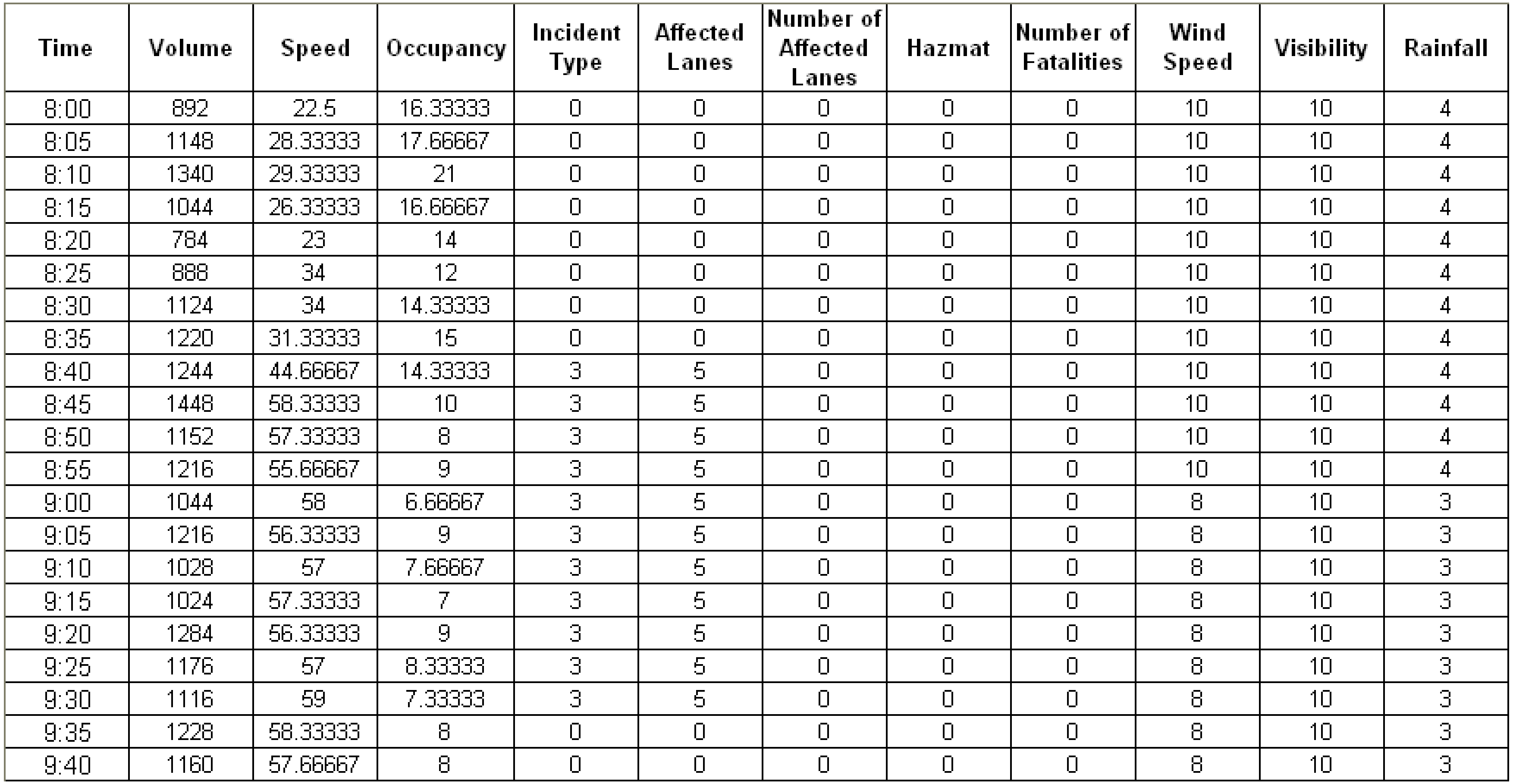
Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
8:00	892	22.5	16.33333								
8:05	1148	28.33333	17.66667								
8:10	1340	29.33333	21								
8:15	1044	26.33333	16.66667								
8:20	784	23	14								
8:25	888	34	12								
8:30	1124	34	14.33333								
8:35	1220	31.33333	15								
8:40	1244	44.66667	14.33333								
8:45	1448	58.33333	10								
8:50	1152	57.33333	8								
8:55	1216	55.66667	9								
9:00	1044	58	6.66667								
9:05	1216	56.33333	9								
9:10	1028	57	7.66667								
9:15	1024	57.33333	7								
9:20	1284	56.33333	9								
9:25	1176	57	8.33333								
9:30	1116	59	7.33333								
9:35	1228	58.33333	8								
9:40	1160	57.66667	8								

Figure D.10 The organized test data – after the first Macro is run



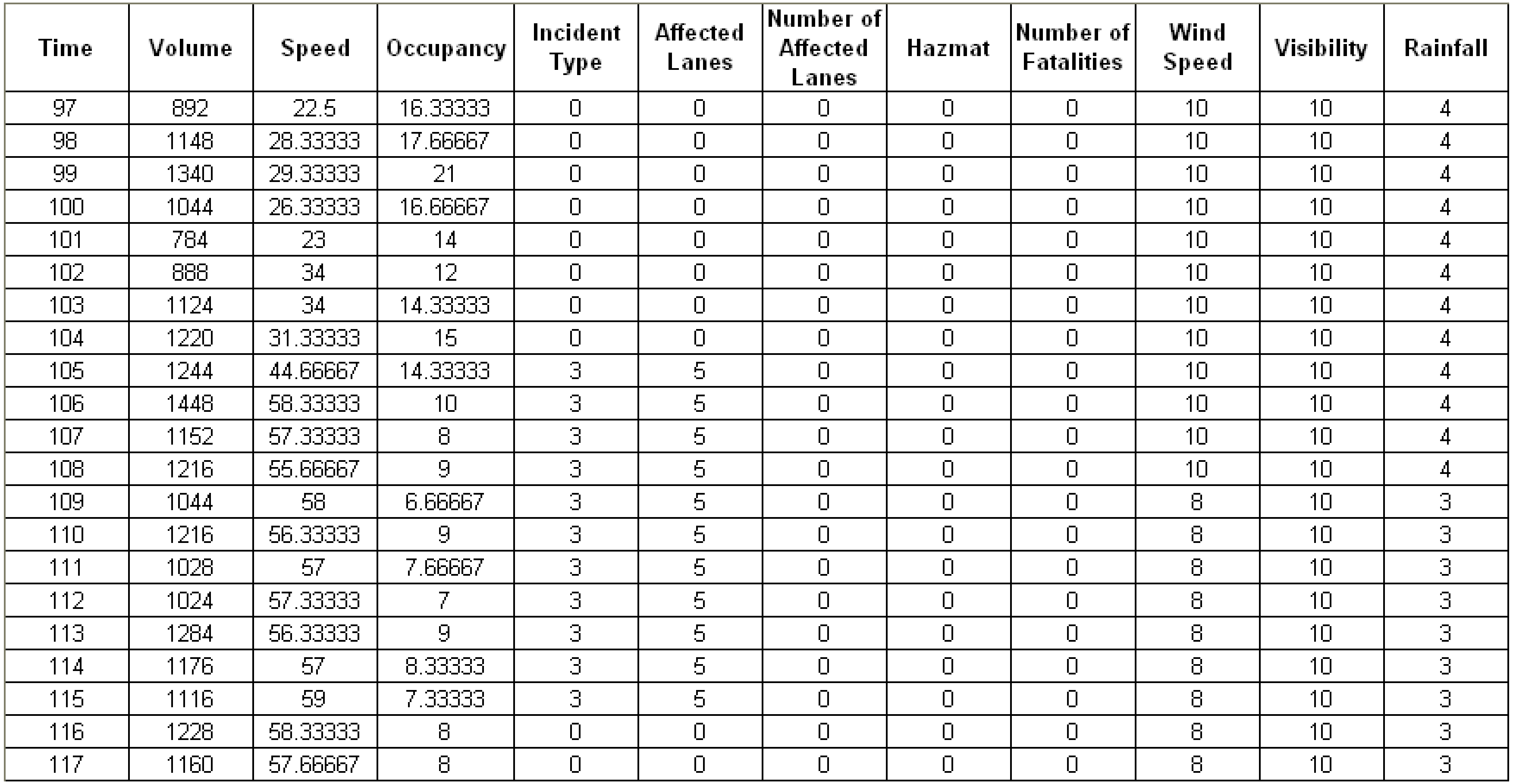
Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
8:00	892	22.5	16.33333	0	0	0	0	0			
8:05	1148	28.33333	17.66667	0	0	0	0	0			
8:10	1340	29.33333	21	0	0	0	0	0			
8:15	1044	26.33333	16.66667	0	0	0	0	0			
8:20	784	23	14	0	0	0	0	0			
8:25	888	34	12	0	0	0	0	0			
8:30	1124	34	14.33333	0	0	0	0	0			
8:35	1220	31.33333	15	0	0	0	0	0			
8:40	1244	44.66667	14.33333	3	5	0	0	0			
8:45	1448	58.33333	10	3	5	0	0	0			
8:50	1152	57.33333	8	3	5	0	0	0			
8:55	1216	55.66667	9	3	5	0	0	0			
9:00	1044	58	6.66667	3	5	0	0	0			
9:05	1216	56.33333	9	3	5	0	0	0			
9:10	1028	57	7.66667	3	5	0	0	0			
9:15	1024	57.33333	7	3	5	0	0	0			
9:20	1284	56.33333	9	3	5	0	0	0			
9:25	1176	57	8.33333	3	5	0	0	0			
9:30	1116	59	7.33333	3	5	0	0	0			
9:35	1228	58.33333	8	0	0	0	0	0			
9:40	1160	57.66667	8	0	0	0	0	0			

Figure D.11 The organized test data – after the second Macro is run



Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
8:00	892	22.5	16.33333	0	0	0	0	0	10	10	4
8:05	1148	28.33333	17.66667	0	0	0	0	0	10	10	4
8:10	1340	29.33333	21	0	0	0	0	0	10	10	4
8:15	1044	26.33333	16.66667	0	0	0	0	0	10	10	4
8:20	784	23	14	0	0	0	0	0	10	10	4
8:25	888	34	12	0	0	0	0	0	10	10	4
8:30	1124	34	14.33333	0	0	0	0	0	10	10	4
8:35	1220	31.33333	15	0	0	0	0	0	10	10	4
8:40	1244	44.66667	14.33333	3	5	0	0	0	10	10	4
8:45	1448	58.33333	10	3	5	0	0	0	10	10	4
8:50	1152	57.33333	8	3	5	0	0	0	10	10	4
8:55	1216	55.66667	9	3	5	0	0	0	10	10	4
9:00	1044	58	6.66667	3	5	0	0	0	8	10	3
9:05	1216	56.33333	9	3	5	0	0	0	8	10	3
9:10	1028	57	7.66667	3	5	0	0	0	8	10	3
9:15	1024	57.33333	7	3	5	0	0	0	8	10	3
9:20	1284	56.33333	9	3	5	0	0	0	8	10	3
9:25	1176	57	8.33333	3	5	0	0	0	8	10	3
9:30	1116	59	7.33333	3	5	0	0	0	8	10	3
9:35	1228	58.33333	8	0	0	0	0	0	8	10	3
9:40	1160	57.66667	8	0	0	0	0	0	8	10	3

Figure D.12 The organized test data – after the third Macro is run



Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
97	892	22.5	16.33333	0	0	0	0	0	10	10	4
98	1148	28.33333	17.66667	0	0	0	0	0	10	10	4
99	1340	29.33333	21	0	0	0	0	0	10	10	4
100	1044	26.33333	16.66667	0	0	0	0	0	10	10	4
101	784	23	14	0	0	0	0	0	10	10	4
102	888	34	12	0	0	0	0	0	10	10	4
103	1124	34	14.33333	0	0	0	0	0	10	10	4
104	1220	31.33333	15	0	0	0	0	0	10	10	4
105	1244	44.66667	14.33333	3	5	0	0	0	10	10	4
106	1448	58.33333	10	3	5	0	0	0	10	10	4
107	1152	57.33333	8	3	5	0	0	0	10	10	4
108	1216	55.66667	9	3	5	0	0	0	10	10	4
109	1044	58	6.66667	3	5	0	0	0	8	10	3
110	1216	56.33333	9	3	5	0	0	0	8	10	3
111	1028	57	7.66667	3	5	0	0	0	8	10	3
112	1024	57.33333	7	3	5	0	0	0	8	10	3
113	1284	56.33333	9	3	5	0	0	0	8	10	3
114	1176	57	8.33333	3	5	0	0	0	8	10	3
115	1116	59	7.33333	3	5	0	0	0	8	10	3
116	1228	58.33333	8	0	0	0	0	0	8	10	3
117	1160	57.66667	8	0	0	0	0	0	8	10	3

Figure D.13 The organized test data – after the fourth Macro is run

- Incident related data: the reorganization of raw incident related data can be automatically processed by the following program except the ID of the incident that occurred at the selected station needs to be appointed to the program by hand. Then the program can use the incident ID input to locate the incident related data in the incident data table (column J to column S) shown in the raw data in Figure D.1. The second Macro containing the following program for the reorganization of raw incident related data is named "Insert_incident."

```
Sub Insert_incident()
    Dim id, i, row_no, duration, lanes_no, type_id, lane_id, j, hazmat_id,
    fatalities_no As Integer
    Dim occur_time As Date
    Dim incident_type, lanes, hazmat As String
    Workbooks(2).Activate       ///Activate the workbook where the raw data are
                                  ///stored in "Sheet 1"
    Worksheets(1).Select             ///Activate the "Sheet 1" of this workbook
    id = Application.InputBox(prompt:="Please type in the incident ID",
    Title:="Incident ID?", Default:=1, Type:=1)        ///Pop out a window asking for
                                                        ///the incident ID
    If id = False Then
        Exit Sub
    End If
    For i = 23 To 35
        If Cells(i, "j").Value = id Then
            row_no = I                  ///Find out the row number of the incident
                                        ///data related to the ID
        End If
    Next i
    occur_time = Cells(row_no, "n").Value
    duration = Cells(row_no, "o").Value
    lanes_no = Cells(row_no, "m").Value
    incident_type = Cells(row_no, "p").Value
    lanes = Cells(row_no, "q").Value
    hazmat = Cells(row_no, "r").Value
    fatalities_no = Cells(row_no, "s").Value   ///Read all the incident related data
                                               ///considered in the test data
    Select Case incident_type
        Case Is = "Crash"
            type_id = 1
        Case Is = "Stall"
            type_id = 2
        Case Is = "Debris"
            type_id = 3
        Case Is = "Construction"
            type_id = 4
        Case Else
            type_id = 5
    End Select                          ///Change the data of the incident type from
```

```
                                        ///text format to integer format
Select Case lanes
    Case Is = "Left Lanes"
        lane_id = 1
    Case Is = "Right Lanes"
        lane_id = 2
    Case Is = "Center Lanes"
        lane_id = 3
    Case Is = "All Lanes"
        lane_id = 4
    Case Is = "Left Shoulder"
        lane_id = 5
    Case Is = "Right Shouler"
        lane_id = 6
    Case Else
        lane_id = 7
End  Select                         ///Change the data of the lane type from text
                                    ///format to integer format
Select Case hazmat
    Case Is = "yes"
        hazmat_id = 1
    Case Is = "no"
        hazmat_id = 0
End Select                          ///Change the data of the hazmat from text format
                                    ///to integer format
```

As shown in Figure D.1, the ID of the incident that occurred at I-205 NB Gladstone station is 530660. When the above program starts running and a window pops out asking for the incident ID, put in 530660 by hand. Then the program can use the incident ID to find and copy the related data of this incident into the organized data as shown in Figure D.11. Because the occurrence time of the incident is 8:40:12 as shown in Figure D.1, all of the incident related data is copied into the organized data, starting at 8:40 as shown in Figure D.11 by rounding the occurrence time into the time of day in 5-minute increments. Since the incident type is debris and the affected lane type is left lanes, the incident type ID and the lane type ID, which are assigned by the program, are used to express the data for these two variables, that is 3 and 5, respectively.

- Weather data: the programs for raw weather data reorganization are saved as the third Macro with the name “Insert_weather.”

```
Sub Insert_weather()
    Dim i As Integer, j As Integer, x As Integer
    x = 2
    Workbooks(2).Activate         ///Activate the workbook where the raw data are
```

```
                                 ///stored in "Sheet 1"
      Worksheets(1).Select         ///Activate the "Sheet 1" of this workbook
      Range("u1:y25").Copy               ///Copy the raw weather data stored in columns
                                          ///U to Y
      Worksheets(2).Select         ///Activate the "Sheet 2" of this workbook
      Range("p1").Select                 ///Temporarily paste the raw weather data from
      ActiveSheet.Paste             ///column P and they will be deleted after the
                                     ///weather data are copied into organized data
      For i = 2 To 25
          For j = x To x + 11
            Cells(j, 10).Value = Cells(i, 18).Value
            Cells(j, 11).Value = Cells(i, 19).Value
            Cells(j, 12).Value = Cells(i, 20).Value
          Next j
            x = x + 12                  ///Every hourly weather data are copied repeatedly
                                         ///for 12 times in the organized data because the
      Next i                              /// data of time of day are in 5-minute increments
      Range("p1:t25").Delete       ///Delete the raw weather data pasted in "Sheet 2"
  End Sub
```

As shown in the raw weather data in Figure D.1, the hourly data for wind speed, visibility and rainfall are 10, 10 and 4 for 8:00, respectively. In Figure D.12, these data are repeatedly copied into all the time points from 8:00 to 8:55.

- Time of day data: the programs for time of day data reorganization are saved as the fourth Macro with the name "Time_of_day." The program uses the consecutive integer numbers from 1 to 288 to substitute the time of day data that is originally in time format, because data in time format cannot be processed by S-PLUS.

```
  Sub Time_of_day()
      Dim i As Integer
      Workbooks(2).Activate        ///Activate the workbook where the raw data are
      stored in "Sheet 1"
      Worksheets(2).Select              ///Activate the "Sheet 2" of this workbook
      Range("b2").Copy
      Range("a2:a289").PasteSpecial xlPasteFormats    ///Paste the format of column B
                                                         ///to column A
      For i = 2 To 289
          Cells(i, 1).Value = i – 1             ///Change the time of day data into the
                                                 ///consecutive integer numbers from 1 to 288
      Next i
  End Sub
```

APPENDIX E. MODEL CONSTRUCTION IN S-PLUS USING EXAMPLE TEST DATA

The regression tree algorithm is implemented in S-PLUS. Thus, in this research, regression tree model construction and validation are both performed in S-PLUS by importing test data and validation data into the software. By using S-PLUS, the regression tree models can be constructed upon the test data sets obtained after data collection and reorganization. The implementation of regression tree algorithm in S-PLUS for model construction is described in the following using the example test data set shown in Table 4.1.

To construct a regression tree model using the example test data using S-PLUS, the test data needs to first be imported into S-PLUS by clicking File>Import Data>From File as shown in Figure E.1.

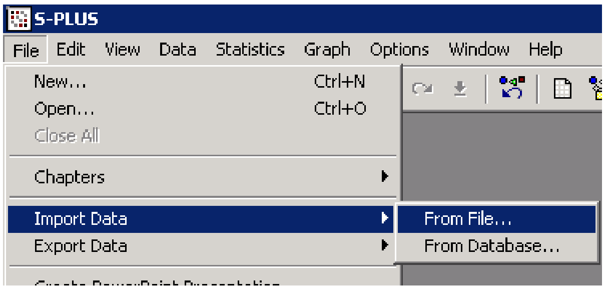
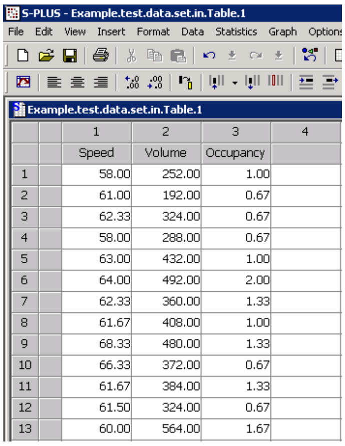


Figure E.1 Import test data set into S-PLUS

Then a window named “Import From File” will appear and the test data file to be imported can be selected by using “Browse.” After clicking OK, the test data set will show up as in Figure E.2, which means this test data set can now be used to construct a regression tree model in S-PLUS. By clicking Statistics>Tree>Tree Models as shown in Figure E.3, the window “Tree Models” is opened and the first three tabs in the “Tree Models” window – Model, Results and Plot – are used to construct the tree model, show the result summary and display the tree plot, respectively, as shown in Figures E.4, E.5 and E.6.


S-PLUS - Example.test.data.set.in.Table.1
File Edit View Insert Format Data Statistics Graph Options
Example.test.data.set.in.Table.1

	1	2	3	4
	Speed	Volume	Occupancy	
1	58.00	252.00	1.00	
2	61.00	192.00	0.67	
3	62.33	324.00	0.67	
4	58.00	288.00	0.67	
5	63.00	432.00	1.00	
6	64.00	492.00	2.00	
7	62.33	360.00	1.33	
8	61.67	408.00	1.00	
9	68.33	480.00	1.33	
10	66.33	372.00	0.67	
11	61.67	384.00	1.33	
12	61.50	324.00	0.67	
13	60.00	564.00	1.67	

Figure E.2 Imported test data set in S-PLUS

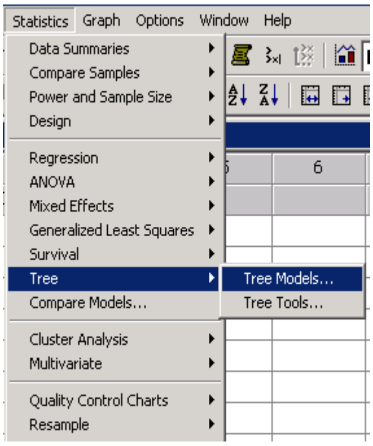
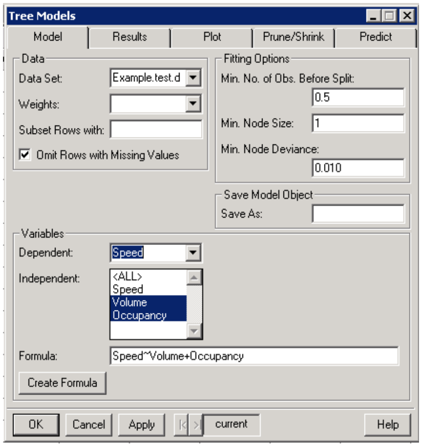


Figure E.3 Open tree models window


Tree Models
Model Results Plot Prune/Shrink Predict
OK Cancel Apply current Help

Data

Data Set: Example.test.d

Weights:

Subset Rows with:

☑ Omit Rows with Missing Values

Fitting Options

Min. No. of Obs. Before Split: 0.5

Min. Node Size: 1

Min. Node Deviance: 0.010

Save Model Object

Save As:

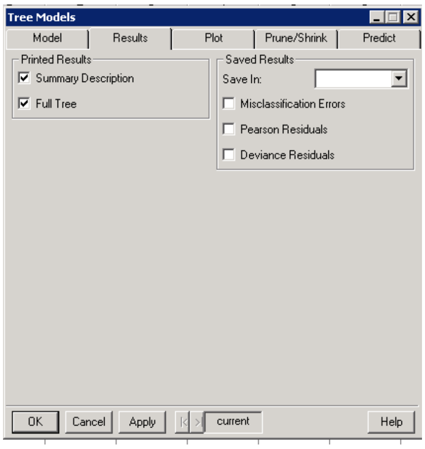
Variables

Dependent: Speed

Independent: <ALL> Speed Volume Occupancy

Formula: Speed~Volume+Occupancy

Figure E.4 “Model” tab


Tree Models
Model Results Plot Prune/Shrink Predict
OK Cancel Apply current Help

Printed Results

☑ Summary Description

☑ Full Tree

Saved Results

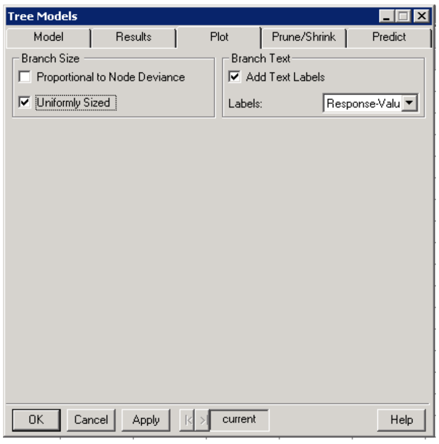
Save In:

☐ Misclassification Errors

☐ Pearson Residuals

☐ Deviance Residuals

Figure E.5 “Results” tab


Tree Models
Model Results Plot Prune/Shrink Predict
OK Cancel Apply current Help

Branch Size

☐ Proportional to Node Deviance

☑ Uniformly Sized

Branch Text

☑ Add Text Labels

Labels: Response-Valu

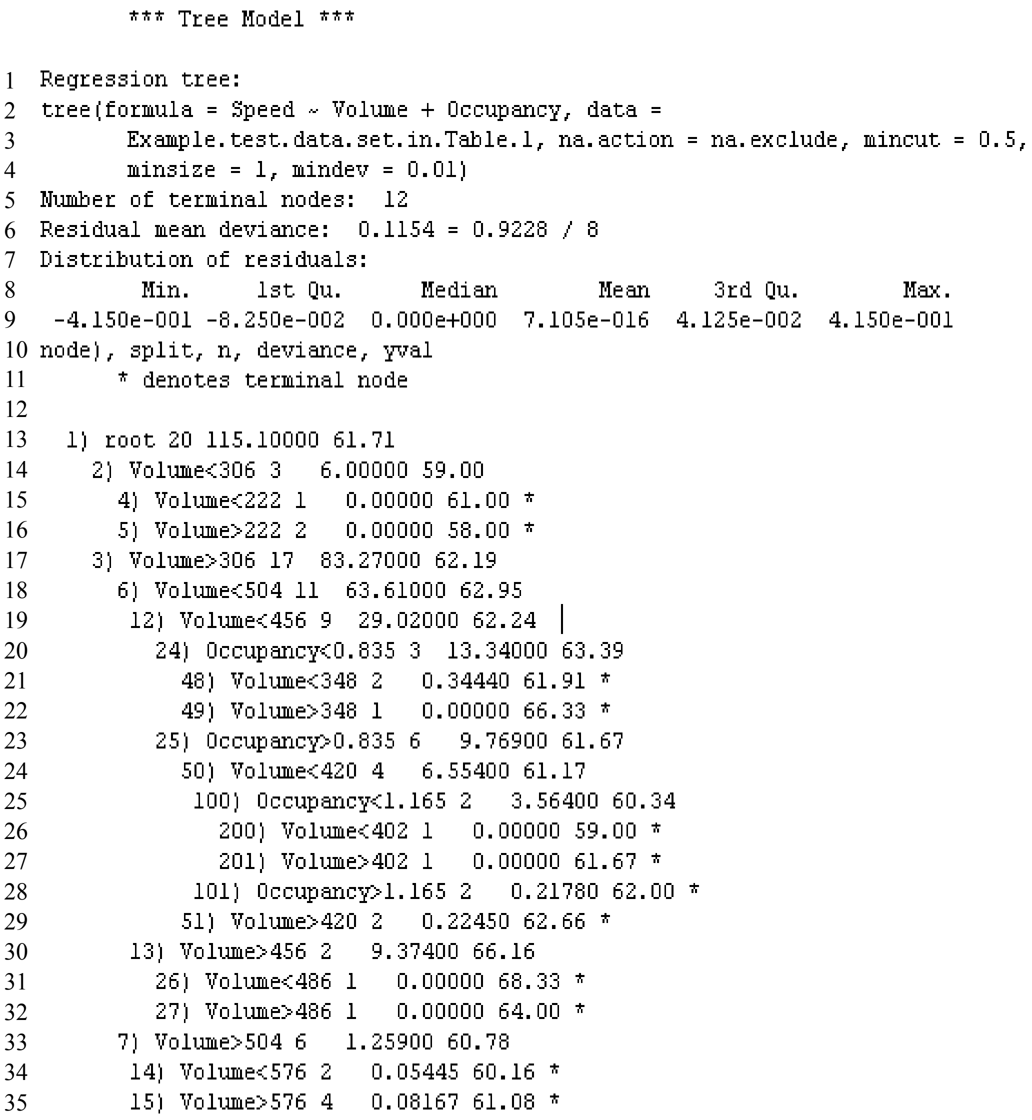
Figure E.6 “Plot” tab

There are four sections in the tab “Model,” Data, Fitting Options, Variables and Save Model Object, as shown in Figure E.4. Only the first three sections are used to construct the tree model. In the “Data” section, select the test data set for constructing the regression tree model in “Data Set.” In the “Fitting Options” section, the three options are used to setup when to stop the regression tree model construction, that is, to stop splitting the test data set. In the “Variables” section, the response variable for the tree model needs to be selected in “Dependent” and the explanatory variables need to be selected in “Independent.” For the example test data shown in Table 4.1, the response variable is speed and the explanatory variables are occupancy and volume.

