

**DEVELOPMENT OF ROADWAY LINK SCREENING CRITERIA
FOR MICROSCALE CARBON MONOXIDE AND PARTICULATE
MATTER CONFORMITY ANALYSES THROUGH APPLICATION
OF CLASSIFICATION TREE MODEL**

A Thesis
Presented to
The Academic Faculty

by

Ghufran Shafi

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE IN ENVIRONMENTAL ENGINEERING in the
School of CIVIL & ENVIRONMENTAL ENGINEERING

Georgia Institute of Technology
May 2008

**DEVELOPMENT OF ROADWAY LINK SCREENING CRITERIA
FOR MICROSCALE CARBON MONOXIDE AND PARTICULATE
MATTER CONFORMITY ANALYSES THROUGH APPLICATION
OF CLASSIFICATION TREE MODEL**

Approved by:

Dr. Randall L. Guensler, Advisor
School of Civil & Environmental Engineering
Georgia Institute of Technology

Dr. Armistead G. Russell
School of Civil & Environmental Engineering
Georgia Institute of Technology

Dr. Michael O. Rodgers
School of Civil & Environmental Engineering
Georgia Institute of Technology

Date Approved: March 31, 2008

ACKNOWLEDGEMENTS

I would like to thank Dr. Randall Guensler, my advisor, whose guidance and advice made this research possible. I would also like to thank the committee members, Dr. Armistead G. Russell and Dr. Michael O. Rodgers for reviewing the thesis. All my colleagues, especially those at Drive Laboratory deserve my gratitude. Finally I am thankful to my parents for being a constant source of encouragement and the Institute of International Education (IIE) for supporting me for MS degree.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF SYMBOLS AND ABBREVIATIONS	viii
SUMMARY	ix
 <u>CHAPTER</u>	
1 INTRODUCTION	1
Objective	3
Approach	3
2 MICROSCALE DISPERSION MODELING	6
CALINE4 – Line Source Dispersion Model	8
3 DECISION TREE MODELS	10
Classification and Regression Tree	11
Construction of Tree Based Model	12
Over-Learning, Over-Fitting and Pruning	14
V-fold Cross Validation	15
Model Validity	17
4 CLASSIFICATION TREE AND RULES	18
Selection of Variable and Construction of Tree	18
Discussion	23
Conclusions and Recommendations	29

APPENDIX A: Programming Rules for CO	30
APPENDIX B: Programming Rules for PM	33
REFERENCES	37

LIST OF TABLES

	Page
Table 1: National Ambient Air Quality Standards	1
Table 2: Sample Cost Matrix	16
Table 3: Meteorological Parameters Used	19
Table 4: Classification Matrices for Tree	21

LIST OF FIGURES

	Page
Figure 1: Gaussian Profile of Air Pollutant Concentration Downwind of Source	7
Figure 2: CALINE4 Equivalent Finite Line Sources (FLS)	9
Figure 3: Dispersion Zone of Emission Source	9
Figure 4: Classification Tree of High and Low Risk Patients	14
Figure 5: Assumed Orientation of Link and Receptor	18
Figure 6: CV Costs and Reclassification Costs Plotted Against Terminal Nodes	20
Figure 7: Distribution of Correctly and Incorrectly Classified Data	22
Figure 8: QUEST Trees	25
Figure 9: Frequency Distribution of Contributed Concentrations of Misclassified Significant Links or False Negative Links	28

LIST OF SYMBOLS AND ABBREVIATIONS

C	concentration
CO	carbon monoxide
h	height of receptor
$i(t)$	impurity of node
l	length of a link
LE	linear emissions
phi	receptor angular distance
PM	particulate matter
$p(j t)$	conditional probability of class j in node t
q	lineal source strength
R	receptor linear distance
s.theta	wind directional variability
t	node index
u	wind velocity
σ	Gaussian dispersion parameter
CART	classification and regression tree
EPA	environmental protection agency
FLS	finite line source
NAAQS	national ambient air quality standards
NEPA	national environmental policy act

SUMMARY

The impacts of emissions sources of carbon monoxide and particulate matter pollution levels for projected level conformity assessment and National Environmental Policy Act (NEPA) analyses are usually estimated through computer-aided models. Because of the involvement and interaction of a large number of variables that affect formation of CO and PM hot spots, exhaustive impact assessment studies can be time consuming. This is especially true for complex urban projects consisting of numerous roadways whose potential CO and PM impacts on surrounding neighborhoods must be disclosed. A highway project may consist of hundreds of roadway links, therefore undertaking project level conformity analysis without screening tools can be computationally resource intensive.

CALINE4, a line source emission modeling tool, is used to predict downwind CO and PM concentrations for various receptors to generate a learning dataset for development of screening rules. This research has developed statistical screening criteria based on Classification and Regression Tree modeling that can be used to eliminate those links from the CALINE4 analysis whose contribution of pollutant concentration to a particular receptor site are insignificant. For the purpose of this study, any link that contributes a concentration of 0 ppm of CO or $0 \mu\text{g}/\text{m}^3$ of PM to a particular receptor site is termed insignificant for the corresponding pollutant. The model uses seven predictor variables, namely wind speed, wind directional variability, linear emission flux, link length and receptor polar coordinates. Response vector has two classes of pollutant

concentrations namely significant and insignificant which are obtained by conversion of numerical values of pollutant concentration according to above mentioned criterion, thereby converting a regression problem into categorical or classification problem.

The developed rules based on constructed model were validated through test samples and can be applied to future dataset to classify and screen out the insignificant links in highway planning analyses. The screening tool also allows analysts to prepare gridded pollution concentration predictions for use in environmental justice analyses.

