DYNAMIC TRAVEL TIME ESTIMATION USING REGRESSION TREES

Final Report

SPR 304-351

Oregon Department of Transportation

DYNAMIC TRAVEL TIME ESTIMATION USING REGRESSION TREES

Final Report

PROJECT 304-351

by

Rasaratnam Logendran
and
Lijuan Wang
Oregon State University
School of Mechanical, Industrial, and Manufacturing Engineering
204 Rogers Hall
Corvallis, OR 97331-6001

for

Oregon Department of Transportation
Research Unit
200 Hawthorne Ave. SE, Suite B-240
Salem OR 97301-5192

and

Federal Highway Administration
400 Seventh Street, SW
Washington, DC 20590-0003

October 2008

Technical Report Documentation Page

1. Report No.
FHWA-OR-RD-09-09

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle

Dynamic Travel Time Estimation Using Regression Trees

5. Report Date
October 2008

6. Performing Organization Code

7. Author(s)

Rasaratnam Logendran and Lijuan Wang

8. Performing Organization Report No.

9. Performing Organization Name and Address

Oregon Department of Transportation
Research Unit
200 Hawthorne Ave. SE, Suite B-240
Salem, OR 97301-5192

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

SPR 304-351

12. Sponsoring Agency Name and Address

Oregon Department of Transportation
Research Unit
200 Hawthorne Ave. SE, Suite B-240
Salem, OR 97301-5192

and

Federal Highway Administration
400 Seventh Street, SW
Washington, DC 20590-0003

13. Type of Report and Period Covered

Final Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report presents a methodology for travel time estimation by using regression trees. The dissemination of travel time information has become crucial for effective traffic management, especially under congested road conditions. In the absence of collected actual observations on travel time, the vehicle speed can be predicted by using regression trees, which in turn is used as a proxy to estimate the travel time. To maintain stable prediction ability in both free flow conditions and near-capacity flow conditions on freeways, the regression tree model developed for this study includes thirteen explanatory variables, categorized in four variable types: traffic flow, incident related, weather data, and time of day. A total of four characterization standards (outliers, weather, incidents, and weekday/weekend) are used to characterize the daily traffic data sets to determine the best regression tree model(s) to predict a day in certain characterization. The results show that not only do the regression tree models have accurate prediction ability of vehicle speed and promising ability to estimate travel time, but also the regression tree models built upon other characterizations are preferred to predict a certain characterization. The loop-detector data on PORTAL (Portland Oregon Regional Transportation Archive Listing) system, for the I5-I205 loop in Portland, Oregon, is used to demonstrate the applicability of regression trees in this report.

17. Key Words

REGRESSION TREES, TRAVEL TIME, MID-POINT ALGORITHM

18. Distribution Statement

Copies available from NTIS, and online at
http://www.oregon.gov/ODOT/TD/TP_RES/

19. Security Classification (of this report)
Unclassified

20. Security Classification (of this page)
Unclassified

21. No. of Pages
115

22. Price

Reproduction of completed page authorized Printed on recycled paper

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS					APPROXIMATE CONVERSIONS FROM SI UNITS				
Symbol	When You Know	Multiply By	To Find	Symbol	Symbol	When You Know	Multiply By	To Find	Symbol
LENGTH					**LENGTH**				
in	inches	25.4	millimeters	mm	mm	millimeters	0.039	inches	in
ft	feet	0.305	meters	m	m	meters	3.28	feet	ft
yd	yards	0.914	meters	m	m	meters	1.09	yards	yd
mi	miles	1.61	kilometers	km	km	kilometers	0.621	miles	mi
AREA					**AREA**				
in^2	square inches	645.2	millimeters squared	mm^2	mm^2	millimeters squared	0.0016	square inches	in^2
ft^2	square feet	0.093	meters squared	m^2	m^2	meters squared	10.764	square feet	ft^2
yd^2	square yards	0.836	meters squared	m^2	m^2	meters squared	1.196	square yards	yd^2
ac	acres	0.405	hectares	ha	ha	hectares	2.47	acres	ac
mi^2	square miles	2.59	kilometers squared	km^2	km^2	kilometers squared	0.386	square miles	mi^2
VOLUME					**VOLUME**				
fl oz	fluid ounces	29.57	milliliters	ml	ml	milliliters	0.034	fluid ounces	fl oz
gal	gallons	3.785	liters	L	L	liters	0.264	gallons	gal
ft^3	cubic feet	0.028	meters cubed	m^3	m^3	meters cubed	35.315	cubic feet	ft^3
yd^3	cubic yards	0.765	meters cubed	m^3	m^3	meters cubed	1.308	cubic yards	yd^3
NOTE: Volumes greater than 1000 L shall be shown in m^3.									
MASS					**MASS**				
oz	ounces	28.35	grams	g	g	grams	0.035	ounces	oz
lb	pounds	0.454	kilograms	kg	kg	kilograms	2.205	pounds	lb
T	short tons (2000 lb)	0.907	megagrams	Mg	Mg	megagrams	1.102	short tons (2000 lb)	T
TEMPERATURE (exact)					**TEMPERATURE (exact)**				
°F	Fahrenheit	(F-32)/1.8	Celsius	°C	°C	Celsius	1.8C+32	Fahrenheit	°F

*SI is the symbol for the International System of Measurement

ACKNOWLEDGEMENTS

We thank Dr. Kristin Tufte with Portland State University for her assistance with questions on data made available on PORTAL (Portland Transportation Archive Listing) system. We thank Mr. Geoffrey Boyer with the Oregon Department of Transportation (ODOT) for his assistance with questions on incident data. Thanks are also due to the members of the Technical Advisory Committee (TAC) from ODOT for their feedback and suggestions throughout the course of the project.

DISCLAIMER

This document is disseminated under the sponsorship of the Oregon Department of Transportation and the United States Department of Transportation in the interest of information exchange. The State of Oregon and the United States Government assume no liability of its contents or use thereof.

The contents of this report reflect the view of the authors who are solely responsible for the facts and accuracy of the material presented. The contents do not necessarily reflect the official views of the Oregon Department of Transportation or the United States Department of Transportation.

The State of Oregon and the United States Government do not endorse products of manufacturers. Trademarks or manufacturers' names appear herein only because they are considered essential to the object of this document.

This report does not constitute a standard, specification, or regulation.

TABLE OF CONTENTS

APPENDICIES

TABLE OF TABLES

APPENDICES TABLES

TABLE OF FIGURES

APPENDICES FIGURES

1.0 INTRODUCTION

Travel time estimation is of increasing importance to the real time travelers' information and route guidance system. Travel time estimation provides valuable information for traveler routing and transportation scheduling. Therefore, the accuracy of travel time estimates has become a high priority. Various techniques and methodologies have been used for estimating travel time on freeways, such as probe vehicles, automatic vehicle identification (AVI) through the use of license plate matching, video detection, artificial neural networks (ANNs), etc. This report examines the use of regression trees in estimating travel time in I5-I205 loop in the Portland Metro area, Oregon. All of the data used in this report are collected from PORTAL (Portland Oregon Regional Transportation Archive Listing) system managed by Portland State University.

A lot of research has been performed on travel time estimation that provides different perspectives. Using prediction methodologies, various time series models (*Al-Deek et al., 1998; Anderson et al., 1994*) and artificial neural network models (*Park et al., 1998; Rilett and Park, 1999*) have been developed as indirect methods for travel time estimation. With regards to input data source, most of these studies used collected traffic data such as volume, occupancy, and vehicle speed to calculate travel time as a function of these parameters. However, the relationships among these parameters might not be valid during near-capacity flow conditions (*Chen and Chien, 2001*). Since all of the above methods relied on the traffic flow data from loop detectors, missing data, and a relatively large amount of outliers, caused by detector errors, it may be infeasible to construct a proper model effectively (*Lee et al., 2006*). Moreover, travel time can be affected by various factors other than speed, volume and occupancy, such as geometric conditions, speed limit, incidents, vehicle composition, weather condition, etc.

In some investigations, travel time data are obtained directly through various sources, such as loop detectors, microwave detectors, radar, etc. However, in reality the highway network is not always covered by such data collection devices. Consequently, probe vehicles, as mobile detectors, are considered as a valuable source of real-time travel time data, if the appropriate probe percentage and the report frequency are applied to ensure reliable travel time estimation (*Chen and Chien, 2000; Sen et al., 1997*). The use of probe reports, as real-time observation, could cause the variance of observations in each time period to vary. For a given probe percentage (e.g., 1%), larger variance of travel times reported by probe vehicles are expected when traffic volume approaches capacity, which would result in larger prediction errors (*Chen and Chien, 2001*).

In this research, regression tree analysis is employed to estimate travel time by using speed as a proxy. Because the regression tree model needs to be built on test data, in this instance, historical data, which is not available in PORTAL system, the regression tree model has been built to predict speed first and then the standard mid-point algorithm is used to estimate travel time.

Regression tree was first introduced by Breiman et al. (*1984*). 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 quite a few studies related to traffic security and accident analysis (*Golias and Karlaftis, 2001; Karlaftis and Golias, 2002; Chang and Chen, 2005; Chang and Wang, 2006*). Golias and Karlaftis (*2001*) applied hierarchical tree-based regression (HTBR), also known as regression trees, to identify which external factors affect the related aspects of self-reported driver behavior and found that regression tree is 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 also concluded that HTBR (non-parametric model) without any assumption of functional form of the model has both theoretical and applied advantages over multiple linear and negative binomial regression models (parametric models) in analyzing highway accident rates. Chang and Chen (*2005*) proposed using 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 the risk factors that can influence the 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 described above, 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. Developed tree models can explain various relationships between variables without sacrificing the prediction accuracy, which is really needed in building models for travel time estimation. Therefore, it is very promising that the regression tree method can overcome the limitation of 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. These variables considered for travel time estimation include traffic flow variables, weather variables, incident variables and time of day variable.

The regression tree models for this study are built based on the daily historical data sets, including not only the traffic flow variables but also the incident related variables, weather data variables and time of day. This ensures the models to maintain stable prediction ability among different flow conditions on freeways. Because the actual historical travel time data is not available in PORTAL system, the regression tree model is 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 set up to track the characteristics of both test data sets and validation data sets. The prediction abilities (the accuracy of the predicted speeds) among 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 LSD method (*Montgomery, 2005*). Regression tree analysis, RCBD and multiple comparisons are performed by use of the statistical software package S-PLUS.

Section 2 introduces the regression tree methodology and how a regression tree model is constructed. Section 3 describes the regression tree model constructed for speed prediction, including data collection and raw data reorganizations, and the implementation of regression trees in S-PLUS. Validation of the constructed regression tree model is illustrated in section 4. Then characterization approach is introduced in section 5, followed by experimental design for comparing the prediction abilities among characterization regression tree models and the full regression tree model in section 6 and analysis of results in section 7. Finally, this report closes with a brief discussion and conclusions.

2.0 REGRESSION TREE METHODOLOGY

Regression tree was first introduced by Breiman *et al.* (*1984*) in their classic text on Classification and Regression Trees. The regression tree-based model and algorithm are somewhat intertwined. Regression 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. The data set used to split to construct the regression tree model is called test data, while the data used to feed in the regression tree model for prediction purpose is called validation data. The regression tree algorithm can be further explained by using the following example.

In the test data shown in Table 2.1, speed is the response variable, while volume and occupancy are two explanatory variables. To construct a regression tree model using the above described procedure, we can start assessing any explanatory variable, i.e. volume or occupancy in this case. Starting with volume as an explanatory variable, for example, the assessment steps can be documented as follows:

- ♦ Select a threshold value; say 306, of the explanatory variable volume (the vertical dotted line in Figure 2.1).
- ♦ 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 2.1).

Table 2.1: 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

Speed	Volume	Occupancy
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

- 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 threshold selected in step 1 (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.

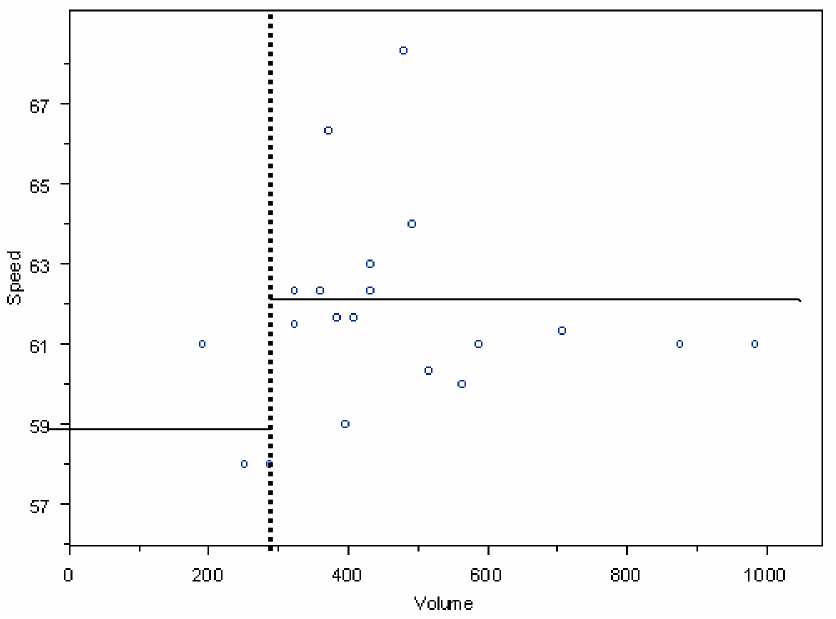


Figure 2.1: Example of Splitting the Test Data Starting from Volume

3.0 REGRESSION TREE MODEL DEVELOPMENT AND ALGORITHM IMPLEMENTATION IN S-PLUS

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

3.1 REGRESSION TREE MODEL DEVELOPMENT

Although the test data can be collected by lengths of time, such as one day, three days, a week, etc., for this study, the test data was collected on a daily basis, which is the shortest time period, in order to better track the traffic pattern at a station. In the test data set for constructing the regression tree model, the response variable was speed, and all explanatory variables considered were classified into four types: traffic flow variables, incident related variables, weather data variables and time of day variable. By using I-205 NB Gladstone as an example of station, we demonstrate the formation of the test data collected at this station on a certain day, say March 23rd, 2005.

3.1.1 Data collection

3.1.1.1 Traffic flow variables

Speed, volume and occupancy were collected on a daily basis in 5-minute increments, which is the smallest time increment possible in PORTAL system to collect traffic flow data in order to track traffic patterns in regression tree model construction more accurately. The traffic flow data at the station I-205 NB Gladstone on March 23rd, 2005, was collected as shown in Appendix A and a part of the collected raw volume data from 9:10 to 10:10 am is shown in Table 3.1 in the interest of space.

Table 3.1: Raw Volume Data at Station I-205 NB Gladstone on 03/23/05 (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

3.1.1.2 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; affected lanes by incident (such as right lanes, left lanes); number of affected lanes; hazard materials (hazmat) and number of fatalities were considered to track the impact of incidents on speed/travel time in the model for this study. The process used to collect the incident data for these seven variables is shown in Appendix B. The raw incident data at the station I-205 NB Gladstone on March 23rd, 2005 collected from PORTAL system is shown in Table 3.2.

Table 3.2: Incident Data at the Station I-205 NB Gladstone on March 23rd, 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

3.1.1.3 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 PORTAL system was found to be incomplete and also because extreme temperature conditions do not occur often in I5-I205 loop in the Portland Metro area. The method used to collect the weather data from PORTAL system is shown in Appendix C, with Table 3.3 showing the partial hourly weather data (3:00 – 11:00 am) for the same station I-205 NB Gladstone on March 23rd, 2005.

Table 3.3: 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

3.1.1.4 Time of day variable

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 PORTAL system, 5 minutes, is used in the final test data set.

3.1.2 Raw data reorganizations

After the raw data are collected for the four types of explanatory variables described above, it must be determined how these variables can be expressed in one test data set in order to construct a regression tree model. Therefore, raw data reorganizations need to be performed and are described below. At the same time, since the regression tree algorithm is implemented in S-PLUS, the test data, after raw data reorganization, has to be compatible in S-PLUS. Because raw data reorganizations are needed for every raw daily data set collected, and to save time and increase accuracy, four macros were written in EXCEL Visual Basic Application (VBA) with the purpose of reorganizing daily raw data saved in EXCEL files. This process is further described in Appendix D.

3.1.2.1 Traffic flow variables

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

3.1.2.2 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, other five data items, incident type, affected lanes, number of affected lanes, hazard materials and number of fatalities, are inserted into the final test data set according to the start time and duration. Since the time frame in final test 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. 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. We will use the incident data collected at the station I-205 NB Gladstone on March 23rd, 2005, shown in Table 3.4, as an example to illustrate how to insert the raw data of the seven incident variables into the final test data set.

It is easy to see in Table 3.4 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) is inserted into the test data to start at 9:35 and end at 9:50, as shown in Table 3.4, because the duration of this incident is 14 minutes and the cleared time of 9:49 can be rounded to 9:50.

Table 3.4: 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

3.1.2.3 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 of weather variables into the final test data in 5-minute increments, the weather data needs to be inserted for one hour into all time points in that hour, as shown in Table 3.5 (partial test data at the station I-205 NB Gladstone on March 23rd, 2005 (9:10 – 10:10 am).

Table 3.5: Test Data with Traffic Flow Data, Incident Data and Weather Data

Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Temp	Wind Speed	Rainfall	Visibility
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

3.1.2.4 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 can not be processed by S-PLUS.

3.2 REGRESSION TREE ALGORITHM IMPLEMENTATION IN S-PLUS

The regression tree algorithm implementation in S-PLUS can be described by use of two examples. The first example is the test data set shown in Table 2.1, with speed as a response variable, occupancy and volume as two explanatory variables. The second example is the final test data set after reorganizing the raw data collected at the station I-205 NB Gladstone on January 10th, 2006, with speed as response variable and all the four types of explanatory variables considered in our regression tree model.

3.2.1 Small test data set in Table 2.1

Before applying the test data set in S-PLUS to construct the regression tree model, test data set needs to be imported into S-PLUS first by clicking File>Import Data>From File in S-PLUS as shown in Figure 3.1.

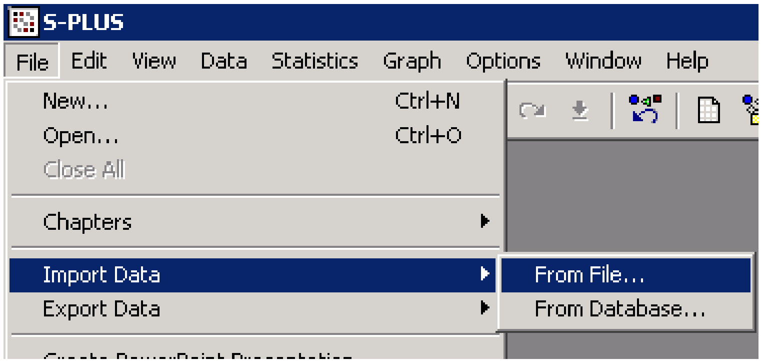
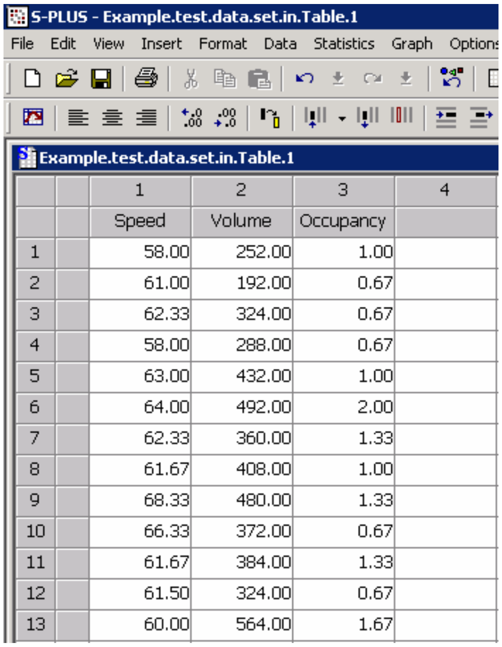


Figure 3.1: Import Test Data Set into S-PLUS

A window titled “Import From File” will appear and the test data file can be imported by selecting “Browse.” After clicking OK, the test data set will appear as shown in Figure 3.2, which means this test data set can now be used to construct the regression tree model in S-PLUS.