In the “Results” tab as shown in Figure E.5, only the section “Printed Results” needs to be used to decide what results would be shown in the results summary. Both of the options in this section need to be checked to view the description of the regression tree model constructed and the complete tree in the results summary as shown in Figure E.7 later.

The “Plot” tab as shown in Figure E.6 is used to decide how the plot of the constructed regression tree model would be displayed. In the “Branch Size” section, selecting the first option can make the lengths of the branches of the regression tree proportional to the node deviance. That is, the larger the node deviance, the longer the branch of that node is. Since the node deviances are all shown in the results summary, “Uniformly Sized” can be just selected for tree plot to make the lengths of the braches all same for clarity. In the “Branch Text” section, by selecting “Add Text Labels,” the text labels will be added to the terminal nodes of the regression tree constructed. For the types of the labels, “Response-Value” is selected here to view the mean values of the response variable on all the terminal nodes of the regression tree.

After finishing all the steps described above in the tabs “Model,” “Results” and “Plot” in the window “Tree Models,” click OK. The results summary and the regression tree plot for the regression tree model constructed upon the example test data will then be displayed as shown in Figures E.7 and E.8.



```
          *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Volume + Occupancy, data =
       Example.test.data.set.in.Table.1, na.action = na.exclude, mincut = 0.5,
       minsize = 1, mindev = 0.01)
Number of terminal nodes:  12
Residual mean deviance:  0.1154 = 0.9228 / 8
Distribution of residuals:
        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
 -4.150e-001 -8.250e-002  0.000e+000  7.105e-016  4.125e-002  4.150e-001
node), split, n, deviance, yval
      * denotes terminal node

  1) root 20 115.10000 61.71
    2) Volume<306 3   6.00000 59.00
      4) Volume<222 1   0.00000 61.00 *
      5) Volume>222 2   0.00000 58.00 *
    3) Volume>306 17  83.27000 62.19
      6) Volume<504 11  63.61000 62.95
       12) Volume<456 9  29.02000 62.24
         24) Occupancy<0.835 3  13.34000 63.39
           48) Volume<348 2   0.34440 61.91 *
           49) Volume>348 1   0.00000 66.33 *
         25) Occupancy>0.835 6   9.76900 61.67
           50) Volume<420 4   6.55400 61.17
            100) Occupancy<1.165 2   3.56400 60.34
              200) Volume<402 1   0.00000 59.00 *
              201) Volume>402 1   0.00000 61.67 *
            101) Occupancy>1.165 2   0.21780 62.00 *
           51) Volume>420 2   0.22450 62.66 *
       13) Volume>456 2   9.37400 66.16
         26) Volume<486 1   0.00000 68.33 *
         27) Volume>486 1   0.00000 64.00 *
      7) Volume>504 6   1.25900 60.78
       14) Volume<576 2   0.05445 60.16 *
       15) Volume>576 4   0.08167 61.08 *
```

Figure E.7 Results summary of the example regression tree model

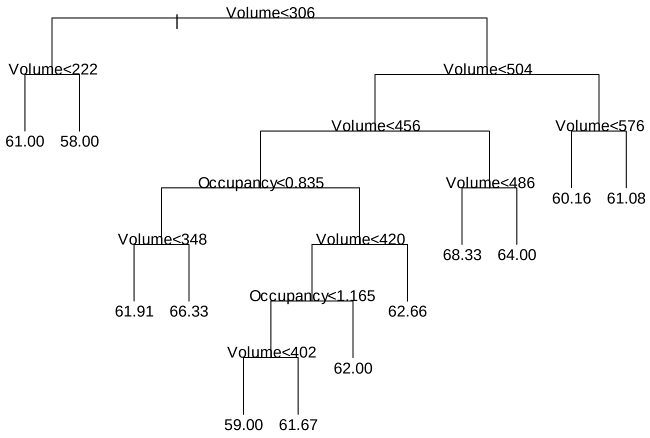


Figure E.8 Regression tree plot of the example test data set

APPENDIX F. RANDOM NUMBER GENERATION PROGRAM

In this research, random number generation is needed in both full regression tree construction and validation data sets random selection for RCBD. The challenge for the random number generation in these two applications is that random numbers in a large percentage of the original numbers need to be generated and at the same time these random numbers cannot be repetitive. For example, in full regression tree model construction, 227 non-repetitive daily test data sets need to be randomly selected out of the total of 342 daily test data sets.

Therefore, a Macro written in Excel VBA is developed to perform the random number generation for our research. The program for this Macro is shown below.

```
Sub Randx()
    Dim xx(1 To AAA) As Integer
    For t = 1 To BBB
    rerand:
    x = Int(Rnd() * AAA + 1)
   If xx(x) > 0 Then GoTo rerand
    r = r + 1
    Cells(r, 1) = x
    xx(x) = r
    Next
End Sub
```

To apply the Macro, an Excel file needs to be created first. In the newly-created Excel file, by clicking Tools>Macro>Visual Basic Editor, the Microsoft Visual Basic Editor is opened. After inserting a Module in this Excel file as shown in Figure D.7, the above program can be copied into the right blank area in Visual Basic Editor. In the second and the fifth line of the program, AAA needs to be substituted with the number that we need to randomly select from. In the third line of the program, BBB needs to be substituted with the number of random numbers needed to be generated. For example, to generate 36 random numbers out of the integer numbers 1 to 51, AAA needs to be 51 and BBB needs to be 36. After saving all the changes, this Macro can be run by opening Macro window and selecting Randx (the name of this Macro) in the Macro window and then clicking Run. Then result of the random generated numbers will be shown in the first column in this Excel file.

APPENDIX G. VALIDATION OF REGRESSION TREE MODEL IN S-PLUS

As an example, the validation of the full regression tree model in S-PLUS will be shown using a randomly-selected daily validation data set. To validate a regression tree model in S-PLUS, the validation data set needs to be imported into S-PLUS first by clicking File>Import Data>From File as shown in Figure G.1. Validation data sets can use any daily data sets collected at the selected I-205 NB Gladstone station, which have been reorganized and are applicable in S-PLUS. Here, to validate the full regression tree model, the daily data set collected at the same station I-205 NB Gladstone on August 2, 2006 is randomly selected as the validation data set, which is only shown between 7:05 and 9:05 pm (between 230 and 254 in time order) in Figure G.2 in the interest of space.

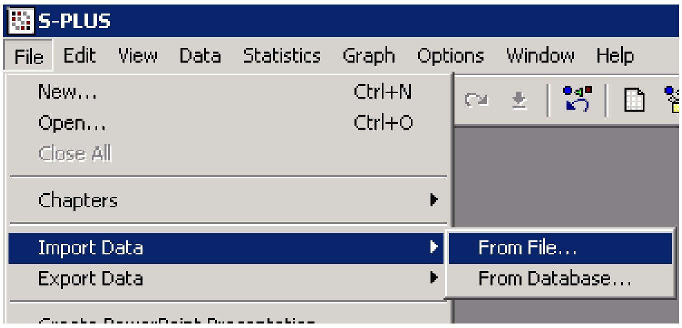
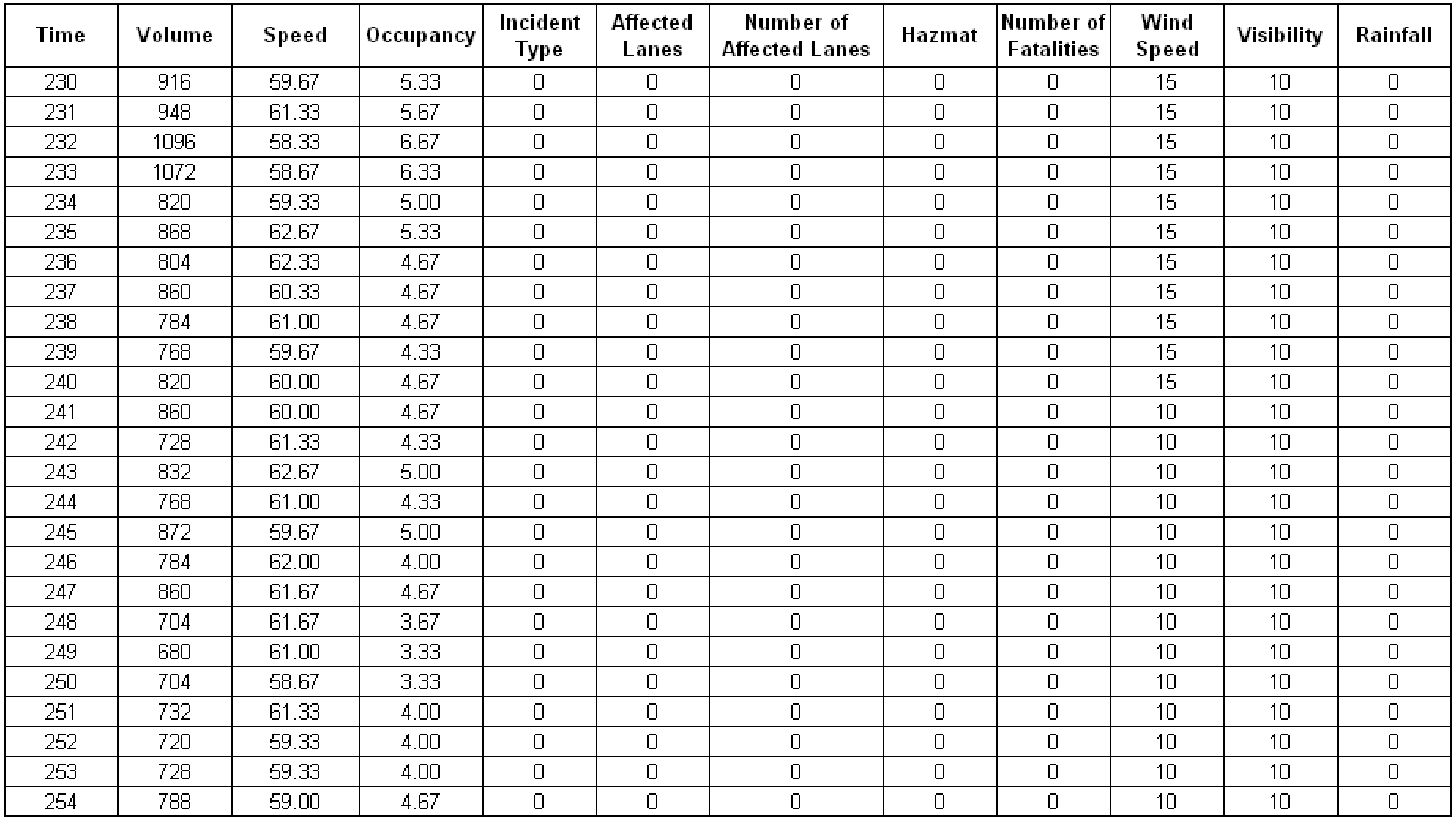


Figure G.1 Importation of validation data set into S-PLUS



Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
230	916	59.67	5.33	0	0	0	0	0	15	10	0
231	948	61.33	5.67	0	0	0	0	0	15	10	0
232	1096	58.33	6.67	0	0	0	0	0	15	10	0
233	1072	58.67	6.33	0	0	0	0	0	15	10	0
234	820	59.33	5.00	0	0	0	0	0	15	10	0
235	868	62.67	5.33	0	0	0	0	0	15	10	0
236	804	62.33	4.67	0	0	0	0	0	15	10	0
237	860	60.33	4.67	0	0	0	0	0	15	10	0
238	784	61.00	4.67	0	0	0	0	0	15	10	0
239	768	59.67	4.33	0	0	0	0	0	15	10	0
240	820	60.00	4.67	0	0	0	0	0	15	10	0
241	860	60.00	4.67	0	0	0	0	0	10	10	0
242	728	61.33	4.33	0	0	0	0	0	10	10	0
243	832	62.67	5.00	0	0	0	0	0	10	10	0
244	768	61.00	4.33	0	0	0	0	0	10	10	0
245	872	59.67	5.00	0	0	0	0	0	10	10	0
246	784	62.00	4.00	0	0	0	0	0	10	10	0
247	860	61.67	4.67	0	0	0	0	0	10	10	0
248	704	61.67	3.67	0	0	0	0	0	10	10	0
249	680	61.00	3.33	0	0	0	0	0	10	10	0
250	704	58.67	3.33	0	0	0	0	0	10	10	0
251	732	61.33	4.00	0	0	0	0	0	10	10	0
252	720	59.33	4.00	0	0	0	0	0	10	10	0
253	728	59.33	4.00	0	0	0	0	0	10	10	0
254	788	59.00	4.67	0	0	0	0	0	10	10	0

Figure G.2 Validation data set at I-205 NB Gladstone on August 2, 2006. (7:05 – 9:05 pm)

After the test data set for full regression tree construction and the selected validation data set are imported into S-PLUS, by clicking Statistics>Tree>Tree Models as shown in Figure G.3, the window “Tree Models” is opened. The first three tabs in “Tree Models” window – Model, Results and Plot – are used to construct the tree model, show the result summary and display the tree plot, respectively.

After the full regression tree model is setup as described in section 6.1.1, the fifth tab “Predict” in “Tree Models” window, as shown in Figure G.4, is used to setup the validation of the tree model. “X080206” is the file name of the validation data collected at I-205 NB Gladstone on August 2, 2006, which is selected in “New Data” in Figure G.4 as the validation data. “response” is selected in “Prediction Type,” since “Speed” is the response variable and will be predicted by the full regression tree model. “Save As” is used to choose where the validation results are saved and here the same validation data set file is chosen to save the predicted speeds. Then by clicking “OK” at the bottom of the “Tree Models” window, the predicted speeds are shown in column 13 of the validation data set “X080206” with column name “fit,” which means fitted values by the regression

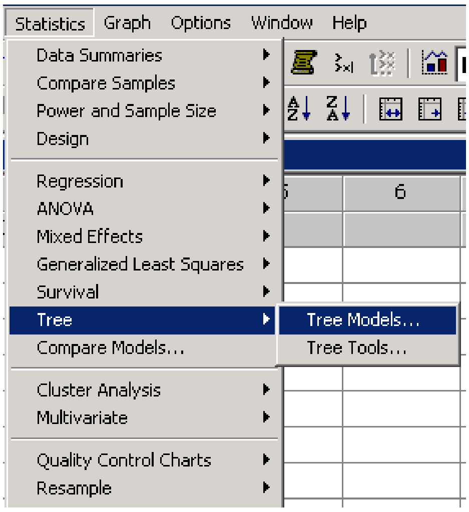
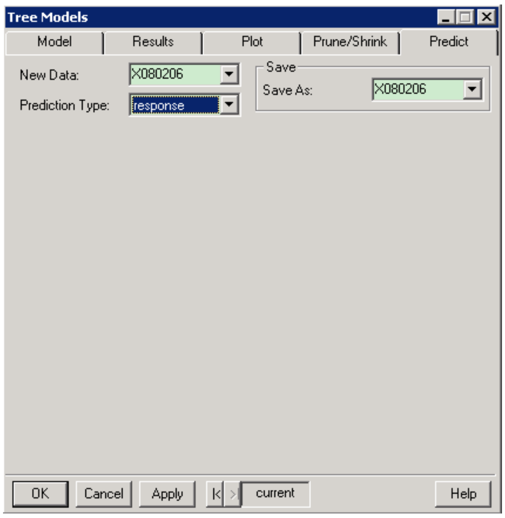


Figure G.3 Open tree models window



New Data: X080206

Prediction Type: response

Save As: X080206

Figure G.4 “Tree Models” window – “Predict” tab tree model.

MSE is used to estimate the accuracy of the predicted speeds by the full regression tree model compared with the actual speeds of the validation data set. For example, if MSE is used to estimate the accuracy of the predicted speeds between 7:05 and 9:05 pm (between 230 and 254 in time order) on August 2, 2006 by the full regression tree model, the MSE result is 1.70 as shown in Table G.1, which is fairly low for the errors of speed values.

Table G.1 MSE result of validation data on August 2, 2006

Time	Actual Speed	Predicted Speed	Squared Error
230	59.67	60.00	0.11
231	61.33	60.00	1.77
232	58.33	58.11	0.05
233	58.67	58.11	0.31
234	59.33	60.00	0.45
235	62.67	60.00	7.10
236	62.33	60.00	5.43
237	60.33	60.00	0.11
238	61.00	60.00	0.99
239	59.67	60.00	0.11
240	60.00	60.00	0.00
241	60.00	60.00	0.00
242	61.33	60.00	1.77
243	62.67	60.00	7.10
244	61.00	60.00	0.99
245	59.67	60.00	0.11
246	62.00	60.00	3.99
247	61.67	60.00	2.77
248	61.67	60.00	2.77
249	61.00	60.00	0.99
250	58.67	60.00	1.78
251	61.33	60.00	1.77
252	59.33	60.00	0.45
253	59.33	60.00	0.45
254	59.00	60.00	1.01
		MSE	1.70

APPENDIX H. MACRO IN EXCEL VBA FOR CHARACTERIZATION

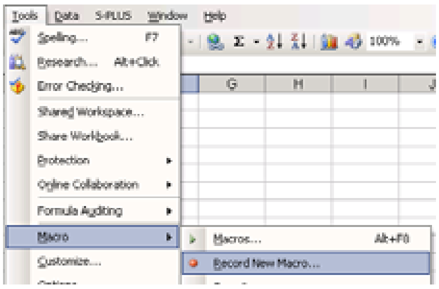
In the characterization approach, four standards are setup to track the characteristics of both test data sets and validation data sets, including “Outliers”, “Good weather”, “Incidents” and “Weekday or Weekend”. “Outliers” is to check if there are missing data or erroneous data in traffic flow data due to detector error. For “Good Weather”, based on published sources, a data set is regarded as having good weather if wind speed is lower than 15 mph, visibility is higher than 8 miles and rainfall is less than 3 mm per hour and no good weather if any of the three conditions is not satisfied. “Incidents” is to check if any incidents existed in the daily data sets we collected. “Weekday or Weekend” is used to track the characteristic of day of week in the data sets, since the traffic flow patterns between weekdays and weekends are surely different.

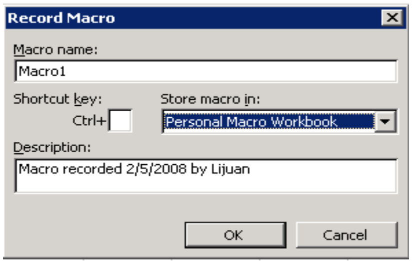
The challenge in applying the characterization approach is to determine which characterization a daily data set belongs to. Although tracking the characteristics of the data sets can be done manually, computer programs can be written to perform the same function accurately and more efficiently. Thus, a Macro written in Excel VBA programs is used to perform characterization for all the collected daily test data sets and validation data sets after raw data reorganizations. The following program can be saved as a Macro in a special Excel file PERSONAL.XLS, which can make Macros applicable for any opened Excel files.

```
Sub Characterization()
     Dim i As Integer, j As Integer, k As Integer, m, n As Integer, day_no As Integer, d As
   Date
     Workbooks(2).Activate
     Worksheets(3).Select
     Range("a1").Value = "Outliers?"
     Range("a2").Value = "Good Weather?"
     Range("a3").Value = "Incidents?"
     Range("a4").Value = "Weekday or Weekend?"
     j = 0
     k = 0
   m = 0
   Worksheets(1).Select
   d = DateValue(Range("u2").Value)
   day_no = Weekday(d, vbMonday)
   Worksheets(2).Select
   For i = 2 To 289
```

```
        If Cells(i, 2).Value = 0 And Cells(i, 3).Value = 0 And Cells(i, 4).Value = 0 Then j = j + 1
        If Cells(i, 10).Value > 15 Or Cells(i, 11).Value < 8 Or Cells(i, 12).Value > 3 Then m = m + 1
        If Not Cells(i, 5).Value = 0 Then k = k + 1
    Next i
    Worksheets(3).Select
    If j > 0 Then Range("b1").Value = "Yes" Else Range("b1").Value = "No"
    If m > 0 Then Range("b2").Value = "No" Else Range("b2").Value = "Yes"
    If k > 0 Then Range("b3").Value = "Yes" Else Range("b3").Value = "No"
    If day_no < 6 Then Range("b4").Value = "Weekday" Else Range("b4").Value = "Weekend"
    Range("b6").Value = d
    Range("a1:b6").Columns.AutoFit
    Range("a1:b6").HorizontalAlignment = xlCenter
    Worksheets(1).Select
    ActiveSheet.Name = "Raw Data"
    ActiveSheet.Tab.ColorIndex = 4
    Worksheets(2).Select
    ActiveSheet.Name = "Organized Data"
    ActiveSheet.Tab.ColorIndex = 22
    Worksheets(3).Select
    ActiveSheet.Name = "Characterization"
    ActiveSheet.Tab.ColorIndex = 45
End Sub
```

To access PERSONAL.XLS, the software Excel needs to be opened first and then followed by clicking Tools>Macro>Record New Macro as shown in Figure H.1. A dialog window will show up and Personal Macro Workbook needs to be selected in “Store macro in:” as shown in Figure H.2. After clicking OK, a new file called PERSONAL.XLS will be created automatically in Excel. Then by clicking Window>Unhide as shown in Figure H.3, “PERSONAL” needs to be selected in a popped out window “Unhide workbook.” Now all the opened Excel files can be closed by clicking Yes in a popped-out confirmation window as shown in Figure H.4. Next time no matter which Excel file is opened, the file PERSONAL.XLS will open automatically. Then programs in Excel VBA can be written as Macros in the opened file PERSONAL.XLS by clicking Tools>Macro>Visual Basic Editor as shown in Figure H.5. After a window named “Microsoft Visual Basic – PERSONAL.XLS” pops out, right click “Sheet 1” under “VBAProject (PERSONAL.XLS)” and then click Insert>Module as shown in Figure H.6. A blank window will then pop out for programs writing (or code imputing) to create Macros in the file PERSONAL.XLS. After the above program is copied into the blank window that popped out for programs writing, the file PERSONAL.XLS can be closed with changes saved.





Macro name:
Macro1

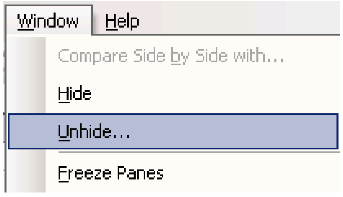
Shortcut key:
Ctrl+

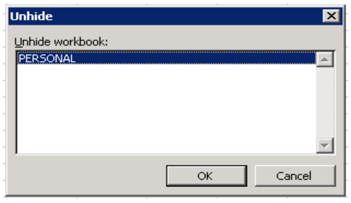
Store macro in:
Personal Macro Workbook

Description:
Macro recorded 2/5/2008 by Lijuan

Figure H.1 Record new Macro

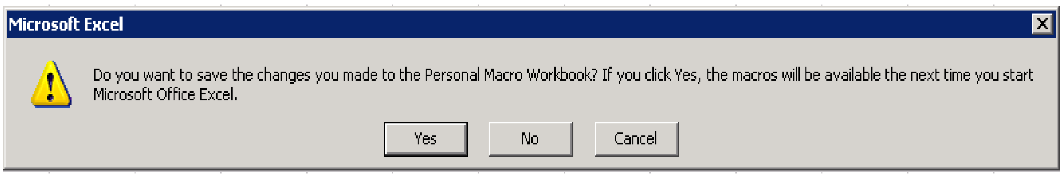
Figure H.2 Personal Macro workbook





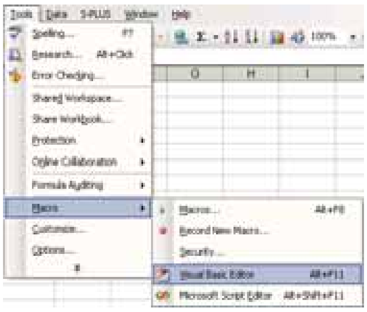
Unhide workbook:
PERSONAL

Figure H.3 Unhide PERSONAL.XLS



Do you want to save the changes you made to the Personal Macro Workbook? If you click Yes, the macros will be available the next time you start Microsoft Office Excel.