CHAPTER 1

INTRODUCTION

The National Environmental Policy Act of 1969 (NEPA) requires that environmental impacts of federal and federalized state projects, including transportation projects, be integrated into the decision making processes. NEPA requires transportation planners to disclose the impacts of proposed projects on the human environment and to conduct environmental impact and environmental justice analyses. For air quality, NEPA analysis is carried out for National Ambient Air Quality Standard (NAAQS) set out by Environmental Protection Agency (EPA) for harmful pollutants. Carbon monoxide (CO) and particulate matter (PM) are two of the six criteria pollutants for which EPA has set ambient air quality standards which are listed in Table 1.

Table 1: National Ambient Air Quality Standards

Pollutant	Primary Standards	Averaging Times
Carbon Monoxide	9 ppm	8-hour
	35 ppm	1-hour
Particulate Matter (PM ₁₀)	Revoked	Annual (Arithmetic mean)
	150 µg/m ³	24-hour
Particulate Matter (PM _{2.5})	15 µg/m ³	Annual (Arithmetic mean)
	35 µg/m ³	24-hour

(Source: <http://www.epa.gov/air/criteria.html>)

Fossil fuels combustion in automobiles and trucks is a major source of carbon monoxide and particulates in urban areas; therefore, transport planners need to ensure that the construction of a new project will not contribute to violation of air quality standards. Federal and state law stipulates that quantitative CO and PM hotspot modeling be carried out for the surrounding locations of transportation projects to satisfy NEPA and project level conformity assessment (*CFR Title 40*).

Such hotspot modeling entails analyzing impacts of emission source for CO and PM concentrations, which are usually estimated through computer-aided models. EPA has developed regulatory guidelines for modeling techniques to accurately assess pollutant concentrations (*EPA: Air quality modeling*). For transportation projects, various line source emission modeling tools are available for estimating CO and PM concentrations, such as CALINE4 (*Benson, 1984*) and CAL3QHC (*EPA-454/R-92-006*). Such tools, though user friendly, are time consuming because complex urban projects can involve hundreds of roadway links and receptors of interest. Estimating worst-case scenario for air pollutants hotspot analysis means running of model over and over again in an iterative fashion so that impacts from every road link can be ascertained for every receptor. Therefore undertaking project level conformity analysis without screening tools can be computationally resource intensive. For example in the Atlantic steel project, an exhaustive scenario run involved more than one million dispersion computations or more than 50 hours of computational time (*Guensler, et al., 1999*). By applying a screening criteria based only on CO emission rates, one of the several factors that determine CO

concentrations, the researchers were able to reduce the computational time by more than half.

Objective

As stated earlier, conformity analysts work with a large number of highway links in complex road network to identify potential hotspots through intensive pollutant modeling. The objective of this research is to establish a set of statistical rules that can be used to classify links as either insignificant or significant based on their contribution to a pollutant concentration at a particular receptor. These rules can be applied to screen out the insignificant links to improve the overall processing time. For the purpose of this study, any link whose contribution under the worst-case scenario to that receptor is greater than 0 ppm for CO and $0 \mu\text{g}/\text{m}^3$ for PM is termed as significant for that particular pollutant¹. By doing so, a numerical decision problem is being converted to classification rule problem. Our decision rules will then simply divide the sample space into two regions such that those links that yield $C = 0$ fall into one region making such links insignificant, while significant links would fall into other region where $C > 0$.

Approach

Many statistical methods have been proposed by researchers for analysis of air pollutant concentrations (*Zoumakis, et al., 1994*). However, unlike those methods where

¹ CALINE4 rounds the concentration estimate to single digit after decimal point. Therefore, any concentration less than 0.05 ppm (or $\mu\text{g}/\text{m}^3$) is treated as insignificant since all lower values are rounded to zero.

emphasis has been placed on calculating concentrations generally by application of probability models, an effort has been made in this research to develop a set of comprehensive statistical screening rules that can be applied to air quality conformity analyses for substantial reduction in modeling workload. The statistical classification approach that has been adopted makes use of actual data generated by using a widely used line source modeling program, CALINE4, developed by Paul Benson of the California Department of Transportation (*Benson, 1984*). Worst-case scenario modeling has been undertaken using the software and classification rules have been developed that can discriminate between links on the basis of their impacts on a receptor.

Modeling requires selection of a modeling technique from a range of available alternatives. Much has been written about statistical modeling techniques and methods available for model selection are numerous (*e.g. Anderson and Burnham 1998*). Classification models fall into the category of predictive models where objective is to accurately predict the class of new data. There are three basic problems in development of statistical classification technique namely selection of variables, selection of a model and evaluation of the effectiveness of a model. These are discussed in the context of this research in following chapters which have been organized as follows: Chapter 2 describes microscale dispersion modeling and working of CALINE4 has also been outlined. Chapter 3 briefly presents the Classification and Regression Tree models, the working of such models and their advantages and disadvantages. Finally, the construction of statistical screening rules has been taken up in chapter 4 and it has been demonstrated

that substantial reduction in modeling workload can be achieved by applying these rules to screen out insignificant links from air quality modeling without compromising the predictive accuracy of pollutant impacts.