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 3.2: Imported Test Data Set in S-PLUS

By clicking Statistics>Tree>Tree Models as shown in Figure 3.3, the window “Tree Models” is opened. When the window “Tree Models” is opened, the first three tabs--Model, Results and Plot--are used to construct the tree model, show the result summary and tree plot, respectively, as shown in Figures 3.4, 3.5 and 3.6

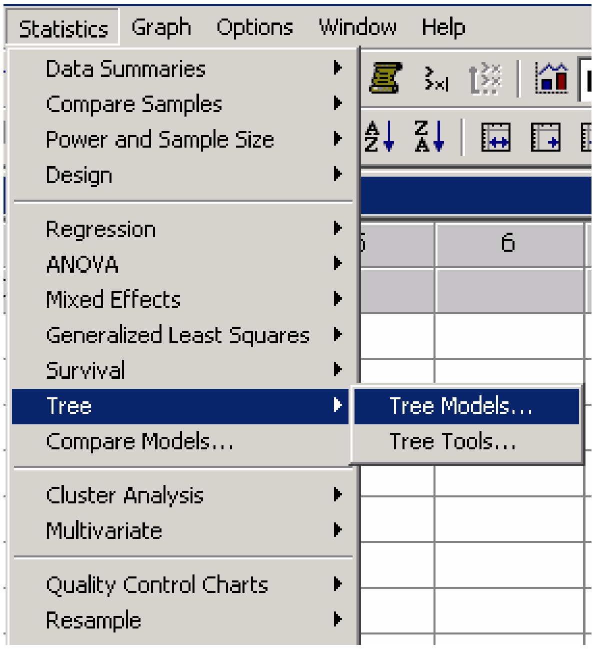
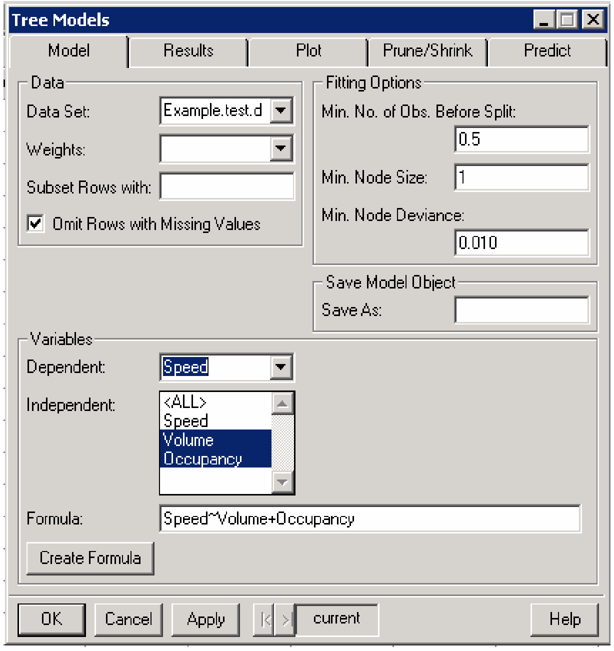


Figure 3.3: Open Tree Models Window



Data Set: Example.test.d

Weights:

Subset Rows with:

☑ Omit Rows with Missing Values

Min. No. of Obs. Before Split: 0.5

Min. Node Size: 1

Min. Node Deviance: 0.010

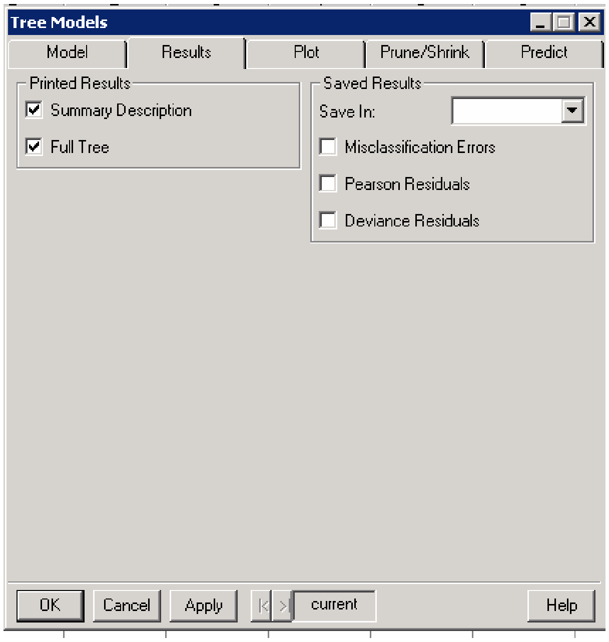
Save As:

Dependent: Speed

Independent: <ALL> Speed Volume Occupancy

Formula: Speed~Volume+Occupancy

Figure 3.4: “Model” Tab



☑ Summary Description

☑ Full Tree

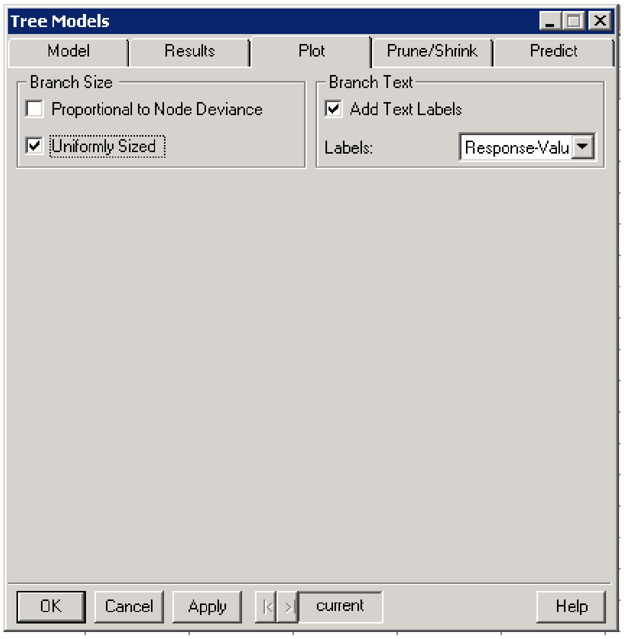
Save In:

☐ Misclassification Errors

☐ Pearson Residuals

☐ Deviance Residuals

Figure 3.5: “Results” Tab



☐ Proportional to Node Deviance

☑ Uniformly Sized

☑ Add Text Labels

Labels: Response-Valu

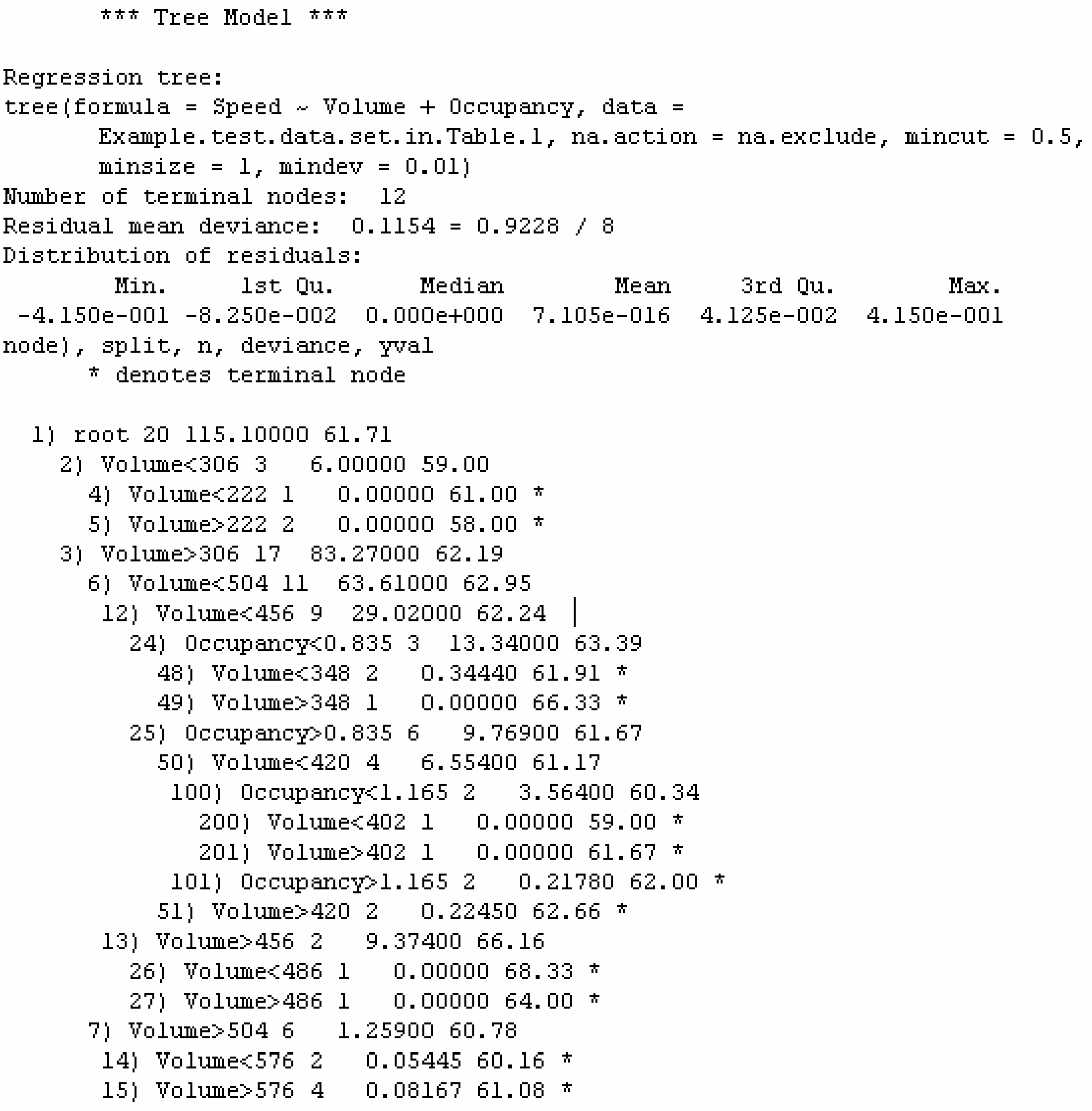
Figure 3.6: “Plot” Tab

There are four sections in the tab “Model,” Data, Fitting Options, Variables and Save Model Object, as shown in Figure 3.4. 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 to set up when to stop the regression tree model construction, that is, stop splitting the test data set. In the “Variables” section, the response variable for the tree model needs to be selected as “Dependent” and the explanatory variables need to be selected as “Independent”.

In the “Results” tab, only the section “Printed Results” needs to be selected for what to view in the results summary as shown in Figure 3.5. Both of the options in this section need to be checked to view the summary description of the regression tree model and the full tree in the results summary, as shown in Figure 3.7 later.

The “Plot” tab as shown in Figure 3.6 is used to decide how to view the plot of the regression tree model based on the selected test data set. 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. Since the node deviances are all shown in the results summary, for tree plot we can just select “Uniformly Sized” 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. For the types of 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” within the window “Tree Models,” click OK. The results summary and the regression tree plot for the model, constructed from the example test data, will then be displayed as shown in Figures 3.7 and 3.8.



```
        *** 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 3.7: Results Summary of the Regression Tree Model

In the results summary shown in Figure 3.7, lines 1 to 9 are summary descriptions of the regression tree model, and lines 10 to 35 show the full tree model, with lines 10 and 11 showing the interpretations of the full tree. In 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 one branch below is 2n and 2n+1, respectively, if the node number of that branch is n. The second term in the line is the split 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 split for the branch with node number 2, is the explanatory variable “Volume” less than 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 equals 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 3.8.

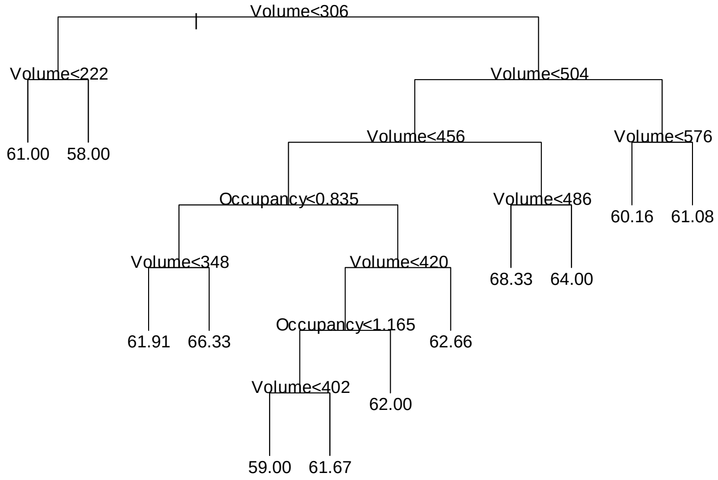


Figure 3.8: Regression Tree Plot of the Example Test Data Set

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 3.8, 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 3.7.

3.2.2 The test data set at the station I-205 NB Gladstone on January 10th, 2006

The final daily test data set after reorganizing the raw data collected at the station I-205 NB Gladstone on January 10th, 2006 is shown partially in Figure 4.1 in the interest of space. In the results summary shown in Figure 4.2, we can see that the formula to construct the regression tree model for this test data set includes all the explanatory variables we proposed in Section 3.1. The tree plot for this regression tree model is shown in Figure 4.3.

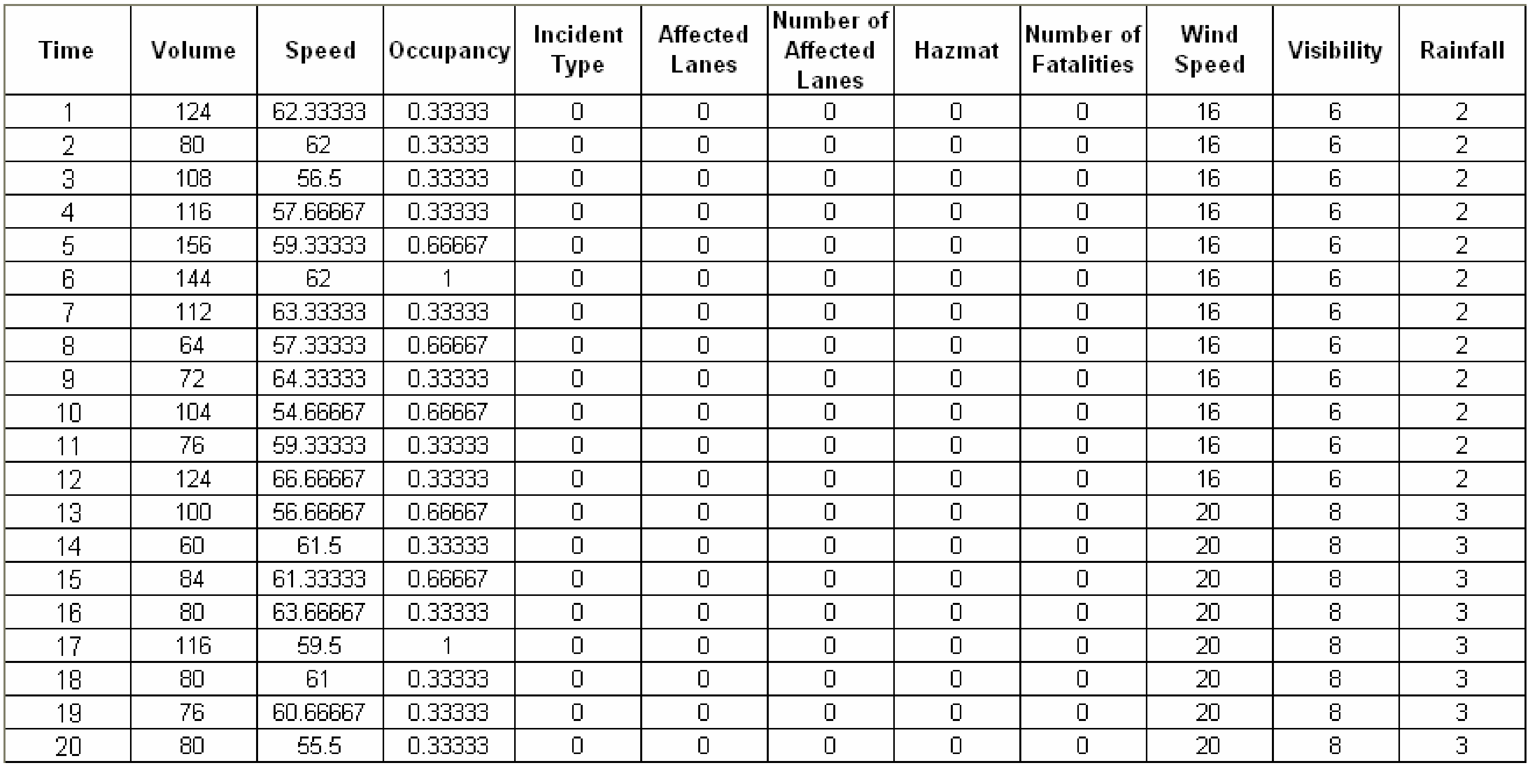
4.0 VALIDATION OF CONSTRUCTED REGRESSION TREE MODEL

After the regression tree model is built on the test data set containing the four types of explanatory variables, the next step is to proceed to regression tree model validation to test the prediction ability of our model. As mentioned earlier, actual travel time data is not available in PORTAL system and speed is used as a proxy to estimate travel time. Therefore, to validate the prediction ability of the regression tree model we built, the model will be used to predict speeds of other daily data sets first by using S-PLUS and the Mean Squared Errors (MSE) will be used to estimate the accuracy of the predicted speeds to the actual speeds. Then predicted speeds will be used to estimate travel time using the Mid-point algorithm, which is also used in PORTAL system to estimate travel time, and the estimated travel time by our regression tree model will be compared with the estimated travel time data stored in PORTAL using the MSEs too.

4.1 REGRESSION TREE MODEL VALIDATION OF SPEED PREDICTION

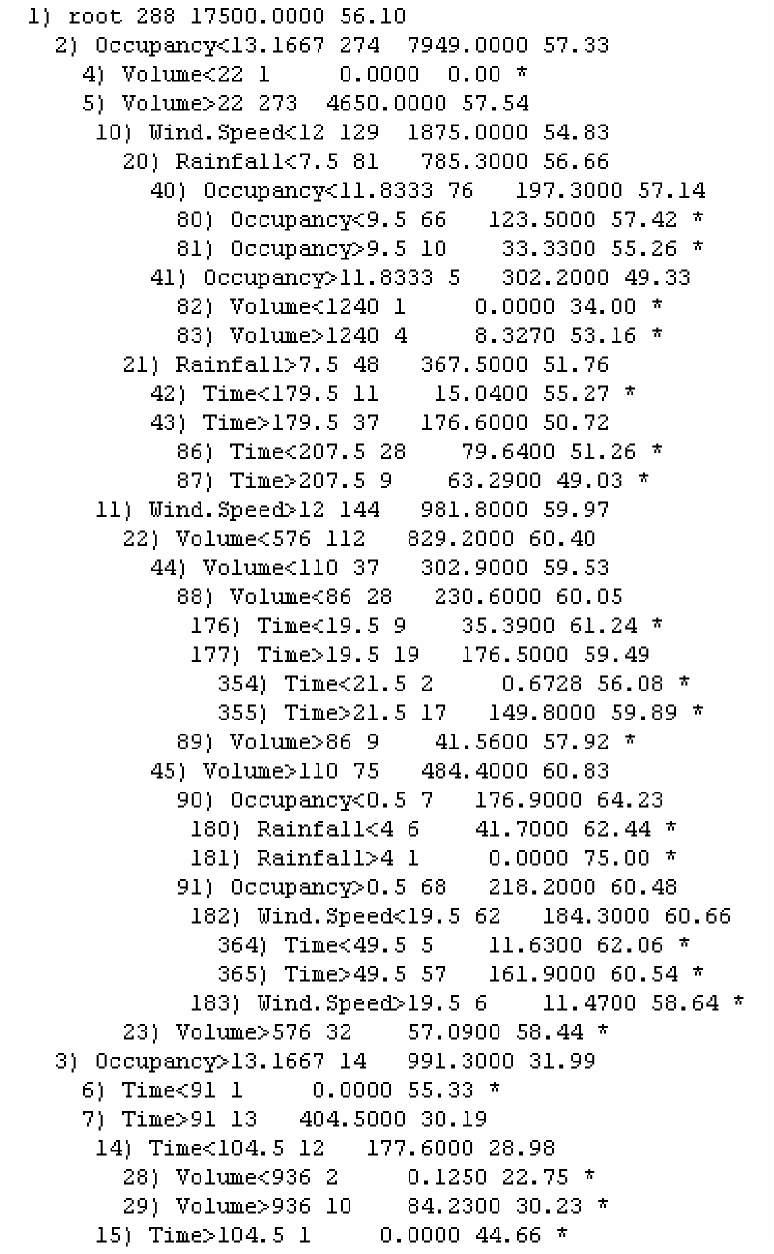
In the previous section, we have shown how the regression tree model is developed and how the regression tree algorithm is implemented in S-PLUS, based on the test data set containing four types of the explanatory variables. The regression tree model validation can also be performed in S-PLUS, which is shown in Appendix E. Here we will use the regression tree, as shown in Figure 3.8, which was built based on the small test data in Table 2.1 in Section 2, and a small validation data, as shown in Table 4.1, to demonstrate the algorithm of the regression tree model validation.

To validate the regression tree model in Figure 3.8 using the validation data in Table 4.1, every row of validation data, including only the data of 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 3.8 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.



Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
1	124	62.33333	0.33333	0	0	0	0	0	16	6	2
2	80	62	0.33333	0	0	0	0	0	16	6	2
3	108	56.5	0.33333	0	0	0	0	0	16	6	2
4	116	57.66667	0.33333	0	0	0	0	0	16	6	2
5	156	59.33333	0.66667	0	0	0	0	0	16	6	2
6	144	62	1	0	0	0	0	0	16	6	2
7	112	63.33333	0.33333	0	0	0	0	0	16	6	2
8	64	57.33333	0.66667	0	0	0	0	0	16	6	2
9	72	64.33333	0.33333	0	0	0	0	0	16	6	2
10	104	54.66667	0.66667	0	0	0	0	0	16	6	2
11	76	59.33333	0.33333	0	0	0	0	0	16	6	2
12	124	66.66667	0.33333	0	0	0	0	0	16	6	2
13	100	56.66667	0.66667	0	0	0	0	0	20	8	3
14	60	61.5	0.33333	0	0	0	0	0	20	8	3
15	84	61.33333	0.66667	0	0	0	0	0	20	8	3
16	80	63.66667	0.33333	0	0	0	0	0	20	8	3
17	116	59.5	1	0	0	0	0	0	20	8	3
18	80	61	0.33333	0	0	0	0	0	20	8	3
19	76	60.66667	0.33333	0	0	0	0	0	20	8	3
20	80	55.5	0.33333	0	0	0	0	0	20	8	3

Figure 4.1: Organized Test Data Set with All the Explanatory Variables Considered at I205 NB Gladstone on January 10th, 2006. (0:00 – 1:35 am)