Figure H.4 Exit Excel



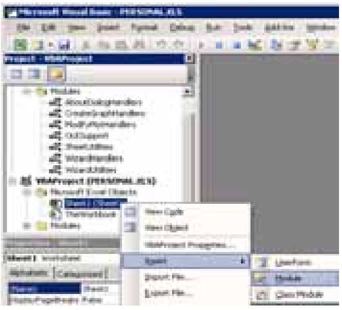


Figure H.5 Open Visual Basic Editor

Figure H.6 Insert a new Module

To apply this Macro saved in PERSONAL.XML file to perform the characterization for a daily data set, the data set has to be first cleaned up and reorganized using the four Macros demonstrated in Appendix D. Then the Excel file containing the data set is opened while PERSONAL.XML file is automatically opened with the data set file. In PERSONAL.XML file, by clicking Tools>Macro>Macros, as shown in Figure H.7, the Macro window is opened, as shown in Figure H.8. By selecting the name of this Macro "Characterization" and then clicking Run, characterization is performed to the opened data set file.

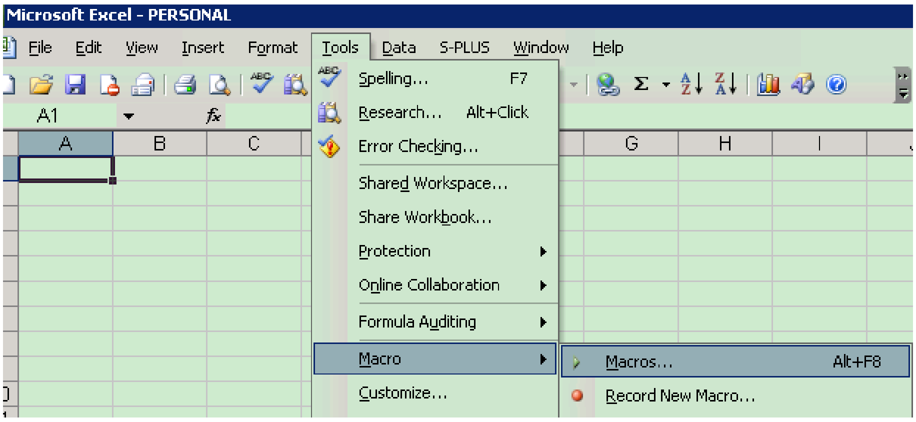
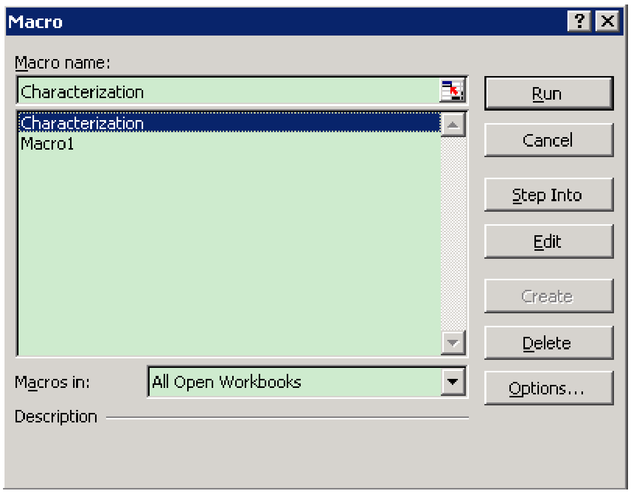


Figure H.7 Open Macro window


Macro name: Characterization

Characterization
Macro1

Macros in: All Open Workbooks

Figure H.8 Macro window

APPENDIX I. REGRESSION TREE MODELS OF TEN CHARACTERIZATIONS

The result summaries and the tree plots of the regression tree models of ten characterizations are shown in the following. Some of the tree plots shown below are crowded and hard to see because those regression tree models are complex ones. However, with the help of the result summaries of these ten regression tree models provided by S-PLUS, it would not be difficult to interpret even a complex tree.

1. Regression tree model of characterization 3

- Results summary

```
    *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.3, na.action = na.exclude, mincut = 0.5, minsize = 1, mindev
     = 0.01)
Variables actually used in tree construction:
[1] "Volume"    "Occupancy"
Number of terminal nodes:  5
Residual mean deviance:  8.893 = 130600 / 14680
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -3.231e+001 -9.795e-001  0.000e+000 -8.972e-014  1.163e+000  2.577e+001
node), split, n, deviance, yval
      * denotes terminal node

 1) root 14688 5782000 51.27
   2) Volume<2 1808      0  0.00 *
   3) Volume>2 12880  361800 58.47
     6) Occupancy<19.1667 12678  169800 58.92
      12) Volume<1146 6846   46350 60.84 *
      13) Volume>1146 5832   69040 56.68
        26) Occupancy<14.1667 4654   34380 57.31 *
        27) Occupancy>14.1667 1178   25410 54.18 *
                   7) Occupancy>19.1667 202   24440 29.90 *
```

- Tree plot

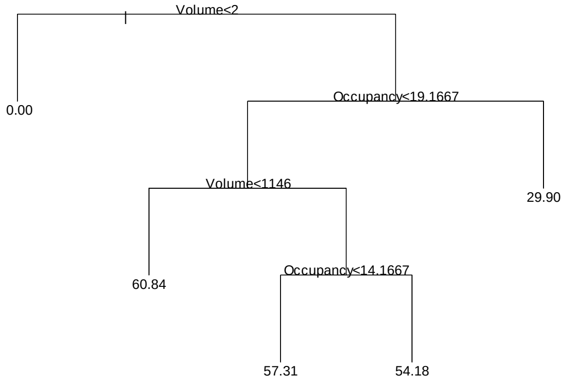


Figure I.1 Tree plot of regression tree model of characterization 3

2. Regression tree model of characterization 4

- Results summary

```
*** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.4, na.action = na.exclude, mincut = 0.5, minsize = 1, mindev
     = 0.01)
Variables actually used in tree construction:
[1] "Volume"    "Occupancy"
Number of terminal nodes:  4
Residual mean deviance:  8.349 = 153800 / 18420
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -3.287e+001 -9.549e-001  0.000e+000  4.411e-014  1.393e+000  2.347e+001
node), split, n, deviance, yval
     * denotes terminal node

 1) root 18422 7428000 50.63
   2) Volume<2 2392      0  0.00 *
   3) Volume>2 16030  381400 58.18
     6) Occupancy<18.8333 15762  180900 58.61
      12) Volume<878 7165   58380 60.62 *
      13) Volume>878 8597   69600 56.94 *
     7) Occupancy>18.8333 268   25800 32.87 *
```

- Tree plot

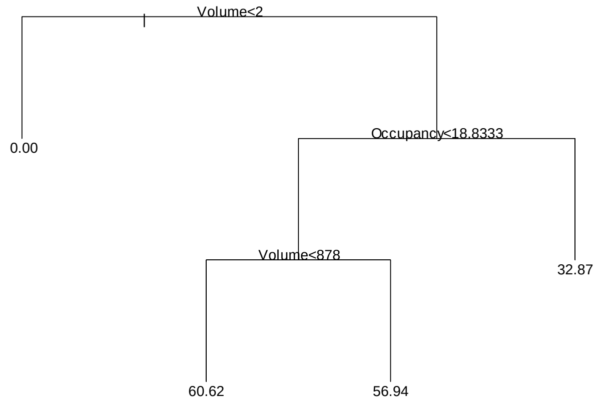


Figure I.2 Tree plot of regression tree model of characterization 4

3. Regression tree model of characterization 5

- Result summary

```
*** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.5, na.action = na.exclude, mincut = 0.5, minsize = 1, mindev
     = 0.01)
Variables actually used in tree construction:
[1] "Occupancy"               "Volume"
[3] "Time"                   "Wind.Speed"
[5] "Number.of.Affected.Lanes"
Number of terminal nodes:  52
Residual mean deviance:  3.351 = 5616 / 1676
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -1.262e+001 -9.098e-001  1.521e-002 -1.114e-015  1.003e+000  1.235e+001
node), split, n, deviance, yval
      * denotes terminal node

    1) root 1728 26720.000 58.90
      2) Occupancy<18.8333 1713 16750.000 59.11
        4) Volume<1078 827  5053.000 61.21
          8) Volume<466 512  3722.000 61.76
           16) Volume<42 2    2.000 65.50 *
           17) Volume>42 510  3692.000 61.75
             34) Time<286.5 498  3622.000 61.71
               68) Time<260.5 376  3162.000 61.87
        136) Time<42.5 250  2285.000 61.60
          272) Wind.Speed<4.5 179  1639.000 61.42
            544) Time<38.5 163  1489.000 61.55
             1088) Time<32.5 135  1073.000 61.33
               2176) Time<23.5 92   632.800 61.74
                 4352) Time<22.5 88   557.700 61.60
                   8704) Volume<86 17   125.800 62.38 *
                   8705) Volume>86 71   418.900 61.41
                    17410) Volume<94 6    32.760 59.72 *
                    17411) Volume>94 65   367.500 61.57
                      34822) Volume<174 58   343.900 61.71
                        69644) Volume<166 56   314.900 61.60
                         139288) Volume<162 55   304.000 61.66
                          278576) Occupancy<0.833335 29   114.500 61.21 *
                          278577) Occupancy>0.833335 26   176.700 62.17 *
                         139289) Volume>162 1     0.000 58.33 *
                        69645) Volume>166 2    10.890 64.67 *
                      34823) Volume>174 7    12.540 60.38 *
                 4353) Time>22.5 4    32.970 64.92 *
               2177) Time>23.5 43   391.400 60.46
                 4354) Volume<86 25   251.100 60.99 *
                 4355) Volume>86 18   123.700 59.72 *
             1089) Time>32.5 28   378.300 62.61
               2178) Volume<66 6    50.870 65.75 *
               2179) Volume>66 22   252.000 61.75 *
            545) Time>38.5 16   119.000 60.09 *
          273) Wind.Speed>4.5 71   625.400 62.06
            546) Volume<74 6   136.800 60.14 *
            547) Volume>74 65   464.300 62.24
             1094) Volume<196 64   443.100 62.31
               2188) Time<26.5 46   370.800 62.64
                 4376) Time<25.5 45   287.500 62.44
                   8752) Volume<90 11    83.660 63.30 *
                   8753) Volume>90 34   193.000 62.16 *
                 4377) Time>25.5 1     0.000 71.67 *
               2189) Time>26.5 18    54.620 61.47 *
             1095) Volume>196 1     0.000 57.67 *
        137) Time>42.5 126   826.600 62.38
          274) Time<49.5 42   509.300 63.06
            548) Volume<122 7    28.600 60.55 *
            549) Volume>122 35   427.600 63.57
             1098) Wind.Speed<6.5 28   352.400 64.04
               2196) Time<46.5 17   190.700 63.19 *
               2197) Time>46.5 11   130.100 65.36 *
             1099) Wind.Speed>6.5 7    43.560 61.67 *
          275) Time>49.5 84   288.200 62.04
            550) Volume<322 39   129.500 61.32 *
            551) Volume>322 45   120.100 62.67 *
       69) Time>260.5 122   423.200 61.24
        138) Time<282.5 98   331.100 61.41
          276) Volume<214 3    12.960 65.11 *
          277) Volume>214 95   275.700 61.29
            554) Occupancy<4.16667 92   253.300 61.36 *
            555) Occupancy>4.16667 3     9.185 59.22 *
        139) Time>282.5 24    77.510 60.54 *
     35) Time>286.5 12    44.070 63.22 *
  9) Volume>466 315   923.600 60.32
   18) Number.of.Affected.Lanes<0.5 288   759.000 60.52
     36) Time<74.5 68   137.300 61.46 *
     37) Time>74.5 220   542.100 60.22
       74) Time<226.5 20    23.420 58.80 *
       75) Time>226.5 200   474.100 60.37
        150) Time<253.5 145   231.200 60.60 *
        151) Time>253.5 55   213.300 59.74 *
   19) Number.of.Affected.Lanes>0.5 27    33.370 58.21 *
5) Volume>1078 886  4621.000 57.15
```

```
   10) Volume<1458 577  1529.000 58.00
     20) Volume<1334 360  1019.000 58.38
       40) Occupancy<12.8333 351   618.700 58.47
         80) Volume<1222 158   280.300 58.75
          160) Occupancy<11.1667 138   208.900 58.85 *
          161) Occupancy>11.1667 20    59.760 58.03 *
         81) Volume>1222 193   315.900 58.24
          162) Time<153.5 135   199.700 58.42 *
          163) Time>153.5 58   100.700 57.80 *
       41) Occupancy>12.8333 9   288.200 54.89
         82) Wind.Speed<7.5 8    16.000 56.83 *
         83) Wind.Speed>7.5 1     0.000 39.33 *
     21) Volume>1334 217   373.000 57.37
       42) Time<167.5 136   161.200 57.65 *
       43) Time>167.5 81   183.000 56.90 *
   11) Volume>1458 309  1885.000 55.55
     22) Occupancy<16.8333 291  1091.000 55.84
       44) Occupancy<15.8333 257   456.500 56.07
         88) Time<186.5 135   214.900 56.43 *
         89) Time>186.5 122   204.100 55.66 *
       45) Occupancy>15.8333 34   517.700 54.10
         90) Volume<1510 1     0.000 37.00 *
         91) Volume>1510 33   216.500 54.62 *
     23) Occupancy>16.8333 18   389.200 50.94
       46) Volume<1556 3   151.200 43.44 *
       47) Volume>1556 15    35.480 52.44 *
3) Occupancy>18.8333 15  1174.000 34.78
  6) Time<183.5 9   151.700 28.96 *
  7) Time>183.5 6   261.700 43.50 *
```

- Tree plot

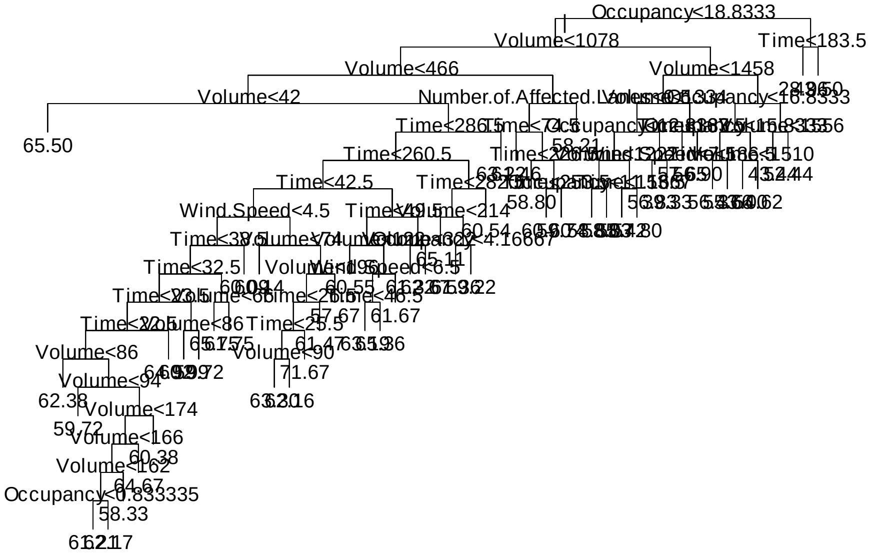


Figure I.3 Tree plot of regression tree model of characterization 5

4. Regression tree model of characterization 6

- Results summary

```
*** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.6, na.action = na.exclude, mincut = 0.5, minsize = 1, mindev
     = 0.01)
Variables actually used in tree construction:
[1] "Occupancy"  "Volume"    "Wind.Speed" "Time"
Number of terminal nodes:  51
Residual mean deviance:  5.464 = 68960 / 12620
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -1.880e+001 -1.020e+000  2.972e-002 -8.024e-015  1.123e+000  2.824e+001
node), split, n, deviance, yval
      * denotes terminal node

    1) root 12672 328500.00 58.540
      2) Occupancy<18.5 12465 143000.00 58.990
        4) Occupancy<7.5 6269  39840.00 60.990
         8) Volume<510 3985  31590.00 61.480
          16) Volume<70 263   3322.00 60.660
            32) Occupancy<0.166665 24   487.50 62.420 *
            33) Occupancy>0.166665 239   2753.00 60.490 *
          17) Volume>70 3722  28080.00 61.540
            34) Volume<430 3386  27200.00 61.600
             68) Wind.Speed<4.5 1517  11990.00 61.830
              136) Time<283.5 1452  11590.00 61.880
                272) Time<41.5 659   7605.00 61.640
                  544) Occupancy<2.33333 658   7561.00 61.650
                  1088) Volume<146 597   7167.00 61.710
                    2176) Time<13.5 139   1865.00 62.080 *
                    2177) Time>13.5 458   5278.00 61.600
                      4354) Volume<118 390   4487.00 61.480
                        8708) Volume<82 102   1914.00 61.970 *
                        8709) Volume>82 288   2541.00 61.310 *
                      4355) Volume>118 68    753.70 62.280 *
                  1089) Volume>146 61    365.80 61.000 *
                  545) Occupancy>2.33333 1     0.00 55.000 *
                273) Time>41.5 793   3916.00 62.080
                  546) Time<263.5 601   3252.00 62.210 *
                  547) Time>263.5 192    620.80 61.660 *
              137) Time>283.5 65    326.30 60.780 *
             69) Wind.Speed>4.5 1869  15060.00 61.410
              138) Time<3.5 87    609.50 60.700 *
              139) Time>3.5 1782  14410.00 61.450
                278) Time<284.5 1659  13680.00 61.470
                  556) Time<45.5 873  10120.00 61.340
                  1112) Occupancy<1.83333 865  10030.00 61.360
                    2224) Time<38.5 745   8910.00 61.430
                      4448) Time<32.5 648   7924.00 61.340
                        8896) Wind.Speed<9.5 583   7159.00 61.410
                        17792) Wind.Speed<7.5 475   5599.00 61.320
                          35584) Volume<134 395   4836.00 61.190
                            71168) Time<11.5 68    557.70 60.170 *
                            71169) Time>11.5 327   4192.00 61.400
                            142338) Time<13.5 29    592.00 63.050 *
                            142339) Time>13.5 298   3513.00 61.240
```

```
                             284678) Volume<106 228   2402.00 61.010 *
                             284679) Volume>106 70   1060.00 62.000 *
                        35585) Volume>134 80    724.70 61.950 *
                      17793) Wind.Speed>7.5 108   1540.00 61.800 *
                     8897) Wind.Speed>9.5 65    735.30 60.690 *
                   4449) Time>32.5 97    943.70 62.040 *
                 2225) Time>38.5 120   1100.00 60.960 *
               1113) Occupancy>1.83333 8     40.66 58.940 *
              557) Time>45.5 786   3525.00 61.620
               1114) Volume<346 562   2905.00 61.780 *
               1115) Volume>346 224    569.20 61.220 *
            279) Time>284.5 123    708.40 61.060 *
       35) Volume>430 336    726.80 60.890 *
    9) Volume>510 2284   5648.00 60.140
     18) Volume<922 1656   3767.00 60.440
       36) Time<72.5 344    680.80 60.980 *
       37) Time>72.5 1312   2961.00 60.300 *
     19) Volume>922 628   1324.00 59.340 *
  5) Occupancy>7.5 6196  52860.00 56.970
   10) Occupancy<14.8333 5733  22440.00 57.360
     20) Volume<1410 3400   7175.00 58.160
       40) Volume<1282 1968   3911.00 58.500
         80) Volume<1190 919   1975.00 58.770 *
         81) Volume>1190 1049   1812.00 58.260 *
       41) Volume>1282 1432   2717.00 57.690 *
     21) Volume>1410 2333   9917.00 56.190
       42) Occupancy<10.8333 780   1482.00 56.990 *
       43) Occupancy>10.8333 1553   7692.00 55.790
         86) Time<184.5 830   3236.00 56.210 *
         87) Time>184.5 723   4146.00 55.310
          174) Volume<1414 8    330.00 51.670 *
          175) Volume>1414 715   3708.00 55.350
            350) Volume<1674 664   3451.00 55.440
              700) Volume<1458 70    511.60 56.130 *
              701) Volume>1458 594   2902.00 55.360 *
            351) Volume>1674 51    182.10 54.180 *
   11) Occupancy>14.8333 463  18970.00 52.190
     22) Volume<1502 50   3460.00 43.360
       44) Time<188.5 26   1572.00 47.540 *
       45) Time>188.5 24    942.40 38.830 *
     23) Volume>1502 413  11140.00 53.260
       46) Wind.Speed<13.5 408  10200.00 53.420
         92) Occupancy<16.1667 259   3829.00 54.430
          184) Volume<1762 248   3509.00 54.560
            368) Time<195.5 159   1651.00 55.040 *
            369) Time>195.5 89   1754.00 53.700 *
          185) Volume>1762 11    220.50 51.480 *
         93) Occupancy>16.1667 149   5657.00 51.670
          186) Volume<1626 24   1775.00 43.810 *
          187) Volume>1626 125   2111.00 53.180 *
       47) Wind.Speed>13.5 5     31.24 39.870 *
3) Occupancy>18.5 207  29840.00 31.350
  6) Volume<1212 48   5517.00 16.740
   12) Volume<948 23    787.40  9.109 *
   13) Volume>948 25   2159.00 23.760 *
  7) Volume>1212 159  10980.00 35.760
   14) Volume<1620 118   5610.00 32.740
     28) Volume<1386 40   1594.00 28.540 *
     29) Volume>1386 78   2951.00 34.890 *
   15) Volume>1620 41   1202.00 44.450 *
```

- Tree plot

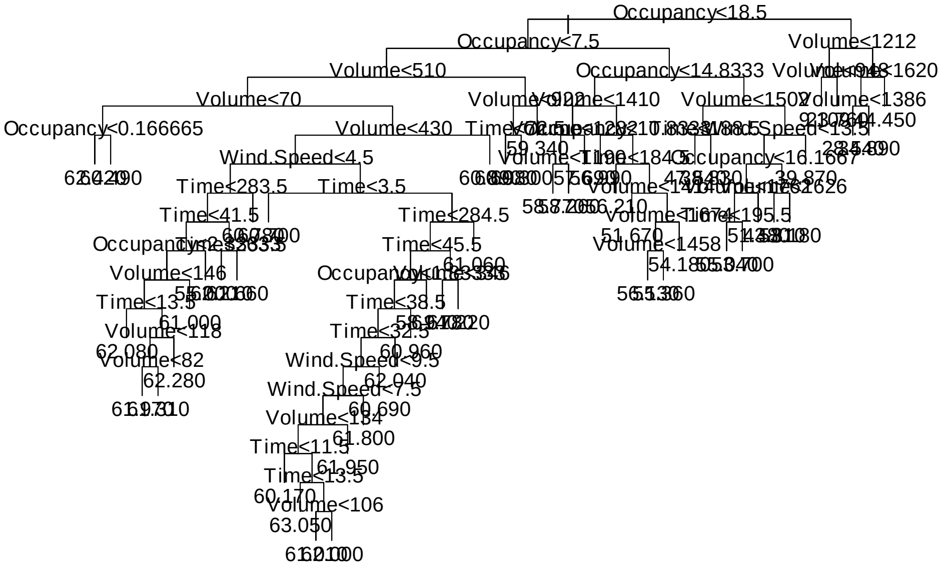


Figure I.4 Tree plot of regression tree model of characterization 6

5. Regression tree model of characterization 7

- Result summary

```
    *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.7, na.action = na.exclude, mincut = 0.5, minsize = 1, mindev
     = 0.01)
Variables actually used in tree construction:
[1] "Occupancy"    "Volume"       "Visibility"   "Time"
[5] "Incident.Type" "Wind.Speed"   "Rainfall"
Number of terminal nodes:  22
Residual mean deviance:  6.691 = 15270 / 2282
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -1.378e+001 -1.424e+000  8.726e-002 -6.396e-015  1.493e+000  1.767e+001
node), split, n, deviance, yval
     * denotes terminal node
```

```
1) root 2304 231200.00 55.8800
  2) Occupancy<18.1667 2176  28390.00 57.9700
    4) Volume<1062 1122   8062.00 60.0200
      8) Volume<462 714   5405.00 60.7000
       16) Visibility<4.5 137    743.50 59.9300 *
       17) Visibility>4.5 577   4561.00 60.8800
         34) Time<3.5 21    199.00 62.4600 *
         35) Time>3.5 556   4308.00 60.8200
           70) Incident.Type<2.5 506   3855.00 60.7200
            140) Occupancy<0.166665 11    115.80 62.4500 *
            141) Occupancy>0.166665 495   3705.00 60.6800
              282) Wind.Speed<9.5 434   3056.00 60.5800
                564) Occupancy<0.833335 212   1879.00 60.2500 *
                565) Occupancy>0.833335 222   1131.00 60.9000 *
              283) Wind.Speed>9.5 61    610.60 61.4300 *
           71) Incident.Type>2.5 50    400.90 61.8000 *
      9) Volume>462 408   1762.00 58.8400 *
    5) Volume>1062 1054  10600.00 55.7900
     10) Occupancy<13.8333 1005   5837.00 56.1300
       20) Time<183.5 730   3156.00 56.6500
         40) Volume<1522 644   2378.00 56.9100
           80) Rainfall<8.5 620   2008.00 57.0300 *
           81) Rainfall>8.5 24    127.50 53.7900 *
         41) Volume>1522 86    415.30 54.7200 *
       21) Time>183.5 275   1966.00 54.7600 *
     11) Occupancy>13.8333 49   2232.00 48.7800 *
  3) Occupancy>18.1667 128  31150.00 20.2800
    6) Volume<274 36   9561.00  4.2500
     12) Occupancy<53.1667 22     43.45  0.5455 *
     13) Occupancy>53.1667 14   8741.00 10.0700
       26) Occupancy<53.6667 1      0.00 60.0000 *
       27) Occupancy>53.6667 13   6056.00  6.2310
         54) Volume<202 9      0.00  0.0000 *
         55) Volume>202 4   4921.00 20.2500
          110) Volume<224 1      0.00 81.0000 *
          111) Volume>224 3      0.00  0.0000 *
    7) Volume>274 92   8709.00 26.5600
     14) Volume<286 1      0.00 75.0000 *
     15) Volume>286 91   6336.00 26.0300
       30) Occupancy<33.3333 84   3063.00 27.7400
         60) Volume<1384 45    983.10 23.8300 *
         61) Volume>1384 39    596.50 32.2600 *
       31) Occupancy>33.3333 7     55.71  5.4290 *
```