CHAPTER 2

MICROSCALE DISPERSION MODELING

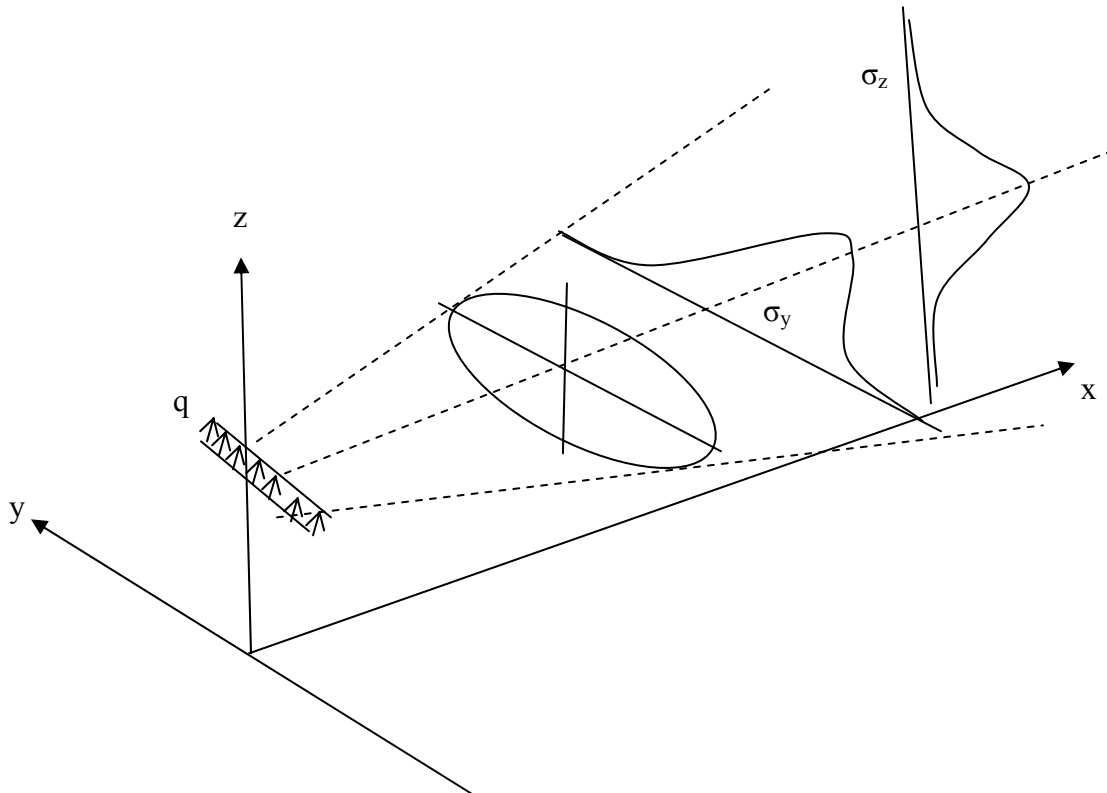
Air quality impacts analyses are generally classified according to pollutant transport range, and are classified at four level of analysis (*Ramaswami et al., 2005*). Contaminant transport that occurs at horizontal range of less than 1 km is treated as a microscale phenomenon. A mesoscale region analysis typically covers pollutant transport usually over the range of a metropolitan area. Synoptic scale transports analyses have a range starting at the upper level of microscale and will take pollutants over multi state region. Global scale phenomena are characterized by distances covering multiple nations. While transportation facilities may impact air quality over a mesoscale area, this research addresses microscale dispersion modeling, which is characteristic of individual transportation and corridor-level transportation project where downwind impacts are typically limited from 500 meters to one kilometer from the roadway.

Carbon monoxide and particulates are major impacts of a highway project over microscale area. Their concentrations are governed by plume rise and its dispersion and can be estimated by Gaussian plume equation which is based on analytical solution of steady state advection-diffusion equation. The equation assumes that advection occurs only in x-axis (perpendicular to line source) and diffusion occurs only in y and z directions. Diffusion in the x direction is negligible in comparison to advection which is the primary mode of downwind transport. Figure 1 shows downwind concentration profile of a line source oriented perpendicular to the wind as given by Gaussian equation.

Incremental concentration from the line source can be determined by integrating with respect to differential line segments:

$$C(x, y) = \int_{y_1}^{y_2} \frac{q}{2\pi u \sigma_y \sigma_z} \exp\left(\frac{-y^2}{2\sigma_y^2}\right) \left(\exp\left[\frac{-(z-h)^2}{2\sigma_z^2}\right] + \exp\left[\frac{-(z+h)^2}{2\sigma_z^2}\right] \right) dy$$

Where q is the lineal source strength, u is the wind speed, σ_y and σ_z are the horizontal and vertical Gaussian dispersion parameters, y_1 and y_2 are the line source end points and z and h are the height of source and receptor respectively.



(Source: Ramaswami et al., 2005)

Figure 1: Gaussian Profile of Air Pollutant Concentration Downwind of Source

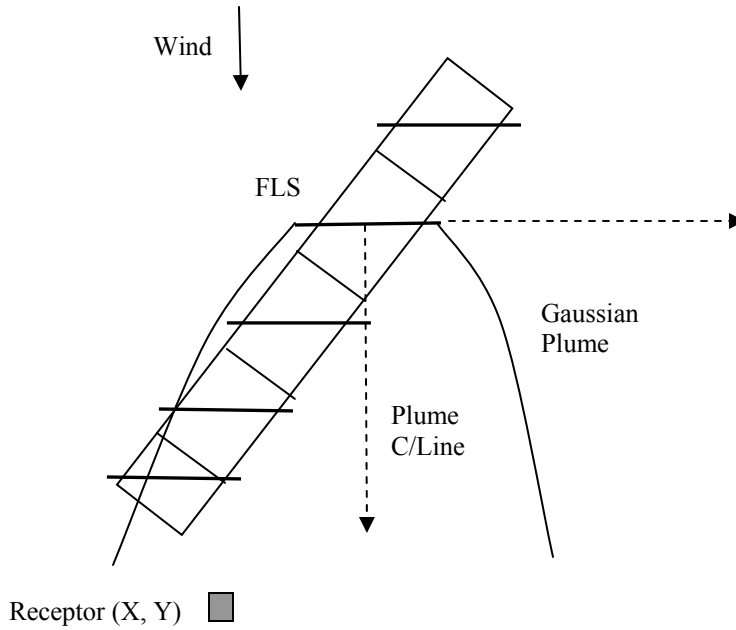
CALINE4 – Line Source Dispersion Model

CALINE4 is a line source emission modeling tool developed by Paul Benson of the California Department of Transportation. CALINE4 makes use of Gaussian dispersion equation and employs a mixing zone concept in which initial horizontal dispersion is imparted to pollutants by mechanical turbulence created by moving vehicle and thermal turbulence created by hot vehicle exhaust. CALINE4 divides individual highway links into series of finite elements. Individual incremental concentration for each element is calculated using Gaussian plume equation for line source; the incremental concentrations are summed to get link concentration.