```
1) root 288 17500.0000 56.10
  2) Occupancy<13.1667 274  7949.0000 57.33
    4) Volume<22 1     0.0000  0.00 *
    5) Volume>22 273  4650.0000 57.54
     10) Wind.Speed<12 129  1875.0000 54.83
       20) Rainfall<7.5 81   785.3000 56.66
         40) Occupancy<11.8333 76   197.3000 57.14
           80) Occupancy<9.5 66   123.5000 57.42 *
           81) Occupancy>9.5 10    33.3300 55.26 *
         41) Occupancy>11.8333 5   302.2000 49.33
           82) Volume<1240 1     0.0000 34.00 *
           83) Volume>1240 4     8.3270 53.16 *
       21) Rainfall>7.5 48   367.5000 51.76
         42) Time<179.5 11    15.0400 55.27 *
         43) Time>179.5 37   176.6000 50.72
           86) Time<207.5 28    79.6400 51.26 *
           87) Time>207.5 9    63.2900 49.03 *
     11) Wind.Speed>12 144   981.8000 59.97
       22) Volume<576 112   829.2000 60.40
         44) Volume<110 37   302.9000 59.53
           88) Volume<86 28   230.6000 60.05
            176) Time<19.5 9    35.3900 61.24 *
            177) Time>19.5 19   176.5000 59.49
              354) Time<21.5 2     0.6728 56.08 *
              355) Time>21.5 17   149.8000 59.89 *
           89) Volume>86 9    41.5600 57.92 *
         45) Volume>110 75   484.4000 60.83
           90) Occupancy<0.5 7   176.9000 64.23
            180) Rainfall<4 6    41.7000 62.44 *
            181) Rainfall>4 1     0.0000 75.00 *
           91) Occupancy>0.5 68   218.2000 60.48
            182) Wind.Speed<19.5 62   184.3000 60.66
              364) Time<49.5 5    11.6300 62.06 *
              365) Time>49.5 57   161.9000 60.54 *
            183) Wind.Speed>19.5 6    11.4700 58.64 *
       23) Volume>576 32    57.0900 58.44 *
  3) Occupancy>13.1667 14   991.3000 31.99
    6) Time<91 1     0.0000 55.33 *
    7) Time>91 13   404.5000 30.19
     14) Time<104.5 12   177.6000 28.98
       28) Volume<936 2     0.1250 22.75 *
       29) Volume>936 10    84.2300 30.23 *
     15) Time>104.5 1     0.0000 44.66 *
```

Figure 4.2: Results Summary for the Test Data Set with All the Explanatory Variables

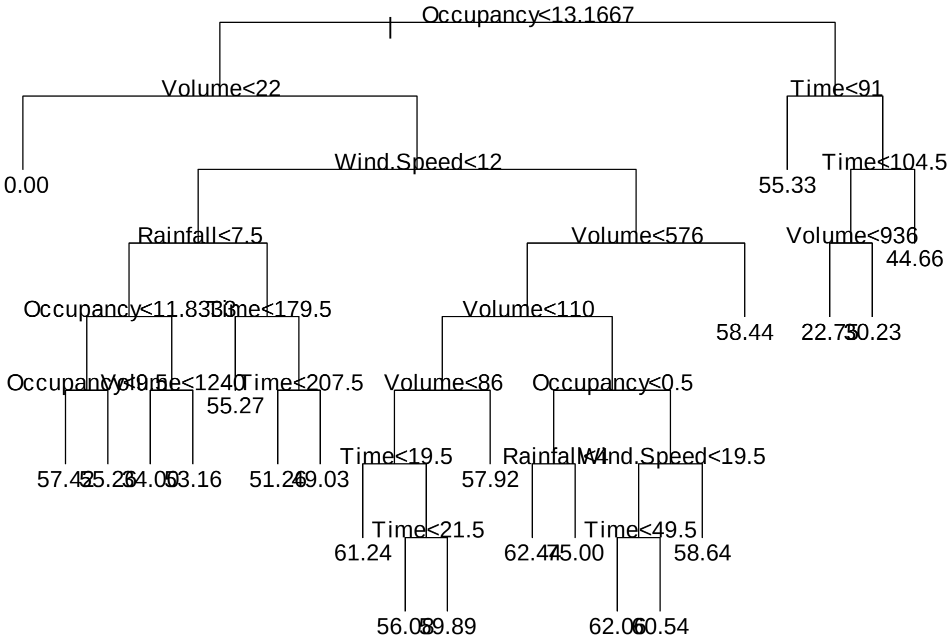


Figure 4.3: Tree Plot of the Regression Tree Model Constructed on the Test Data Set with All the Explanatory Variables

Table 4.1: 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 will be obtained for the validation data set as shown in Table 4.2, which are same as the results given by S-PLUS.

After the fitted speed values are obtained, the MSE is used to evaluate the validation results as also shown in Table 4.2.

Table 4.2: Validation Data Set with Fitted Speed 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 reason why the MSE value shown in Table 4.2 appears high (9.55) is that here the regression tree model is only based on the example test data shown in Table 2.1 including only two explanatory variables, volume and occupancy, instead of the regression tree model including the four types of explanatory variables, which is the model used to predict speed in our research. In contrast, when the regression tree model including four types of explanatory variables is used, the MSE value of predicted speed evaluation is fairly low as shown in Figure E7 in Appendix E.

4.2 REGRESSION TREE MODEL VALIDATION OF TRAVEL TIME ESTIMATION

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

The standard midpoint 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 4.4 (*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 is comparable to the predicted speed by use of regression tree model in the current study.

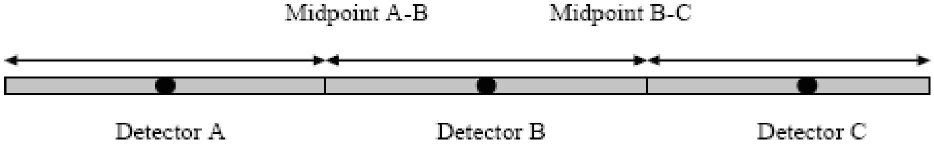
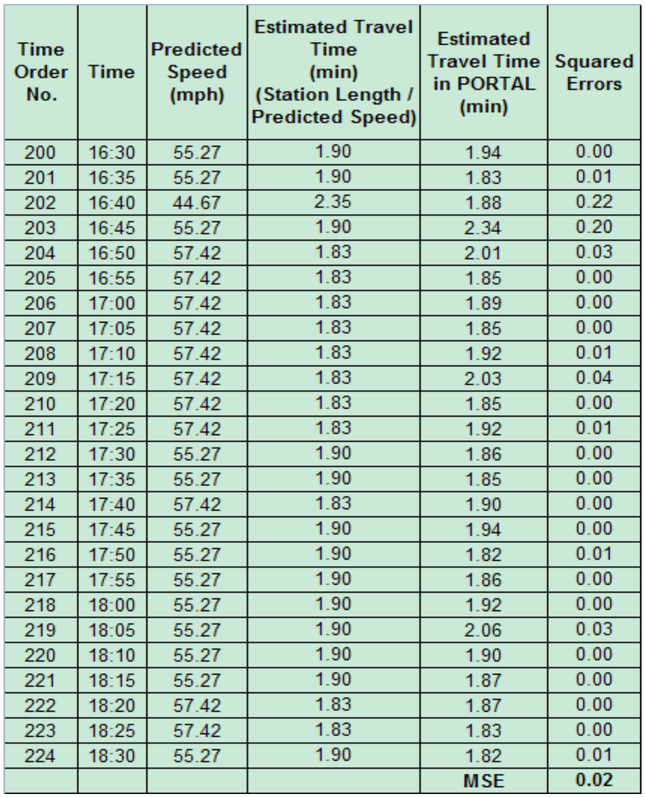


Figure 4.4: Influence Area around the Detector Station

For example, if we obtain the predicted speed between 4:30 and 6:30 pm on August 2nd, 2006 by applying the regression tree model to the test data set for January 10th, 2006 (as shown in Figure 4.3), we can estimate the travel time in this time period and compare that with the estimated travel time data stored in PORTAL. The length of station I-205 NB Gladstone is 1.75 miles on PORTAL system. Therefore, in order to use the Mid-point algorithm to estimate travel time at I-205 NB Gladstone, we need to divide the station length (1.75 miles) by the predicted speeds. The estimated travel time at I-205 NB Gladstone between 4:30 and 6:30 pm on August 2nd, 2006 and the MSE are shown in Figure 4.5. The MSE between the estimated travel time by using the predicted speeds of this study's regression tree model and the estimated travel time in PORTAL is 0.02, which is also fairly low for the travel time errors.



Time Order No.	Time	Predicted Speed (mph)	Estimated Travel Time (min) (Station Length / Predicted Speed)	Estimated Travel Time in PORTAL (min)	Squared Errors
200	16:30	55.27	1.90	1.94	0.00
201	16:35	55.27	1.90	1.83	0.01
202	16:40	44.67	2.35	1.88	0.22
203	16:45	55.27	1.90	2.34	0.20
204	16:50	57.42	1.83	2.01	0.03
205	16:55	57.42	1.83	1.85	0.00
206	17:00	57.42	1.83	1.89	0.00
207	17:05	57.42	1.83	1.85	0.00
208	17:10	57.42	1.83	1.92	0.01
209	17:15	57.42	1.83	2.03	0.04
210	17:20	57.42	1.83	1.85	0.00
211	17:25	57.42	1.83	1.92	0.01
212	17:30	55.27	1.90	1.86	0.00
213	17:35	55.27	1.90	1.85	0.00
214	17:40	57.42	1.83	1.90	0.00
215	17:45	55.27	1.90	1.94	0.00
216	17:50	55.27	1.90	1.82	0.01
217	17:55	55.27	1.90	1.86	0.00
218	18:00	55.27	1.90	1.92	0.00
219	18:05	55.27	1.90	2.06	0.03
220	18:10	55.27	1.90	1.90	0.00
221	18:15	55.27	1.90	1.87	0.00
222	18:20	57.42	1.83	1.87	0.00
223	18:25	57.42	1.83	1.83	0.00
224	18:30	55.27	1.90	1.82	0.01
				MSE	**0.02**

Figure 4.5: MSE Result of the Estimated Travel Times

In the above we have demonstrated how predicted speeds can be used to estimate travel time at a station. Now we will briefly illustrate how the travel time in a segment of highway can be estimated. Since the traffic flow data at 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 time in a segment of highway, we only need to estimate travel time at every station in a segment of highway by using the predicted speed at that station by the regression tree model, and add up the travel time estimates of all the stations in this segment of highway. Then we will have the estimated travel time for the segment of highway.

5.0 CHARACTERIZATION APPROACH

After the daily data sets are collected 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. Four standards are set up to track different 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 of traffic flow data due to detector error.

Preliminary research showed that the regression tree model is not only robust to outliers in the test data sets, but also may have more stable prediction ability than that of test data containing no outliers. For “Good Weather”, based on published sources, a data set was 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 in “Good Weather” is that the temperature data in weather data in PORTAL system is not complete. And it is also because that in Portland Metro area (I5-I205 loop) extreme temperatures is not common. “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. 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 will be distributed.

5.1 DATA COLLECTION AND CHARACTERIZATION RESULTS

To carry out the characterization approach and test the prediction ability of test data sets with different characteristics, data needs to be collected as test data and validation data to build regression tree model and perform regression tree analysis. In Section 4, we have demonstrated that if regression tree model is able to accurately predict speed and then estimate travel time at all stations, similar results can also be obtained for segments of highway. Thus, the following research will focus on stations in I5-I205 loop.

The station I-205 Northbound (NB) Gladstone at milepost 11.05 is randomly selected to collect the daily test data sets and validation data sets. To capture all the characteristics over an entire year in the regression tree models, test data sets are collected by collecting all the daily data sets in 2005. Thus 342 daily test data sets were collected (23 days of data were not complete 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 number 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, the data at the same station, for all of 2006 and the first half of 2007 (1.5 years), were collected as validation data sets, as the weather data and incidents data was not available for the second half of 2007 in the PORTAL system. Thus, 532 daily validation data sets were collected (14 days of data were not complete for unknown reasons). All the data sets were collected manually by copying from the PORTAL system into the Excel files. Raw data, of four types in the same day, are kept in the same excel file and are applied with the four Macros written in Excel Visual Basic Application (VBA) language to organize and adjust the raw data, making sure all daily data sets are suitable for further regression tree analysis, which is mentioned in Section 3 and demonstrated in Appendix D.

For the collected 342 test data sets and the 532 validation data sets, a Macro written in Excel VBA (Appendix F) is used to characterize all these data sets *automatically* with characterization results (Table 5.1).

Table 5.1: 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

There are no test data sets and validation data sets in characterization 9. There are less than or equal to 3 validation data sets in characterizations 1, 2, 10, 13 and 15. Since the test data sets and validation data sets both cover an extended time period, the results show that characterization 1, 2, 9, 10, 13 and 15 are not worthwhile for us to further consider in characterization analysis. Thus, only 10 characterizations are considered: 3, 4, 5, 6, 7, 8, 11, 12, 14 and 16.

5.2 REGRESSION TREE MODELS IN CHARACTERIZATIONS AND FULL MODEL

The ultimate challenge for the research was to determine what kind of regression tree model should be selected to predict speed or estimate travel time for a certain day. Thus, by bringing in the characterization approach, it can be determined if there is a specific characterization of the regression tree based model, capable of better predicting speeds/travel times for existing conditions on the road such as weather, incident, etc. The approach can also be used to determine if the full regression tree model, which contains all the collected daily test data sets, outperforms the regression tree models representing the specific characterizations. Therefore, the prediction abilities for speeds/travel times of 10 characterizations of the regression tree based models and the full regression tree model need to be compared. Before that, the regression tree models, representing characterizations, and the full regression tree model need to be constructed.

To build regression tree model representing a specific characterization, all of the daily test data sets in the same characterization are combined as one test data set in order to construct a regression tree model. For example, to form a test data set representing characterization 3, the 51 daily test data sets in that characterization (Table 5.1) need to be combined into one test data set. The full regression tree model is then built on all the daily test data sets collected. 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 65,536 rows of data, the test data set for the full model only can include 227 daily test data sets (227*288=65,376). These have to be randomly selected from the total of 321 daily test data sets of 10 characterizations.

To make sure these 227 daily test data sets are representative of the 10 characterizations equally in the full model, the same ratio (227/321=70.7%) is used to determine the number of daily test data sets randomly selected out of every characterization. The number of daily test data sets, which is randomly selected from each of the 10 characterizations to construct the full model, is shown in Table 5.2. For example, for characterization 3, 36 daily test data sets need to be randomly selected out of the total 51 to represent characterization 3 in the full model. The random number generation is completed by use of a Macro written in Excel VBA, as shown in Appendix G.

Table 5.2: Number of Test Data Sets to be Randomly Selected for Full Model

Characterization No.	Outliers	Good Weather	Incidents	Weekday or Weekend	Number of Test Data Sets	Number of Test Data Sets to be Randomly Selected
3	Yes	Yes	No	Weekday	51	36
4	Yes	No	No	Weekday	64	45
5	No	Yes	Yes	Weekday	6	4
6	No	Yes	No	Weekday	44	31
7	No	No	Yes	Weekday	8	6
8	No	No	No	Weekday	63	44
11	Yes	Yes	No	Weekend	8	6
12	Yes	No	No	Weekend	7	5
14	No	Yes	No	Weekend	32	23

Characterization No.	Outliers	Good Weather	Incidents	Weekday or Weekend	Number of Test Data Sets	Number of Test Data Sets to be Randomly Selected
16	No	No	No	Weekend	38	27
				Total	321	227

6.0 EXPERIMENTAL DESIGN

After the test data sets for the 10 regression tree models, representing 10 characterizations and the full regression tree model, are imported into S-PLUS, 11 regression tree models in total can be constructed. Then the validation data sets can be imported into S-PLUS to validate the eleven constructed regression tree models, as has been demonstrated in Section 4. The MSEs can be calculated using the actual speeds and the fitted speeds of the validation data sets by the regression tree models. To analyze the MSEs obtained from the validation results, a randomized complete block design (RCBD) is used with a significance level of $\alpha=0.05$.

6.1 INTRODUCTION TO RANDOMIZED COMPLETE BLOCK 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 treatments. In this situation, blocks are the daily validation data sets. Each daily validation data set constitutes a "day" block, in which the validation results (MSEs) of the 11 regression tree models, are kept. Compared with a completely randomized design (CRD), RCBD effectively improves the accuracy of the comparisons among the 11 regression tree models by eliminating the variability among different daily validation data sets.

6.2 RCBD EXPERIMENTAL DESIGN

With the daily validation data sets of 10 characterizations, 10 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 11 levels, because 11 regression tree models are to be compared in RCBD. The response variable is the 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 10 characterizations). Since sample sizes are strongly related to the effect size and the power of the experimental designs, operating characteristic (OC) curves are usually used to determine a reasonable sample size. However, for the study it was difficult to apply OC curves to determine the sample sizes because the number of treatment levels (11) could not be found on the OC curves. Thus, a software program G*Power was used to determine the sample sizes for each RCBD.

G*Power is a general power analysis program developed by Erdfelder, E., Faul, F., and Buchner, A., which can be downloaded from the following website: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/. The use of G*Power to determine sample size is introduced in Appendix H. 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 [6] conventions and a relatively high power test of 95%, a sample size of 407 for each RCBD is required. It means that at least 37 blocks are needed for each RCBD (37*11 = 407). In each of the 10 characterizations, at least 37 daily validation data sets need to be randomly selected for validation and further construct a RCBD. Therefore, for this study 40 daily validation data sets were randomly selected for characterizations 3, 4, 6, 8, 14 and 16, which contain more than 40 daily validation data sets. For characterizations 5, 7, 11 and 12, there are less than 40 daily validation data sets available. Characterization 11 has the least number of daily validation data sets of 11. Therefore, to obtain meaningful conclusions, the effect size needs to be sacrificed (use a higher effect size) in order to reach a higher test power, or the power of the test needs to be sacrificed in order to get a lower effect size. By testing in G*Power, it is found that even for characterization 11, with only 11 daily validation data sets, the large effect size of F-test of 0.40, according to Cohen's conventions of effect size measures and a power of test of 90%, are still guaranteed. Thus, for characterization 5, 7, 11 and 12, we use all the daily validation data sets available in these characterizations in constructing RCBDs. Table 6.1 shows the number of daily validation data sets needed for each of the 10 RCBDs. The program shown in Appendix G can also be used here to randomly select 40 daily validation data sets for characterizations 3, 4, 6, 8, 14 and 16. All validation data sets used for RCBD are imported in S-PLUS to validate the 11 constructed regression tree models, as demonstrated in Section 4.

Table 6.1: Number of Validation Data Sets Used for RCBD

Characterization No.	Outliers	Good Weather	Incidents	Weekday or Weekend	Number of Validation Data Sets	Number of Validation Data Sets 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

RCBD is used to compare both the prediction abilities of speed/travel time of a characterization regression tree model to each of the other characterization regression tree models; and compare each of the 10 characterization regression tree models to the full regression tree model. The response variable used in the RCBD for this study 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, are used to validate 11 different regression tree models, 10 of which represent the 10 characterizations, and one of which represents the full regression tree model. 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 I. The 10 RCBDs constructed using MSEs are shown in Appendix J.

7.0 ANALYSIS OF RESULTS

The study has shown that RCBD can be used to compare both the prediction abilities of speed/travel time of a characterization regression tree model to each of the other characterization regression tree models; and to compare each of the 10 characterization regression tree models to the full regression tree model. Following the construction of 10 RCBDs for validation data sets in 10 characterizations (Appendix J), the analysis of variance (ANOVA) and multiple comparisons are performed for each RCBD in S-PLUS (Appendix K) to analyze the prediction abilities of characterization regression tree models and the full regression tree model. The results of ANOVA and multiple comparisons are shown in Table 7.1.

Table 7.1: Results of ANOVA and Multiple Comparisons for Ten RCBDs

Characterization No. of Validation Data Sets	**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 (barely)	6 (barely)	4, 7, 8	11, 12, 14, 16, 3, 5

*Fisher LSD method is used for multiple comparisons.

The above table shows: the results of ANOVA and comparisons for 10 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 for that characterization is compared with those of regression tree models representing characterizations 4, 5, 6, 7, 8, 11, 12, 14 and 16. The first column in this section shows if there is significant difference among these 10 characterization models with a significance level of α=0.05. Columns 2-4 show 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 for characterization 14, where the Fisher LSD method was used. 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 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 10 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 10 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 portion of Table 7.1 shows the multiple comparison results between the full regression tree model and the regression tree models representing the 10 characterizations. 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 full model and characterization 11, 12, 16, 3 and 4 models, which means these six models are equally good to predict validation data sets in characterization 3.

8.0 CONCLUSIONS

The research reported here focuses on dynamically and accurately estimating travel times in I5-I205 loop in the Portland Metro area of Oregon. To accomplish this, a regression tree methodology was employed, using speed as a proxy for travel time.

Following the introduction of the regression tree methodology, the development of the regression tree model has been demonstrated to accurately predict speed. Four types of explanatory variables, traffic flow variables, incident related variables, weather related variables and time of day variable, are considered in the test data sets for regression tree model construction. This ensures that the regression tree models in this study have the same prediction ability among different flow conditions on a freeway. The collection and reorganization of raw data for these explanatory variables have been described. Four macros, written in Excel VBA, have been developed to increase the efficiency and accuracy of reorganizing the collected raw daily data sets. Following these reorganizations, the daily test data sets are ready to be imported into the statistical software package S-PLUS to build regression tree models. The implementation of a regression tree algorithm in S-PLUS is then illustrated using two test data sets, one of which includes only two explanatory variables, and the other of which is a complete daily test data set collected at a randomly selected station, including all four types of explanatory variables. For the purpose of this study the daily test data collected at the randomly selected station was the daily data set of January 10th, 2006 at I-205 NB Gladstone. By importing the test data sets into S-PLUS, the regression tree model can be constructed using the built-in functions in S-PLUS and the regression tree plot can also be obtained.

The validation of the constructed regression tree models using S-PLUS has then been demonstrated. To compare the predicted speeds from the regression tree models with the actual speeds, the MSEs are used. As described in Section 4, the MSE of predicted speeds using the regression tree model is fairly low, as demonstrated by an example shown in Appendix E. Both the estimation of travel times using predicted speeds as a proxy and Mid-point algorithm and the validation of the estimated travel time have been described in this report. Because historical travel time data are unavailable in PORTAL, and travel time is estimated by using the predicted speed obtained from the regression tree models, the MSEs have to be calculated by comparing the estimated travel time by predicted speed with the estimated travel time data stored in PORTAL. As noted in Figure 4.1 in Section 4, the MSE value for the estimated travel time in our example is fairly low. This estimation is based on the use of predicted speeds by one randomly selected regression tree model. It shows that the regression tree model indeed has promising potential to accurately estimate travel time.

To dynamically estimate travel time for a random day using regression tree models, we have addressed the characterization approach and how it is applied in the regression tree analysis for travel time estimation. A RCBD has been used to compare both the prediction abilities of speed/travel time of a characterization regression tree model to each of the other characterization regression tree models; and to compare each of the 10 characterization regression tree models to the full regression tree model. The analysis of results of RCBD reveals three promising findings:

- 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 are not limited to 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, several characterization regression tree models 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 equally good prediction ability as the characterization regression tree models. The full model covers the test data sets of all characterizations and should have more stable prediction ability. However, our 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 above three 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 randomly collected validation data sets need to be run through the full regression tree model in order to get the predicted values. The average value of the predicted values of all randomly collected 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 different features in the randomly collected validation data sets than the desired day. However, using 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, increasing the accuracy of 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 study, 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 in PORTAL. In the future, if the actual travel time measurements are made available by ODOT, the current regression tree models, which have been demonstrated in this report, can be adjusted to estimate travel time directly. To make the adjustments, we need to first collect the daily travel time data in five-minute increments and incorporate them 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.

9.0 REFERENCES

Al-Deek, H., D'Angelo, M. and Wang, M. (1998) Travel Time Prediction with Non-Linear Time Series. *Proceedings of the ASCE 5th International Conference on Applications of Advanced Technologies in Transportation*, Newport Beach, CA, 317-324.

Anderson, J., Bell, M., Sayers, T., Busch, F. and Heymann, G. (1994) The Short-Term Prediction of Link Travel Time in Signal Controlled Road Networks. *Proceedings of the IFAC/IFORS 7th Symposium on Transportation Systems: Theory and Application of Advanced Technology*, Tianjin, China, 621-626.

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

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. (2000) Determining the Number of Probe Vehicles for Freeway Travel Time Estimation Using Microscopic Simulation. *Transportation Research Record*, (1719), 61-68.

Chen, M. and Chien, S. (2001) Dynamic Freeway Travel Time Prediction with Probe Vehicle Data: Link-based vs. Path-based. *Proceedings of the 80th Transportation Research Board Annual Meeting*, Washington DC, January 7-11, (1768), 157-161.

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

Golias, I. and Karlaftis, M. G. (2001) An International Comparative Study of Self-reported Driver Behavior. *Transportation Research Part F: Psychology and Behavior*, 4(4), 243-256.

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.

Kothuri, S., Tufte, K., Ahn, S. and Bertini, R. (2006) Development of an ITS Data Archive Application for Improving Freeway Travel Time Estimation. *Proceedings of IEEE*, 1263-1268.

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.

Loh, W.-Y. (2002) Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica*, 12, 361-386.

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

Park, D., Rilett, L. and Han, G. (1998) Forecasting Multiple-Period Freeway Link Travel Times Using Neural Networks with Expanded Input Nodes. *Proceedings of the ASCE 5th International Conference on Applications of Advanced Technologies in Transportation*, Newport Beach, CA, 325-332.

Rilett, L. and Park, D. (1999) Direct Forecasting of Freeway Corridor Travel Times Using Spectral Basis Neural Networks. *Presented at the 78th TRB Annual Meeting (CD-ROM)*, Washington, DC.

Sen, A., Thakuriah, P., Zhu, X. and Karr, A. (1997) Frequency of Probe Reports and Variance of Travel Time Estimates. *ASCE Journal of Transportation Engineering*, 123(4), 290-297.

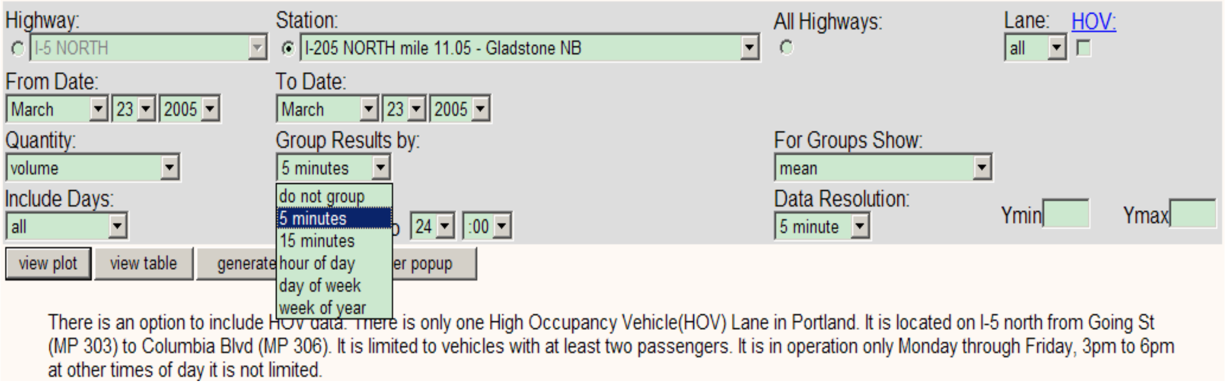
APPENDIX A:
DATA COLLECTION FOR TRAFFIC FLOW VARIABLES IN PORTAL

In free flow condition, only the traffic flow data needs to be considered in the test data. The following traffic flow data shown in Table A1 was collected at the station I-205 NB Gladstone on March 23rd, 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 A1, which is the screen shot taken from 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 PORTAL System for Traffic Flow Data Collection

After clicking the archive "Grouped Data" on the homepage of PORTAL system, the screen like that shown in Figure A1 can be seen. Different stations or segments of highway can be selected in Station or Highway as shown in Figure A1. 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, we only need to select volume, speed and then occupancy 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 A1, the volume data in Table A2 is obtained.

Table A-2: Raw Volume Data at Station I-205 NB Gladstone on 03/23/05. (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

Similarly, speed and occupancy data at the station I-205 NB Gladstone on March 23rd, 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 A1.

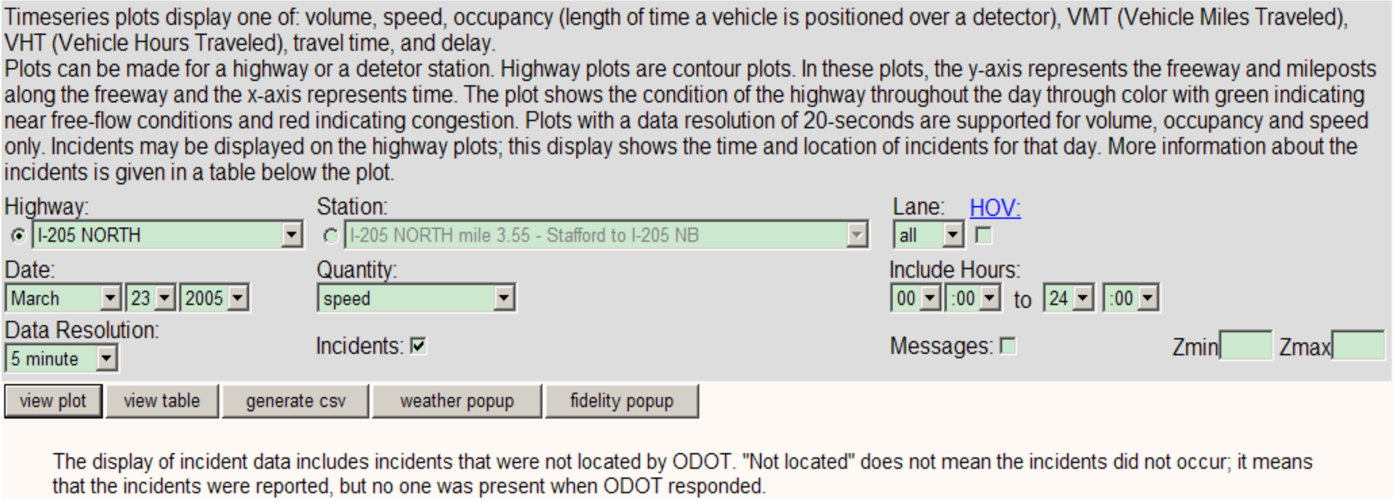
APPENDIX B:
DATA COLLECTION FOR INCIDENT RELATED VARIABLES IN PORTAL

An incident would typically result in a reduced speed between detector stations on the I-5/I-205 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 for us to comprehensively analyze the impact of incident data on the traffic flow. The raw incident data at the station I-205 NB Gladstone on March 23rd, 2005 collected from PORTAL system is shown in Table B1.

Table B-1: Incident Data at the Station I-205 NB Gladstone on March 23rd, 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 B1 is a screen shot of PORTAL system for incident data from which the incident data above in Table B1 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: Incident data portal system screen shot

After clicking the archive “Timeseries” on the homepage of PORTAL system, the screen as in Figure B1 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 B2 and the incident data table for the whole segment of highway as shown in Figure B3 will be seen.

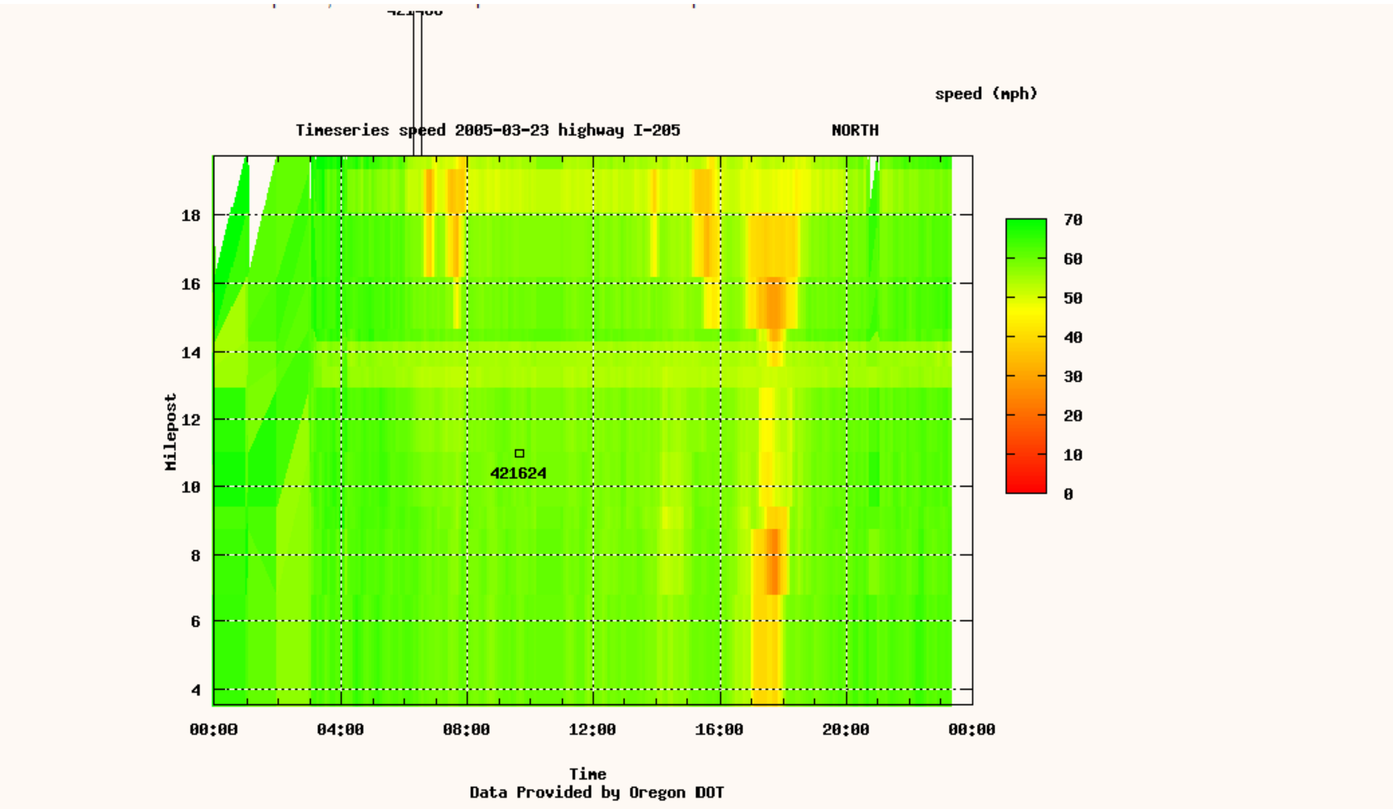
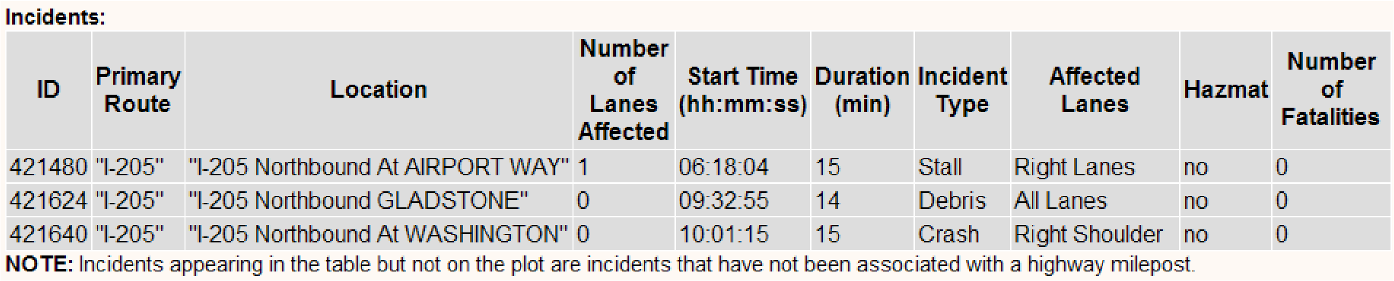


Figure B-2: Incident Data Graph on I-205 NB on March 23rd, 2005



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 Segment of Highway I-205 NB on March 23rd, 2005

Since the station I-205 NB Gladstone (milepost 11.04) is the station we are interested, as in Figure B2, we could see the incident with ID 421624 occurred around milepost 11.04 and its detailed information can be found in Figure B3 with its ID.

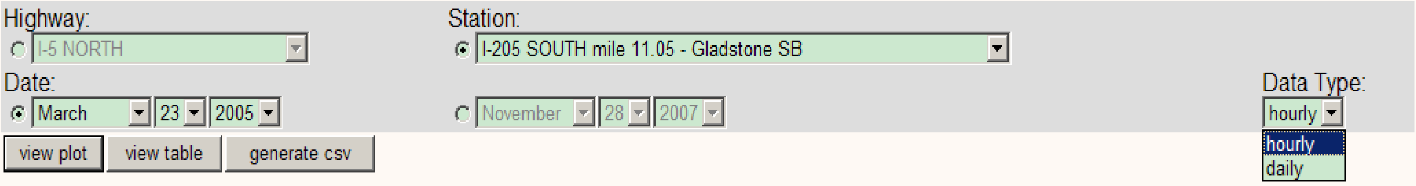
APPENDIX C:
DATA COLLECTION OF WEATHER DATA VARIABLES IN PORTAL

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 23rd, 2005 is shown in Table C1.

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 C1, which is a screen shot of PORTAL system for weather data collection.



Highway: ○ I-5 NORTH

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

Date: ● March 23 2005

○ November 28 2007

Data Type: hourly

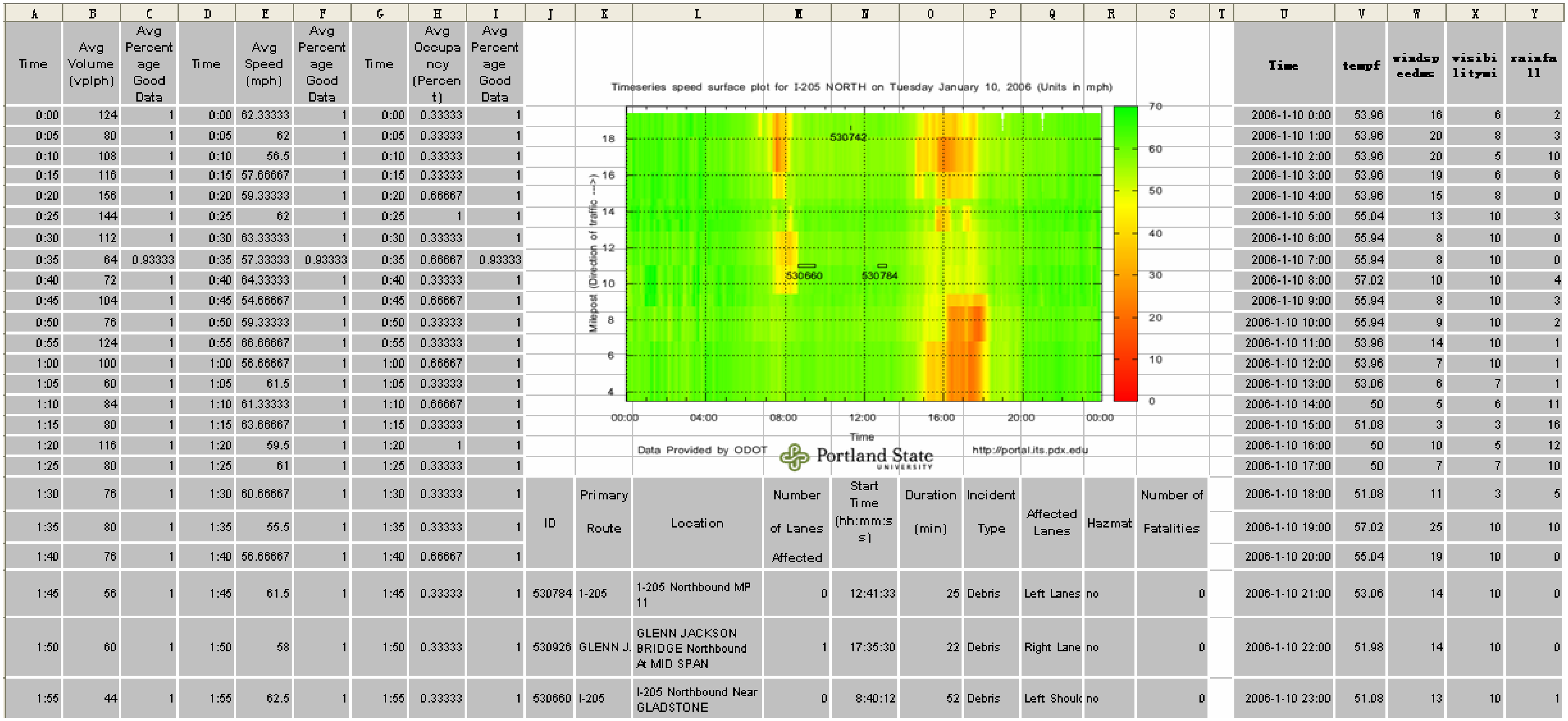
Figure C-1: Screen Shot of PORTAL System for Weather Data Collection

After clicking the archive “Weather” on the homepage of PORTAL system, the screen as shown in Figure C1 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 C1 can be obtained.

APPENDIX D:
EXCEL VBA PROGRAMS FOR RAW DATA REORGANIZATIONS

As described in section 3.1., 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 need to be performed for every raw daily data set collected, to save time and increase accuracy, EXCEL VBA programs are employed to reorganize daily raw data saved in EXCEL files. Before describing the programs, the raw daily data set collected at the station I-205 NB Gladstone on January 10th, 2006 is shown in Figure D1 as an example of the raw daily data sets. Due to space limitations, the example raw daily data set is only shown from 0:00 to 1:55 am for traffic flow data in Figure D1. 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 D1.

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 D2. A dialog window will show up and Personal Macro Workbook needs to be selected in "Store macro in:" as shown in Figure D3. After clicking OK, a new file called PERSONAL.XLS will be created automatically in EXCEL. Then by clicking Window>Unhide as shown in Figure D4, "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 D5. 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 D6. 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 D7. A blank window will then pop out for programs writing (or code imputing) to create Macros in the file PERSONAL.XLS.



A	B	C	D	E	F	G	H	I
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

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
70
60
50
40
30
20
10
0
00:00
04:00
08:00
12:00
16:00
20:00
00:00
Time
Data Provided by ODOT
Portland State UNIVERSITY
http://portal.its.pdx.edu

J	K	L	M	N	O	P	Q	R	S
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 Shoulc	no	0

U	V	W	X	Y
Time	tempf	windspeedmс	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 January 10th, 2006 (0:00 – 1:55 am)

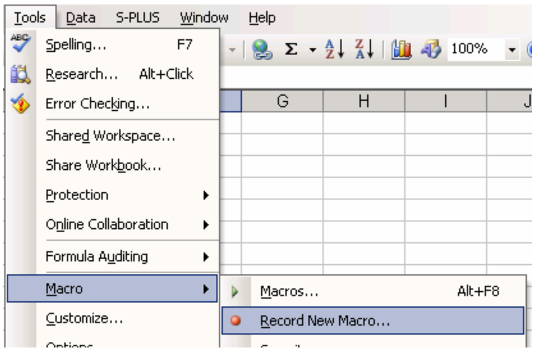
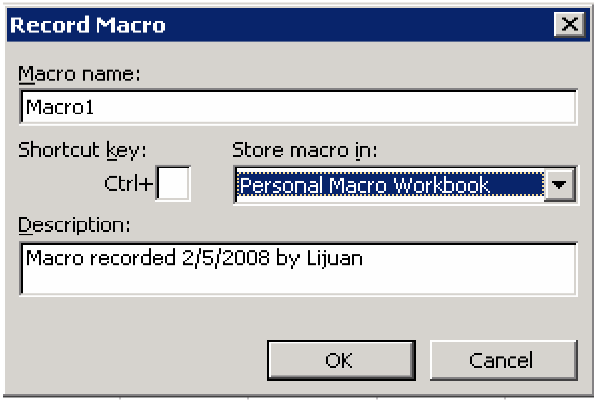


Figure D-2: Record New Macro


Record Macro

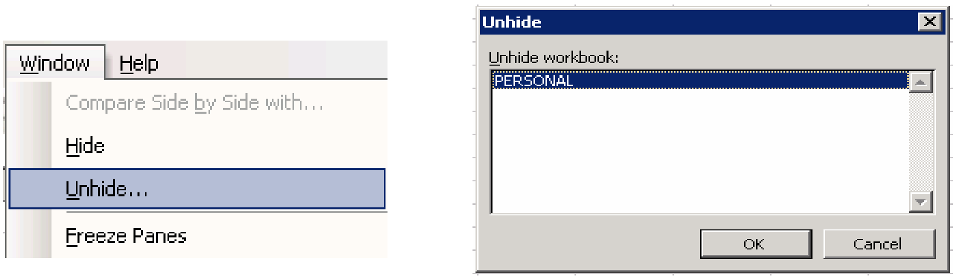
Macro name: Macro1

Shortcut key: Ctrl+

Store macro in: Personal Macro Workbook

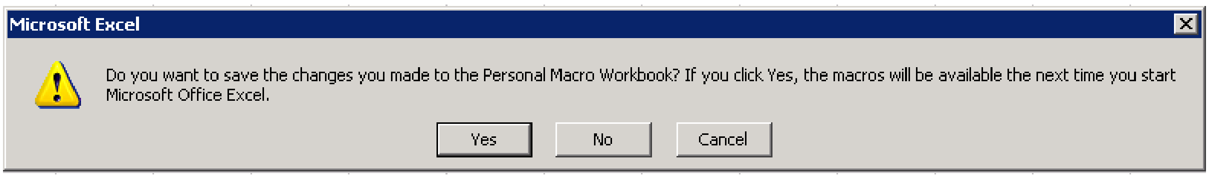
Description: Macro recorded 2/5/2008 by Lijuan

Figure D-3: Personal Macro Workbook



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

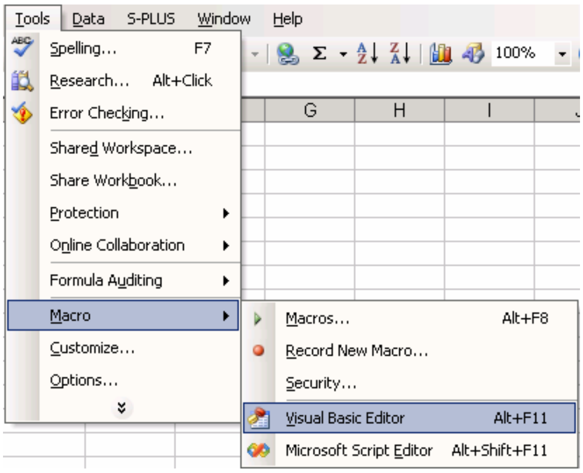


Figure D-6: Open Visual Basic Editor

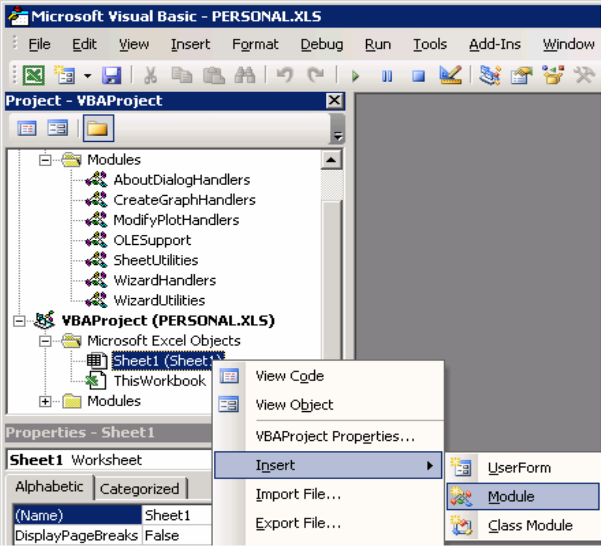


Figure D-7: Insert a New Module

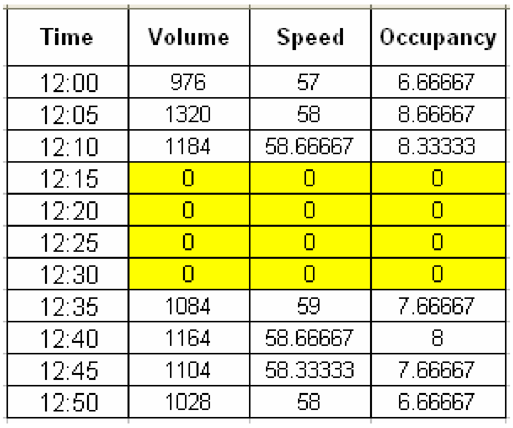
The EXCEL VBA programs are written as four Macros to reorganize the raw data of four types, which are described in the following.

- Traffic flow data: the following program can be copied to the new module created as shown in Figure D7 as a Macro with the name “Traffic_flow_data.” The raw data shown in Figure D1 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
```