- Tree plot

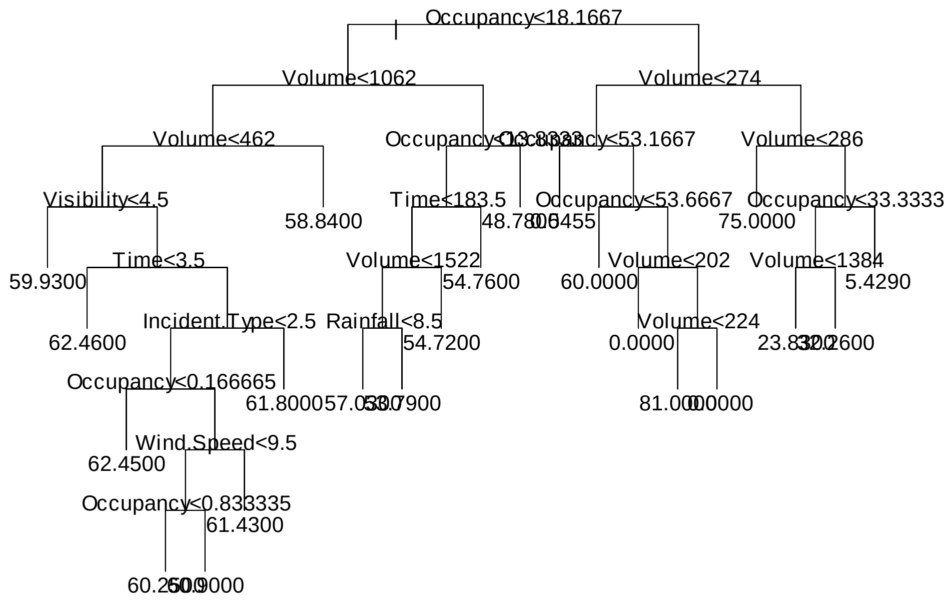


Figure I.5 Tree plot of regression tree model of characterization 7

6. Regression tree model of characterization 8

- Result summary

```
    *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.8, na.action = na.exclude, mincut = 0.5, minsize = 1, mindev
     = 0.01)
Variables actually used in tree construction:
[1] "Occupancy"  "Volume"     "Rainfall"   "Visibility" "Time"
[6] "Wind.Speed"
Number of terminal nodes:  50
Residual mean deviance:  6.252 = 113000 / 18080
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -2.033e+001 -1.226e+000  1.077e-001 -3.498e-014  1.335e+000  1.999e+001
node), split, n, deviance, yval
      * denotes terminal node

    1) root 18132 453500.0 57.92
```

```
2) Occupancy<18.1667 17826 223800.0 58.36
  4) Volume<970 8563  61070.0 60.41
    8) Volume<474 5707  46830.0 60.86
     16) Rainfall<4.5 5278  42890.0 60.96
       32) Visibility<8.5 1244   9892.0 60.52
         64) Time<43.5 700   6651.0 60.25
          128) Time<27.5 444   4633.0 60.50
            256) Volume<82 137   2136.0 61.02 *
            257) Volume>82 307   2444.0 60.27 *
          129) Time>27.5 256   1944.0 59.83 *
         65) Time>43.5 544   3127.0 60.86 *
       33) Visibility>8.5 4034  32690.0 61.10
         66) Wind.Speed<9.5 3175  26470.0 61.25
          132) Time<64.5 2168  22090.0 61.40
            264) Volume<122 1221  15850.0 61.10
              528) Rainfall<3.5 1213  15660.0 61.08
               1056) Time<40.5 1072  13560.0 61.01
                 2112) Volume<30 2     0.0 56.50 *
                 2113) Volume>30 1070  13520.0 61.02
                   4226) Volume<54 63   1381.0 61.79 *
                   4227) Volume>54 1007  12100.0 60.97
                     8454) Volume<66 122   1380.0 60.15 *
                     8455) Volume>66 885  10630.0 61.08
                      16910) Time<36.5 788   9505.0 61.18
                        33820) Time<30.5 644   7173.0 61.06
                          67640) Visibility<9.5 41    407.6 60.05 *
                          67641) Visibility>9.5 603   6721.0 61.13
                           135282) Occupancy<1.16667 585   6445.0 61.09
                             270564) Occupancy<0.5 193   2082.0 61.33 *
                             270565) Occupancy>0.5 392   4346.0 60.97 *
                           135283) Occupancy>1.16667 18    248.4 62.34 *
                        33821) Time>30.5 144   2279.0 61.73 *
                      16911) Time>36.5 97   1055.0 60.30 *
               1057) Time>40.5 141   2048.0 61.65 *
              529) Rainfall>3.5 8    121.1 64.08 *
            265) Volume>122 947   5990.0 61.78
              530) Wind.Speed<3.5 358   2085.0 61.53 *
              531) Wind.Speed>3.5 589   3868.0 61.94 *
          133) Time>64.5 1007   4220.0 60.92 *
         67) Wind.Speed>9.5 859   5883.0 60.54
          134) Time<23.5 195   1911.0 60.10 *
          135) Time>23.5 664   3922.0 60.67 *
     17) Rainfall>4.5 429   3275.0 59.67 *
    9) Volume>474 2856  10690.0 59.50
     18) Rainfall<3.5 2638   8461.0 59.64
       36) Volume<738 1735   4964.0 59.91
         72) Time<72.5 249    733.1 60.64 *
         73) Time>72.5 1486   4077.0 59.79 *
       37) Volume>738 903   3113.0 59.11 *
     19) Rainfall>3.5 218   1564.0 57.82 *
  5) Volume>970 9263  93400.0 56.46
   10) Occupancy<15.5 8856  61300.0 56.71
     20) Volume<1334 4751  26750.0 57.61
       40) Occupancy<13.5 4718  21520.0 57.67
         80) Rainfall<1.5 3988  13150.0 57.97
          160) Time<178.5 3310   8577.0 58.15
            320) Volume<1174 1195   2974.0 58.63 *
            321) Volume>1174 2115   5176.0 57.88
              642) Time<105.5 244    673.2 56.99 *
              643) Time>105.5 1871   4282.0 58.00 *
          161) Time>178.5 678   3921.0 57.07 *
         81) Rainfall>1.5 730   6123.0 56.06
          162) Rainfall<27.5 721   4694.0 56.20
            324) Time<196.5 557   2202.0 56.74 *
            325) Time>196.5 164   1800.0 54.40 *
          163) Rainfall>27.5 9    238.1 44.63 *
    41) Occupancy>13.5 33   2709.0 48.89 *
   21) Volume>1334 4105  26190.0 55.67
    42) Time<188.5 2599  13350.0 56.15
     84) Volume<1570 2076   7622.0 56.56
      168) Visibility<9.5 488   2590.0 55.50 *
      169) Visibility>9.5 1588   4305.0 56.89 *
     85) Volume>1570 523   3942.0 54.50 *
    43) Time>188.5 1506  11190.0 54.83
     86) Wind.Speed<8.5 904   6046.0 55.27
      172) Visibility<5.5 18    359.8 51.39 *
      173) Visibility>5.5 886   5410.0 55.35
       346) Occupancy<14.8333 829   4069.0 55.43 *
       347) Occupancy>14.8333 57   1245.0 54.09 *
     87) Wind.Speed>8.5 602   4711.0 54.18
      174) Rainfall<1.5 516   3469.0 54.57 *
      175) Rainfall>1.5 86    691.7 51.83 *
  11) Occupancy>15.5 407  19710.0 51.06
   22) Volume<1486 49   3606.0 40.23 *
   23) Volume>1486 358   9564.0 52.55
    46) Rainfall<5.5 349   7426.0 52.90
     92) Volume<1574 46   1949.0 49.66 *
     93) Volume>1574 303   4919.0 53.40
      186) Visibility<6.5 38   1117.0 50.80 *
      187) Visibility>6.5 265   3509.0 53.77 *
    47) Rainfall>5.5 9    368.1 38.70 *
3) Occupancy>18.1667 306  30100.0 32.60
 6) Volume<1410 138   5172.0 25.15
  12) Volume<1314 83   2102.0 22.20 *
  13) Volume>1314 55   1255.0 29.61 *
 7) Volume>1410 168  10980.0 38.72
  14) Volume<1666 109   3610.0 34.67 *
  15) Volume>1666 59   2289.0 46.19 *
```

- Tree plot

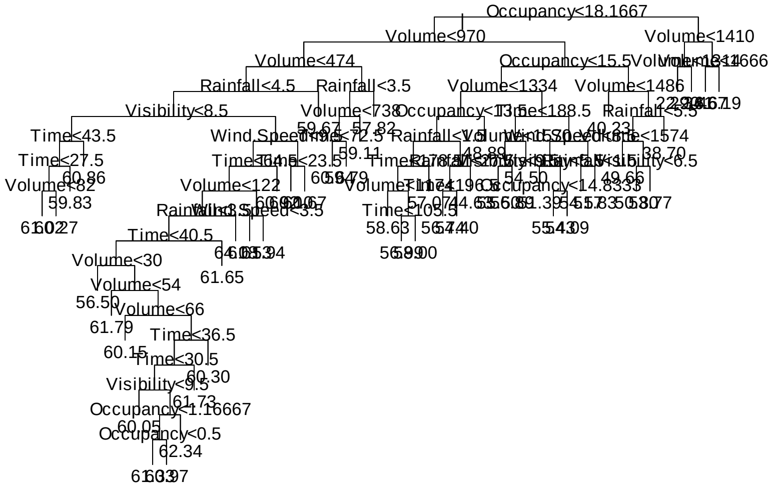


Figure I.6 Tree plot of regression tree model of characterization 8

7. Regression tree model of characterization 11

- Result summary

```
    *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.11, na.action = na.exclude, mincut = 0.5, minsize = 1,
    mindev = 0.01)
Variables actually used in tree construction:
[1] "Volume"
Number of terminal nodes:  2
Residual mean deviance:  3.484 = 8021 / 2302
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -7.796e+000 -4.631e-001  0.000e+000  2.372e-014  2.036e-001  1.454e+001
node), split, n, deviance, yval
      * denotes terminal node

1) root 2304 2085000 33.75
  2) Volume<6 1018      0  0.00 *
  3) Volume>6 1286   8021 60.46 *
```

- Tree plot

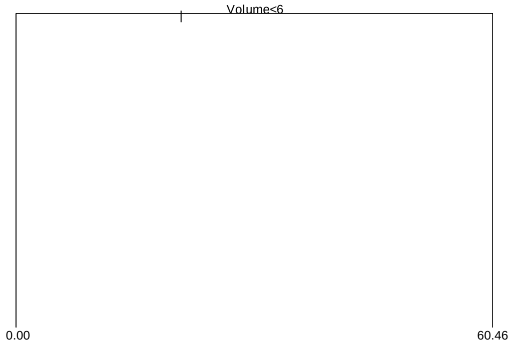


Figure I.7 Tree plot of regression tree model of characterization 11

8. Regression tree model of characterization 12

- Result summary

```
    *** Tree Model ***
```

```
Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.12, na.action = na.exclude, mincut = 0.5, minsize = 1,
    mindev = 0.01)
Variables actually used in tree construction:
[1] "Volume"
Number of terminal nodes:  2
Residual mean deviance:  9.505 = 19030 / 2002
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -3.677e+001 -7.724e-001 -9.412e-002  3.299e-014  5.609e-001  6.024e+001
node), split, n, deviance, yval
     * denotes terminal node

1) root 2004 1624000 41.36000
  2) Volume<2 641   3634  0.09412 *
  3) Volume>2 1363   15390 60.77000 *
```

- Tree plot

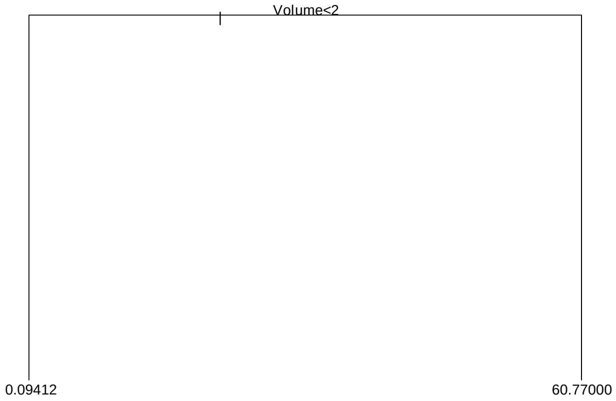


Figure I.8 Tree plot of regression tree model of characterization 12

9. Regression tree model of characterization 14

- Result summary

```
    *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type +
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat +
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data =
    tree.char.14, na.action = na.exclude, mincut = 0.5, minsize = 1,
    mindev = 0.01)
Variables actually used in tree construction:
[1] "Volume"     "Time"       "Wind.Speed" "Occupancy"  "Rainfall"  
[6] "Visibility"
Number of terminal nodes:  134 
Residual mean deviance:  3.711 = 33700 / 9082 
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
 -1.159e+001 -1.084e+000 -6.061e-003  1.533e-015  1.074e+000  1.251e+001 
node), split, n, deviance, yval
     * denotes terminal node

   1) root 9216 56990.000 60.68  
     2) Volume<874 5471 35200.000 61.65  
       4) Time<105.5 3320 26820.000 62.14  
        8) Time<46.5 1472 14120.000 61.40  
         16) Wind.Speed<9.5 1330 12270.000 61.56  
           32) Volume<74 109  1323.000 60.97  
             64) Occupancy<0.5 90   981.500 61.33  
             128) Wind.Speed<3.5 22   210.900 59.90 *
             129) Wind.Speed>3.5 68   711.100 61.79  
               258) Wind.Speed<6.5 44   536.200 62.54 *
               259) Wind.Speed>6.5 24   105.900 60.43 *
             65) Occupancy>0.5 19   275.400 59.27 *
           33) Volume>74 1221 10910.000 61.61  
             66) Volume<114 483  5951.000 61.87  
             132) Occupancy<0.166665 11   136.700 63.77 *
             133) Occupancy>0.166665 472  5774.000 61.83  
               266) Wind.Speed<5.5 257  2872.000 61.58  
                 532) Time<43.5 226  2592.000 61.47  
                  1064) Volume<106 187  2292.000 61.62  
                   2128) Time<35.5 103  1221.000 61.98  
                     4256) Time<14.5 3    22.300 64.56 *
                     4257) Time>14.5 100  1178.000 61.90  
                       8514) Time<19.5 8   101.500 60.15 *
                       8515) Time>19.5 92  1050.000 62.05  
                        17030) Occupancy<0.5 48   493.300 61.65 *
                        17031) Occupancy>0.5 44   540.500 62.49 *
                   2129) Time>35.5 84  1042.000 61.19  
                     4258) Volume<102 78   996.400 61.05  
                       8516) Time<40.5 47   349.300 60.60 *
                       8517) Time>40.5 31   623.000 61.74  
                        17034) Volume<78 4    45.670 57.50 *
                        17035) Volume>78 27   494.900 62.36 *
                     4259) Volume>102 6    25.650 62.94 *
                  1065) Volume>106 39   275.400 60.75 *
                 533) Time>43.5 31   255.500 62.41 *
               267) Wind.Speed>5.5 215  2868.000 62.12  
                 534) Rainfall<0.5 211  2763.000 62.08  
                  1068) Time<11.5 4    42.080 59.92 *
                  1069) Time>11.5 207  2702.000 62.12  
                   2138) Occupancy<0.833335 180  2509.000 62.23  
                     4276) Time<37.5 126  1941.000 62.52  
                       8552) Wind.Speed<8.5 125  1800.000 62.42  
                        17104) Time<12.5 1     0.000 52.33 *
                        17105) Time>12.5 124  1697.000 62.50  
                         34210) Time<28.5 68   836.700 62.05  
                           68420) Wind.Speed<7.5 54   569.100 61.59 *
                           68421) Wind.Speed>7.5 14   211.700 63.83 *
                         34211) Time>28.5 56   830.300 63.05  
                           68422) Time<30.5 10   219.300 65.83 *
                           68423) Time>30.5 46   516.500 62.44 *
                       8553) Wind.Speed>8.5 1     0.000 74.33 *
                     4277) Time>37.5 54   534.600 61.57 *
                   2139) Occupancy>0.833335 27   174.300 61.35 *
                 535) Rainfall>0.5 4    78.890 64.67 *
             67) Volume>114 738  4902.000 61.43
             134) Wind.Speed<3.5 181  1199.000 61.81
               268) Occupancy<0.5 35   349.000 62.97 *
               269) Occupancy>0.5 146   791.600 61.53
                 538) Time<29.5 117   554.000 61.68 *
                 539) Time>29.5 29   224.600 60.93 *
             135) Wind.Speed>3.5 557  3669.000 61.31
               270) Volume<190 402  2874.000 61.12
                 540) Volume<158 298  2397.000 61.23
                  1080) Volume<126 109  1052.000 60.89
                   2160) Volume<122 65   572.400 61.39
                     4320) Volume<118 29   160.700 60.55 *
                     4321) Volume>118 36   374.600 62.07 *
                   2161) Volume>122 44   438.800 60.14 *
                  1081) Volume>126 189  1324.000 61.43
                   2162) Volume<138 84   597.400 61.74
                     4324) Time<16.5 29   228.600 60.99 *
                     4325) Time>16.5 55   343.900 62.14 *
                   2163) Volume>138 105   712.100 61.18
                     4326) Time<21.5 70   464.400 61.53 *
                     4327) Time>21.5 35   221.900 60.48 *
                 541) Volume>158 104   463.700 60.80 *
               271) Volume>190 155   743.000 61.80
                 542) Occupancy<1.83333 137   691.000 61.94
                  1084) Time<18 136   668.500 61.91
                   2168) Time<7.5 93   432.100 62.15 *
                   2169) Time>7.5 43   219.800 61.40 *
                  1085) Time>18 1    0.000 66.67 *
                 543) Occupancy>1.83333 18    27.530 60.70 *
          17) Wind.Speed>9.5 142  1520.000 59.96
           34) Volume<210 118  1239.000 60.40
             68) Wind.Speed<12.5 94   714.900 60.03
             136) Volume<70 6    60.150 61.78 *
             137) Volume>70 88   635.200 59.91
               274) Volume<110 30   122.600 59.23 *
               275) Volume>110 58   491.600 60.26 *
             69) Wind.Speed>12.5 24   462.600 61.83 *
           35) Volume>210 24   151.800 57.85 *
        9) Time>46.5 1848 11250.000 62.73
          18) Volume<450 1365  9450.000 63.01
           36) Time<63.5 544  5109.000 62.47
             72) Occupancy<0.166665 7   176.200 64.64 *
             73) Occupancy>0.166665 537  4900.000 62.44
             146) Time<53.5 220  1901.000 62.13
               292) Occupancy<1.16667 211  1845.000 62.22
                 584) Volume<70 19   173.000 62.97 *
                 585) Volume>70 192  1660.000 62.14
                  1170) Volume<82 22   136.100 60.98 *
                  1171) Volume>82 170  1490.000 62.29
                   2342) Volume<122 115  1166.000 62.52
                     4684) Time<50.5 67   605.400 62.18
                       9368) Volume<118 58   444.000 61.93 *
                       9369) Volume>118 9   133.600 63.81 *
                     4685) Time>50.5 48   542.200 62.99 *
                   2343) Volume>122 55   306.400 61.82 *
               293) Occupancy>1.16667 9    19.800 60.15 *
             147) Time>53.5 317  2964.000 62.65
               294) Wind.Speed<3.5 108   815.100 62.19
                 588) Time<62.5 95   719.500 62.38
                  1176) Volume<186 84   658.000 62.59
                   2352) Occupancy<1.16667 75   567.500 62.79 *
                   2353) Occupancy>1.16667 9    61.330 60.89 *
                  1177) Volume>186 11    28.910 60.76 *
                 589) Time>62.5 13    68.800 60.85 *
               295) Wind.Speed>3.5 209  2114.000 62.89
                 590) Time<60.5 153  1575.000 63.08
```

```
             1180) Wind.Speed<13.5 146  1520.000 63.16
               2360) Occupancy<0.5 28   332.000 64.07 *
               2361) Occupancy>0.5 118  1159.000 62.95
                 4722) Volume<110 15   251.100 64.49 *
                 4723) Volume>110 103   867.400 62.72
                   9446) Volume<118 5    13.240 59.80 *
                   9447) Volume>118 98   809.200 62.87
                    18894) Time<58.5 68   642.200 62.56
                      37788) Volume<198 58   601.100 62.82
                        75576) Occupancy<0.833335 23   156.700 61.91 *
                        75577) Occupancy>0.833335 35   413.100 63.41 *
                      37789) Volume>198 10    13.880 61.03 *
                    18895) Time>58.5 30   144.800 63.59 *
             1181) Wind.Speed>13.5 7    31.650 61.29 *
            591) Time>60.5 56   518.300 62.36 *
     37) Time>63.5 821  4075.000 63.37
       74) Wind.Speed<9.5 726  3532.000 63.55
        148) Wind.Speed<1.5 229  1166.000 64.01
          296) Volume<234 70   581.000 64.69
            592) Volume<222 63   474.800 64.38 *
            593) Volume>222 7    43.750 67.52 *
          297) Volume>234 159   537.900 63.71 *
        149) Wind.Speed>1.5 497  2294.000 63.33
          298) Volume<370 388  1945.000 63.48
            596) Visibility<9 9    19.360 61.81 *
            597) Visibility>9 379  1900.000 63.52
             1194) Time<91.5 363  1794.000 63.58
               2388) Time<64.5 21   145.900 62.58 *
               2389) Time>64.5 342  1626.000 63.64
                 4778) Wind.Speed<6.5 282  1264.000 63.75
                   9556) Volume<298 181   857.300 63.89
                    19112) Volume<124 5     8.311 62.40 *
                    19113) Volume>124 176   837.500 63.94
                      38226) Occupancy<0.5 15   102.900 64.62 *
                      38227) Occupancy>0.5 161   726.900 63.87
                        76454) Time<71.5 78   354.000 64.10 *
                        76455) Time>71.5 83   365.100 63.66 *
                   9557) Volume>298 101   395.900 63.49 *
                 4779) Wind.Speed>6.5 60   343.100 63.14 *
             1195) Time>91.5 16    79.830 62.27 *
          299) Volume>370 109   309.100 62.79 *
       75) Wind.Speed>9.5 95   344.800 62.01 *
   19) Volume>450 483  1402.000 61.96
     38) Wind.Speed<12.5 477  1288.000 61.99
       76) Volume<742 410  1089.000 62.11
        152) Wind.Speed<11.5 388  1034.000 62.16
          304) Volume<466 32   115.900 61.63 *
          305) Volume>466 356   908.100 62.20
            610) Volume<542 133   375.900 62.44 *
            611) Volume>542 223   520.500 62.06 *
        153) Wind.Speed>11.5 22    40.200 61.30 *
       77) Volume>742 67   155.900 61.24 *
     39) Wind.Speed>12.5 6    71.780 59.33 *
5) Time>105.5 2151  6334.000 60.89
 10) Volume<398 535  2046.000 61.57
   20) Volume<214 103   614.900 62.17
     40) Visibility<9.5 20    72.420 61.03 *
     41) Visibility>9.5 83   510.500 62.44 *
   21) Volume>214 432  1386.000 61.43
     42) Wind.Speed<1.5 76   278.200 61.95 *
     43) Wind.Speed>1.5 356  1082.000 61.32
       86) Time<265.5 49    98.910 61.84 *
       87) Time>265.5 307   967.500 61.23
        174) Volume<218 7     5.968 60.14 *
        175) Volume>218 300   953.000 61.26
          350) Volume<226 15    29.080 62.16 *
```

```
            351) Volume>226 285   911.200 61.21  
              702) Visibility<9.5 12   30.780 60.17 *
              703) Visibility>9.5 273   866.800 61.26  
               1406) Wind.Speed<11 264   854.900 61.29  
                 2812) Wind.Speed<7.5 177   519.800 61.15 *
                 2813) Wind.Speed>7.5 87   325.200 61.56 *
               1407) Wind.Speed>11 9    4.000 60.33 *
   11) Volume>398 1616  3959.000 60.67  
     22) Time<135.5 272   817.100 61.44  
       44) Occupancy<9.33333 270   752.500 61.48  
         88) Volume<770 160   430.300 61.80 *
         89) Volume>770 110   281.100 61.01 *
       45) Occupancy>9.33333 2    1.389 55.83 *
     23) Time>135.5 1344  2948.000 60.51  
       46) Occupancy<6.83333 1318  2876.000 60.53  
         92) Volume<490 214   440.600 60.86 *
         93) Volume>490 1104  2408.000 60.47  
          186) Time<249.5 593  1314.000 60.63  
            372) Occupancy<3.16667 36   87.740 59.94 *
            373) Occupancy>3.16667 557  1208.000 60.68  
              746) Time<241.5 358   735.600 60.54  
               1492) Time<219.5 24   60.590 59.94 *
               1493) Time>219.5 334   665.900 60.58  
                 2986) Wind.Speed<6.5 130   230.600 60.37 *
                 2987) Wind.Speed>6.5 204   426.000 60.72 *
              747) Time>241.5 199   453.500 60.92 *
          187) Time>249.5 511  1061.000 60.29  
            374) Volume<582 208   503.200 60.56 *
            375) Volume>582 303   530.700 60.10 *
       47) Occupancy>6.83333 26   30.740 59.26 *
3) Volume>874 3745  9058.000 59.26  
  6) Volume<1246 2313  5267.000 59.68  
   12) Volume<1090 1011  2621.000 60.08  
     24) Occupancy<8.5 914  2371.000 60.15  
       48) Time<207.5 540  1158.000 60.33  
         96) Volume<986 254   599.600 60.59  
          192) Time<119.5 123   309.100 60.30 *
          193) Time>119.5 131   270.700 60.86 *
         97) Volume>986 286   527.500 60.10 *
       49) Time>207.5 374  1172.000 59.90  
         98) Occupancy<6.16667 180   571.500 60.14  
          196) Time<248 175   550.100 60.19 *
          197) Time>248 5    3.422 58.27 *
         99) Occupancy>6.16667 194   581.200 59.68  
          198) Wind.Speed<10.5 170   387.000 59.86 *
          199) Wind.Speed>10.5 24  149.300 58.40 *
     25) Occupancy>8.5 97   191.800 59.34 *
   13) Volume>1090 1302  2364.000 59.37  
     26) Occupancy<10.1667 1222  2121.000 59.42  
       52) Time<178.5 646   974.200 59.56  
        104) Volume<1146 220   329.800 59.82 *
        105) Volume>1146 426   622.300 59.43  
          210) Occupancy<8.5 260   416.000 59.31 *
          211) Occupancy>8.5 166   197.600 59.61 *
       53) Time>178.5 576  1118.000 59.25  
        106) Wind.Speed<9.5 409   718.000 59.36  
          212) Occupancy<9.16667 297   520.600 59.47 *
          213) Occupancy>9.16667 112   184.900 59.08 *
        107) Wind.Speed>9.5 167   382.600 58.98 *
     27) Occupancy>10.1667 80   208.100 58.73 *
  7) Volume>1246 1432  2708.000 58.57  
   14) Occupancy<11.8333 1288  2087.000 58.68  
     28) Volume<1314 604   956.000 58.96  
       56) Wind.Speed<8.5 408   701.400 59.09  
        112) Wind.Speed<1.5 35   45.020 59.57 *
        113) Wind.Speed>1.5 373   647.500 59.04  
```

```
          226) Occupancy<11.5 369   636.600 59.06
            452) Time<143.5 66   116.700 58.75 *
            453) Time>143.5 303   512.200 59.13 *
          227) Occupancy>11.5 4    1.222 57.50 *
       57) Wind.Speed>8.5 196   235.500 58.71 *
     29) Volume>1314 684  1042.000 58.44
       58) Volume<1462 633   916.200 58.49
        116) Wind.Speed<10.5 561   808.100 58.53
          232) Volume<1402 447   641.600 58.59
            464) Occupancy<9.5 248   362.000 58.50 *
            465) Occupancy>9.5 199   275.200 58.70 *
          233) Volume>1402 114   160.200 58.32 *
        117) Wind.Speed>10.5 72    98.870 58.15 *
       59) Volume>1462 51   103.400 57.80 *
   15) Occupancy>11.8333 144   466.800 57.59 *
```