As shown in Figure 2, each element is modeled as equivalent finite line source (FLS) positioned normal to the wind direction and centered at the element mid point. The length and orientation of the FLS are the functions of element size and roadway wind angle. Element size increases with distance from the receptor to improve computational accuracy.

For constant meteorological parameters, pollutant concentrations at a receptor are a function of source strength, distance to receptor and wind speed and its variability. Transportation links impact the receptors. However, the impact of each link on the receptor is a function of the dispersion zone of that link. Transportation links that are located a great distance from the receptor, or from which very little pollutant emissions are generated, have no significant impact on the pollutant concentrations at that receptor. Hence, any link in the transportation system can be defined as having either a significant

or insignificant impact on a receptor under worst-case meteorological conditions. In Figure 3, the pollutant concentration at the receptor is only impacted by two links, and as such only these two links need be treated as significant.



(Source: Benson, 1984)

Figure 2: CALINE4 Equivalent Finite Line Sources (FLS)

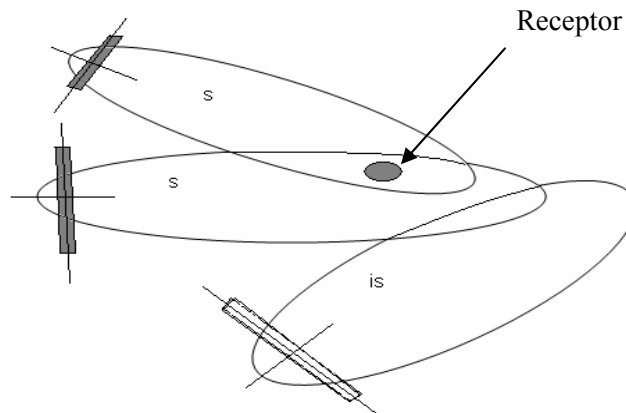


Figure 3 – Dispersion Zone of Emission Source Determines Significant or Insignificant Impact. Receptor Falls in the Dispersion Zone of Significant Links (Gray) and Outside of Insignificant Link (White)

CHAPTER 3

DECISION TREE MODELS

Tree-based modeling is a widely used technique in data mining. The modeling technique generally uses a binary classification system to construct a ‘decision tree’, which is a tree-structured classification system consisting of set of attributes that can be used to assign class membership of the dependent variable (classification tree) or predict its numerical value (regression tree).

A decision tree model is constructed by using historical data and it generates set of rules that can then be used to predict results from new data. Tree-based models have much in common with the traditional classification methods in statistical studies. However because of its recursive hierarchical form as compared to simultaneous decisive forms of traditional methods, it is more flexible. The methodology is inherently non-parametric and is not affected by underlying distribution of independent variables. It is fairly robust with outliers, which can be isolated in distinct node or nodes. Another property which makes this tool very flexible is that it is invariant to monotone transformation of independent variables. Decision trees are more adept in identifying non-additive and non-linear effects in the data (*Chambers and Hastie (1993), Washington, S (2000)*). Once developed, tree-based rules can be validated through what is known as “dropping” - cases from a new independent dataset are dropped through the tree and the results are observed. Testing of tree on a new dataset ensure the validity of the model tree.

Classification and Regression Tree

Classification and Regression Tree (CART) Algorithm was made popular by Breiman et al., (1984) and is considered the reference on decision trees. For classification, the CART algorithm predicts class membership by generating a set of classification rules, also known as classifiers, in the form of if-then statements. As an example of classification tree problem, Breiman et al, (1984) developed a decision tree for those heart attack patients which immediately after hospitalization need to be identified as high risk patients. Their tree shown in Figure 4 is a simple three question tree. The high risk level is then identified as one which satisfies the condition “if the minimum systolic blood pressure is greater than 91 and if the age is greater than 62.5 and if the patient displays sinus tachycardia, then and only then the patient is at high risk.”

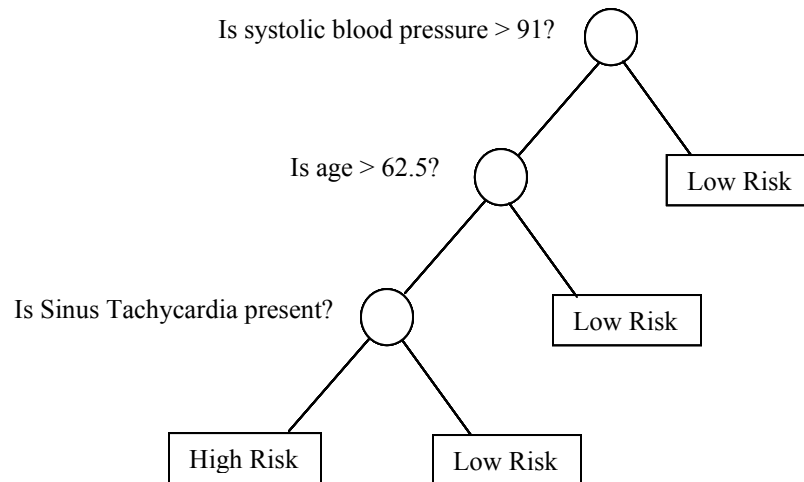


Figure 4: Classification Tree of High and Low Risk Patients.

Construction of Tree-Based Model

A range of different algorithms have been proposed by researchers for construction of decision trees. All of them have to answer two basic questions to develop univariate split classifiers: 1) which predictor variable should be selected? and 2) what value or class of variable should split a node? The CART methodology (*Breiman et al, 1984*) examines all possible binary splits of data along each predictor variable before selecting the split that most significantly reduces some measure of node impurity in child nodes. Several node impurity functions have been defined in literature. Gini index is one of the most common splitting functions and is defined as sum of products of all pairs of class proportions for classes present at the node. Mathematically, the Gini index measure of impurity, $i(t)$ of a node t containing number of j classes can be expressed as $i(t) = \sum_j p^2(j|t)$ where $p(j|t)$ is the conditional probability of class j in node t . Impurity is at maximum for a node when all the class sizes are equal and minimum when there is only one class present. The node with only one class is said to be pure, or homogenous. Therefore at a node, CART will exhaustively search through all possible values of all variables in the dataset and will select a split such that the child nodes have minimum impurity.