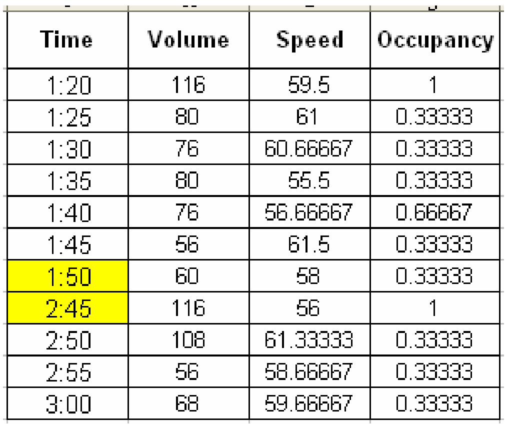
```
  Workbooks(2).Activate          ///Activate the workbook where the raw data are stored
  Worksheets(2).Select           ///Activate the “Sheet 2” of this workbook
  Range("a1").Select
  ActiveSheet.Paste              ///Paste the copied data into the first column
  Worksheets(1).Select           ///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    ///Set up 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 D10
  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 can not 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 our collected data are mainly the missing data and the erroneous data in traffic flow data due to detector errors, as shown in Figure D8 and Figure D9, respectively. In Figure D8, 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 D9, 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
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-8: Missing 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

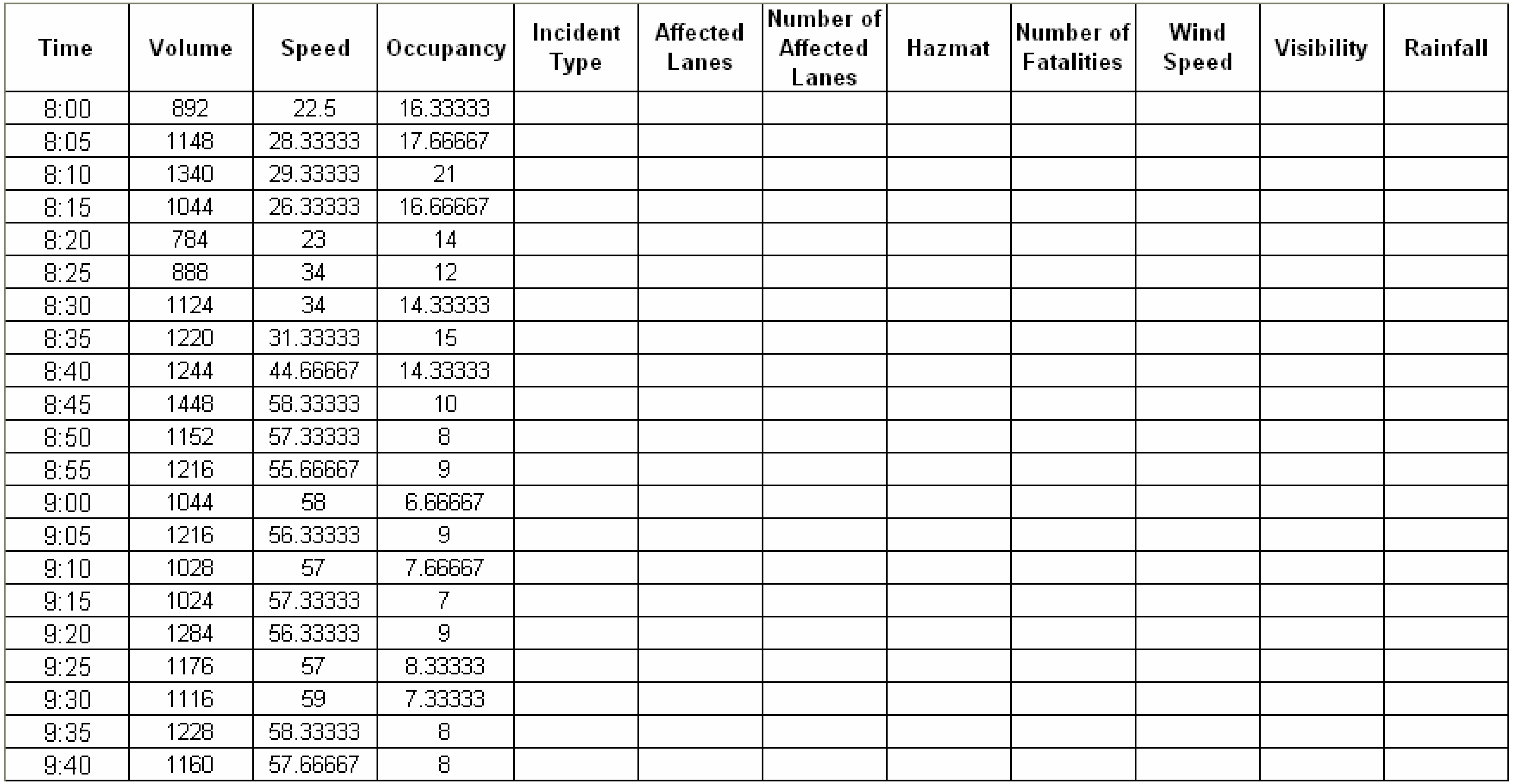
Figure D-9: Erroneous Data

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 set up 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 D10 to D13. 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 D10 to D13 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 I205 NB Gladstone on January 10th, 2006, as shown in the raw data in Figure D1.

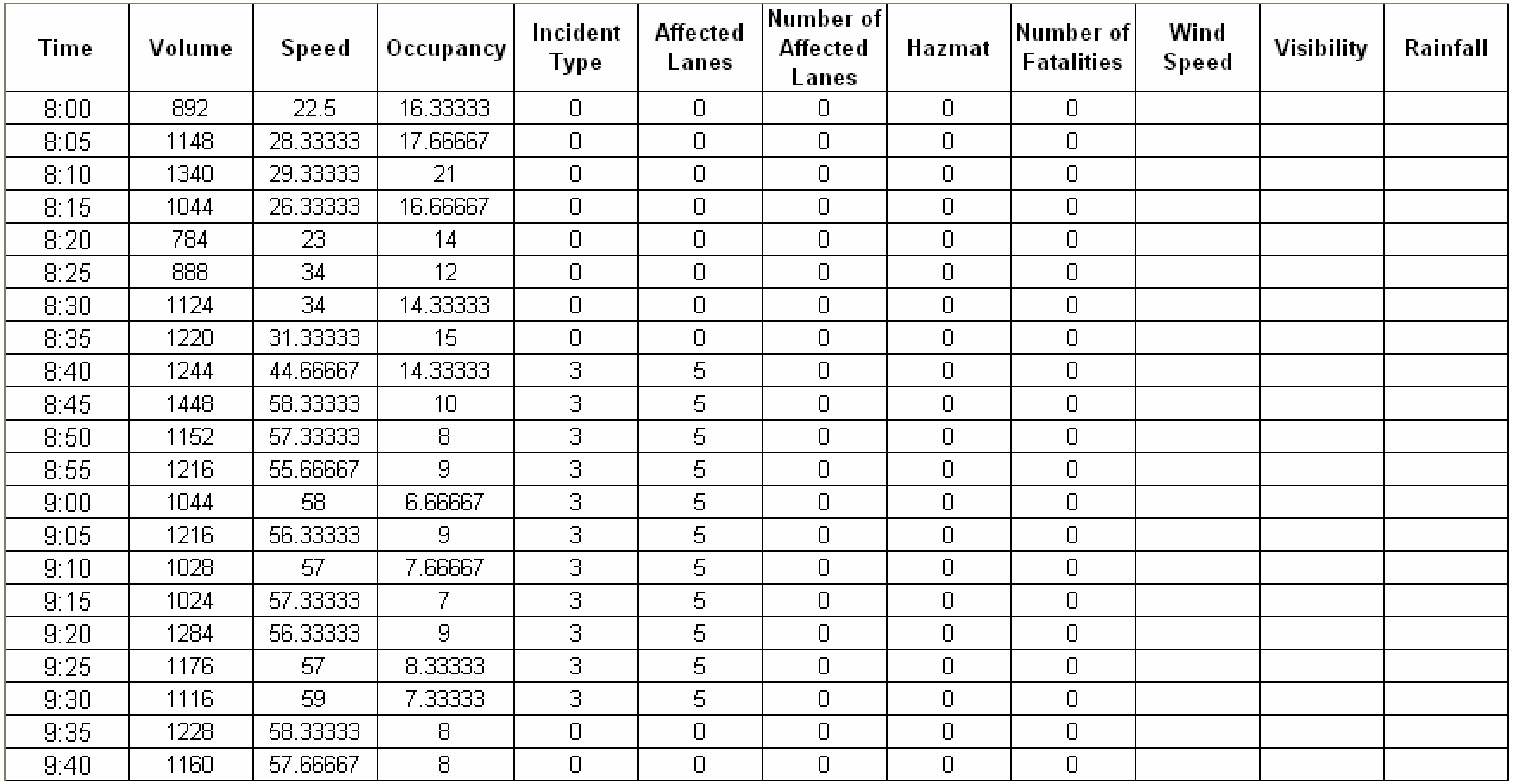
- Incident related data: the reorganization of raw incident related data can be automatically processed by the following program except the ID of the incident that occurred at the selected station needs to be appointed to the program by hand. Then the program can use the incident ID input by hand to locate the incident related data in the incident data table (column J to column S) shown in the raw data in Figure D1. 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    ///To be continued on page 50
```