- Tree plot

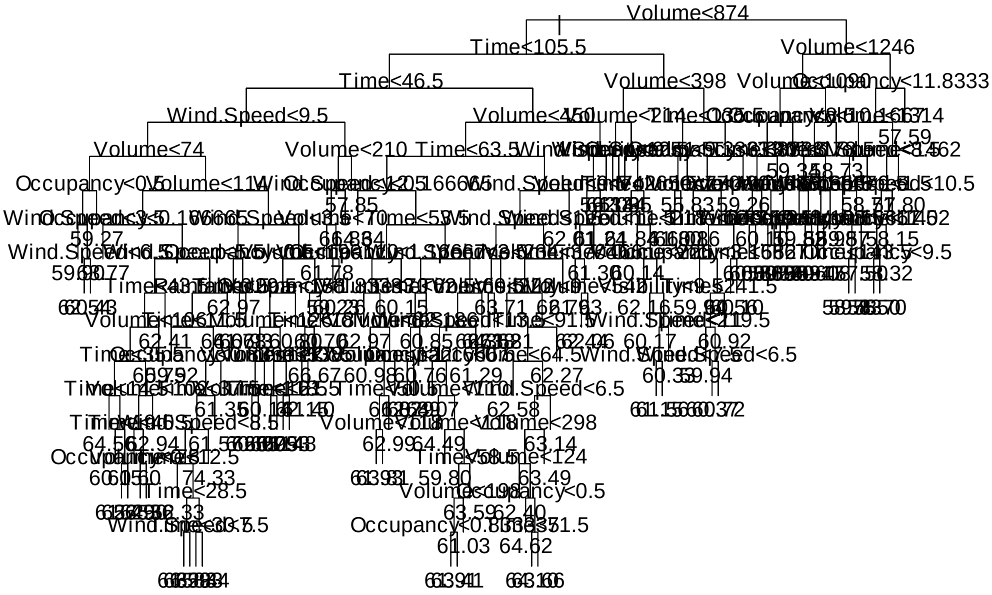


Figure I.9 Tree plot of regression tree model of characterization 14

10. Regression tree model of characterization 16

- Result summary

```
    *** Tree Model ***

Regression tree:
tree(formula = Speed ~ Time + Volume + Occupancy + Incident.Type + 
    Affected.Lanes + Number.of.Affected.Lanes + Hazmat + 
    Number.of.Fatalities + Wind.Speed + Visibility + Rainfall, data = 
    tree.char.16, na.action = na.exclude, mincut = 0.5, minsize = 1, 
    mindev = 0.01)
Variables actually used in tree construction:
[1] "Occupancy"  "Time"       "Volume"     "Wind.Speed" "Rainfall"  
[6] "Visibility"
Number of terminal nodes:  81 
Residual mean deviance:  5.227 = 56780 / 10860 
Distribution of residuals:
       Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
 -3.367e+001 -1.185e+000 -2.260e-002 -8.440e-018  1.193e+000  2.182e+001 
node), split, n, deviance, yval
     * denotes terminal node

    1) root 10944 172200.000 59.86  
      2) Occupancy<15.1667 10927 132100.000 59.93  
        4) Time<115.5 4370  44750.000 61.29  
          8) Volume<38 43   4308.000 53.06  
           16) Volume<14 3   1701.000 33.67 *
           17) Volume>14 40   1394.000 54.51 *
          9) Volume>38 4327  37510.000 61.37  
           18) Volume<786 4065  36140.000 61.48  
             36) Volume<50 53   1503.000 58.35 *
             37) Volume>50 4012  34110.000 61.52  
               74) Wind.Speed<18.5 3791  32510.000 61.61  
                148) Time<28.5 977   8253.000 61.11  
                  296) Volume<62 7    169.300 55.50 *
                  297) Volume>62 970   7863.000 61.15  
                    594) Rainfall<1.5 871   7157.000 61.29  
                     1188) Volume<66 4     28.560 65.50 *
                     1189) Volume>66 867   7057.000 61.27  
                       2378) Wind.Speed<10.5 769   6543.000 61.35  
                         4756) Wind.Speed<9.5 753   6257.000 61.32  
                           9512) Visibility<5.5 52    582.400 62.08 *
                           9513) Visibility>5.5 701   5642.000 61.26  
                            19026) Wind.Speed<4.5 470   3677.000 61.12  
                              38052) Rainfall<0.5 458   3584.000 61.07  
                                76104) Time<11.5 220   1526.000 61.38 *
                                76105) Time>11.5 238   2019.000 60.79  
                                 152210) Volume<174 222   1950.000 60.72  
                                   304420) Time<12.5 14     75.150 59.25 *
                                   304421) Time>12.5 208   1842.000 60.82  
                                     608842) Time<13.5 12    132.800 62.19 *
                                     608843) Time>13.5 196   1685.000 60.73 *
                                 152211) Volume>174 16     49.770 61.85 *
                              38053) Rainfall>0.5 12     47.340 63.04 *
                            19027) Wind.Speed>4.5 231   1939.000 61.54  
                              38054) Volume<102 30    365.200 62.79 *
                              38055) Volume>102 201   1520.000 61.35 *
                         4757) Wind.Speed>9.5 16    239.100 63.04 *
                       2379) Wind.Speed>10.5 98    471.700 60.65 *
                    595) Rainfall>1.5 99    521.000 59.85 *
                149) Time>28.5 2814  23920.000 61.78  
                  298) Volume<514 2325  22120.000 61.91  
                    596) Volume<142 1064  14300.000 61.47  
                     1192) Time<71.5 1037  13100.000 61.59  
                       2384) Volume<82 267   4001.000 61.03  
                         4768) Occupancy<0.5 180   2727.000 61.56  
                           9536) Volume<58 32    783.300 63.13 *
```

```
                   9537) Volume>58 148   1847.000 61.22
                    19074) Visibility<7.5 46    311.400 60.05 *
                    19075) Visibility>7.5 102   1446.000 61.74 *
                 4769) Occupancy>0.5 87   1121.000 59.94 *
               2385) Volume>82 770   8986.000 61.78
                 4770) Volume<94 191   3175.000 62.45
                   9540) Occupancy<1.16667 189   2928.000 62.57
                    19080) Wind.Speed<1.5 37    552.100 63.89 *
                    19081) Wind.Speed>1.5 152   2296.000 62.25
                      38162) Time<29.5 5    103.900 66.60 *
                      38163) Time>29.5 147   2094.000 62.10
                        76326) Time<64.5 146   2011.000 62.16
                         152652) Occupancy<0.833335 140   1859.000 62.05
                           305304) Time<30.5 3     4.963 57.22 *
                           305305) Time>30.5 137   1783.000 62.15
                             610610) Time<54.5 129   1586.000 62.03 *
                             610611) Time>54.5 8    163.000 64.15 *
                         152653) Occupancy>0.833335 6    106.600 64.83 *
                        76327) Time>64.5 1     0.000 53.00 *
                   9541) Occupancy>1.16667 2     4.500 51.50 *
                 4771) Volume>94 579   5697.000 61.56
                   9542) Visibility<0.625 18    110.700 58.89 *
                   9543) Visibility>0.625 561   5454.000 61.65
                    19086) Time<53.5 376   3465.000 61.40
                      38172) Wind.Speed<12.5 345   2892.000 61.27
                        76344) Volume<98 38    204.400 60.30 *
                        76345) Volume>98 307   2647.000 61.39
                         152690) Time<48.5 232   2226.000 61.59
                           305380) Volume<122 177   1700.000 61.34 *
                           305381) Volume>122 55    479.100 62.40 *
                         152691) Time>48.5 75    383.300 60.77 *
                      38173) Wind.Speed>12.5 31    504.200 62.82 *
                    19087) Time>53.5 185   1917.000 62.16
                      38174) Time<65.5 163   1553.000 62.41 *
                      38175) Time>65.5 22    278.900 60.31 *
             1193) Time>71.5 27    598.200 56.81 *
            597) Volume>142 1261   7451.000 62.27
             1194) Time<102.5 1219   6952.000 62.34
               2388) Time<82.5 783   4565.000 62.17
                 4776) Wind.Speed<4.5 385   1862.000 61.91
                   9552) Volume<250 245   1329.000 62.19 *
                   9553) Volume>250 140    480.100 61.42 *
                 4777) Wind.Speed>4.5 398   2650.000 62.43
                   9554) Wind.Speed<6.5 130    814.400 63.01 *
                   9555) Wind.Speed>6.5 268   1771.000 62.15
                    19110) Time<47.5 4    12.970 59.25 *
                    19111) Time>47.5 264   1723.000 62.19
                      38222) Time<75.5 179   1039.000 62.41 *
                      38223) Time>75.5 85    659.000 61.75 *
               2389) Time>82.5 436   2328.000 62.63
                 4778) Volume<194 21    224.100 58.78 *
                 4779) Volume>194 415   1776.000 62.83
                   9558) Volume<290 84    379.400 63.68 *
                   9559) Volume>290 331   1321.000 62.61 *
             1195) Time>102.5 42    354.100 60.44 *
          299) Volume>514 489   1592.000 61.20 *
       75) Wind.Speed>18.5 221   1054.000 59.98 *
   19) Volume>786 262    591.200 59.70 *
5) Time>115.5 6557  73980.000 59.03
 10) Wind.Speed<17.5 6192  24190.000 59.37
   20) Volume<1022 3082  11760.000 60.13
     40) Volume<466 1001   5303.000 60.76
       80) Volume<68 2     1.125 49.25 *
       81) Volume>68 999   5036.000 60.79
        162) Wind.Speed<13.5 946   3951.000 60.87
          324) Volume<238 261   1384.000 61.51 *
```

```
            325) Volume>238 685  2423.000 60.63
              650) Visibility<8.5 179   634.300 60.01 *
              651) Visibility>8.5 506  1695.000 60.85 *
          163) Wind.Speed>13.5 53   948.200 59.22 *
       41) Volume>466 2081  5872.000 59.83
         82) Time<199.5 504  1075.000 60.48 *
         83) Time>199.5 1577  4519.000 59.62
          166) Volume<774 1058  2591.000 59.86
            332) Occupancy<3.83333 529  1251.000 59.65 *
            333) Occupancy>3.83333 529  1291.000 60.07 *
          167) Volume>774 519  1747.000 59.14
            334) Occupancy<6.16667 289  1011.000 58.81 *
            335) Occupancy>6.16667 230   664.300 59.56 *
     21) Volume>1022 3110  8857.000 58.61
       42) Volume<1258 1955  5311.000 58.92
         84) Occupancy<8.16667 1015  3173.000 58.64
          168) Wind.Speed<9.5 769  2101.000 58.90
            336) Rainfall<0.5 693  1748.000 59.02
              672) Volume<1118 253   693.600 59.40 *
              673) Volume>1118 440   998.100 58.80 *
            337) Rainfall>0.5 76   255.400 57.83 *
          169) Wind.Speed>9.5 246   843.400 57.80 *
         85) Occupancy>8.16667 940  1965.000 59.23
          170) Occupancy<9.83333 669  1330.000 59.44 *
          171) Occupancy>9.83333 271   534.400 58.71 *
       43) Volume>1258 1155  3047.000 58.09
         86) Occupancy<14.8333 1154  2766.000 58.11
          172) Wind.Speed<8.5 755  1609.000 58.40 *
          173) Wind.Speed>8.5 399   971.400 57.55 *
         87) Occupancy>14.8333 1     0.000 41.33 *
   11) Wind.Speed>17.5 365 36780.000 53.22
     22) Visibility<1.5 31  1777.000 36.29
       44) Volume<828 22   563.700 32.92 *
       45) Volume>828 9   354.900 44.52 *
     23) Visibility>1.5 334 25280.000 54.80
       46) Volume<316 59 10690.000 46.13
         92) Time<230.5 9  1438.000 22.30 *
         93) Time>230.5 50  3219.000 50.42
          186) Time<237.5 7  2056.000 41.74
            372) Time<236.5 6    23.800 48.69 *
            373) Time>236.5 1     0.000  0.00 *
          187) Time>237.5 43   548.700 51.84 *
       47) Volume>316 275  9215.000 56.66
         94) Occupancy<12.1667 272  8220.000 56.85
          188) Volume<588 38  6654.000 53.14
            376) Occupancy<3.66667 31   121.700 59.34 *
            377) Occupancy>3.66667 7    46.360 25.64 *
          189) Volume>588 234   958.600 57.45 *
         95) Occupancy>12.1667 3    84.220 39.33 *
3) Occupancy>15.1667 17  1595.000 12.30 *
```

- Tree plot

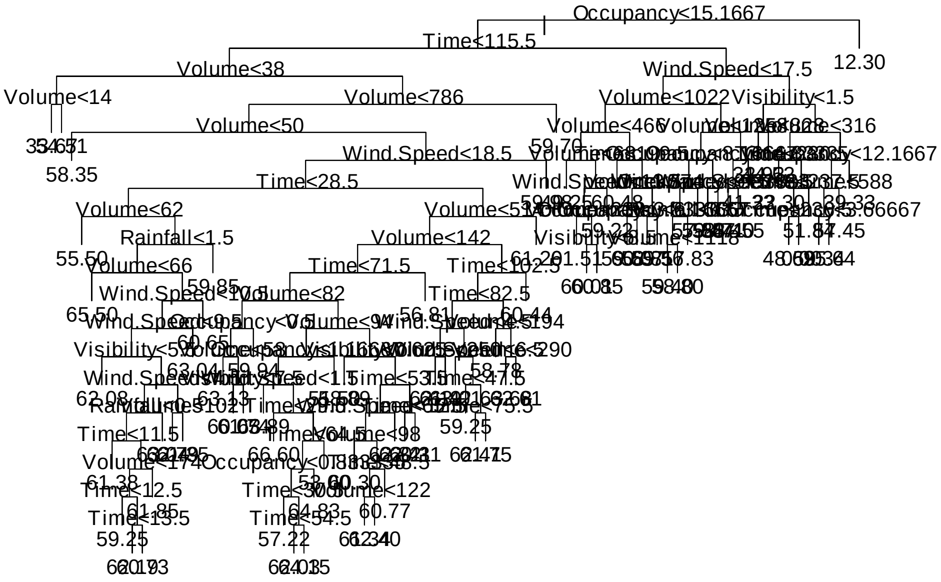
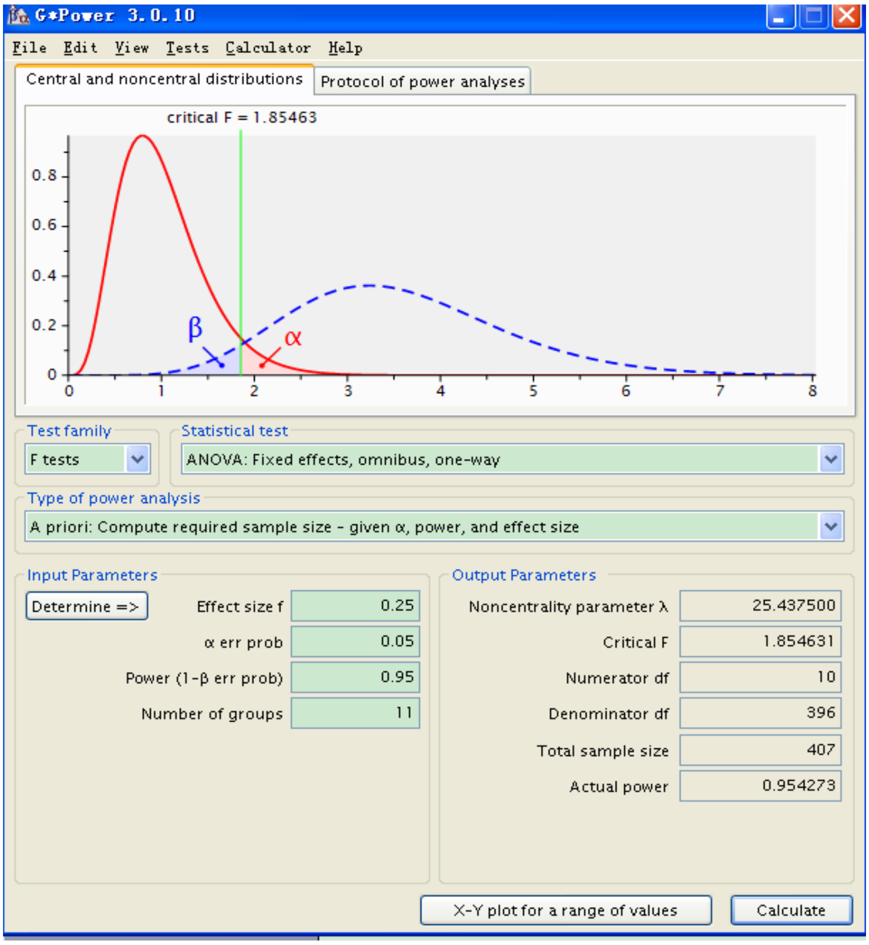


Figure I.10 Tree plot of regression tree model of characterization 16

APPENDIX J. THE USE OF G*POWER TO DETERMINE SAMPLE SIZES

After downloading the installing files of G*Power from the website http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/ and installing them on the computer, by clicking Start>Programs>G*Power, the software G*Power is opened, as shown in Figure J.1. For both of the experimental designs used in this research, one sample t-test and RCBD, G*Power is used to determine the sample sizes for them. By using RCBD as an example, the use of G*Power to determine sample sizes is demonstrated.



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Effect size f: 0.25

α err prob: 0.05

Power (1-β err prob): 0.95

Number of groups: 11

Output Parameters

Noncentrality parameter λ: 25.437500

Critical F: 1.854631

Numerator df: 10

Denominator df: 396

Total sample size: 407

Actual power: 0.954273

Figure J.1 G*Power software

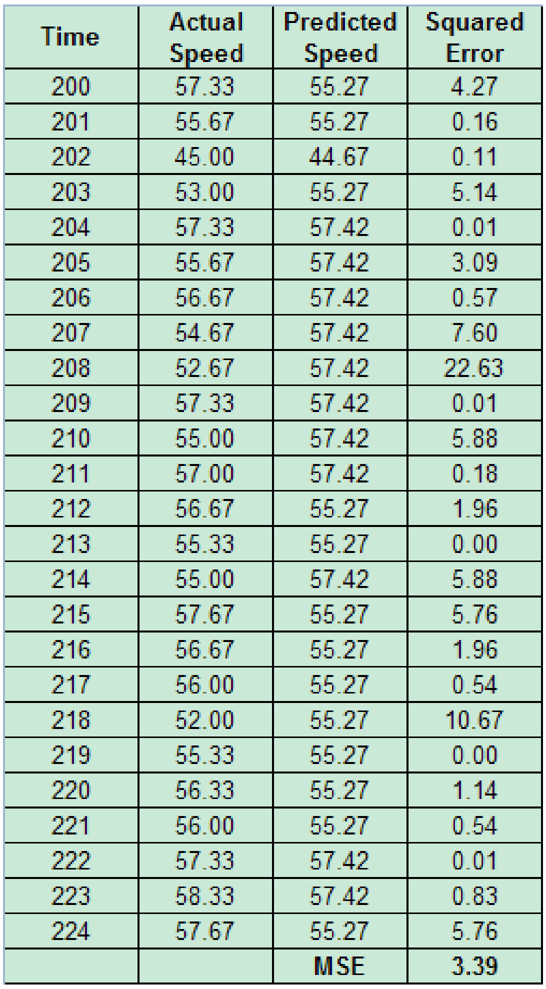
To use G*Power to determine sample size, “Test family”, “Statistical test”, “Type of power analysis” and “Input parameters” have to be appropriately selected first. Since RCBD is analyzed using ANOVA in the statistical software package S-PLUS, “F tests” and “ANOVA: Fixed effects, omnibus, one-way” need to be selected in “Test family” and “Statistical test” in G*Power. For “Type of power analysis,” “A priori: Compute required sample size – given α, power, and effect size” needs to be selected, since the priori analysis is the analysis used to decide the sample size. In the input parameters section, the effect size of 0.25 is used, which is the medium effect size for F-test according to Cohen’s (1988) conventions of effect size measures. And the power of test 95% and the number of groups 11 are used. The number of groups here refers to the number of levels of the single factor in ANOVA. The single factor in RCBD is regression tree models, which include 11 regression tree models considered in this experimental design, ten characterization regression tree models and the full regression tree model. Thus, 11 needs to be typed in for “Number of groups”.

Then, by clicking Calculate, the results for the sample size are shown in the output parameters section. As shown in Figure J.1, the total sample size of 407 is needed to guarantee the effect size of 0.25 and the power of test of 95% for 11 levels of the factor in the RCBD. Therefore, for each level of factor, at least 407÷11 = 37 blocks are needed.

APPENDIX K. CALCULATION OF MSE VALUES USED IN RCBD

Since validation of regression tree models in S-PLUS has been demonstrated in Appendix G, here the focus is to explain how to calculate the MSE values used in randomized complete block design (RCBD). RCBD is used to compare the prediction abilities of speed/travel time of a characterization regression tree model vs. each of the other characterization regression tree models and each of the ten characterization regression tree models vs. full regression tree model. The response variable in each RCBD is MSE values from validation of regression tree models by using validation data sets. The Mean Squared Error (MSE) is used to estimate the accuracy of the predicted speeds by the regression tree model compared with the actual speeds of the validation data set.