The classification tree algorithm which has been used in this research is known as QUEST (Quick, Unbiased, Efficient Statistical Trees). QUEST is a binary split tree-structured classification model, developed by Loh and Shih (1997-2005) which deals with the problem of variable selection and split point selection separately. The algorithm

- gives univariate split, although an option of linear combination split is also available,
- uses an unbiased variable selection technique, which is important because some classification trees such as C&RT (*Breiman et al., 1984*) that employs exhaustive search for variable selection have a bias toward selecting variables offering more levels of splits and such bias can affect the predictive accuracy in independent samples though not necessarily in learning sample (*Cameron-Jones et al., 1995*)
- includes a family of splitting criteria,

QUEST involves use of fairly technical algorithms. However, unlike exhaustive search, QUEST does not combine the problem of variable selection and split point selection. QUEST uses statistical tests of significance for relationship of class membership with each predictor, using a suggested or user-specified alpha value. If the data consist of both numerical and categorical variables, the algorithm calculates p-value from ANOVA F-test and Chi-Squared test for numerical and categorical variable respectively. Thus, the main idea is to select a variable to split a node on the basis of statistical significance test instead of impurity function. Once a variable is selected, QUEST employs a modification of recursive quadratic discriminant analysis to determine the best split point. Complete details of the methodology can be found in Loh and Shih (1997). If the number of classes is two then QUEST offers exhaustive search for split selection as the default option. For the purpose of this research both methods of split selection (discriminant analysis and exhaustive search) were tried and the exhaustive search was found to construct a tree with better predictive accuracy.

Over-Learning, Over-Fitting and Pruning

The objective of classification tree analysis is to obtain maximum predictive accuracy. The predictive accuracy of a tree model is defined in terms of “misclassification cost”, which in simplified terms is the proportion of misclassified cases. A major problem with construction of a classification tree using a learning sample is that the model will always “over-learn” or “over-fit” the data. That is, the model will try to accommodate every random noise that is present in the sample. If not stopped, the tree algorithm will extract all information, both useful as well as redundant, from the data until a maximal tree has been built. Such a tree will perform nicely on the learning sample, though not necessarily on an independent test sample. To tackle this problem, some researchers have suggested the implementation of a direct stopping rule. A tree is built to a desired size by specifying the maximum number of observations in the terminal node or when improvement due to additional splits is less than some specified value of misclassification. Breiman, et al., (1984) pointed out many problems with the direct stopping rule and introduced an innovative concept for selection of right-sized trees. They suggested that instead of stopping the tree in mid-growth, one should grow an over-large tree and then prune it back to optimal size. An automatic procedure known as minimum cost-complexity pruning was suggested by Breiman, et al., (1984) for selection of right-size trees.

With the increase in tree complexity (defined as tree size) there is a corresponding decrease in misclassification error. The idea behind minimum cost-complexity pruning is an optimization between complexity and misclassification and the technique works through the complexity function (defined as costs for the tree plus a

complexity parameter times the tree size). The complexity parameter is a real number with initial value of 0. Starting with the terminal node, the value is increased continuously until the value of complexity function for the largest tree exceeds the value of the function for a smaller sized tree. Take the smaller-sized tree to be the new largest tree, continue increasing the complexity parameter continuously until the value of the function for the largest tree exceeds the value of the function for a smaller-sized tree, and continue the process until the root node is the largest tree. Thus the function is a linear combination of cost of the tree and its complexity. As the complexity parameter is increased, the larger trees are penalized more and more for their complexity until a point when a larger tree's complexity outweighs a smaller tree's cost.

This automatic procedure of pruning generates a sequence of simpler and simpler trees, each of which is a candidate for the final tree. It is then only a question of selection of the right-sized tree. Breiman et al., (1984) suggested 1-SE rule whereby the best tree is selected as the smallest-sized tree (least complex tree) whose costs do not exceed the minimum costs plus 1 times the standard error of the costs of the minimum cost tree.

V-fold Cross-Validation

Cross validation is a preferred method of estimating the predictive accuracy of a model. The learning sample is split into specified V random sections, as nearly equal in size as possible. Classification tree is constructed V times, each time leaving out one section which is treated as test sample for cross-validation so that each section is used $V-1$ times for learning sample and only once as test sample. The misclassification costs computed for each of the V test samples are then averaged to give the V -fold estimate of

the CV costs. Cross-validation gives reasonably accurate estimate of performance of the model built from learning sample on an independent testing sample.

Another useful option that a classification tree offers is the decision cost matrix. Misclassification cost is incorporated as a parameter in tree models because sometimes accurate predictions for some classes are more desirable than accurate predictions for other classes. For example if a conformity project belongs to such borderline area where the probability of its impacts violating the air quality standards is high, the analyst might want to minimize the misclassification of significant links. In other words it would be much more serious an error to classify a link as insignificant when it actually impacts a receptor than it would be to classify a link as significant when it is not impacting a receptor. Such a sample variable cost matrix is shown in Table 2. This variable cost matrix $\{C(ij)\}$ – cost of misclassification of class j case as class i case can be incorporated into node impurity function.

Table 2: Sample Cost Matrix

		Predicted by QUEST	
		Significant (S)	Insignificant (IS)
Actual Class	Significant (S)	0	2
	Insignificant (IS)	1	0

We ran our models incorporating equal as well as variable cost matrix and observed that by making $\{\text{Cost (IS|S)}\} = \{2*\text{Cost (S|IS)}\}$, the misclassification of

significant links dropped considerably in both training and test sample. Nonetheless, overall misclassification rate was minimal when equal misclassification costs were assigned and the trees in Figure 8 (Chapter 4) were constructed by assigning equal misclassification costs.