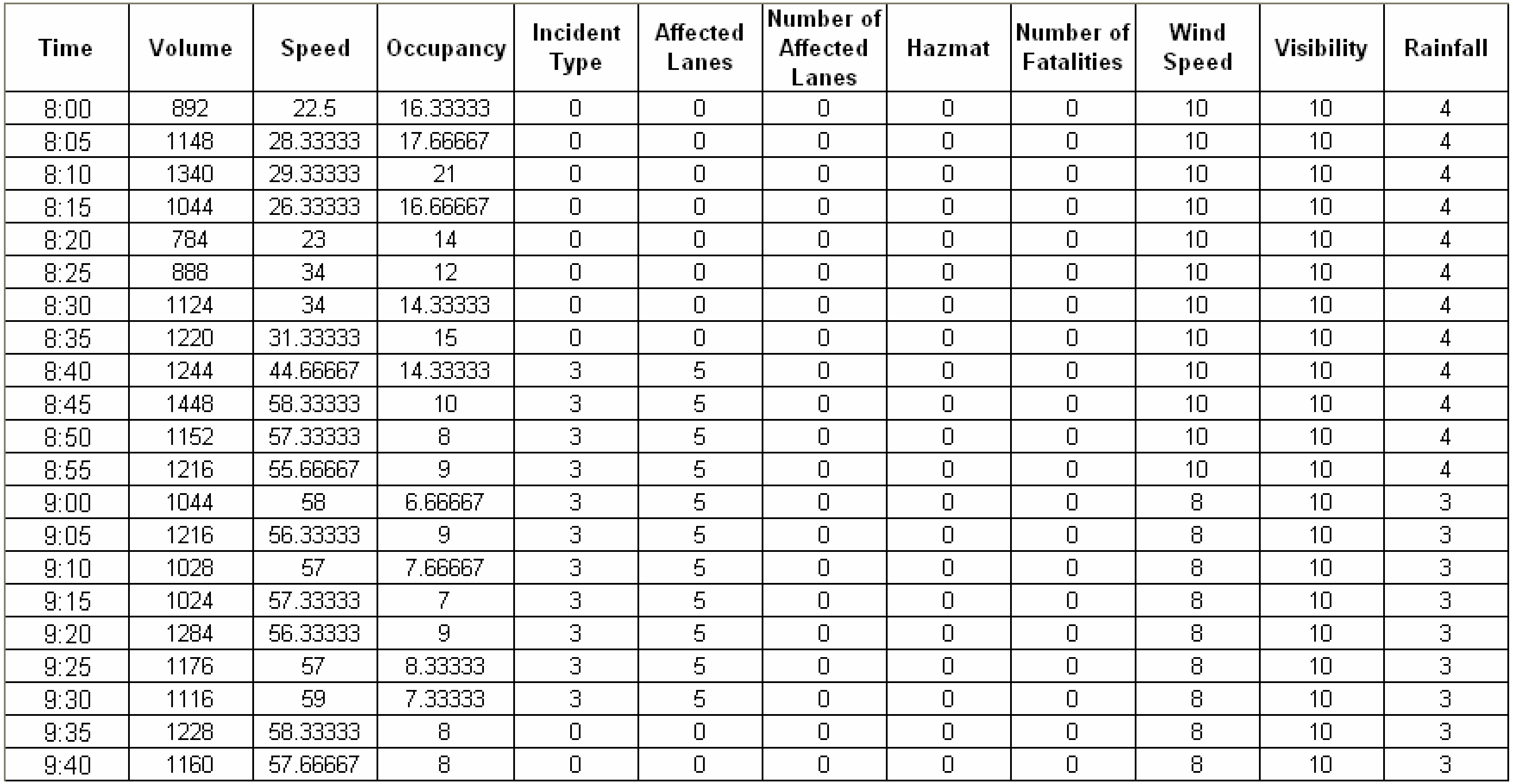
Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
8:00	892	22.5	16.33333								
8:05	1148	28.33333	17.66667								
8:10	1340	29.33333	21								
8:15	1044	26.33333	16.66667								
8:20	784	23	14								
8:25	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



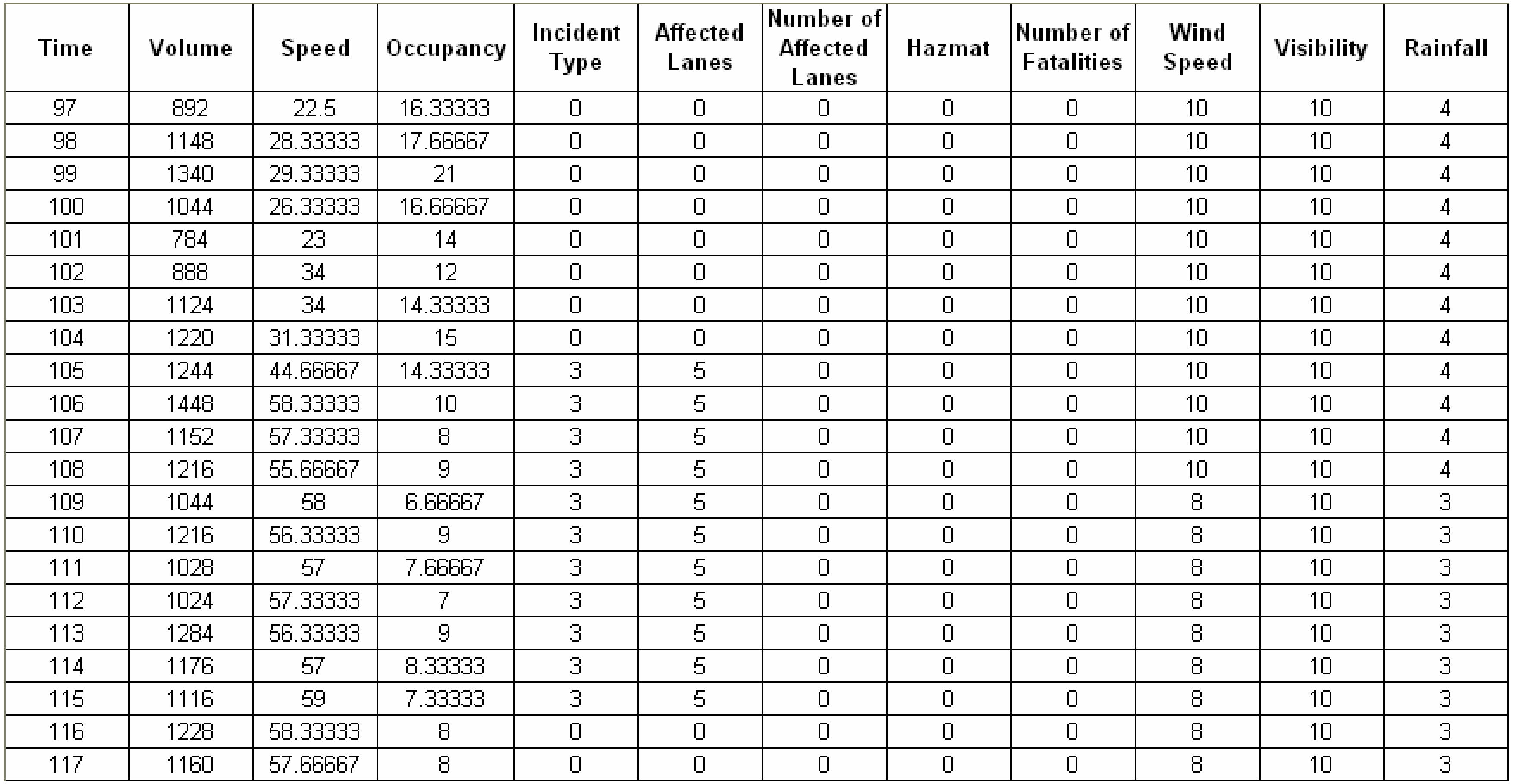
Time	Volume	Speed	Occupancy	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
8:00	892	22.5	16.33333	0	0	0	0	0			
8:05	1148	28.33333	17.66667	0	0	0	0	0			
8:10	1340	29.33333	21	0	0	0	0	0			
8:15	1044	26.33333	16.66667	0	0	0	0	0			
8:20	784	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_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
Select Case incident_type                                      ///test data
    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

Worksheets(2).Select
For i = 2 To 289
    Cells(i, 5).Value = 0
    Cells(i, 6).Value = 0
    Cells(i, 7).Value = 0
    Cells(i, 8).Value = 0
    Cells(i, 9).Value = 0
Next I              ///Default the data of five incident related variables with zeros first
For i = 2 To 289
    If occur_time < Cells(i, 1).Value And occur_time > Cells(i - 1, 1).Value Then
        If Cells(i, 1).Value - occur_time < occur_time - Cells(i - 1, 1).Value Then
            For j = i To i + (duration \ 5)    ///Decide the beginning and ending time points
                Cells(j, 5).Value = type_id    ///between which the incident related data
                Cells(j, 6).Value = lane_id     ///should be copied by rounding the occur time
                Cells(j, 7).Value = lanes_no   ///and the duration of the incident according to
                Cells(j, 8).Value = hazmat_id ///the data of time of day in 5-minute
        increments
                Cells(j, 9).Value = fatalities_no
            Next j
        Else
            For j = i - 1 To i - 1 + (duration \ 5)
                Cells(j, 5).Value = type_id
                Cells(j, 6).Value = lane_id
                Cells(j, 7).Value = lanes_no
                Cells(j, 8).Value = hazmat_id
                Cells(j, 9).Value = fatalities_no
            Next j
        End If
      End If
    Next i
End Sub
```

As shown in Figure D1, the ID of the incident that occurred at I205 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 D11. Because the occurrence time of the incident is 8:40:12 as shown in Figure D1, all of the incident related data is copied into the organized data, starting at 8:40 as shown in Figure D11 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 of 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 column p and
      they will
    ActiveSheet.Paste         ///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
    Next i          ///data because the 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 D1, the hourly data of wind speed, visibility and rainfall are 10, 10 and 4 for 8:00, respectively. In Figure D12, these data are repeatedly copied 12 times 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 can not 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
  Next I                         ///numbers from 1 to 288
End Sub
```

APPENDIX E:
VALIDATION OF REGRESSION TREE MODEL IN S-PLUS

As an example, we will show how to validate the regression tree model constructed on the test data set at I205 NB Gladstone on January 10th, 2006 in S-PLUS. To validate a regression tree model in S-PLUS, we need to import the validation data set into S-PLUS first by clicking File>Import Data>From File in S-PLUS as shown in Figure E1. Validation data sets can use any daily data sets collected at the stations including all the explanatory variables, which have been reorganized and are applicable in S-PLUS. Here, to validate the regression tree model built on the test data at I-205 NB Gladstone on January 10th, 2006, we randomly choose the daily data set collected at the same station I-205 NB Gladstone on August 2nd, 2006 as the validation data set, which is shown partially in Figure E2 due to space limitations.

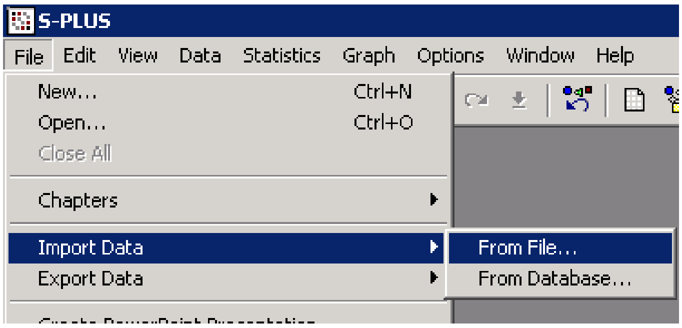
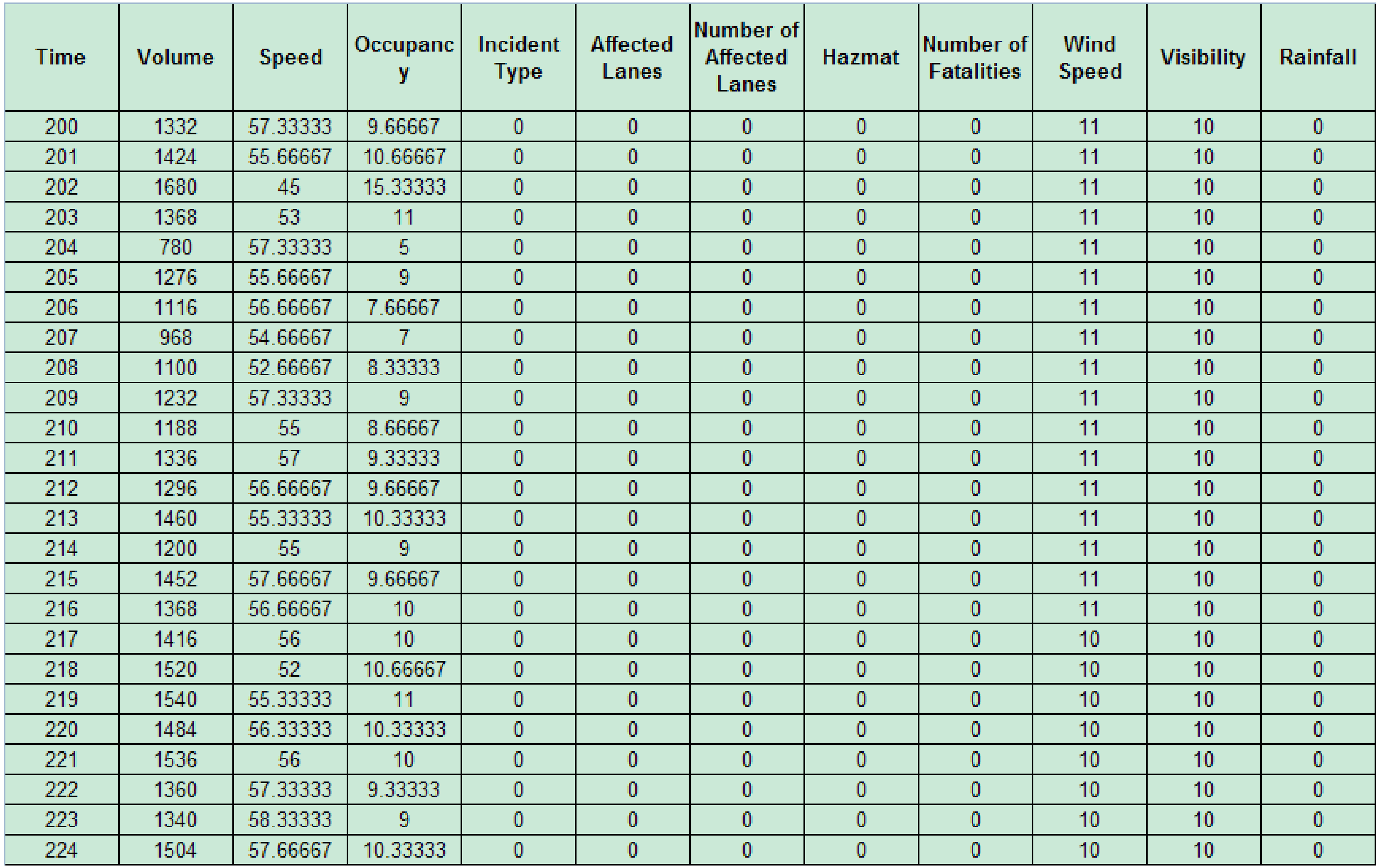


Figure E-1: Importation of Validation Data Set into S-PLUS



Time	Volume	Speed	Occupanc y	Incident Type	Affected Lanes	Number of Affected Lanes	Hazmat	Number of Fatalities	Wind Speed	Visibility	Rainfall
200	1332	57.33333	9.66667	0	0	0	0	0	11	10	0
201	1424	55.66667	10.66667	0	0	0	0	0	11	10	0
202	1680	45	15.33333	0	0	0	0	0	11	10	0
203	1368	53	11	0	0	0	0	0	11	10	0
204	780	57.33333	5	0	0	0	0	0	11	10	0
205	1276	55.66667	9	0	0	0	0	0	11	10	0
206	1116	56.66667	7.66667	0	0	0	0	0	11	10	0
207	968	54.66667	7	0	0	0	0	0	11	10	0
208	1100	52.66667	8.33333	0	0	0	0	0	11	10	0
209	1232	57.33333	9	0	0	0	0	0	11	10	0
210	1188	55	8.66667	0	0	0	0	0	11	10	0
211	1336	57	9.33333	0	0	0	0	0	11	10	0
212	1296	56.66667	9.66667	0	0	0	0	0	11	10	0
213	1460	55.33333	10.33333	0	0	0	0	0	11	10	0
214	1200	55	9	0	0	0	0	0	11	10	0
215	1452	57.66667	9.66667	0	0	0	0	0	11	10	0
216	1368	56.66667	10	0	0	0	0	0	11	10	0
217	1416	56	10	0	0	0	0	0	10	10	0
218	1520	52	10.66667	0	0	0	0	0	10	10	0
219	1540	55.33333	11	0	0	0	0	0	10	10	0
220	1484	56.33333	10.33333	0	0	0	0	0	10	10	0
221	1536	56	10	0	0	0	0	0	10	10	0
222	1360	57.33333	9.33333	0	0	0	0	0	10	10	0
223	1340	58.33333	9	0	0	0	0	0	10	10	0
224	1504	57.66667	10.33333	0	0	0	0	0	10	10	0

Figure E-2: Validation Data Set at I-205 NB Gladstone on August 2nd, 2006. (4:30 – 6:30 pm)

After the test data set and the validation data set are imported into S-PLUS, by clicking Statistics>Tree>Tree Models as shown in Figure E3, the window "Tree Models" is opened as shown in Figure E4. The first three tabs in "Tree Models" window--Model, Results and Plot--are used to construct the tree model, show the result summary and tree plot, respectively.