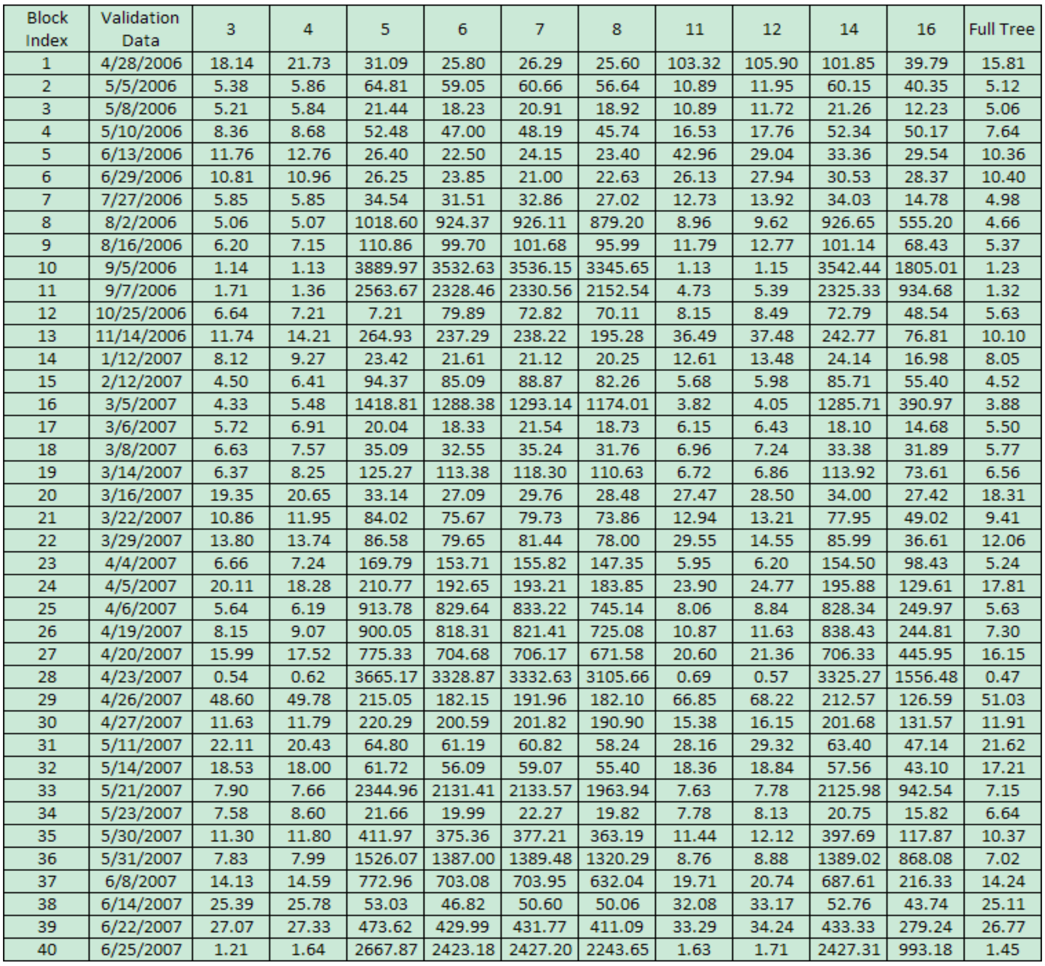
Each of the validation data sets, serving as one block in RCBD, is used to validate 11 different regression tree models, including the ten characterization regression tree models and the one full regression tree model. Thus, 11 MSE values will be calculated for each block (each of the validation data sets). By using the predicted speed from the validation implemented in S-PLUS for the validation data set, MSE is calculated as shown in Figure K.1 from time index 200 to 224, which refers to the time period between 4:30 and 6:30 pm (time index 1-288 is used for 24 hours in 5-minute increments). Different from MSE based on only two hours of data shown in Figure K.1, the MSE values used in RCBD are calculated based on squared errors between the predicted speeds and the actual speeds in 24 hours for each daily validation data set.



Time	Actual Speed	Predicted Speed	Squared Error
200	57.33	55.27	4.27
201	55.67	55.27	0.16
202	45.00	44.67	0.11
203	53.00	55.27	5.14
204	57.33	57.42	0.01
205	55.67	57.42	3.09
206	56.67	57.42	0.57
207	54.67	57.42	7.60
208	52.67	57.42	22.63
209	57.33	57.42	0.01
210	55.00	57.42	5.88
211	57.00	57.42	0.18
212	56.67	55.27	1.96
213	55.33	55.27	0.00
214	55.00	57.42	5.88
215	57.67	55.27	5.76
216	56.67	55.27	1.96
217	56.00	55.27	0.54
218	52.00	55.27	10.67
219	55.33	55.27	0.00
220	56.33	55.27	1.14
221	56.00	55.27	0.54
222	57.33	57.42	0.01
223	58.33	57.42	0.83
224	57.67	55.27	5.76
		MSE	**3.39**

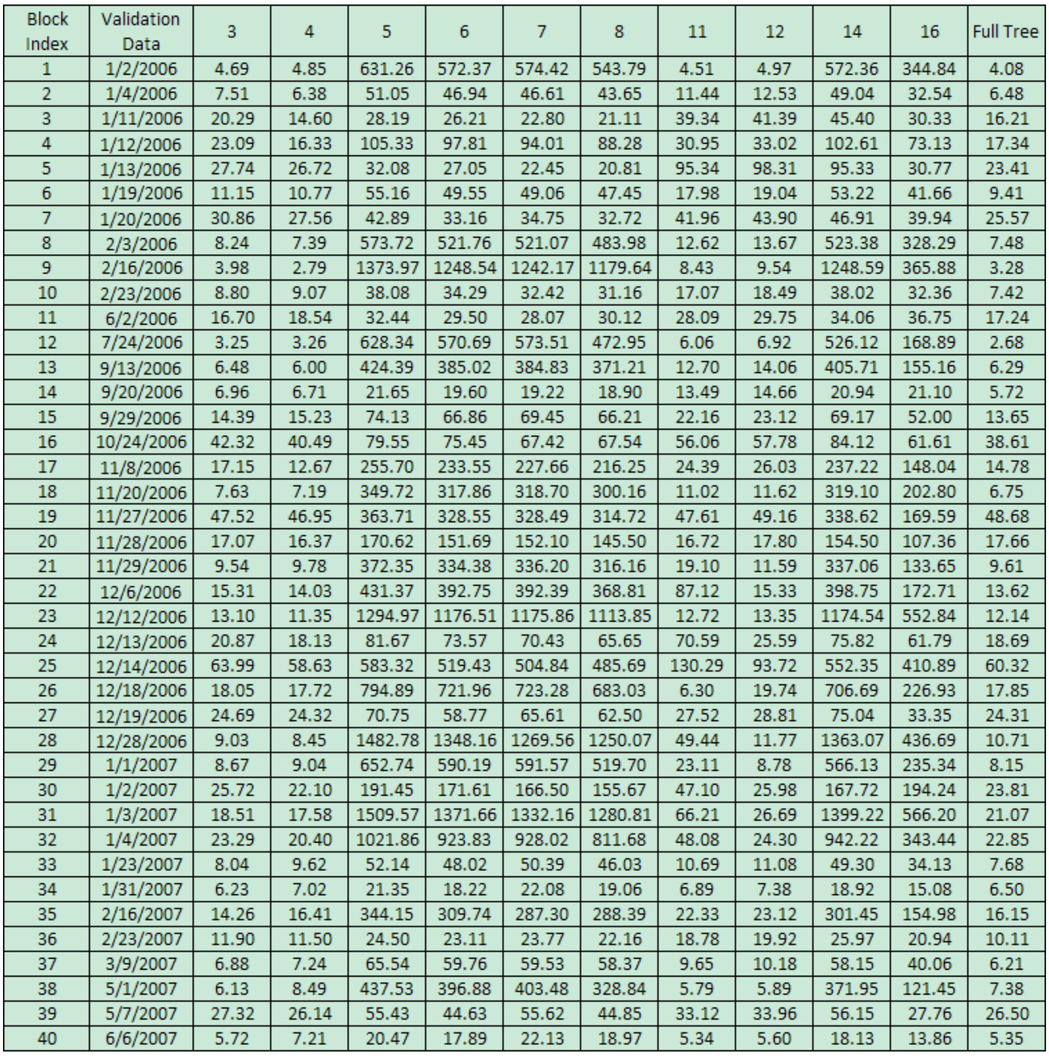
Figure K.1 MSE calculation

APPENDIX L. RCBD FOR VALIDATION DATA SETS IN TEN CHARACTERIZATIONS



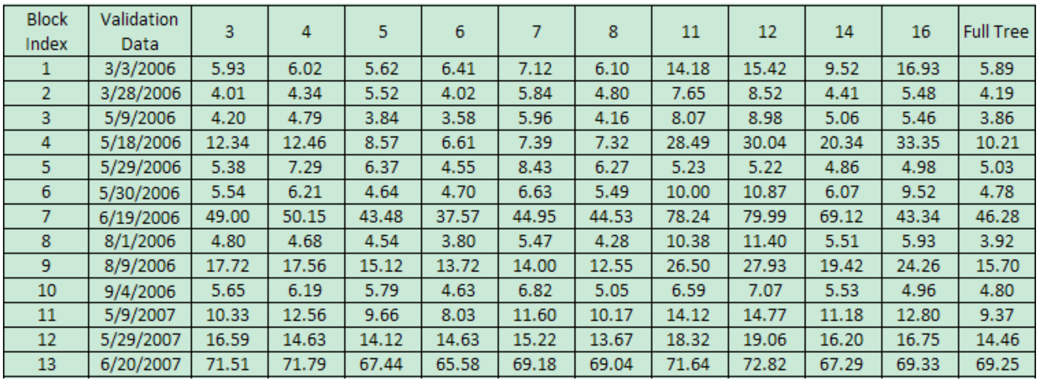
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	4/28/2006	18.14	21.73	31.09	25.80	26.29	25.60	103.32	105.90	101.85	39.79	15.81
2	5/5/2006	5.38	5.86	64.81	59.05	60.66	56.64	10.89	11.95	60.15	40.35	5.12
3	5/8/2006	5.21	5.84	21.44	18.23	20.91	18.92	10.89	11.72	21.26	12.23	5.06
4	5/10/2006	8.36	8.68	52.48	47.00	48.19	45.74	16.53	17.76	52.34	50.17	7.64
5	6/13/2006	11.76	12.76	26.40	22.50	24.15	23.40	42.96	29.04	33.36	29.54	10.36
6	6/29/2006	10.81	10.96	26.25	23.85	21.00	22.63	26.13	27.94	30.53	28.37	10.40
7	7/27/2006	5.85	5.85	34.54	31.51	32.86	27.02	12.73	13.92	34.03	14.78	4.98
8	8/2/2006	5.06	5.07	1018.60	924.37	926.11	879.20	8.96	9.62	926.65	555.20	4.66
9	8/16/2006	6.20	7.15	110.86	99.70	101.68	95.99	11.79	12.77	101.14	68.43	5.37
10	9/5/2006	1.14	1.13	3889.97	3532.63	3536.15	3345.65	1.13	1.15	3542.44	1805.01	1.23
11	9/7/2006	1.71	1.36	2563.67	2328.46	2330.56	2152.54	4.73	5.39	2325.33	934.68	1.32
12	10/25/2006	6.64	7.21	7.21	79.89	72.82	70.11	8.15	8.49	72.79	48.54	5.63
13	11/14/2006	11.74	14.21	264.93	237.29	238.22	195.28	36.49	37.48	242.77	76.81	10.10
14	1/12/2007	8.12	9.27	23.42	21.61	21.12	20.25	12.61	13.48	24.14	16.98	8.05
15	2/12/2007	4.50	6.41	94.37	85.09	88.87	82.26	5.68	5.98	85.71	55.40	4.52
16	3/5/2007	4.33	5.48	1418.81	1288.38	1293.14	1174.01	3.82	4.05	1285.71	390.97	3.88
17	3/6/2007	5.72	6.91	20.04	18.33	21.54	18.73	6.15	6.43	18.10	14.68	5.50
18	3/8/2007	6.63	7.57	35.09	32.55	35.24	31.76	6.96	7.24	33.38	31.89	5.77
19	3/14/2007	6.37	8.25	125.27	113.38	118.30	110.63	6.72	6.86	113.92	73.61	6.56
20	3/16/2007	19.35	20.65	33.14	27.09	29.76	28.48	27.47	28.50	34.00	27.42	18.31
21	3/22/2007	10.86	11.95	84.02	75.67	79.73	73.86	12.94	13.21	77.95	49.02	9.41
22	3/29/2007	13.80	13.74	86.58	79.65	81.44	78.00	29.55	14.55	85.99	36.61	12.06
23	4/4/2007	6.66	7.24	169.79	153.71	155.82	147.35	5.95	6.20	154.50	98.43	5.24
24	4/5/2007	20.11	18.28	210.77	192.65	193.21	183.85	23.90	24.77	195.88	129.61	17.81
25	4/6/2007	5.64	6.19	913.78	829.64	833.22	745.14	8.06	8.84	828.34	249.97	5.63
26	4/19/2007	8.15	9.07	900.05	818.31	821.41	725.08	10.87	11.63	838.43	244.81	7.30
27	4/20/2007	15.99	17.52	775.33	704.68	706.17	671.58	20.60	21.36	706.33	445.95	16.15
28	4/23/2007	0.54	0.62	3665.17	3328.87	3332.63	3105.66	0.69	0.57	3325.27	1556.48	0.47
29	4/26/2007	48.60	49.78	215.05	182.15	191.96	182.10	66.85	68.22	212.57	126.59	51.03
30	4/27/2007	11.63	11.79	220.29	200.59	201.82	190.90	15.38	16.15	201.68	131.57	11.91
31	5/11/2007	22.11	20.43	64.80	61.19	60.82	58.24	28.16	29.32	63.40	47.14	21.62
32	5/14/2007	18.53	18.00	61.72	56.09	59.07	55.40	18.36	18.84	57.56	43.10	17.21
33	5/21/2007	7.90	7.66	2344.96	2131.41	2133.57	1963.94	7.63	7.78	2125.98	942.54	7.15
34	5/23/2007	7.58	8.60	21.66	19.99	22.27	19.82	7.78	8.13	20.75	15.82	6.64
35	5/30/2007	11.30	11.80	411.97	375.36	377.21	363.19	11.44	12.12	397.69	117.87	10.37
36	5/31/2007	7.83	7.99	1526.07	1387.00	1389.48	1320.29	8.76	8.88	1389.02	868.08	7.02
37	6/8/2007	14.13	14.59	772.96	703.08	703.95	632.04	19.71	20.74	687.61	216.33	14.24
38	6/14/2007	25.39	25.78	53.03	46.82	50.60	50.06	32.08	33.17	52.76	43.74	25.11
39	6/22/2007	27.07	27.33	473.62	429.99	431.77	411.09	33.29	34.24	433.33	279.24	26.77
40	6/25/2007	1.21	1.64	2667.87	2423.18	2427.20	2243.65	1.63	1.71	2427.31	993.18	1.45

Figure L.1 RCBD for validation data sets in characterization 3



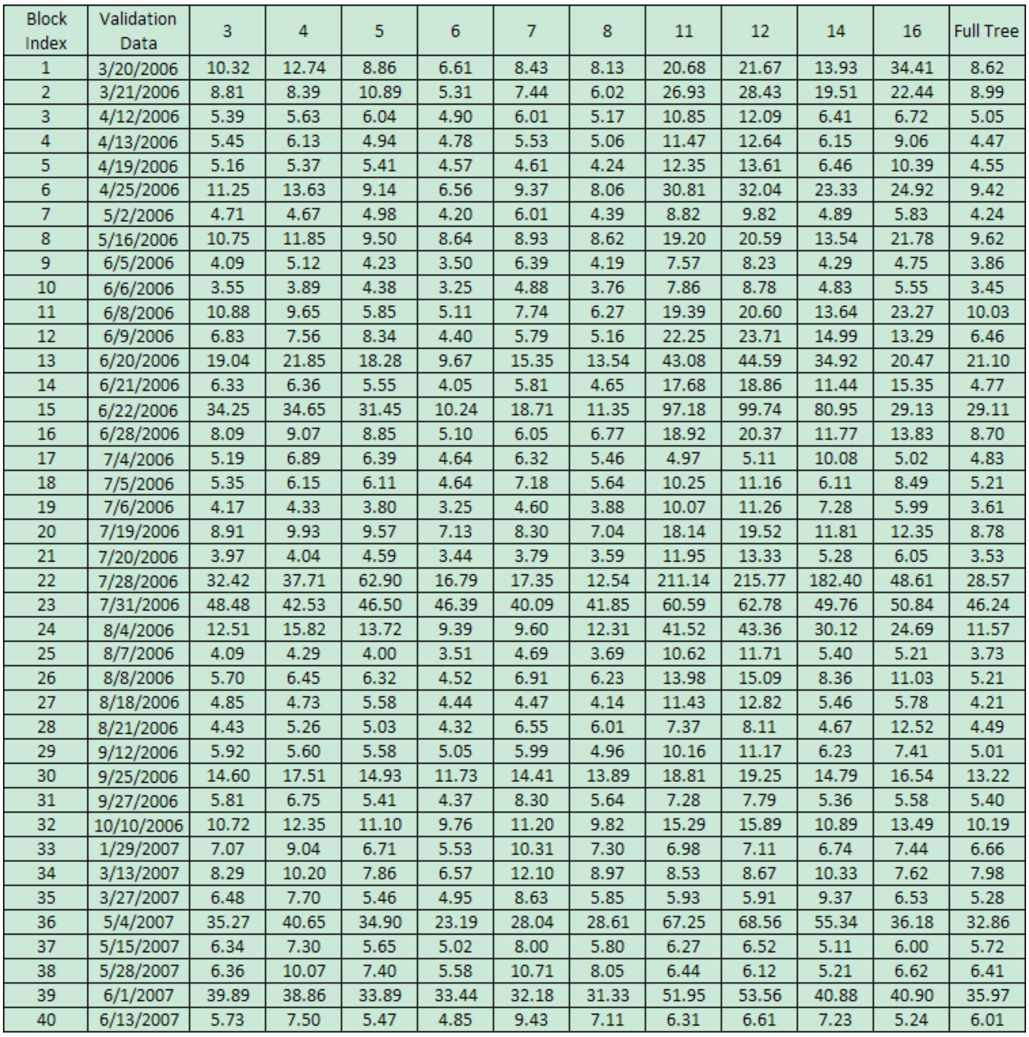
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/2/2006	4.69	4.85	631.26	572.37	574.42	543.79	4.51	4.97	572.36	344.84	4.08
2	1/4/2006	7.51	6.38	51.05	46.94	46.61	43.65	11.44	12.53	49.04	32.54	6.48
3	1/11/2006	20.29	14.60	28.19	26.21	22.80	21.11	39.34	41.39	45.40	30.33	16.21
4	1/12/2006	23.09	16.33	105.33	97.81	94.01	88.28	30.95	33.02	102.61	73.13	17.34
5	1/13/2006	27.74	26.72	32.08	27.05	22.45	20.81	95.34	98.31	95.33	30.77	23.41
6	1/19/2006	11.15	10.77	55.16	49.55	49.06	47.45	17.98	19.04	53.22	41.66	9.41
7	1/20/2006	30.86	27.56	42.89	33.16	34.75	32.72	41.96	43.90	46.91	39.94	25.57
8	2/3/2006	8.24	7.39	573.72	521.76	521.07	483.98	12.62	13.67	523.38	328.29	7.48
9	2/16/2006	3.98	2.79	1373.97	1248.54	1242.17	1179.64	8.43	9.54	1248.59	365.88	3.28
10	2/23/2006	8.80	9.07	38.08	34.29	32.42	31.16	17.07	18.49	38.02	32.36	7.42
11	6/2/2006	16.70	18.54	32.44	29.50	28.07	30.12	28.09	29.75	34.06	36.75	17.24
12	7/24/2006	3.25	3.26	628.34	570.69	573.51	472.95	6.06	6.92	526.12	168.89	2.68
13	9/13/2006	6.48	6.00	424.39	385.02	384.83	371.21	12.70	14.06	405.71	155.16	6.29
14	9/20/2006	6.96	6.71	21.65	19.60	19.22	18.90	13.49	14.66	20.94	21.10	5.72
15	9/29/2006	14.39	15.23	74.13	66.86	69.45	66.21	22.16	23.12	69.17	52.00	13.65
16	10/24/2006	42.32	40.49	79.55	75.45	67.42	67.54	56.06	57.78	84.12	61.61	38.61
17	11/8/2006	17.15	12.67	255.70	233.55	227.66	216.25	24.39	26.03	237.22	148.04	14.78
18	11/20/2006	7.63	7.19	349.72	317.86	318.70	300.16	11.02	11.62	319.10	202.80	6.75
19	11/27/2006	47.52	46.95	363.71	328.55	328.49	314.72	47.61	49.16	338.62	169.59	48.68
20	11/28/2006	17.07	16.37	170.62	151.69	152.10	145.50	16.72	17.80	154.50	107.36	17.66
21	11/29/2006	9.54	9.78	372.35	334.38	336.20	316.16	19.10	11.59	337.06	133.65	9.61
22	12/6/2006	15.31	14.03	431.37	392.75	392.39	368.81	87.12	15.33	398.75	172.71	13.62
23	12/12/2006	13.10	11.35	1294.97	1176.51	1175.86	1113.85	12.72	13.35	1174.54	552.84	12.14
24	12/13/2006	20.87	18.13	81.67	73.57	70.43	65.65	70.59	25.59	75.82	61.79	18.69
25	12/14/2006	63.99	58.63	583.32	519.43	504.84	485.69	130.29	93.72	552.35	410.89	60.32
26	12/18/2006	18.05	17.72	794.89	721.96	723.28	683.03	6.30	19.74	706.69	226.93	17.85
27	12/19/2006	24.69	24.32	70.75	58.77	65.61	62.50	27.52	28.81	75.04	33.35	24.31
28	12/28/2006	9.03	8.45	1482.78	1348.16	1269.56	1250.07	49.44	11.77	1363.07	436.69	10.71
29	1/1/2007	8.67	9.04	652.74	590.19	591.57	519.70	23.11	8.78	566.13	235.34	8.15
30	1/2/2007	25.72	22.10	191.45	171.61	166.50	155.67	47.10	25.98	167.72	194.24	23.81
31	1/3/2007	18.51	17.58	1509.57	1371.66	1332.16	1280.81	66.21	26.69	1399.22	566.20	21.07
32	1/4/2007	23.29	20.40	1021.86	923.83	928.02	811.68	48.08	24.30	942.22	343.44	22.85
33	1/23/2007	8.04	9.62	52.14	48.02	50.39	46.03	10.69	11.08	49.30	34.13	7.68
34	1/31/2007	6.23	7.02	21.35	18.22	22.08	19.06	6.89	7.38	18.92	15.08	6.50
35	2/16/2007	14.26	16.41	344.15	309.74	287.30	288.39	22.33	23.12	301.45	154.98	16.15
36	2/23/2007	11.90	11.50	24.50	23.11	23.77	22.16	18.78	19.92	25.97	20.94	10.11
37	3/9/2007	6.88	7.24	65.54	59.76	59.53	58.37	9.65	10.18	58.15	40.06	6.21
38	5/1/2007	6.13	8.49	437.53	396.88	403.48	328.84	5.79	5.89	371.95	121.45	7.38
39	5/7/2007	27.32	26.14	55.43	44.63	55.62	44.85	33.12	33.96	56.15	27.76	26.50
40	6/6/2007	5.72	7.21	20.47	17.89	22.13	18.97	5.34	5.60	18.13	13.86	5.35

Figure L.2 RCBD for validation data sets in characterization 4



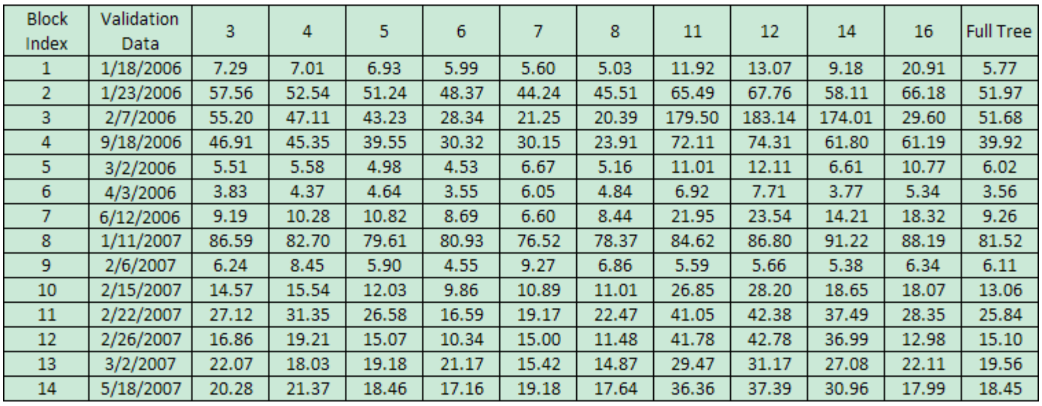
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	3/3/2006	5.93	6.02	5.62	6.41	7.12	6.10	14.18	15.42	9.52	16.93	5.89
2	3/28/2006	4.01	4.34	5.52	4.02	5.84	4.80	7.65	8.52	4.41	5.48	4.19
3	5/9/2006	4.20	4.79	3.84	3.58	5.96	4.16	8.07	8.98	5.06	5.46	3.86
4	5/18/2006	12.34	12.46	8.57	6.61	7.39	7.32	28.49	30.04	20.34	33.35	10.21
5	5/29/2006	5.38	7.29	6.37	4.55	8.43	6.27	5.23	5.22	4.86	4.98	5.03
6	5/30/2006	5.54	6.21	4.64	4.70	6.63	5.49	10.00	10.87	6.07	9.52	4.78
7	6/19/2006	49.00	50.15	43.48	37.57	44.95	44.53	78.24	79.99	69.12	43.34	46.28
8	8/1/2006	4.80	4.68	4.54	3.80	5.47	4.28	10.38	11.40	5.51	5.93	3.92
9	8/9/2006	17.72	17.56	15.12	13.72	14.00	12.55	26.50	27.93	19.42	24.26	15.70
10	9/4/2006	5.65	6.19	5.79	4.63	6.82	5.05	6.59	7.07	5.53	4.96	4.80
11	5/9/2007	10.33	12.56	9.66	8.03	11.60	10.17	14.12	14.77	11.18	12.80	9.37
12	5/29/2007	16.59	14.63	14.12	14.63	15.22	13.67	18.32	19.06	16.20	16.75	14.46
13	6/20/2007	71.51	71.79	67.44	65.58	69.18	69.04	71.64	72.82	67.29	69.33	69.25