Model Validity

One of the fundamental issues for a classification technique is verifying the reliability of the model. Some sort of “goodness of fit” must be defined for the model in order to verify it for both the training and test samples. For a classification tree with equal misclassification costs assigned to classes, such goodness of fit can simply be measured by proportion of misclassified cases. The classification power of a tree is inversely proportional to impurity of terminal nodes. Thus a misclassification of 0 for a node represents a completely homogenous node in which there are links of one class only. The important factor in reliability of classification trees is to validate it by test sample that is to apply the developed rules to independent data and check the performance of the tree. The model should only be accepted when the classification rules hold both for learning data set as well as test data set.

CHAPTER 4

CLASSIFICATION TREE AND RULES

Selection of Variables and Construction of Tree

To develop the classification tree model for identifying link significance in microscale modeling, a data set was developed that contained observations consisting of two components, namely a categorical outcome or response variable and a set of predictors or independent variables. The training dataset was developed by running CALINE4 to generate observations on 6 predictors (all ordered variables), namely, link length (m), linear hourly emissions (g/hr/mi), wind speed (m/s), wind directional variability (theta) and receptor polar coordinates (m & theta). As shown in Figure 5, the link was oriented in North-South direction and the center of the link was treated as the origin. Symmetry of the problem allowed that receptor be positioned only in 1st (NE) coordinate. Pollutant concentration was the numeric dependent variable which was converted to a categorical value of significant or insignificant class as outlined earlier, i.e. for predicted marginal impact $C = 0$ ppm (for CO) or $0 \mu\text{g}/\text{m}^3$ (for PM) the link was deemed insignificant and significant otherwise.

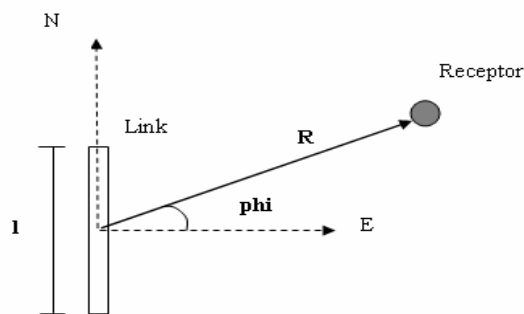


Figure 5: Assumed Orientation of Link and Receptor

Because conformity analysis is carried out for worst-case pollution concentrations, worst-case job and meteorological parameters have been used for model runs listed in Table 3 and are based on studies conducted by Guensler, et al., (1999) simulating the January conditions for Atlanta, GA.

Table 3: Meteorological Parameters Used

Aerodynamic roughness coefficient	Suburban
Stability Class	7
Mixing height	40m
Ambient temperature	0°C

Classification trees in Figure 8 were constructed for a learning sample which consisted of 1800 observations on previously mentioned predictors. CO and PM were response variables in Figure 8a and Figure 8b respectively. Both models were trained by assigning equal costs for misclassification of classes i.e. $\text{Cost}(I|S) = \text{Cost}(S|I) = 1$; where $\text{Cost}(i|j)$ is cost of misclassifying class j as class i . A maximal tree is built after which it is pruned back by 10-fold cross validation. CV costs and learning sample cost of successively pruned trees have been shown in Figure 6 from which desired size tree can be selected. Use of this automatic method helps avoid over-fitting and under-fitting of data.

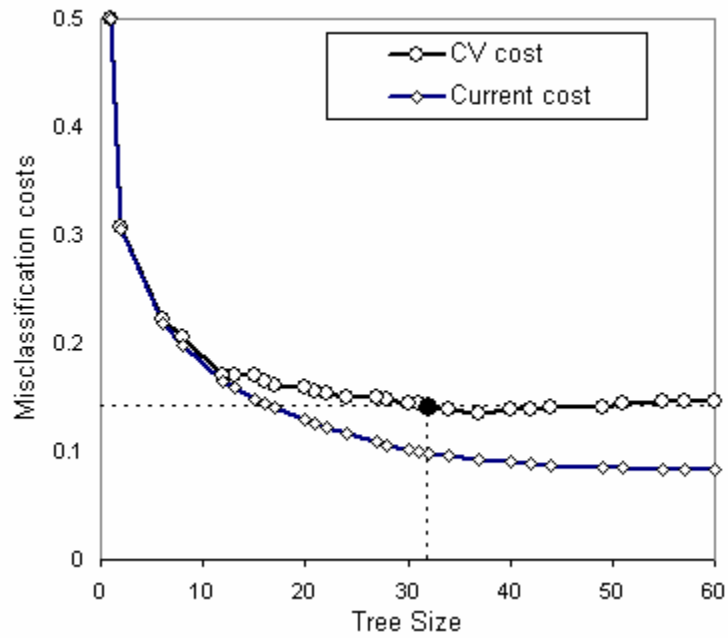


Figure 6a – By 1-SE Rule, a Tree with CV Costs of .14 and 32 Terminal Nodes is Selected. 0-SE would have Produced a Tree with CV Costs of .136 and 37 Terminal Nodes.