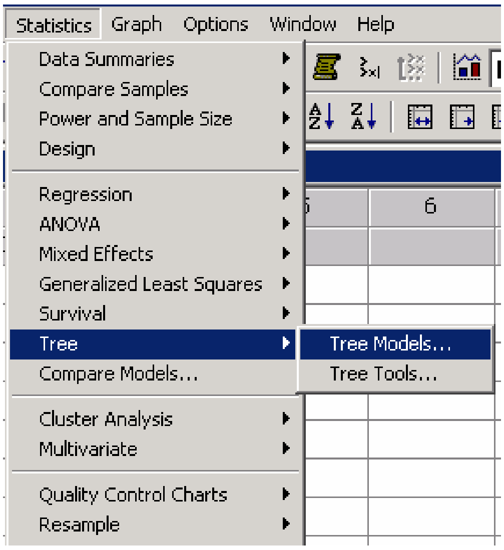
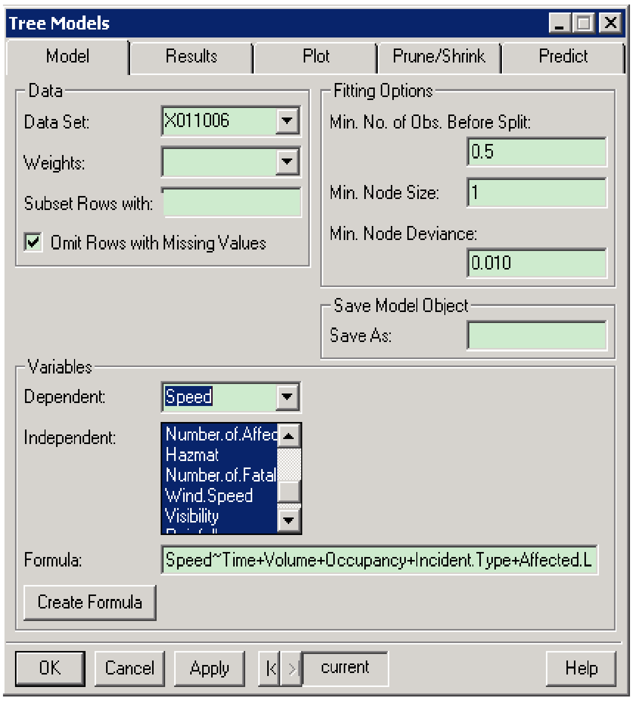


Figure E-3: Open Tree Models Window



Data Set: X011006

Weights:

Subset Rows with:

☑ Omit Rows with Missing Values

Min. No. of Obs. Before Split: 0.5

Min. Node Size: 1

Min. Node Deviance: 0.010

Save As:

Dependent: Speed

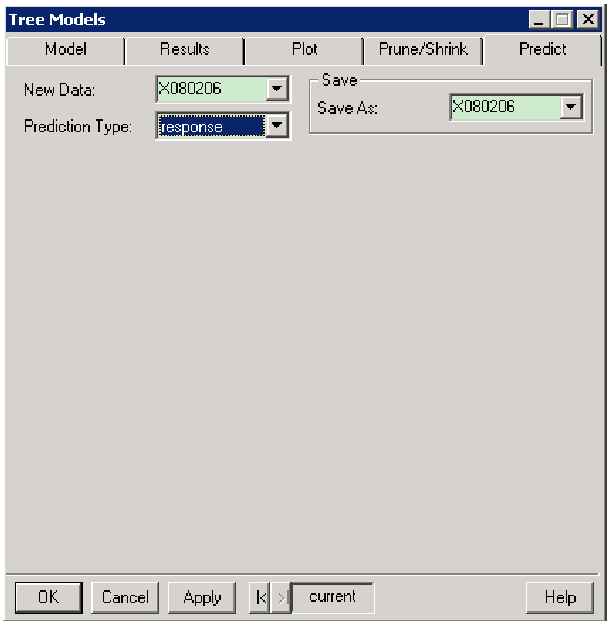
Independent: Number.of.Affec
Hazmat
Number.of.Fatal
Wind.Speed
Visibility

Formula: Speed~Time+Volume+Occupancy+Incident.Type+Affected.L

Figure E-4: "Tree Models" Window – "Model" Tab

As shown in Figure E4, the test data set is "X011006," which is the file name of the test data set collected at I-205 NB Gladstone on January 10th, 2006. As introduced in Section 2, the response variable (or dependent variable) is "Speed" and all of the thirteen explanatory variables are selected as independent variables. Then the construction of the regression tree model on the test data set of January 10th, 2006 is appropriately set up as shown in Figure E4.

After the regression tree model is set up, the fifth tab "Predict" in "Tree Models" window, as shown in Figure E5, is used to set up the validation of the tree model. "X080206" is the file name of the validation data collected at I-205 NB Gladstone on August 2nd, 2006, which is selected in "New Data" in Figure E5 as the validation data. "response" is selected in "Prediction Type," since "Speed" is the response variable and will be predicted by the model we built. "Save As" is used to choose where the validation results are saved and here we choose to save in the validation data set file itself. 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 tree model, as shown in Figure E6.

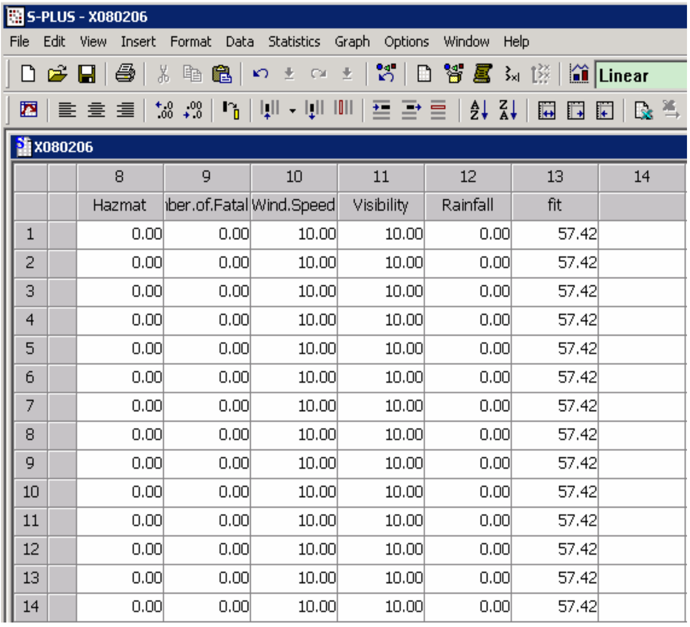


New Data: X080206

Prediction Type: response

Save As: X080206

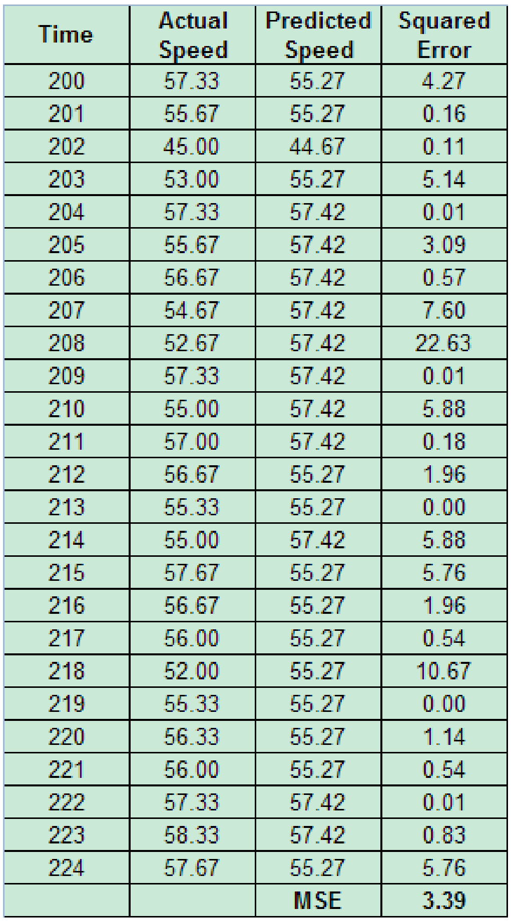
Figure E-5: “Tree Models” Window


S-PLUS - X080206
File Edit View Insert Format Data Statistics Graph Options Window Help
Linear
X080206

	8	9	10	11	12	13	14
	Hazmat	ıber.of.Fatal	Wind.Speed	Visibility	Rainfall	fit	
1	0.00	0.00	10.00	10.00	0.00	57.42	
2	0.00	0.00	10.00	10.00	0.00	57.42	
3	0.00	0.00	10.00	10.00	0.00	57.42	
4	0.00	0.00	10.00	10.00	0.00	57.42	
5	0.00	0.00	10.00	10.00	0.00	57.42	
6	0.00	0.00	10.00	10.00	0.00	57.42	
7	0.00	0.00	10.00	10.00	0.00	57.42	
8	0.00	0.00	10.00	10.00	0.00	57.42	
9	0.00	0.00	10.00	10.00	0.00	57.42	
10	0.00	0.00	10.00	10.00	0.00	57.42	
11	0.00	0.00	10.00	10.00	0.00	57.42	
12	0.00	0.00	10.00	10.00	0.00	57.42	
13	0.00	0.00	10.00	10.00	0.00	57.42	
14	0.00	0.00	10.00	10.00	0.00	57.42	

Figure E-6: Validation Data Set at I-205 NB – “Predict” Tab Gladstone on August 2nd, 2006 with Fitted Values of Speed

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. For example, if we use MSE to estimate the accuracy of the predicted speeds between 4:30 and 6:30 pm on August 2nd, 2006 by the regression tree model on the test data set of January 10th, 2006, the MSE result is 3.39 as shown in Figure E7, which is fairly low for the errors of speed values.



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 E-7: MSE Result of Validation Data on August 2nd, 2006 by Test Data on January 10th, 2006

APPENDIX F:

MACRO IN EXCEL VBA FOR CHARACTERIZATION

In the characterization approach, we set up four standards to track different 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 of traffic flow data due to detector error. For “Good Weather”, based on published sources, we regard a data set 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 we are faced with 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 test data sets and validation data sets after raw data clean-up and reorganization. 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 F1. A dialog window will show up and Personal Macro Workbook needs to be selected in “Store macro in:” as shown in Figure F2. After clicking OK, a new file called PERSONAL.XLS will be created automatically in EXCEL. Then by clicking Window>Unhide as shown in Figure F3, “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 F4. 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 F5. 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 F6. 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.

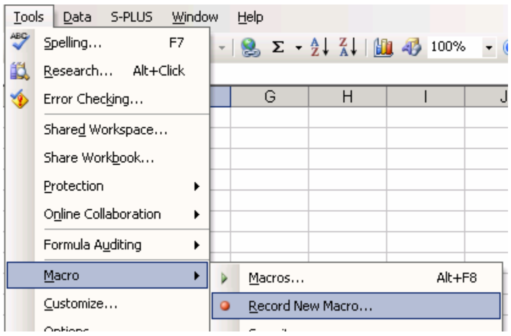
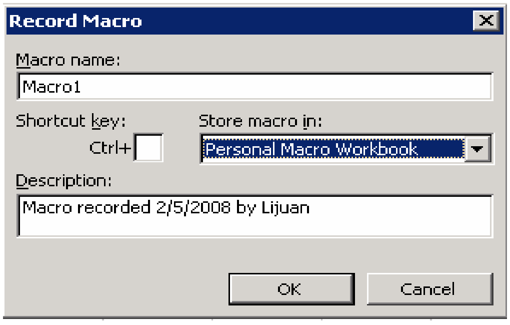


Figure F-1: Record New Macro


Record Macro

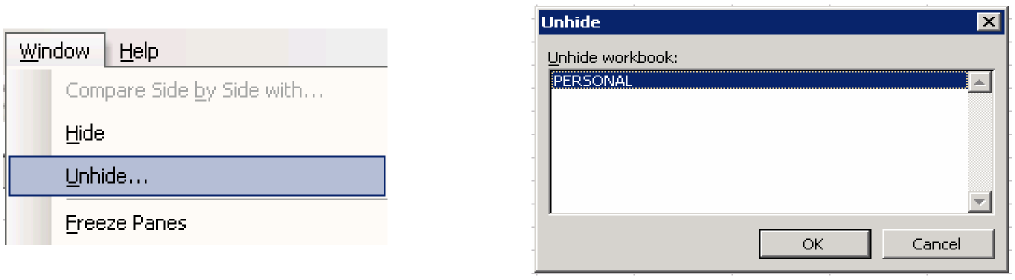
Macro name:
Macro1

Shortcut key:
Ctrl+

Store macro in:
Personal Macro Workbook

Description:
Macro recorded 2/5/2008 by Lijuan

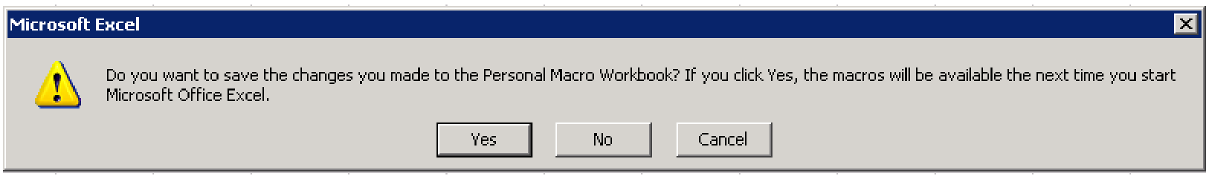
Figure F-2: Personal Macro Workbook



Unhide

Unhide workbook:
PERSONAL

Figure F-3: 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 F-4: Exit EXCEL

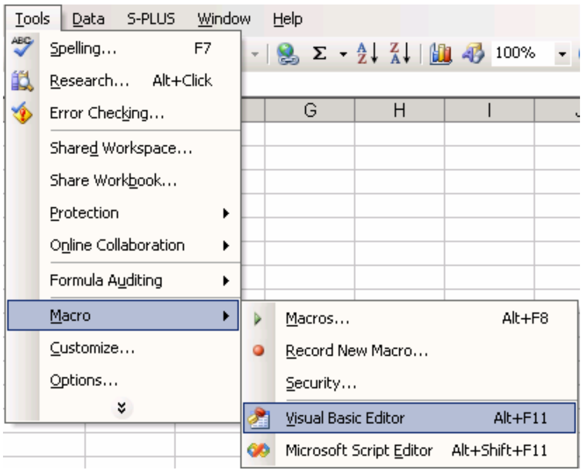


Figure F-5: Open Visual Basic Editor

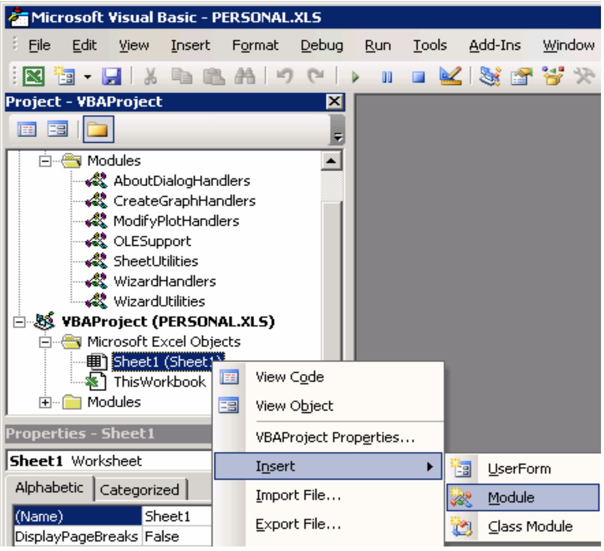


Figure F-6: Insert a New Module

To apply this Macro saved in PERSONAL.XML file to perform the characterization for a 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 F7, the Macro window is opened, as shown in Figure F8. By selecting the name of this Macro “Characterization” and then clicking Run, characterization is performed to the opened data set file.

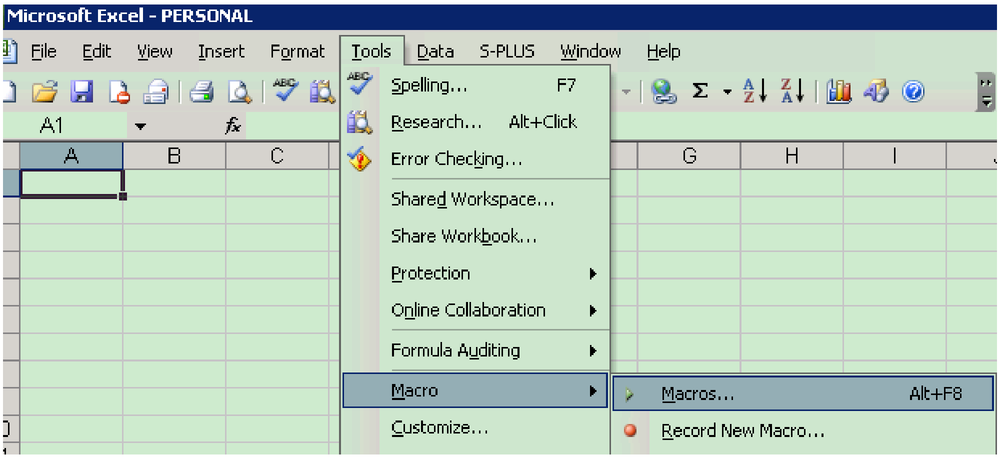
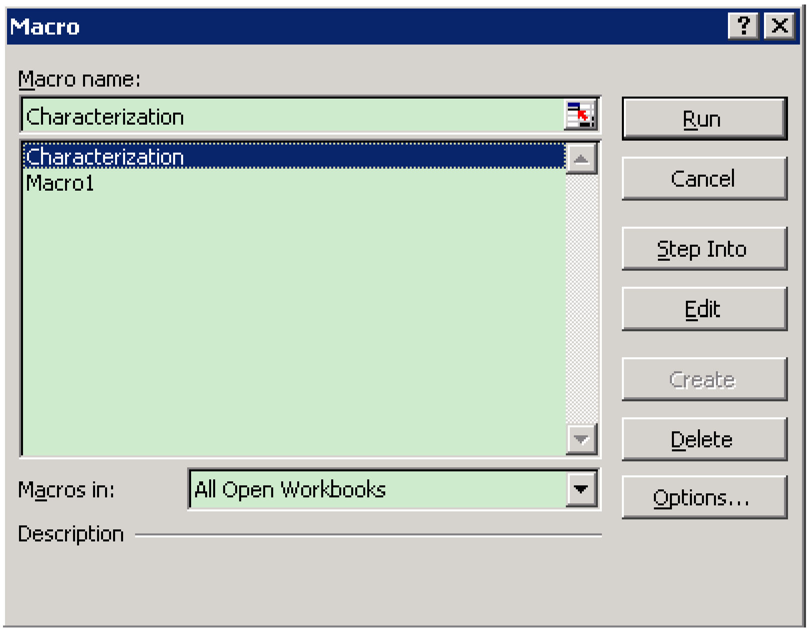


Figure F-7: Open Macro Window



Macro name: Characterization

Macros in: All Open Workbooks

Figure F-8: Macro Window

APPENDIX G:

RANDOM NUMBER GENERATION PROGRAM

In our research described in this report, 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 we need to generate random numbers in a large percentage of the original numbers and at the same time these random numbers can not be repetitive. For example, in full regression tree model construction, we need to randomly select 36 non-repetitive daily test data sets out of the total of 51 daily test data sets in characterization 3 to represent this characterization in the full model.

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.

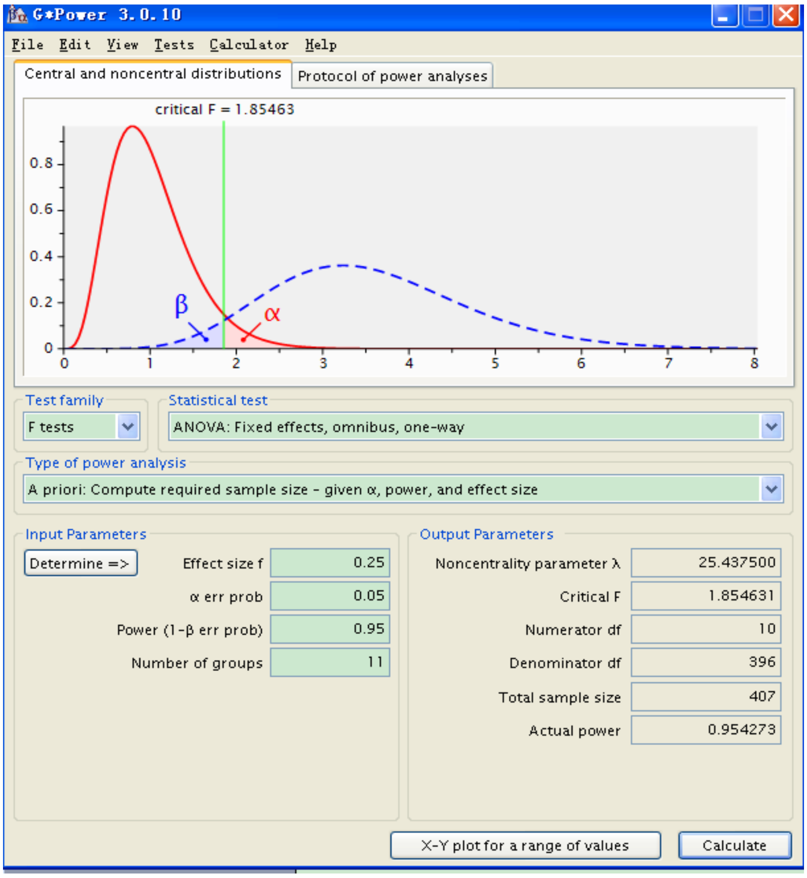
```
Sub Randx()
    Dim xx(1 To AAA) As Integer
    For t = 1 To BBB
    rerand:
    x = Int(Rnd() * AAA + 1)
   If xx(x) > 0 Then GoTo rerand
    r = r + 1
    Cells(r, 1) = x
    xx(x) = r
    Next
End Sub
```

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 A6, 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 as shown in Figure A8 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 H:

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 H1. For the randomized complete block design used in our research, the meaningful sample size needs to be decided. 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 randomized complete block design 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 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 of 95% and the number of groups of 11 are used. The number of groups here refers to the number of levels of the single factor in ANOVA. The single factor in our experimental design is regression tree models, which include 11 regression tree models considered in our experimental design, ten of them representing the ten characterizations and one full regression tree model. Thus, we need to type 11 for number of groups.