Figure L.3 RCBD for validation data sets in characterization 5



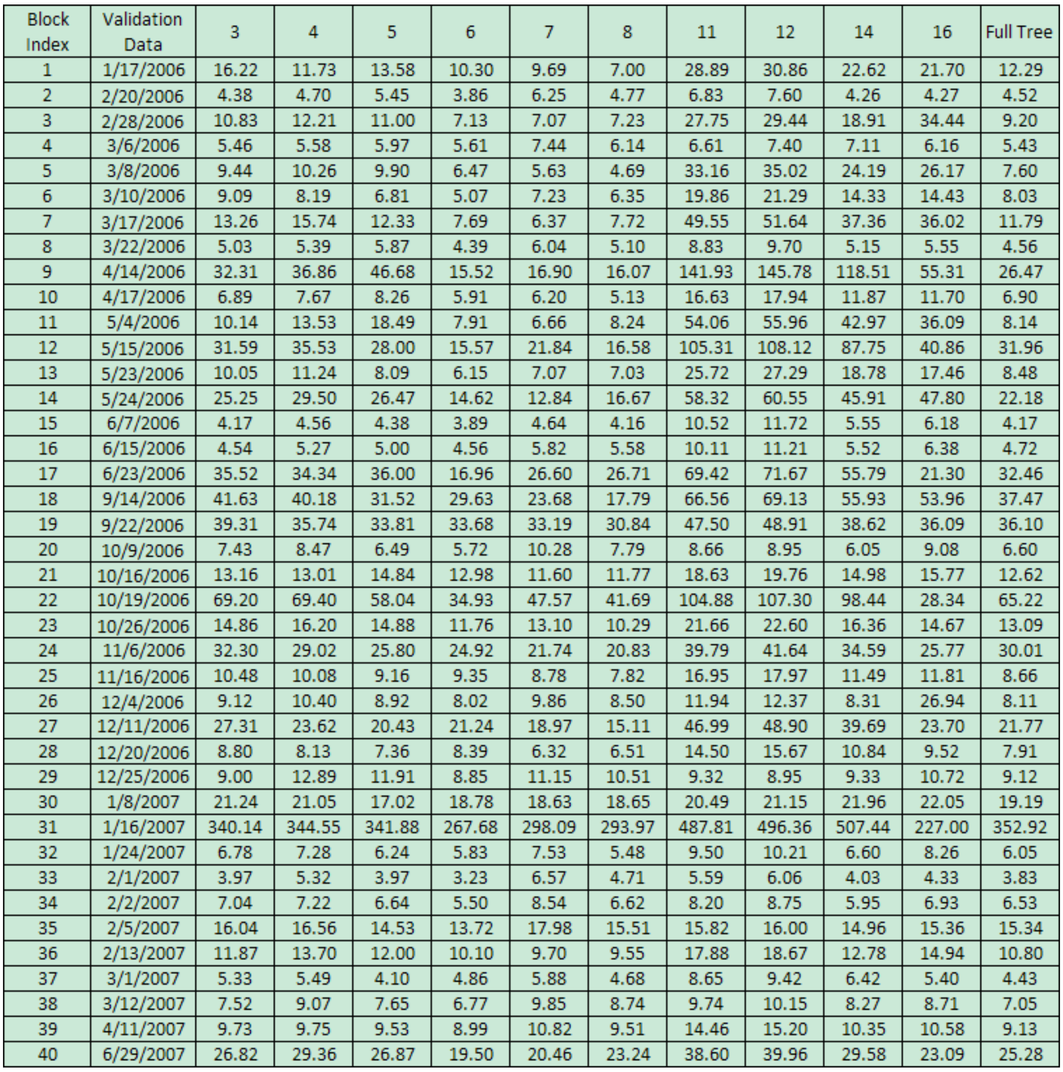
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	3/20/2006	10.32	12.74	8.86	6.61	8.43	8.13	20.68	21.67	13.93	34.41	8.62
2	3/21/2006	8.81	8.39	10.89	5.31	7.44	6.02	26.93	28.43	19.51	22.44	8.99
3	4/12/2006	5.39	5.63	6.04	4.90	6.01	5.17	10.85	12.09	6.41	6.72	5.05
4	4/13/2006	5.45	6.13	4.94	4.78	5.53	5.06	11.47	12.64	6.15	9.06	4.47
5	4/19/2006	5.16	5.37	5.41	4.57	4.61	4.24	12.35	13.61	6.46	10.39	4.55
6	4/25/2006	11.25	13.63	9.14	6.56	9.37	8.06	30.81	32.04	23.33	24.92	9.42
7	5/2/2006	4.71	4.67	4.98	4.20	6.01	4.39	8.82	9.82	4.89	5.83	4.24
8	5/16/2006	10.75	11.85	9.50	8.64	8.93	8.62	19.20	20.59	13.54	21.78	9.62
9	6/5/2006	4.09	5.12	4.23	3.50	6.39	4.19	7.57	8.23	4.29	4.75	3.86
10	6/6/2006	3.55	3.89	4.38	3.25	4.88	3.76	7.86	8.78	4.83	5.55	3.45
11	6/8/2006	10.88	9.65	5.85	5.11	7.74	6.27	19.39	20.60	13.64	23.27	10.03
12	6/9/2006	6.83	7.56	8.34	4.40	5.79	5.16	22.25	23.71	14.99	13.29	6.46
13	6/20/2006	19.04	21.85	18.28	9.67	15.35	13.54	43.08	44.59	34.92	20.47	21.10
14	6/21/2006	6.33	6.36	5.55	4.05	5.81	4.65	17.68	18.86	11.44	15.35	4.77
15	6/22/2006	34.25	34.65	31.45	10.24	18.71	11.35	97.18	99.74	80.95	29.13	29.11
16	6/28/2006	8.09	9.07	8.85	5.10	6.05	6.77	18.92	20.37	11.77	13.83	8.70
17	7/4/2006	5.19	6.89	6.39	4.64	6.32	5.46	4.97	5.11	10.08	5.02	4.83
18	7/5/2006	5.35	6.15	6.11	4.64	7.18	5.64	10.25	11.16	6.11	8.49	5.21
19	7/6/2006	4.17	4.33	3.80	3.25	4.60	3.88	10.07	11.26	7.28	5.99	3.61
20	7/19/2006	8.91	9.93	9.57	7.13	8.30	7.04	18.14	19.52	11.81	12.35	8.78
21	7/20/2006	3.97	4.04	4.59	3.44	3.79	3.59	11.95	13.33	5.28	6.05	3.53
22	7/28/2006	32.42	37.71	62.90	16.79	17.35	12.54	211.14	215.77	182.40	48.61	28.57
23	7/31/2006	48.48	42.53	46.50	46.39	40.09	41.85	60.59	62.78	49.76	50.84	46.24
24	8/4/2006	12.51	15.82	13.72	9.39	9.60	12.31	41.52	43.36	30.12	24.69	11.57
25	8/7/2006	4.09	4.29	4.00	3.51	4.69	3.69	10.62	11.71	5.40	5.21	3.73
26	8/8/2006	5.70	6.45	6.32	4.52	6.91	6.23	13.98	15.09	8.36	11.03	5.21
27	8/18/2006	4.85	4.73	5.58	4.44	4.47	4.14	11.43	12.82	5.46	5.78	4.21
28	8/21/2006	4.43	5.26	5.03	4.32	6.55	6.01	7.37	8.11	4.67	12.52	4.49
29	9/12/2006	5.92	5.60	5.58	5.05	5.99	4.96	10.16	11.17	6.23	7.41	5.01
30	9/25/2006	14.60	17.51	14.93	11.73	14.41	13.89	18.81	19.25	14.79	16.54	13.22
31	9/27/2006	5.81	6.75	5.41	4.37	8.30	5.64	7.28	7.79	5.36	5.58	5.40
32	10/10/2006	10.72	12.35	11.10	9.76	11.20	9.82	15.29	15.89	10.89	13.49	10.19
33	1/29/2007	7.07	9.04	6.71	5.53	10.31	7.30	6.98	7.11	6.74	7.44	6.66
34	3/13/2007	8.29	10.20	7.86	6.57	12.10	8.97	8.53	8.67	10.33	7.62	7.98
35	3/27/2007	6.48	7.70	5.46	4.95	8.63	5.85	5.93	5.91	9.37	6.53	5.28
36	5/4/2007	35.27	40.65	34.90	23.19	28.04	28.61	67.25	68.56	55.34	36.18	32.86
37	5/15/2007	6.34	7.30	5.65	5.02	8.00	5.80	6.27	6.52	5.11	6.00	5.72
38	5/28/2007	6.36	10.07	7.40	5.58	10.71	8.05	6.44	6.12	5.21	6.62	6.41
39	6/1/2007	39.89	38.86	33.89	33.44	32.18	31.33	51.95	53.56	40.88	40.90	35.97
40	6/13/2007	5.73	7.50	5.47	4.85	9.43	7.11	6.31	6.61	7.23	5.24	6.01

Figure L.4 RCBD for validation data sets in characterization 6



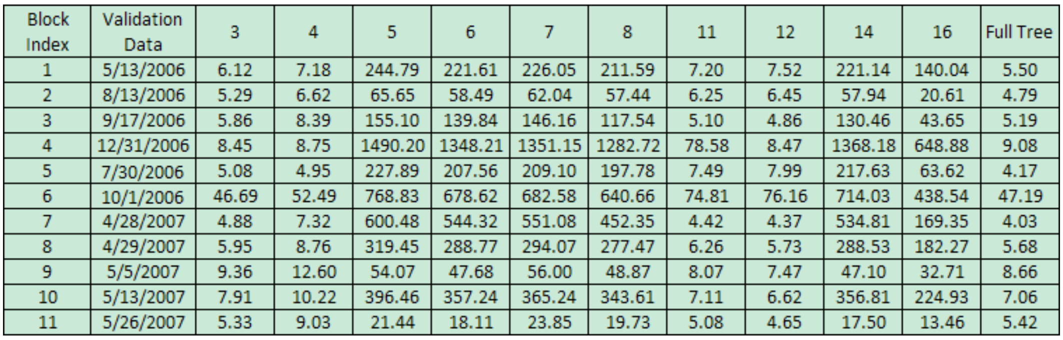
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/18/2006	7.29	7.01	6.93	5.99	5.60	5.03	11.92	13.07	9.18	20.91	5.77
2	1/23/2006	57.56	52.54	51.24	48.37	44.24	45.51	65.49	67.76	58.11	66.18	51.97
3	2/7/2006	55.20	47.11	43.23	28.34	21.25	20.39	179.50	183.14	174.01	29.60	51.68
4	9/18/2006	46.91	45.35	39.55	30.32	30.15	23.91	72.11	74.31	61.80	61.19	39.92
5	3/2/2006	5.51	5.58	4.98	4.53	6.67	5.16	11.01	12.11	6.61	10.77	6.02
6	4/3/2006	3.83	4.37	4.64	3.55	6.05	4.84	6.92	7.71	3.77	5.34	3.56
7	6/12/2006	9.19	10.28	10.82	8.69	6.60	8.44	21.95	23.54	14.21	18.32	9.26
8	1/11/2007	86.59	82.70	79.61	80.93	76.52	78.37	84.62	86.80	91.22	88.19	81.52
9	2/6/2007	6.24	8.45	5.90	4.55	9.27	6.86	5.59	5.66	5.38	6.34	6.11
10	2/15/2007	14.57	15.54	12.03	9.86	10.89	11.01	26.85	28.20	18.65	18.07	13.06
11	2/22/2007	27.12	31.35	26.58	16.59	19.17	22.47	41.05	42.38	37.49	28.35	25.84
12	2/26/2007	16.86	19.21	15.07	10.34	15.00	11.48	41.78	42.78	36.99	12.98	15.10
13	3/2/2007	22.07	18.03	19.18	21.17	15.42	14.87	29.47	31.17	27.08	22.11	19.56
14	5/18/2007	20.28	21.37	18.46	17.16	19.18	17.64	36.36	37.39	30.96	17.99	18.45

Figure L.5 RCBD for validation data sets in characterization 7



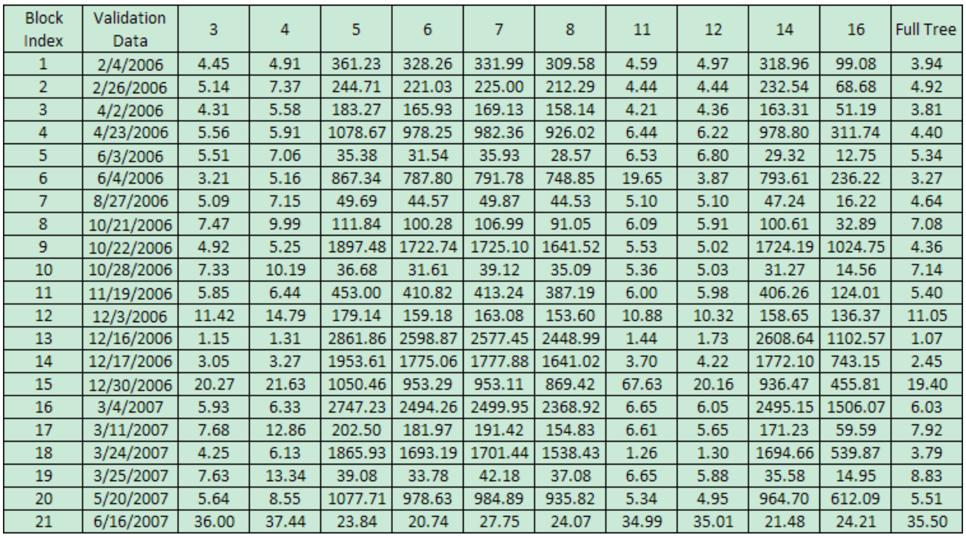
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/17/2006	16.22	11.73	13.58	10.30	9.69	7.00	28.89	30.86	22.62	21.70	12.29
2	2/20/2006	4.38	4.70	5.45	3.86	6.25	4.77	6.83	7.60	4.26	4.27	4.52
3	2/28/2006	10.83	12.21	11.00	7.13	7.07	7.23	27.75	29.44	18.91	34.44	9.20
4	3/6/2006	5.46	5.58	5.97	5.61	7.44	6.14	6.61	7.40	7.11	6.16	5.43
5	3/8/2006	9.44	10.26	9.90	6.47	5.63	4.69	33.16	35.02	24.19	26.17	7.60
6	3/10/2006	9.09	8.19	6.81	5.07	7.23	6.35	19.86	21.29	14.33	14.43	8.03
7	3/17/2006	13.26	15.74	12.33	7.69	6.37	7.72	49.55	51.64	37.36	36.02	11.79
8	3/22/2006	5.03	5.39	5.87	4.39	6.04	5.10	8.83	9.70	5.15	5.55	4.56
9	4/14/2006	32.31	36.86	46.68	15.52	16.90	16.07	141.93	145.78	118.51	55.31	26.47
10	4/17/2006	6.89	7.67	8.26	5.91	6.20	5.13	16.63	17.94	11.87	11.70	6.90
11	5/4/2006	10.14	13.53	18.49	7.91	6.66	8.24	54.06	55.96	42.97	36.09	8.14
12	5/15/2006	31.59	35.53	28.00	15.57	21.84	16.58	105.31	108.12	87.75	40.86	31.96
13	5/23/2006	10.05	11.24	8.09	6.15	7.07	7.03	25.72	27.29	18.78	17.46	8.48
14	5/24/2006	25.25	29.50	26.47	14.62	12.84	16.67	58.32	60.55	45.91	47.80	22.18
15	6/7/2006	4.17	4.56	4.38	3.89	4.64	4.16	10.52	11.72	5.55	6.18	4.17
16	6/15/2006	4.54	5.27	5.00	4.56	5.82	5.58	10.11	11.21	5.52	6.38	4.72
17	6/23/2006	35.52	34.34	36.00	16.96	26.60	26.71	69.42	71.67	55.79	21.30	32.46
18	9/14/2006	41.63	40.18	31.52	29.63	23.68	17.79	66.56	69.13	55.93	53.96	37.47
19	9/22/2006	39.31	35.74	33.81	33.68	33.19	30.84	47.50	48.91	38.62	36.09	36.10
20	10/9/2006	7.43	8.47	6.49	5.72	10.28	7.79	8.66	8.95	6.05	9.08	6.60
21	10/16/2006	13.16	13.01	14.84	12.98	11.60	11.77	18.63	19.76	14.98	15.77	12.62
22	10/19/2006	69.20	69.40	58.04	34.93	47.57	41.69	104.88	107.30	98.44	28.34	65.22
23	10/26/2006	14.86	16.20	14.88	11.76	13.10	10.29	21.66	22.60	16.36	14.67	13.09
24	11/6/2006	32.30	29.02	25.80	24.92	21.74	20.83	39.79	41.64	34.59	25.77	30.01
25	11/16/2006	10.48	10.08	9.16	9.35	8.78	7.82	16.95	17.97	11.49	11.81	8.66
26	12/4/2006	9.12	10.40	8.92	8.02	9.86	8.50	11.94	12.37	8.31	26.94	8.11
27	12/11/2006	27.31	23.62	20.43	21.24	18.97	15.11	46.99	48.90	39.69	23.70	21.77
28	12/20/2006	8.80	8.13	7.36	8.39	6.32	6.51	14.50	15.67	10.84	9.52	7.91
29	12/25/2006	9.00	12.89	11.91	8.85	11.15	10.51	9.32	8.95	9.33	10.72	9.12
30	1/8/2007	21.24	21.05	17.02	18.78	18.63	18.65	20.49	21.15	21.96	22.05	19.19
31	1/16/2007	340.14	344.55	341.88	267.68	298.09	293.97	487.81	496.36	507.44	227.00	352.92
32	1/24/2007	6.78	7.28	6.24	5.83	7.53	5.48	9.50	10.21	6.60	8.26	6.05
33	2/1/2007	3.97	5.32	3.97	3.23	6.57	4.71	5.59	6.06	4.03	4.33	3.83
34	2/2/2007	7.04	7.22	6.64	5.50	8.54	6.62	8.20	8.75	5.95	6.93	6.53
35	2/5/2007	16.04	16.56	14.53	13.72	17.98	15.51	15.82	16.00	14.96	15.36	15.34
36	2/13/2007	11.87	13.70	12.00	10.10	9.70	9.55	17.88	18.67	12.78	14.94	10.80
37	3/1/2007	5.33	5.49	4.10	4.86	5.88	4.68	8.65	9.42	6.42	5.40	4.43
38	3/12/2007	7.52	9.07	7.65	6.77	9.85	8.74	9.74	10.15	8.27	8.71	7.05
39	4/11/2007	9.73	9.75	9.53	8.99	10.82	9.51	14.46	15.20	10.35	10.58	9.13
40	6/29/2007	26.82	29.36	26.87	19.50	20.46	23.24	38.60	39.96	29.58	23.09	25.28

Figure L.6 RCBD for validation data sets in characterization 8



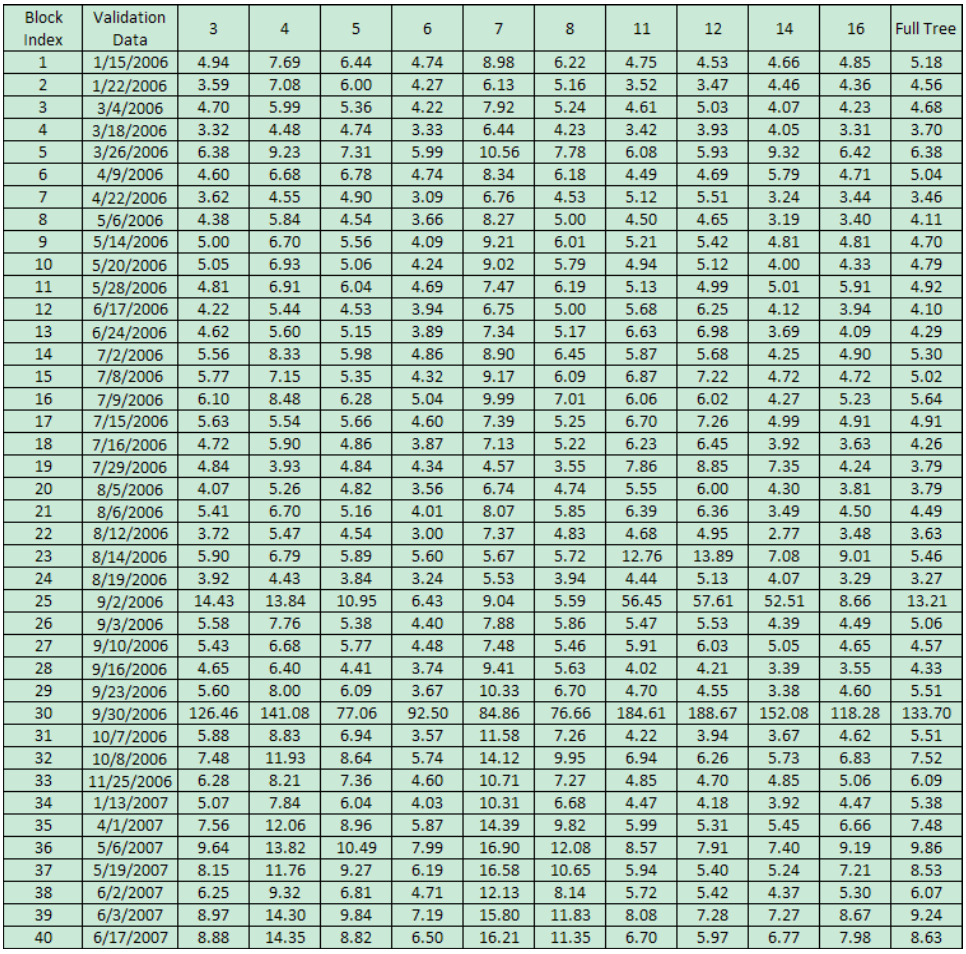
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	5/13/2006	6.12	7.18	244.79	221.61	226.05	211.59	7.20	7.52	221.14	140.04	5.50
2	8/13/2006	5.29	6.62	65.65	58.49	62.04	57.44	6.25	6.45	57.94	20.61	4.79
3	9/17/2006	5.86	8.39	155.10	139.84	146.16	117.54	5.10	4.86	130.46	43.65	5.19
4	12/31/2006	8.45	8.75	1490.20	1348.21	1351.15	1282.72	78.58	8.47	1368.18	648.88	9.08
5	7/30/2006	5.08	4.95	227.89	207.56	209.10	197.78	7.49	7.99	217.63	63.62	4.17
6	10/1/2006	46.69	52.49	768.83	678.62	682.58	640.66	74.81	76.16	714.03	438.54	47.19
7	4/28/2007	4.88	7.32	600.48	544.32	551.08	452.35	4.42	4.37	534.81	169.35	4.03
8	4/29/2007	5.95	8.76	319.45	288.77	294.07	277.47	6.26	5.73	288.53	182.27	5.68
9	5/5/2007	9.36	12.60	54.07	47.68	56.00	48.87	8.07	7.47	47.10	32.71	8.66
10	5/13/2007	7.91	10.22	396.46	357.24	365.24	343.61	7.11	6.62	356.81	224.93	7.06
11	5/26/2007	5.33	9.03	21.44	18.11	23.85	19.73	5.08	4.65	17.50	13.46	5.42

Figure L.7 RCBD for validation data sets in characterization 11



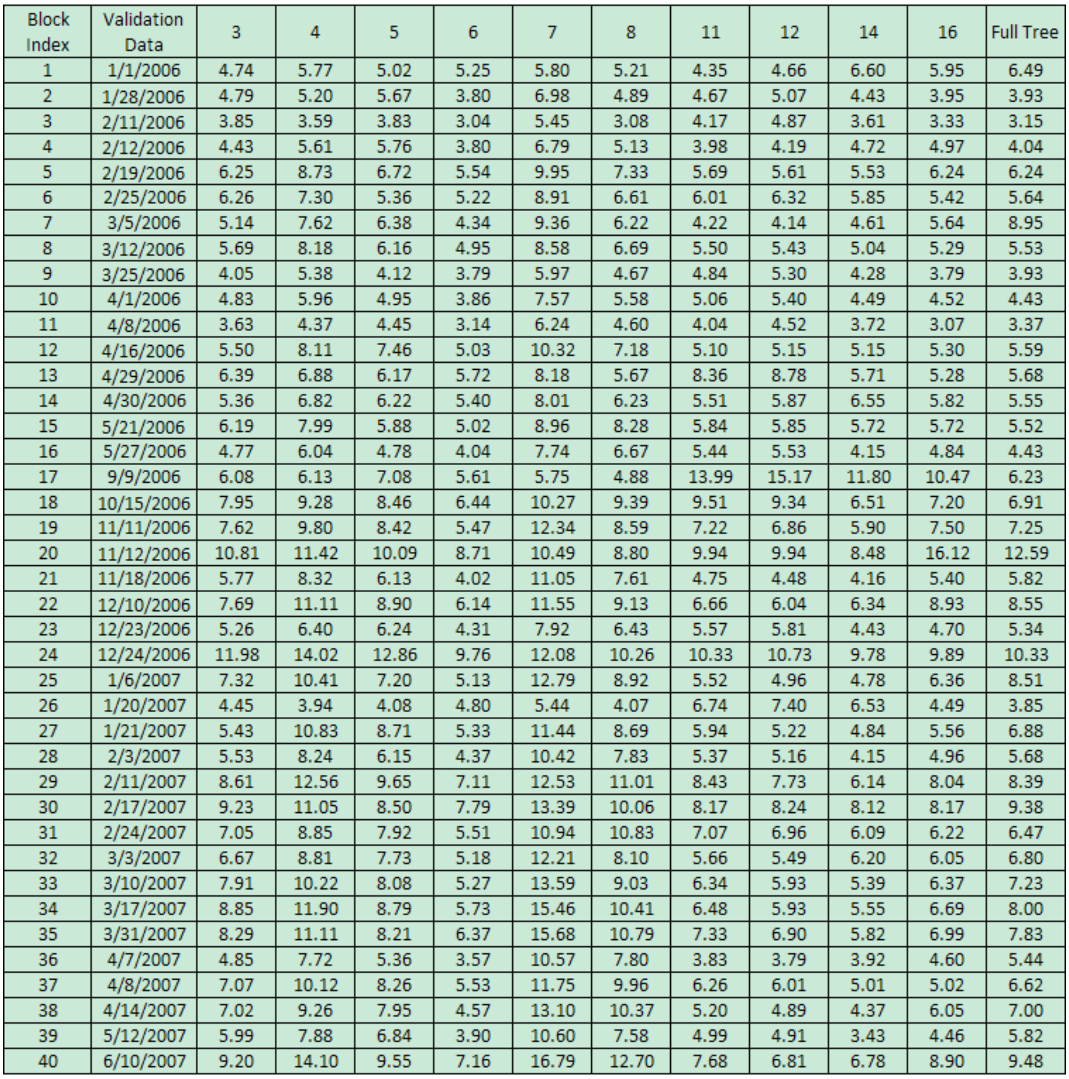
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	2/4/2006	4.45	4.91	361.23	328.26	331.99	309.58	4.59	4.97	318.96	99.08	3.94
2	2/26/2006	5.14	7.37	244.71	221.03	225.00	212.29	4.44	4.44	232.54	68.68	4.92
3	4/2/2006	4.31	5.58	183.27	165.93	169.13	158.14	4.21	4.36	163.31	51.19	3.81
4	4/23/2006	5.56	5.91	1078.67	978.25	982.36	926.02	6.44	6.22	978.80	311.74	4.40
5	6/3/2006	5.51	7.06	35.38	31.54	35.93	28.57	6.53	6.80	29.32	12.75	5.34
6	6/4/2006	3.21	5.16	867.34	787.80	791.78	748.85	19.65	3.87	793.61	236.22	3.27
7	8/27/2006	5.09	7.15	49.69	44.57	49.87	44.53	5.10	5.10	47.24	16.22	4.64
8	10/21/2006	7.47	9.99	111.84	100.28	106.99	91.05	6.09	5.91	100.61	32.89	7.08
9	10/22/2006	4.92	5.25	1897.48	1722.74	1725.10	1641.52	5.53	5.02	1724.19	1024.75	4.36
10	10/28/2006	7.33	10.19	36.68	31.61	39.12	35.09	5.36	5.03	31.27	14.56	7.14
11	11/19/2006	5.85	6.44	453.00	410.82	413.24	387.19	6.00	5.98	406.26	124.01	5.40
12	12/3/2006	11.42	14.79	179.14	159.18	163.08	153.60	10.88	10.32	158.65	136.37	11.05
13	12/16/2006	1.15	1.31	2861.86	2598.87	2577.45	2448.99	1.44	1.73	2608.64	1102.57	1.07
14	12/17/2006	3.05	3.27	1953.61	1775.06	1777.88	1641.02	3.70	4.22	1772.10	743.15	2.45
15	12/30/2006	20.27	21.63	1050.46	953.29	953.11	869.42	67.63	20.16	936.47	455.81	19.40
16	3/4/2007	5.93	6.33	2747.23	2494.26	2499.95	2368.92	6.65	6.05	2495.15	1506.07	6.03
17	3/11/2007	7.68	12.86	202.50	181.97	191.42	154.83	6.61	5.65	171.23	59.59	7.92
18	3/24/2007	4.25	6.13	1865.93	1693.19	1701.44	1538.43	1.26	1.30	1694.66	539.87	3.79
19	3/25/2007	7.63	13.34	39.08	33.78	42.18	37.08	6.65	5.88	35.58	14.95	8.83
20	5/20/2007	5.64	8.55	1077.71	978.63	984.89	935.82	5.34	4.95	964.70	612.09	5.51
21	6/16/2007	36.00	37.44	23.84	20.74	27.75	24.07	34.99	35.01	21.48	24.21	35.50