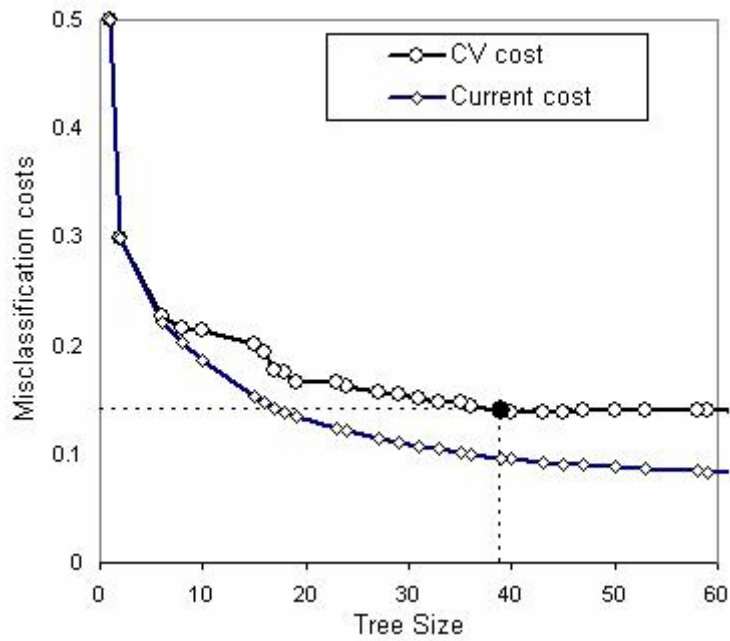


Figure 6b – By 1-SE Rule, a Tree with CV Costs of .1397 and 39 Terminal Nodes is Selected. 0-SE would have Produced a Tree with CV Costs of .1377 and 45 Terminal Nodes.

Figure 6: CV Costs and Reclassification Costs Plotted Against Terminal Nodes

Final trees that have been selected using 1-SE rule are shown in Figure 8. The final tree consisted of 32 terminal nodes for CO and 39 terminal nodes for PM; each terminal node representing a classification rule in the form of logical if-then statement (Screening rules for CO and PM are listed in Appendix A and Appendix B respectively). Thus a set of classification rules were obtained on which future datasets could be tested.

The validity of trees was analyzed through the classification rate, both for learning and future data. Classification matrices for learning and future dataset are shown in Table 4 and depicted in Figure 7. Overall misclassification rates for the learning and future dataset are 10.0% (11% for PM) and 16.1% (13% for PM), the rate of misclassifying an insignificant link as significant is 9.3% (6.5% for PM) and 16.8% (17.2% for PM) and rate of misclassifying a significant link as insignificant is 10.4% (12.6% for PM) and 15.9% (11.7% for PM) respectively.

Table 4a: Classification Matrices for CO Tree

Learn sample (1800 Observations)				Test sample (360 Observations)			
Actual Class	Predicted Class		% Correct	Actual Class	Predicted Class		% Correct
	IS	S			IS	S	
IS	584	60	91 %	IS	74	15	83 %
S	121	1035	90 %	S	43	228	84 %

Table 4b: Classification Matrices for PM Tree

Learn sample (1800 Observations)				Test sample (360 Observations)			
Actual Class	Predicted Class		% Correct	Actual Class	Predicted Class		% Correct
	IS	S			IS	S	
IS	567	40	93 %	IS	72	15	83 %
S	151	1042	87 %	S	32	241	88 %

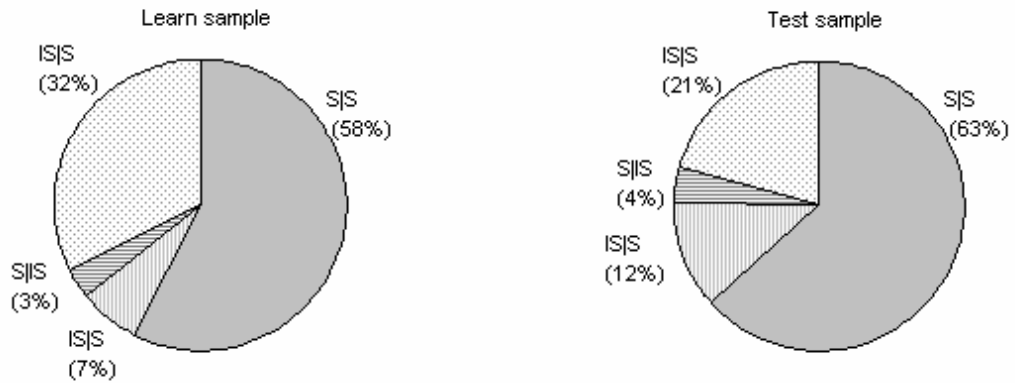


Figure 7a – Distribution of Correctly and Incorrectly Classified CO Data

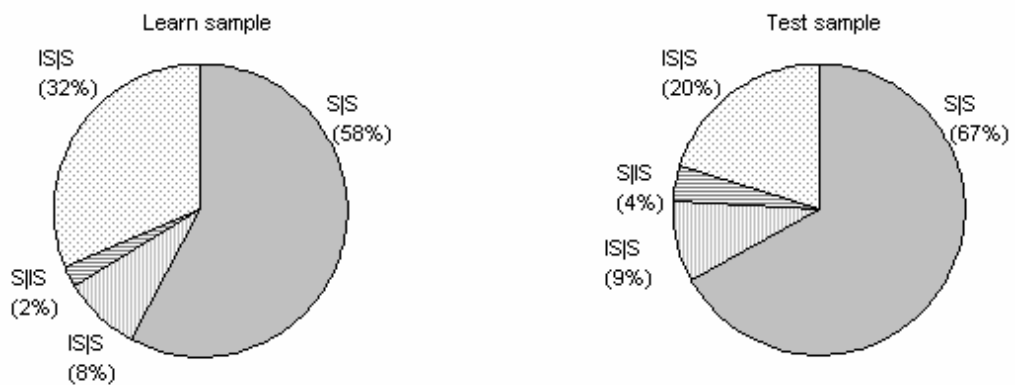


Figure 7b – Distribution of Correctly and Incorrectly Classified PM Data

Figure 7: Solid Gray and Dotted Gray Represents Correctly Classified Significant And Insignificant Links Respectively. Horizontal and Vertical Gray Lines Represent Incorrectly Classified Significant and Insignificant Links Respectively.