G*Power 3.0.10
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 1.85463
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8
β
α
Determine =>
X-Y plot for a range of values
Calculate

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 H-1: G*Power Software

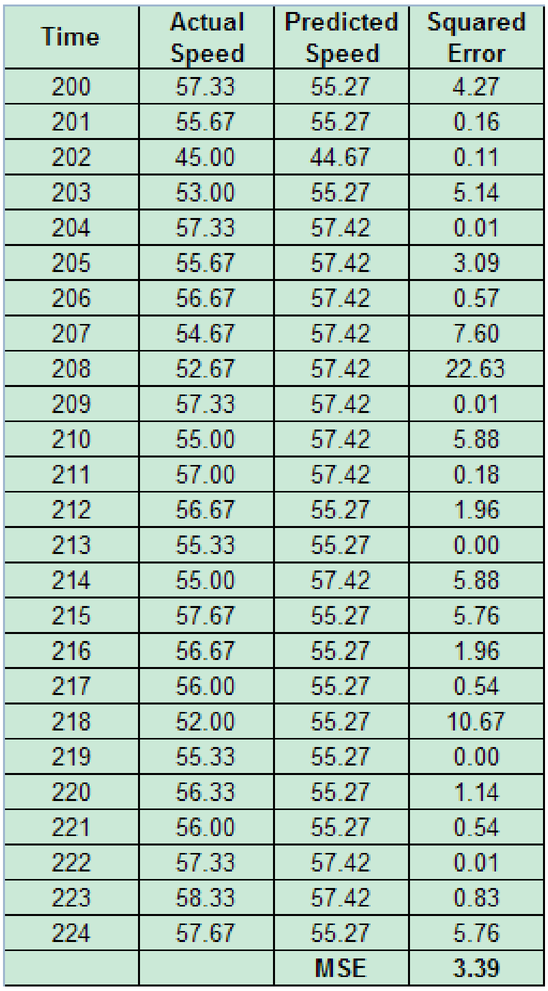
Then, by clicking Calculate, the results for the sample size are shown in the output parameters section. As shown in Figure H1, the total sample size of 407 is needed to guarantee the effect size of 0.25 and the power of test of 95% for 11 levels of the factor in our experimental design. Therefore, for each level of factor, at least 407÷11 = 37 blocks are needed.

APPENDIX I:

CALCULATION OF MSE VALUES USED IN RCBD

Since validation of regression tree models in S-PLUS has been demonstrated in Appendix E of this report, we focus on explaining 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 our RCBD is MSE values from validation of regression tree models by using validation data sets. 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.

Each of the validation data sets, serving as one block in RCBD, is used to validate 11 different regression tree models, 10 of which representing the ten characterizations 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 I1 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 two hours of data shown in Figure I1, 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.

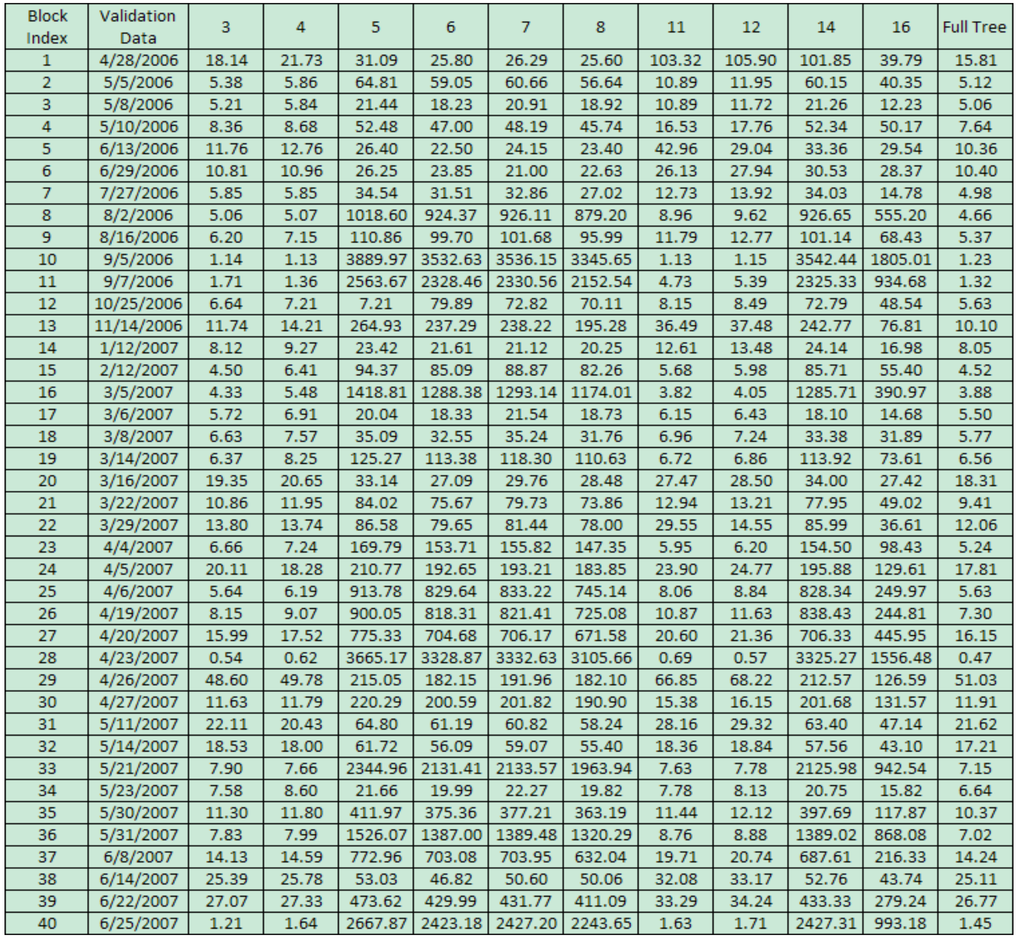


Time	Actual Speed	Predicted Speed	Squared Error
200	57.33	55.27	4.27
201	55.67	55.27	0.16
202	45.00	44.67	0.11
203	53.00	55.27	5.14
204	57.33	57.42	0.01
205	55.67	57.42	3.09
206	56.67	57.42	0.57
207	54.67	57.42	7.60
208	52.67	57.42	22.63
209	57.33	57.42	0.01
210	55.00	57.42	5.88
211	57.00	57.42	0.18
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 I-1: MSE Calculation

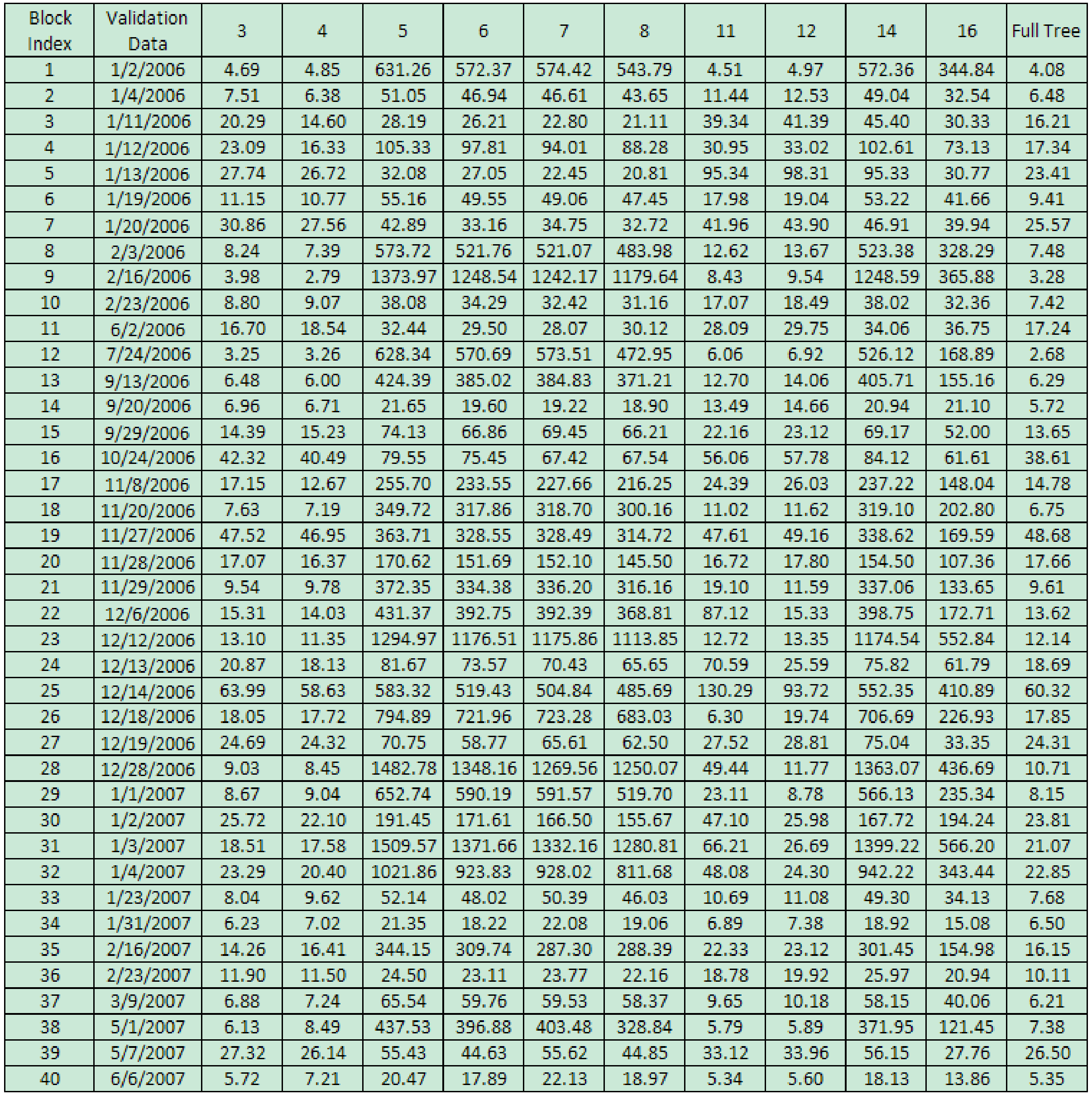
APPENDIX J:

RCBD FOR VALIDATION DATA SETS IN TEN CHARACTERIZATIONS



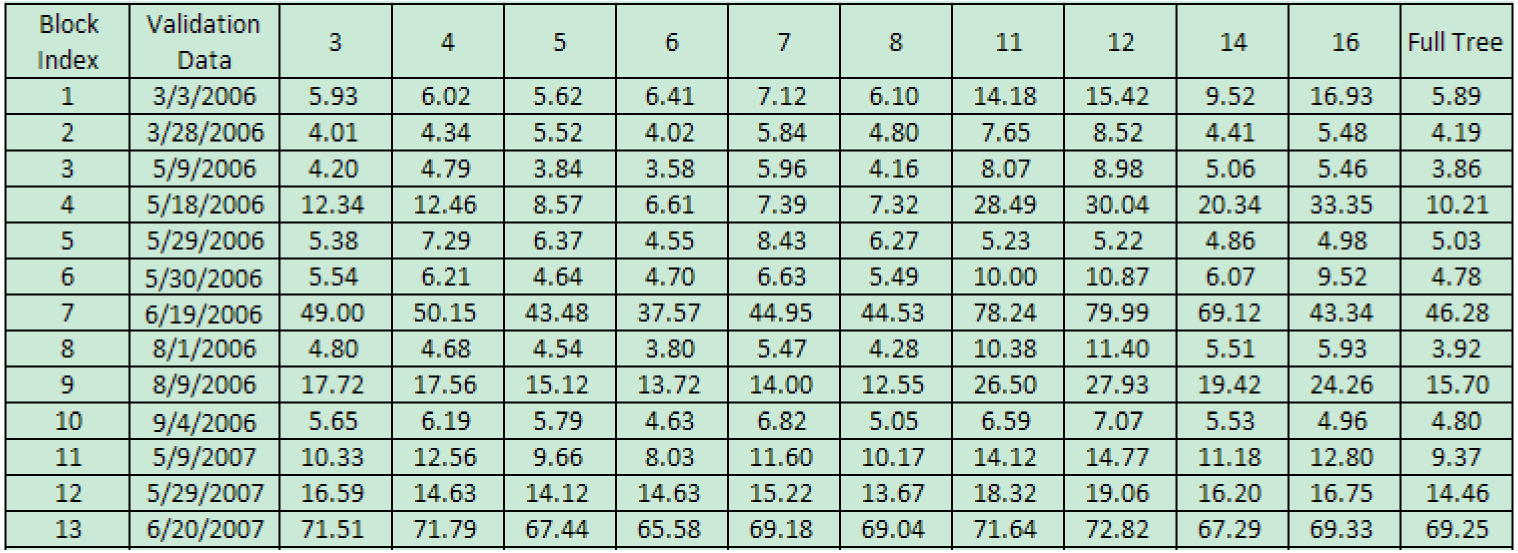
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	4/28/2006	18.14	21.73	31.09	25.80	26.29	25.60	103.32	105.90	101.85	39.79	15.81
2	5/5/2006	5.38	5.86	64.81	59.05	60.66	56.64	10.89	11.95	60.15	40.35	5.12
3	5/8/2006	5.21	5.84	21.44	18.23	20.91	18.92	10.89	11.72	21.26	12.23	5.06
4	5/10/2006	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 J-1: RCBD for Validation Data Sets in Characterization 3



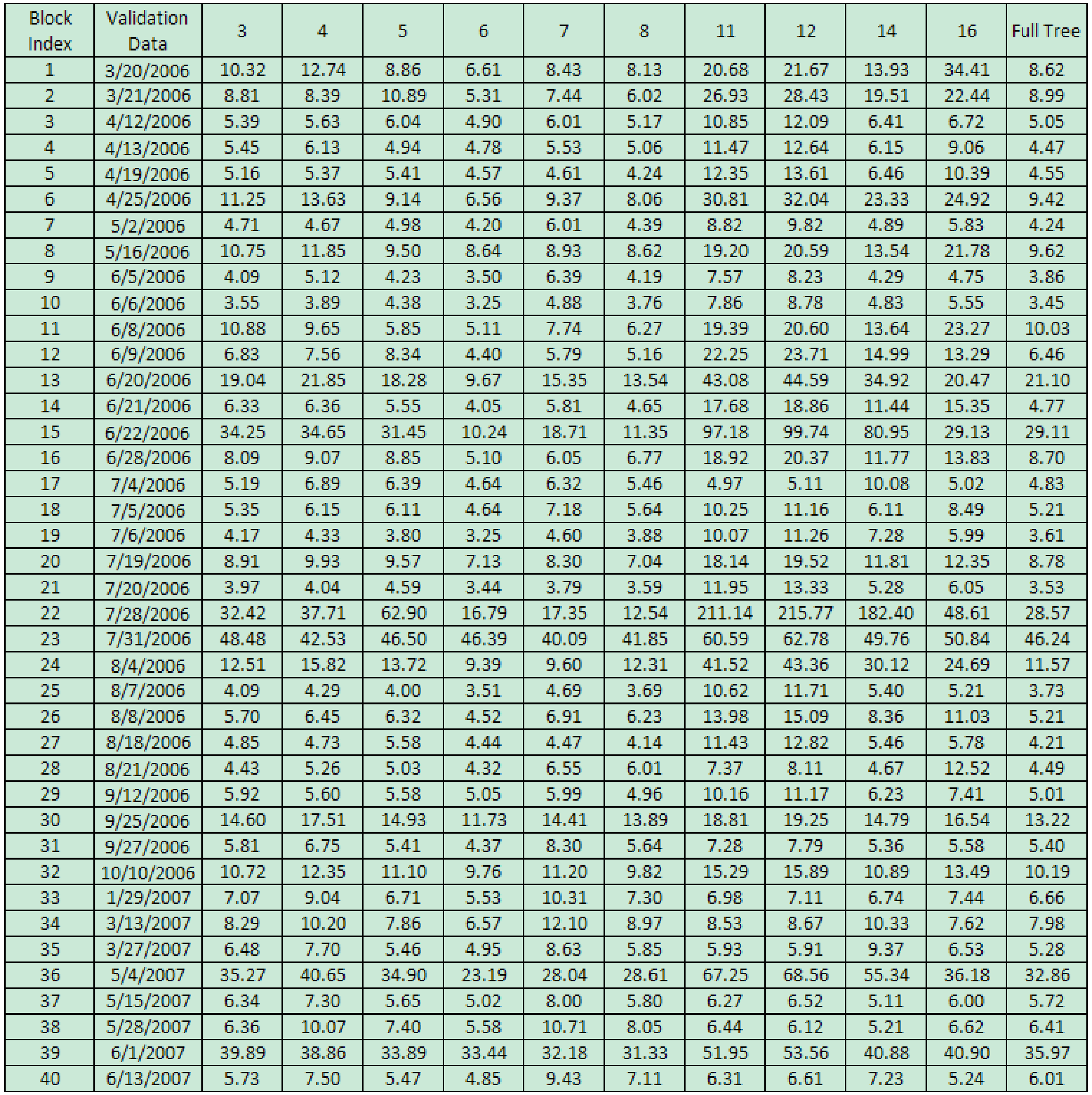
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/2/2006	4.69	4.85	631.26	572.37	574.42	543.79	4.51	4.97	572.36	344.84	4.08
2	1/4/2006	7.51	6.38	51.05	46.94	46.61	43.65	11.44	12.53	49.04	32.54	6.48
3	1/11/2006	20.29	14.60	28.19	26.21	22.80	21.11	39.34	41.39	45.40	30.33	16.21
4	1/12/2006	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 J-2: RCBD for Validation Data Sets in Characterization 4



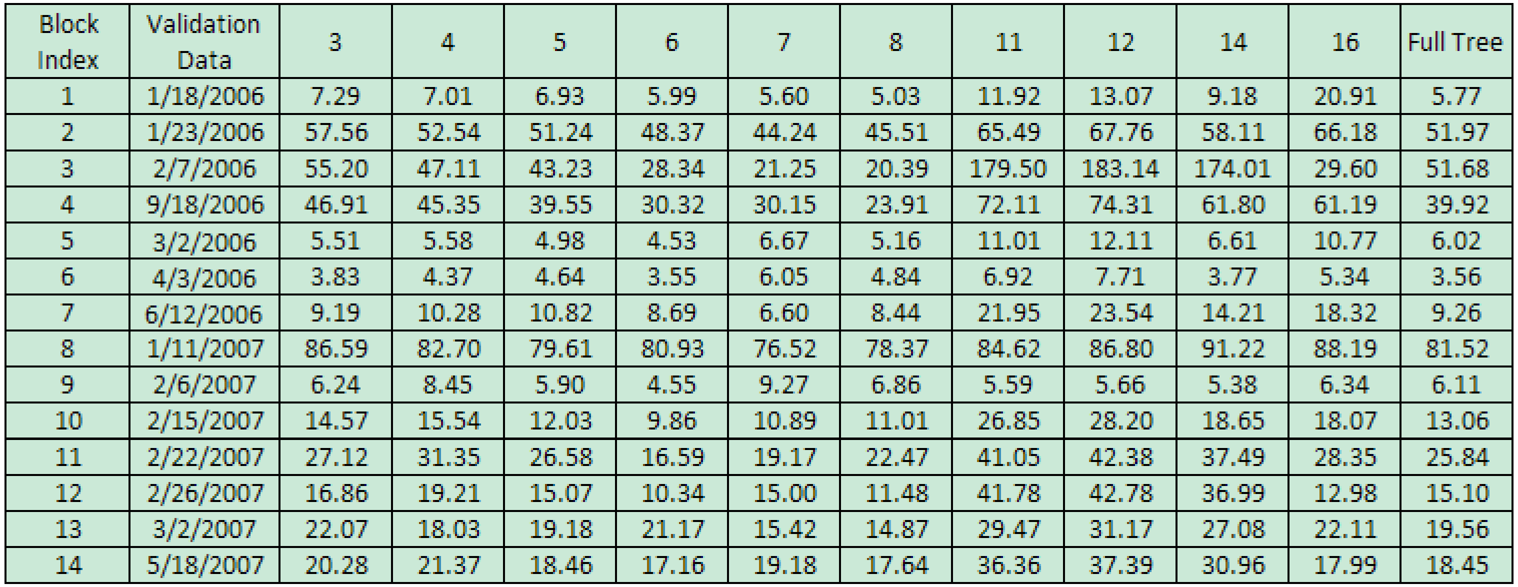
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	3/3/2006	5.93	6.02	5.62	6.41	7.12	6.10	14.18	15.42	9.52	16.93	5.89
2	3/28/2006	4.01	4.34	5.52	4.02	5.84	4.80	7.65	8.52	4.41	5.48	4.19
3	5/9/2006	4.20	4.79	3.84	3.58	5.96	4.16	8.07	8.98	5.06	5.46	3.86
4	5/18/2006	12.34	12.46	8.57	6.61	7.39	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 J-3: RCBD for Validation Data Sets in Characterization 5