Figure L.8 RCBD for validation data sets in characterization 12



Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/15/2006	4.94	7.69	6.44	4.74	8.98	6.22	4.75	4.53	4.66	4.85	5.18
2	1/22/2006	3.59	7.08	6.00	4.27	6.13	5.16	3.52	3.47	4.46	4.36	4.56
3	3/4/2006	4.70	5.99	5.36	4.22	7.92	5.24	4.61	5.03	4.07	4.23	4.68
4	3/18/2006	3.32	4.48	4.74	3.33	6.44	4.23	3.42	3.93	4.05	3.31	3.70
5	3/26/2006	6.38	9.23	7.31	5.99	10.56	7.78	6.08	5.93	9.32	6.42	6.38
6	4/9/2006	4.60	6.68	6.78	4.74	8.34	6.18	4.49	4.69	5.79	4.71	5.04
7	4/22/2006	3.62	4.55	4.90	3.09	6.76	4.53	5.12	5.51	3.24	3.44	3.46
8	5/6/2006	4.38	5.84	4.54	3.66	8.27	5.00	4.50	4.65	3.19	3.40	4.11
9	5/14/2006	5.00	6.70	5.56	4.09	9.21	6.01	5.21	5.42	4.81	4.81	4.70
10	5/20/2006	5.05	6.93	5.06	4.24	9.02	5.79	4.94	5.12	4.00	4.33	4.79
11	5/28/2006	4.81	6.91	6.04	4.69	7.47	6.19	5.13	4.99	5.01	5.91	4.92
12	6/17/2006	4.22	5.44	4.53	3.94	6.75	5.00	5.68	6.25	4.12	3.94	4.10
13	6/24/2006	4.62	5.60	5.15	3.89	7.34	5.17	6.63	6.98	3.69	4.09	4.29
14	7/2/2006	5.56	8.33	5.98	4.86	8.90	6.45	5.87	5.68	4.25	4.90	5.30
15	7/8/2006	5.77	7.15	5.35	4.32	9.17	6.09	6.87	7.22	4.72	4.72	5.02
16	7/9/2006	6.10	8.48	6.28	5.04	9.99	7.01	6.06	6.02	4.27	5.23	5.64
17	7/15/2006	5.63	5.54	5.66	4.60	7.39	5.25	6.70	7.26	4.99	4.91	4.91
18	7/16/2006	4.72	5.90	4.86	3.87	7.13	5.22	6.23	6.45	3.92	3.63	4.26
19	7/29/2006	4.84	3.93	4.84	4.34	4.57	3.55	7.86	8.85	7.35	4.24	3.79
20	8/5/2006	4.07	5.26	4.82	3.56	6.74	4.74	5.55	6.00	4.30	3.81	3.79
21	8/6/2006	5.41	6.70	5.16	4.01	8.07	5.85	6.39	6.36	3.49	4.50	4.49
22	8/12/2006	3.72	5.47	4.54	3.00	7.37	4.83	4.68	4.95	2.77	3.48	3.63
23	8/14/2006	5.90	6.79	5.89	5.60	5.67	5.72	12.76	13.89	7.08	9.01	5.46
24	8/19/2006	3.92	4.43	3.84	3.24	5.53	3.94	4.44	5.13	4.07	3.29	3.27
25	9/2/2006	14.43	13.84	10.95	6.43	9.04	5.59	56.45	57.61	52.51	8.66	13.21
26	9/3/2006	5.58	7.76	5.38	4.40	7.88	5.86	5.47	5.53	4.39	4.49	5.06
27	9/10/2006	5.43	6.68	5.77	4.48	7.48	5.46	5.91	6.03	5.05	4.65	4.57
28	9/16/2006	4.65	6.40	4.41	3.74	9.41	5.63	4.02	4.21	3.39	3.55	4.33
29	9/23/2006	5.60	8.00	6.09	3.67	10.33	6.70	4.70	4.55	3.38	4.60	5.51
30	9/30/2006	126.46	141.08	77.06	92.50	84.86	76.66	184.61	188.67	152.08	118.28	133.70
31	10/7/2006	5.88	8.83	6.94	3.57	11.58	7.26	4.22	3.94	3.67	4.62	5.51
32	10/8/2006	7.48	11.93	8.64	5.74	14.12	9.95	6.94	6.26	5.73	6.83	7.52
33	11/25/2006	6.28	8.21	7.36	4.60	10.71	7.27	4.85	4.70	4.85	5.06	6.09
34	1/13/2007	5.07	7.84	6.04	4.03	10.31	6.68	4.47	4.18	3.92	4.47	5.38
35	4/1/2007	7.56	12.06	8.96	5.87	14.39	9.82	5.99	5.31	5.45	6.66	7.48
36	5/6/2007	9.64	13.82	10.49	7.99	16.90	12.08	8.57	7.91	7.40	9.19	9.86
37	5/19/2007	8.15	11.76	9.27	6.19	16.58	10.65	5.94	5.40	5.24	7.21	8.53
38	6/2/2007	6.25	9.32	6.81	4.71	12.13	8.14	5.72	5.42	4.37	5.30	6.07
39	6/3/2007	8.97	14.30	9.84	7.19	15.80	11.83	8.08	7.28	7.27	8.67	9.24
40	6/17/2007	8.88	14.35	8.82	6.50	16.21	11.35	6.70	5.97	6.77	7.98	8.63

Figure L.9 RCBD for validation data sets in characterization 14



Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/1/2006	4.74	5.77	5.02	5.25	5.80	5.21	4.35	4.66	6.60	5.95	6.49
2	1/28/2006	4.79	5.20	5.67	3.80	6.98	4.89	4.67	5.07	4.43	3.95	3.93
3	2/11/2006	3.85	3.59	3.83	3.04	5.45	3.08	4.17	4.87	3.61	3.33	3.15
4	2/12/2006	4.43	5.61	5.76	3.80	6.79	5.13	3.98	4.19	4.72	4.97	4.04
5	2/19/2006	6.25	8.73	6.72	5.54	9.95	7.33	5.69	5.61	5.53	6.24	6.24
6	2/25/2006	6.26	7.30	5.36	5.22	8.91	6.61	6.01	6.32	5.85	5.42	5.64
7	3/5/2006	5.14	7.62	6.38	4.34	9.36	6.22	4.22	4.14	4.61	5.64	8.95
8	3/12/2006	5.69	8.18	6.16	4.95	8.58	6.69	5.50	5.43	5.04	5.29	5.53
9	3/25/2006	4.05	5.38	4.12	3.79	5.97	4.67	4.84	5.30	4.28	3.79	3.93
10	4/1/2006	4.83	5.96	4.95	3.86	7.57	5.58	5.06	5.40	4.49	4.52	4.43
11	4/8/2006	3.63	4.37	4.45	3.14	6.24	4.60	4.04	4.52	3.72	3.07	3.37
12	4/16/2006	5.50	8.11	7.46	5.03	10.32	7.18	5.10	5.15	5.15	5.30	5.59
13	4/29/2006	6.39	6.88	6.17	5.72	8.18	5.67	8.36	8.78	5.71	5.28	5.68
14	4/30/2006	5.36	6.82	6.22	5.40	8.01	6.23	5.51	5.87	6.55	5.82	5.55
15	5/21/2006	6.19	7.99	5.88	5.02	8.96	8.28	5.84	5.85	5.72	5.72	5.52
16	5/27/2006	4.77	6.04	4.78	4.04	7.74	6.67	5.44	5.53	4.15	4.84	4.43
17	9/9/2006	6.08	6.13	7.08	5.61	5.75	4.88	13.99	15.17	11.80	10.47	6.23
18	10/15/2006	7.95	9.28	8.46	6.44	10.27	9.39	9.51	9.34	6.51	7.20	6.91
19	11/11/2006	7.62	9.80	8.42	5.47	12.34	8.59	7.22	6.86	5.90	7.50	7.25
20	11/12/2006	10.81	11.42	10.09	8.71	10.49	8.80	9.94	9.94	8.48	16.12	12.59
21	11/18/2006	5.77	8.32	6.13	4.02	11.05	7.61	4.75	4.48	4.16	5.40	5.82
22	12/10/2006	7.69	11.11	8.90	6.14	11.55	9.13	6.66	6.04	6.34	8.93	8.55
23	12/23/2006	5.26	6.40	6.24	4.31	7.92	6.43	5.57	5.81	4.43	4.70	5.34
24	12/24/2006	11.98	14.02	12.86	9.76	12.08	10.26	10.33	10.73	9.78	9.89	10.33
25	1/6/2007	7.32	10.41	7.20	5.13	12.79	8.92	5.52	4.96	4.78	6.36	8.51
26	1/20/2007	4.45	3.94	4.08	4.80	5.44	4.07	6.74	7.40	6.53	4.49	3.85
27	1/21/2007	5.43	10.83	8.71	5.33	11.44	8.69	5.94	5.22	4.84	5.56	6.88
28	2/3/2007	5.53	8.24	6.15	4.37	10.42	7.83	5.37	5.16	4.15	4.96	5.68
29	2/11/2007	8.61	12.56	9.65	7.11	12.53	11.01	8.43	7.73	6.14	8.04	8.39
30	2/17/2007	9.23	11.05	8.50	7.79	13.39	10.06	8.17	8.24	8.12	8.17	9.38
31	2/24/2007	7.05	8.85	7.92	5.51	10.94	10.83	7.07	6.96	6.09	6.22	6.47
32	3/3/2007	6.67	8.81	7.73	5.18	12.21	8.10	5.66	5.49	6.20	6.05	6.80
33	3/10/2007	7.91	10.22	8.08	5.27	13.59	9.03	6.34	5.93	5.39	6.37	7.23
34	3/17/2007	8.85	11.90	8.79	5.73	15.46	10.41	6.48	5.93	5.55	6.69	8.00
35	3/31/2007	8.29	11.11	8.21	6.37	15.68	10.79	7.33	6.90	5.82	6.99	7.83
36	4/7/2007	4.85	7.72	5.36	3.57	10.57	7.80	3.83	3.79	3.92	4.60	5.44
37	4/8/2007	7.07	10.12	8.26	5.53	11.75	9.96	6.26	6.01	5.01	5.02	6.62
38	4/14/2007	7.02	9.26	7.95	4.57	13.10	10.37	5.20	4.89	4.37	6.05	7.00
39	5/12/2007	5.99	7.88	6.84	3.90	10.60	7.58	4.99	4.91	3.43	4.46	5.82
40	6/10/2007	9.20	14.10	9.55	7.16	16.79	12.70	7.68	6.81	6.78	8.90	9.48

Figure L.10 RCBD for validation data sets in characterization 16

APPENDIX M. ANOVA AND MULTIPLE COMPARISONS FOR RCBDS IN S-PLUS

After ten RCBDs for validation data sets in ten characterizations are constructed, as shown in Appendix M, ANOVA and multiple comparisons are performed for each RCBD in S-PLUS, to compare the prediction abilities of ten characterization regression trees model and full regression tree model.

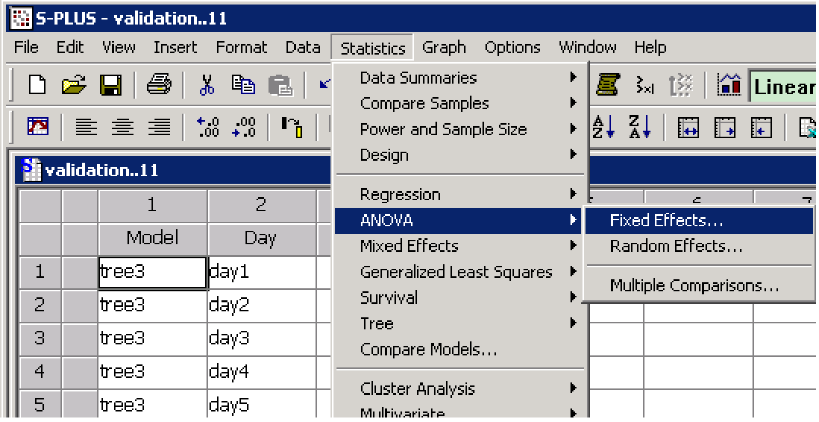
Before importing ten RCBDs into S-PLUS to perform ANOVA and multiple comparisons, the RCBDs need to be adjusted to make sure the data sets are compatible in S-PLUS. For example, the RCBD for validation data set in characterization 11, as shown in Figure L.7 in Appendix L, needs to be adjusted as shown in Table M.1. In Table M.1, there are three columns, “Model” referring to which regression tree model is used, “Day” referring to which daily validation data set is used and “MSE” referring to what the MSE is for the certain “Model” and “Day”. Due to space limitations, Table M.1 only shows the MSEs for regression tree models representing characterization 3, 4 and 5 to predict the eleven daily validation data sets in characterization 11. Clearly, the first eleven rows of MSEs in Table M.1 are just the first column of MSEs in Figure L.7, the second eleven rows of MSEs in Table M.1 are the second column of MSEs in Figure L.7, etc.

Table M.1 Adjusted data set of RCBD for characterization 11

Model	Day	MSE
tree3	day1	6.12
tree3	day2	5.29
tree3	day3	5.86
tree3	day4	8.45
tree3	day5	5.08
tree3	day6	46.69
tree3	day7	4.88
tree3	day8	5.95
tree3	day9	9.36
tree3	day10	7.91
tree3	day11	5.33
tree4	day1	7.18
tree4	day2	6.62
tree4	day3	8.39
tree4	day4	8.75
tree4	day5	4.95
tree4	day6	52.49
tree4	day7	7.32
tree4	day8	8.76
tree4	day9	12.60
tree4	day10	10.22
tree4	day11	9.03
tree5	day1	244.79
tree5	day2	65.65
tree5	day3	155.10
tree5	day4	1490.20
tree5	day5	227.89
tree5	day6	768.83
tree5	day7	600.48
tree5	day8	Model
tree5	day9	54.07
tree5	day10	396.46
tree5	day11	21.44

After all the ten RCBDs are adjusted into the data sets in the manner shown in Table M.1, these ten data sets containing ten RCBDs can be imported into S-PLUS. The following will explain how to perform ANOVA and multiple comparisons for the ten RCBDs, in which the data set containing RCBD for validation data sets in characterization 11 is still used as an example.

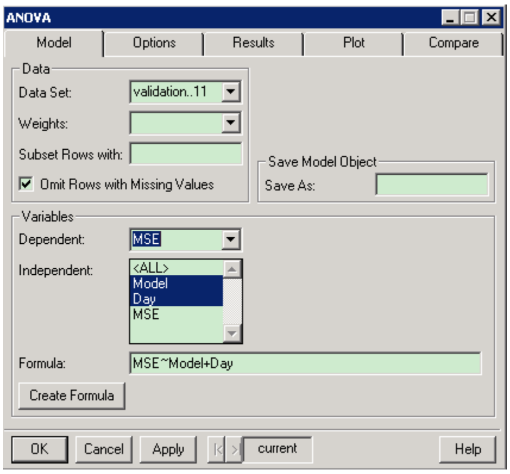
After importing the data set containing RCBD for validation data sets in characterization 11 into S-PLUS, by clicking Statistics>ANOVA>Fixed Effects, as shown in Figure M.1, the ANOVA window for fixed effects is opened.



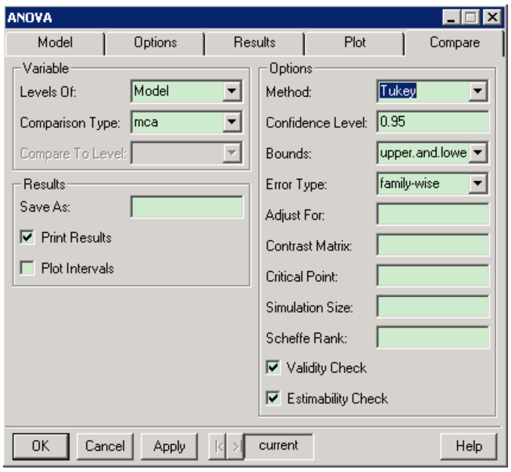
	1	2
	Model	Day
1	tree3	day1
2	tree3	day2
3	tree3	day3
4	tree3	day4
5	tree3	day5

Figure M.1 Open ANOVA window

In the opened ANOVA window, there are five tabs—Model, Options, Results, Plot and Compare, in which only "Model" tab, as shown in Figure M.2, and "Compare" tab, as shown in Figure M.3, need to be used.



Data Set: validation..11
Weights:
Subset Rows with:
☑ Omit Rows with Missing Values
Save As:
Dependent: MSE
Independent: <ALL> Model Day MSE
Formula: MSE~Model+Day



Levels Of: Model
Comparison Type: mca
Compare To Level:
Save As:
☑ Print Results
☐ Plot Intervals
Method: Tukey
Confidence Level: 0.95
Bounds: upper.and.lowe
Error Type: family-wise
Adjust For:
Contrast Matrix:
Critical Point:
Simulation Size:
Scheffe Rank:
☑ Validity Check
☑ Estimability Check

Figure M.2 "Model" tab in ANOVA Figure M.3 "Compare" tab in ANOVA

In “Model” tab, dependent and independent variables need to be selected. In RCBD, MSE is the response variable, Model is the single factor with 11 levels and Day is the day block using daily validation data sets. Thus, the dependent and independent variables are selected as shown in Figure M.2. “Compare” tab is used for multiple comparisons, in which only “Levels Of” in Variable section in the top left corner and “Method” and “Error Type” in Options section in the right need to be appropriately selected. Since the purpose of multiple comparisons is to compare the prediction abilities of speed/travel time of characterization regression tree models and full model, in Variable section, Model needs to be selected for “Levels Of”, as shown in Figure M.3. For comparison method in Options section, the conservative method Tukey’s method is used for the multiple comparisons in all RCBDs except the RCBD for characterization 14, in which a less conservative method, Fisher LSD, is used. For the Error Type, family-wise needs to be selected for Tukey’s method, while comparison-wise needs to be selected for Fisher LSD method.

After all the options are appropriately selected as shown in Figure M.2 and Figure M.3 for RCBD of characterization 11, by clicking OK in ANOVA window, the following result is shown, in which the first part is the results for ANOVA and the second part is the results for multiple comparisons using Tukey’s method. As shown in the ANOVA results, the p-value for Model is 2.815096e-0102, which means that there is significant difference among all the eleven regression tree models to predict validation data sets in characterization 11. In the multiple comparison results using Tukey’s method, any comparison pair flagged by “****” means that that pair of regression tree models are significantly different.

```
*** Analysis of Variance Model ***

Short Output:
Call:
   aov(formula = MSE ~ Model + Day, data = validation..11, na.action = na.exclude
    )

Terms:
               Model     Day Residuals
 Sum of Squares 3347373 4441787   3777971
Deg. of Freedom      10      10       100

Residual standard error: 194.37
Estimated effects are balanced
```

```
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
    Model  10   3347373 334737.3  8.86024 2.815096e-010
      Day  10   4441787 444178.7 11.75707 3.804000e-013
Residuals 100   3777971  37779.7


95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 3.2945
response variable: MSE

intervals excluding 0 are flagged by '****'

                Estimate Std.Error Lower Bound Upper Bound
full tree-tree11   -9.410     82.9     -282.0      264.0
full tree-tree12   -3.040     82.9     -276.0      270.0
full tree-tree14 -350.000     82.9     -623.0      -76.7 ****
full tree-tree16 -170.000     82.9     -443.0      103.0
 full tree-tree3   -0.375     82.9     -273.0      273.0
 full tree-tree4   -2.690     82.9     -276.0      270.0
 full tree-tree5 -385.000     82.9     -658.0     -112.0 ****
 full tree-tree6 -346.000     82.9     -619.0      -72.7 ****
 full tree-tree7 -351.000     82.9     -624.0      -77.9 ****
 full tree-tree8 -322.000     82.9     -595.0      -49.0 ****
   tree11-tree12    6.370     82.9     -267.0      279.0
   tree11-tree14 -340.000     82.9     -613.0      -67.3 ****
   tree11-tree16 -161.000     82.9     -434.0      112.0
    tree11-tree3    9.040     82.9     -264.0      282.0
    tree11-tree4    6.730     82.9     -266.0      280.0
    tree11-tree5 -376.000     82.9     -649.0     -103.0 ****
    tree11-tree6 -336.000     82.9     -609.0      -63.3 ****
    tree11-tree7 -342.000     82.9     -615.0      -68.5 ****
    tree11-tree8 -313.000     82.9     -586.0      -39.6 ****
   tree12-tree14 -347.000     82.9     -620.0      -73.7 ****
   tree12-tree16 -167.000     82.9     -440.0      106.0
    tree12-tree3    2.670     82.9     -270.0      276.0
    tree12-tree4    0.360     82.9     -273.0      273.0
             Estimate Std.Error Lower Bound Upper Bound
 tree12-tree5 -382.000     82.9     -655.0     -109.0 ****
 tree12-tree6 -343.000     82.9     -616.0      -69.7 ****
 tree12-tree7 -348.000     82.9     -621.0      -74.9 ****
 tree12-tree8 -319.000     82.9     -592.0      -46.0 ****
tree14-tree16  180.000     82.9      -93.4      453.0
 tree14-tree3  349.000     82.9       76.3      622.0 ****
 tree14-tree4  347.000     82.9       74.0      620.0 ****
 tree14-tree5  -35.500     82.9     -309.0      238.0
 tree14-tree6    3.970     82.9     -269.0      277.0
 tree14-tree7   -1.200     82.9     -274.0      272.0
 tree14-tree8   27.700     82.9     -245.0      301.0
 tree16-tree3  170.000     82.9     -103.0      443.0
 tree16-tree4  167.000     82.9     -106.0      440.0
 tree16-tree5 -215.000     82.9     -488.0       57.9
 tree16-tree6 -176.000     82.9     -449.0       97.4
 tree16-tree7 -181.000     82.9     -454.0       92.2
 tree16-tree8 -152.000     82.9     -425.0      121.0
  tree3-tree4   -2.310     82.9     -275.0      271.0
  tree3-tree5 -385.000     82.9     -658.0     -112.0 ****
  tree3-tree6 -345.000     82.9     -618.0      -72.4 ****
  tree3-tree7 -351.000     82.9     -624.0      -77.5 ****
  tree3-tree8 -322.000     82.9     -595.0      -48.7 ****
  tree4-tree5 -383.000     82.9     -656.0     -110.0 ****
            Estimate Std.Error Lower Bound Upper Bound
tree4-tree6 -343.000     82.9     -616.0      -70.1 ****
tree4-tree7 -348.000     82.9     -621.0      -75.2 ****
```

```
tree4-tree8 -319.000     82.9     -592.0      -46.4 ****
tree5-tree6   39.400     82.9     -234.0      312.0
tree5-tree7   34.300     82.9     -239.0      307.0
tree5-tree8   63.100     82.9     -210.0      336.0
tree6-tree7   -5.170     82.9     -278.0      268.0
tree6-tree8   23.700     82.9     -249.0      297.0
tree7-tree8   28.900     82.9     -244.0      302.0
```