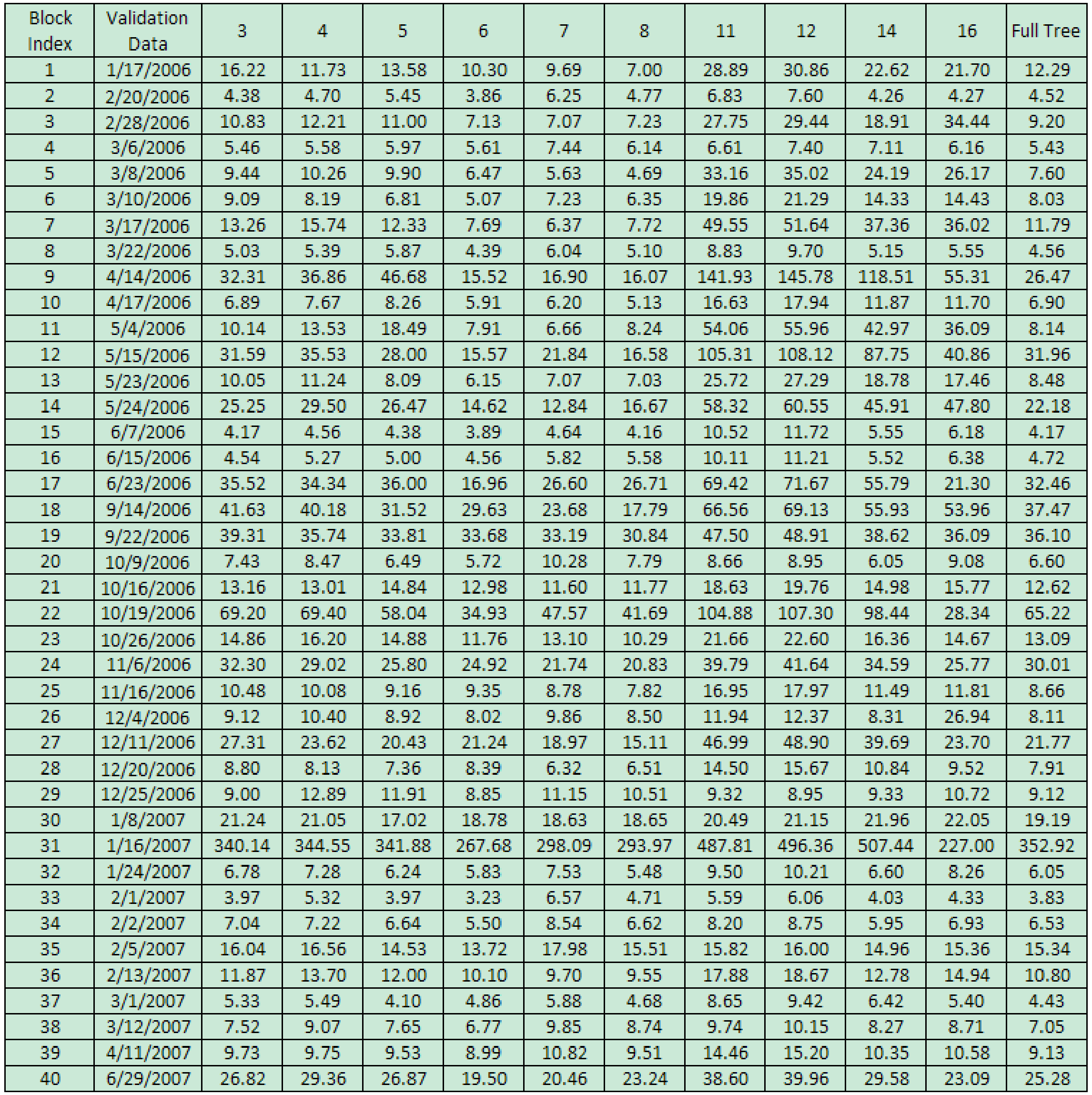
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	3/20/2006	10.32	12.74	8.86	6.61	8.43	8.13	20.68	21.67	13.93	34.41	8.62
2	3/21/2006	8.81	8.39	10.89	5.31	7.44	6.02	26.93	28.43	19.51	22.44	8.99
3	4/12/2006	5.39	5.63	6.04	4.90	6.01	5.17	10.85	12.09	6.41	6.72	5.05
4	4/13/2006	5.45	6.13	4.94	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 J-4: RCBD for Validation Data Sets in Characterization 6



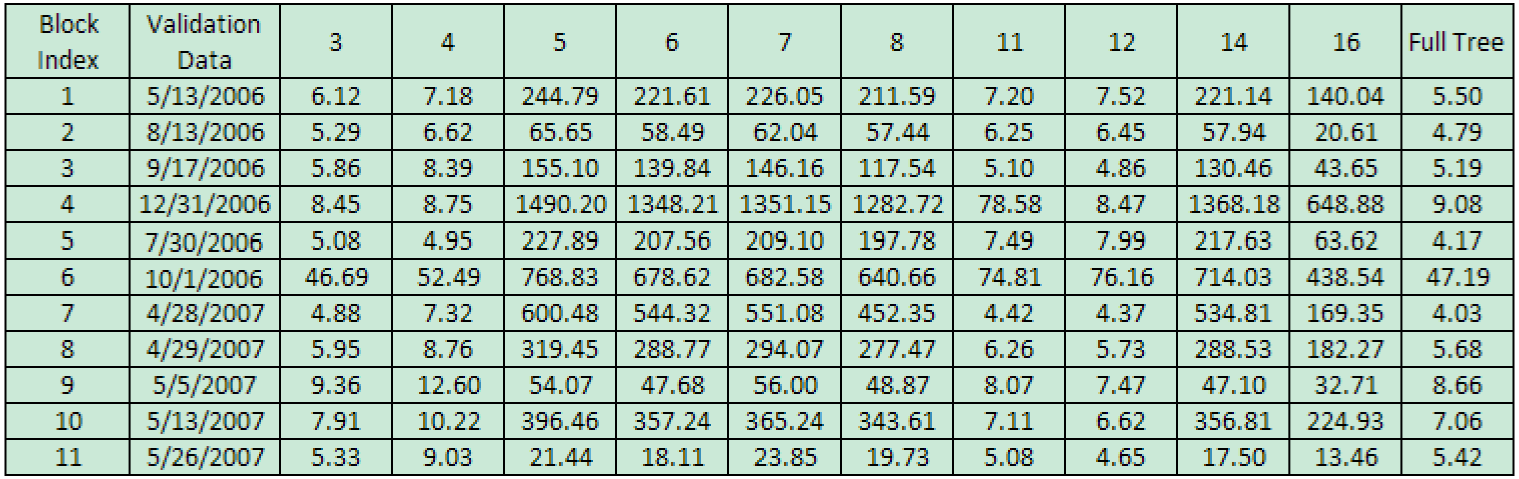
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/18/2006	7.29	7.01	6.93	5.99	5.60	5.03	11.92	13.07	9.18	20.91	5.77
2	1/23/2006	57.56	52.54	51.24	48.37	44.24	45.51	65.49	67.76	58.11	66.18	51.97
3	2/7/2006	55.20	47.11	43.23	28.34	21.25	20.39	179.50	183.14	174.01	29.60	51.68
4	9/18/2006	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 J-5: RCBD for Validation Data Sets in Characterization 7



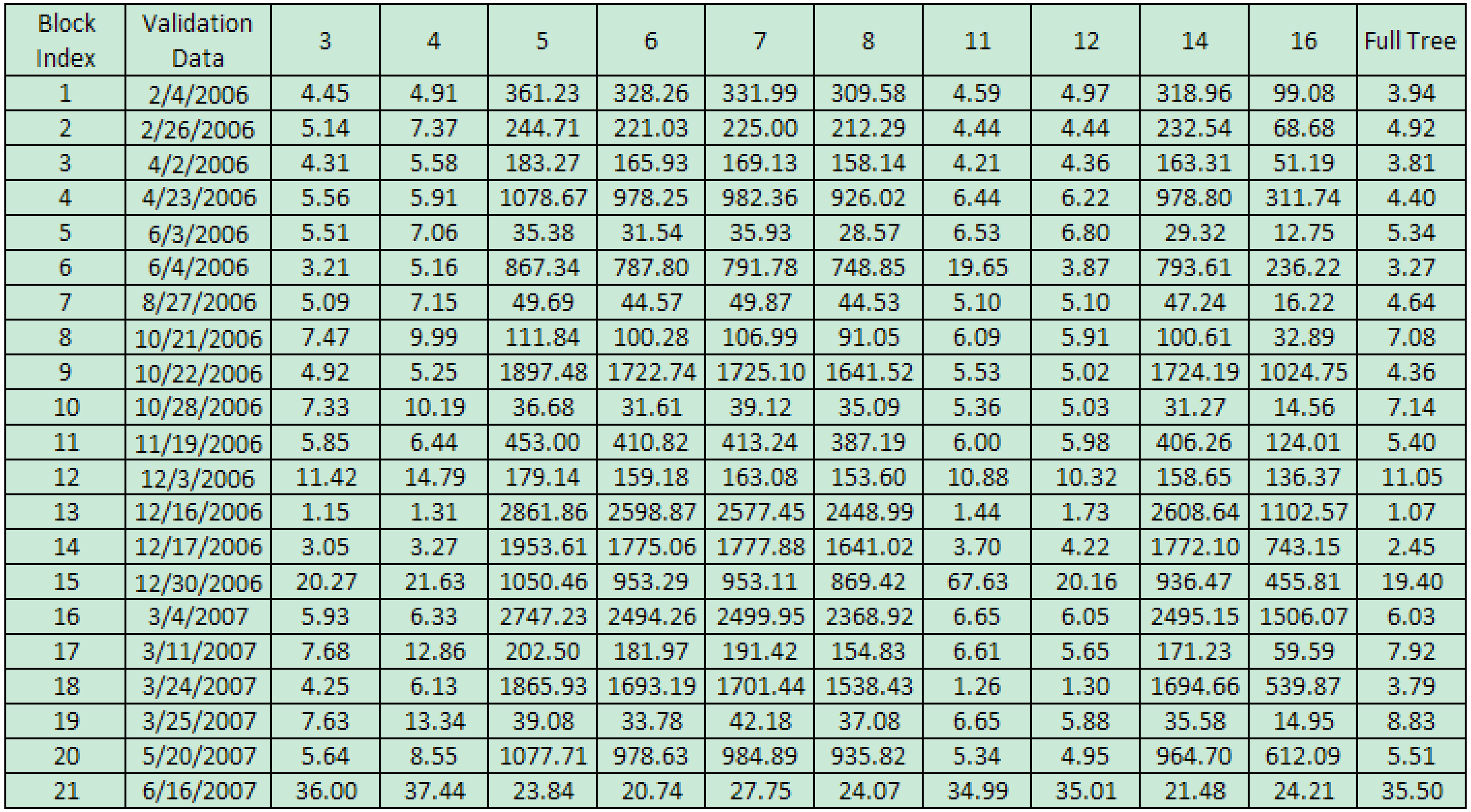
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	1/17/2006	16.22	11.73	13.58	10.30	9.69	7.00	28.89	30.86	22.62	21.70	12.29
2	2/20/2006	4.38	4.70	5.45	3.86	6.25	4.77	6.83	7.60	4.26	4.27	4.52
3	2/28/2006	10.83	12.21	11.00	7.13	7.07	7.23	27.75	29.44	18.91	34.44	9.20
4	3/6/2006	5.46	5.58	5.97	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 J-6: RCBD for Validation Data Sets in Characterization 8



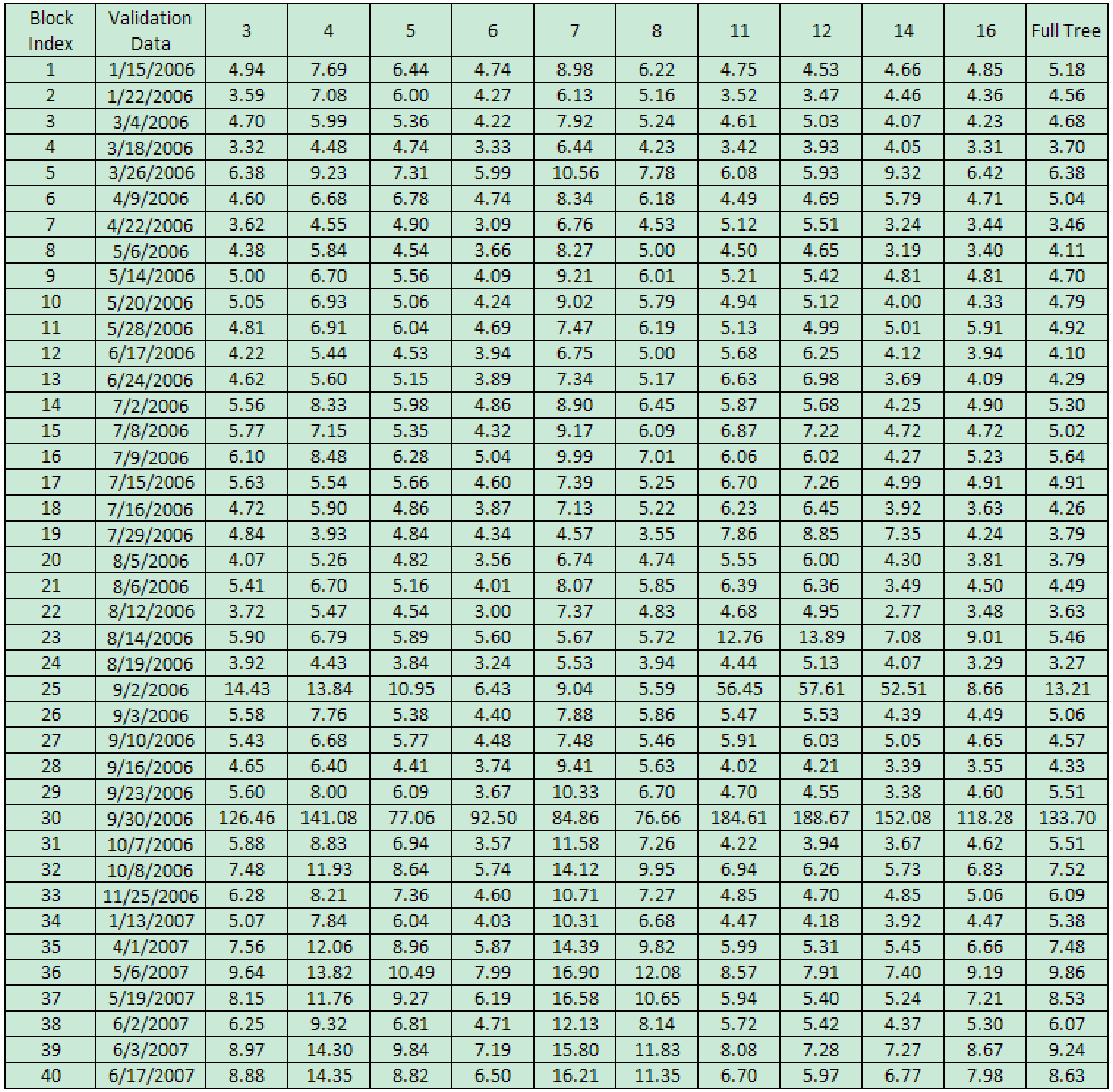
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	5/13/2006	6.12	7.18	244.79	221.61	226.05	211.59	7.20	7.52	221.14	140.04	5.50
2	8/13/2006	5.29	6.62	65.65	58.49	62.04	57.44	6.25	6.45	57.94	20.61	4.79
3	9/17/2006	5.86	8.39	155.10	139.84	146.16	117.54	5.10	4.86	130.46	43.65	5.19
4	12/31/2006	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 J-7: RCBD for Validation Data Sets in Characterization 11



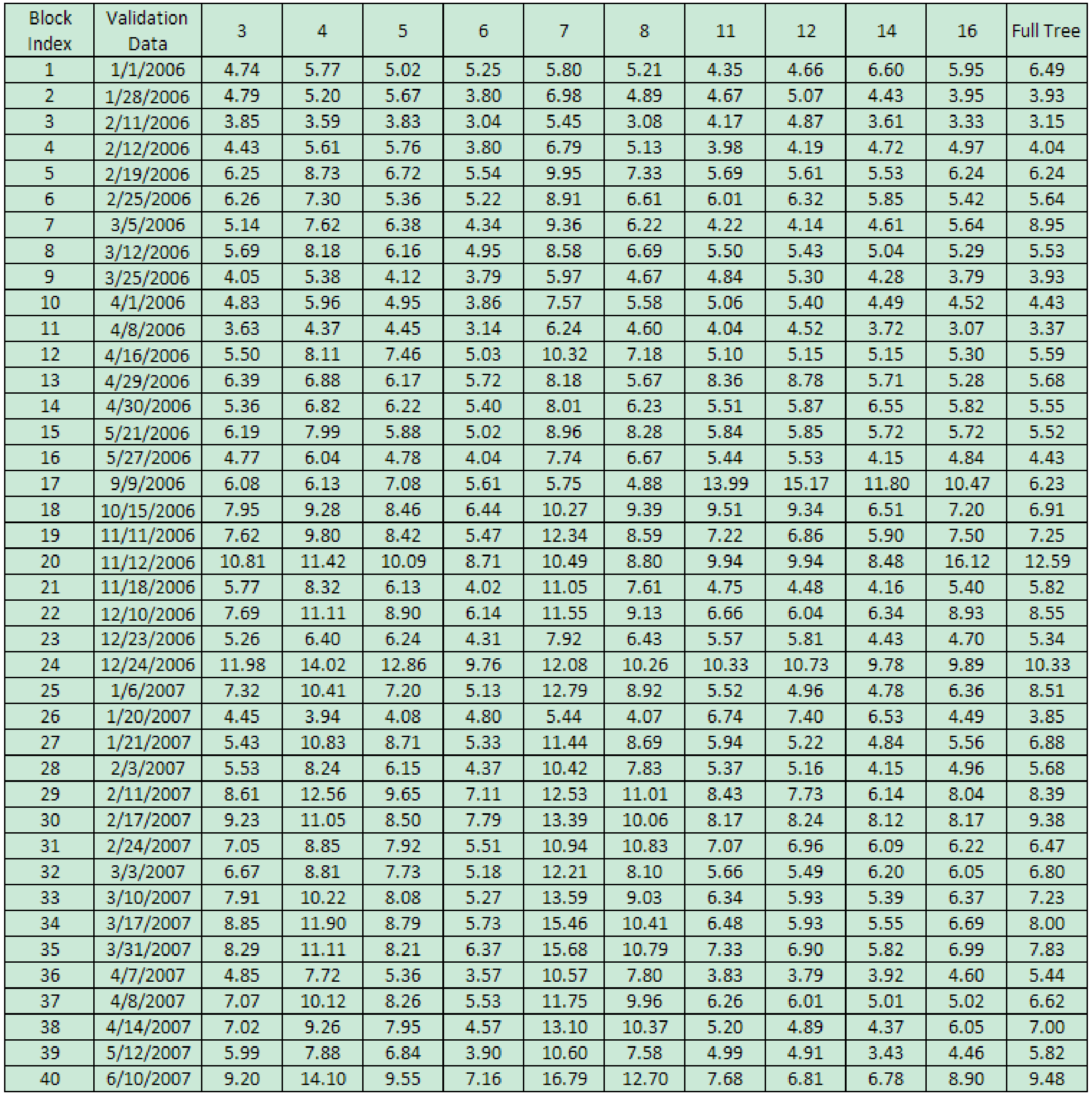
Block Index	Validation Data	3	4	5	6	7	8	11	12	14	16	Full Tree
1	2/4/2006	4.45	4.91	361.23	328.26	331.99	309.58	4.59	4.97	318.96	99.08	3.94
2	2/26/2006	5.14	7.37	244.71	221.03	225.00	212.29	4.44	4.44	232.54	68.68	4.92
3	4/2/2006	4.31	5.58	183.27	165.93	169.13	158.14	4.21	4.36	163.31	51.19	3.81
4	4/23/2006	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 J-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 J-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 J-10: RCBD for Validation Data Sets in Characterization 16

APPENDIX K:
ANOVA AND MULTIPLE COMPARISONS FOR RCBDS IN S-PLUS

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

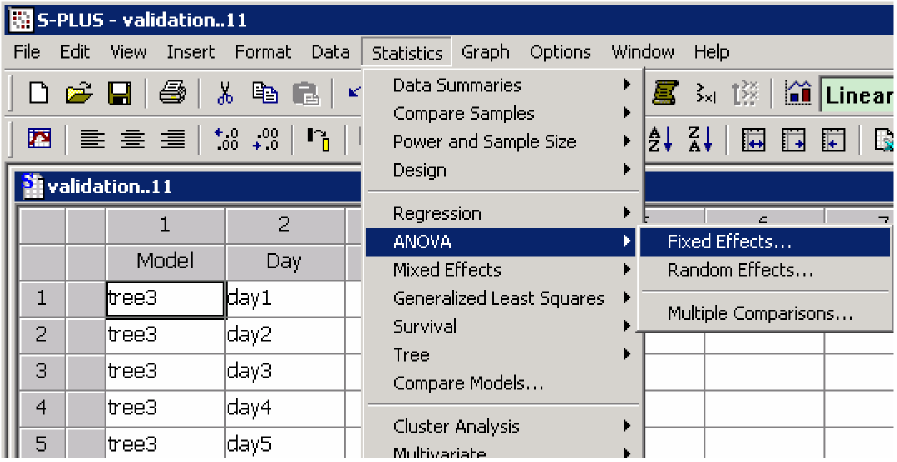
Table K-1: Adjusted Data Set of RCBD for Characterization 11

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	319.45
tree5	day9	54.07
tree5	day10	396.46
tree5	day11	21.44

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 J7 in Appendix J in this report, needs to be adjusted as shown in Table K1. In Table K1, 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 K1 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 K1 are just the first column of MSEs in Figure J7, the second eleven rows of MSEs in Table K1 are the second column of MSEs in Figure J7, etc.

After all the ten RCBDs are adjusted into the data sets in the manner shown in Table K1, 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.

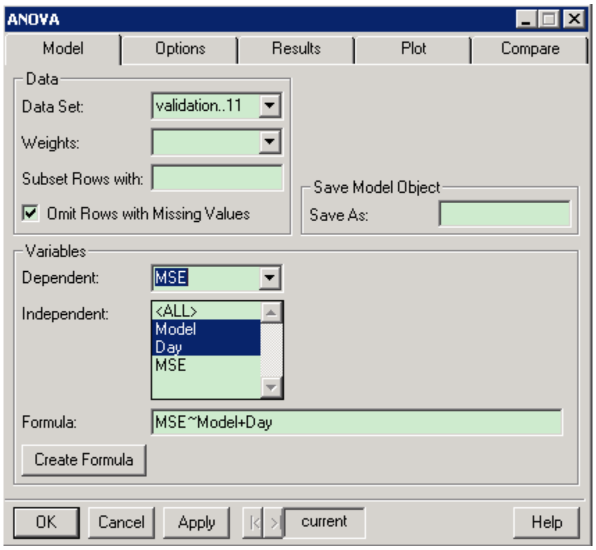
After importing the data set containing RCBD for validation data sets in characterization 11 into S-PLUS, by clicking Statistics>ANOVA>Fixed Effects, as shown in Figure K1, the ANOVA window for fixed effects is opened.



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

Figure K-1: Open ANOVA Window

In the opened ANOVA window, there are five tabs—Model, Options, Results, Plot and Compare, in which we only need to use Model tab, as shown in Figure K2, and Compare tab, as shown in Figure K3.



Data Set: validation..11

Weights:

Subset Rows with:

☑ Omit Rows with Missing Values

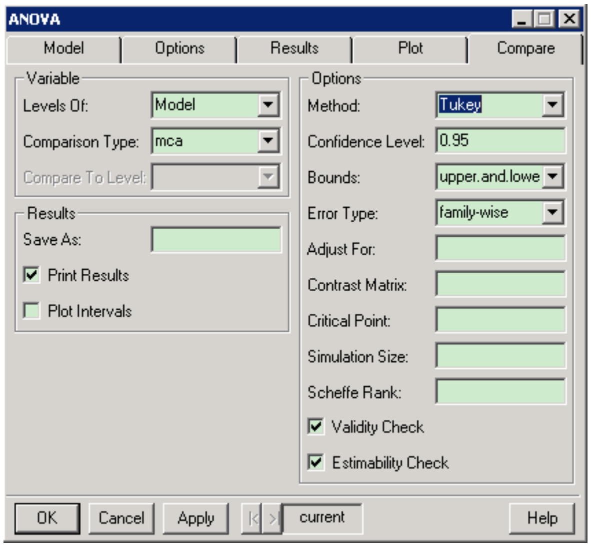
Save As:

Dependent: MSE

Independent: <ALL> Model Day MSE

Formula: MSE~Model+Day

Figure K-2: Model Tab in ANOVA



Variable

Levels Of: Model

Comparison Type: mca

Compare To Level:

Results

Save As:

☑ Print Results

□ Plot Intervals

Options

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 K-3: Compare Tab in ANOVA

In Model tab, dependent and independent variables need to be selected. In our RCBD, MSE is the response variable, Model is the single factor with 11 levels and Day is the day blocks using daily validation data sets. Thus, the dependent and independent variables are selected as shown in Figure K2. Compare tab is used for multiple comparisons, in which we only need to appropriately select “Levels Of” in Variable section in the top left corner and “Method” and “Error Type” in Options section in the right. 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 K3. For comparison method in Options section, we choose to use the conservative method Tukey’s method 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 K2 and Figure K3 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
```