We will discuss the impacts of misclassification on net predicted CO and PM concentrations in the next section. It will be demonstrated that whereas computational workload for hotspot modeling can appreciably be reduced by using the screening rules, the impact on predicted emissions using the screening rules and without using the screening rules are not significant for any given scenario.

Discussion

Tree-based models are useful exploratory techniques and reveal important patterns and relationships in data. Here we have advantage of comparing our model side by side with mathematical model which helps cross checking the validity of statistical results. QUEST uses significance test to select the split variables and the highly significant variables are selected first. Splits near the top reflect main effects and/or lower-order interactions. Conversely, splits at the bottom reflect higher-order interaction and apply only to small subsets of the sample. Figure 8 shows that first split (in both models) is made on phi followed by R, which are receptor polar coordinates with respect to link centroid confirming that distance and direction are the most important factors in deciding the impacts of linear emission source. For CO, CALINE4 is more sensitive to directional variability for parallel winds than cross winds (*Benson, 1984*). This is confirmed by our CO model where it treats s.theta to be more significant for a region with $\text{phi} > 43.50$. Note that phi is direction of receptor and not of wind but since tree has been constructed for worst-case; it is safe to assume that for $\text{phi} > 45$, model is generating parallel wind conditions. The only variable which QUEST is unable to capture is wind speed which has been used as split variables at the lower nodes only. CO concentration is sensitive to wind speed in more than one way. Wind speed has an inverse relationship with concentration in the Gaussian formula. CALINE4 uses mixing zone concept and wind speed determines the mixing zone residence time used for computing the initial vertical dispersion parameter. It is also used in horizontal dispersion parameter. Why QUEST does not select u as split variable at top nodes is inexplicable except that the afore mentioned relationship holds when the concentration is treated numerically. Here

we have converted our response variable to a categorical one on a totally theoretical criterion rather than a statistical one and in doing so the classification tree might not have been able to completely reveal the significant relationship of wind speed.

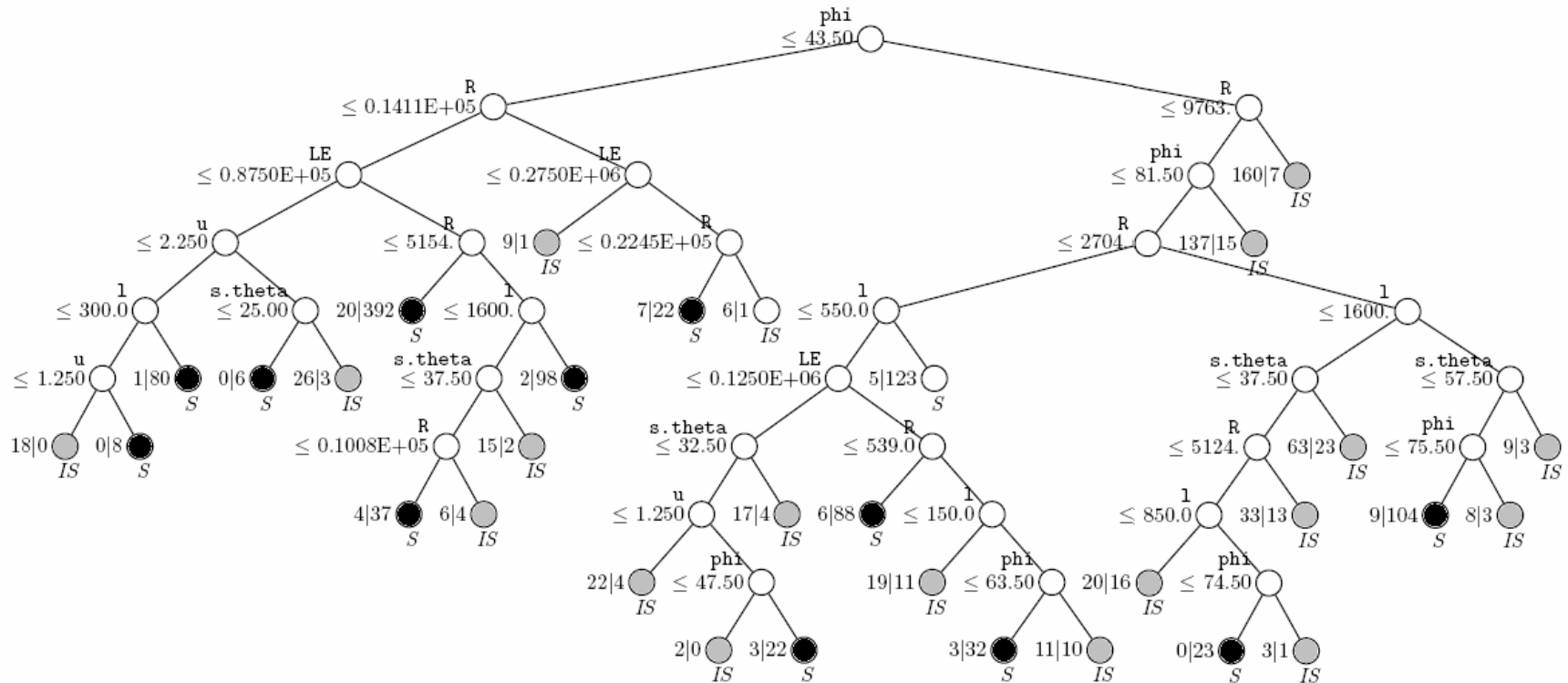


Figure 8a - QUEST Tree for CO. Sold gray represents a predicted insignificant and solid black represents a predicted significant terminal node. The list of numbers beside a terminal node gives the number of learning samples for each class in the node. Their class labels, from left to right are IS, S. Splitting rule for each intermediate node is given beside the node. Variables are R & phi = receptor polar coordinates, LE = linear emissions, u = wind speed, s.theta = wind directional variability and l = link length. Every terminal node represents a logical if-then rule. For instance the right most terminal node corresponds to if phi > 43.5° and if R > 9763m then link is insignificant.

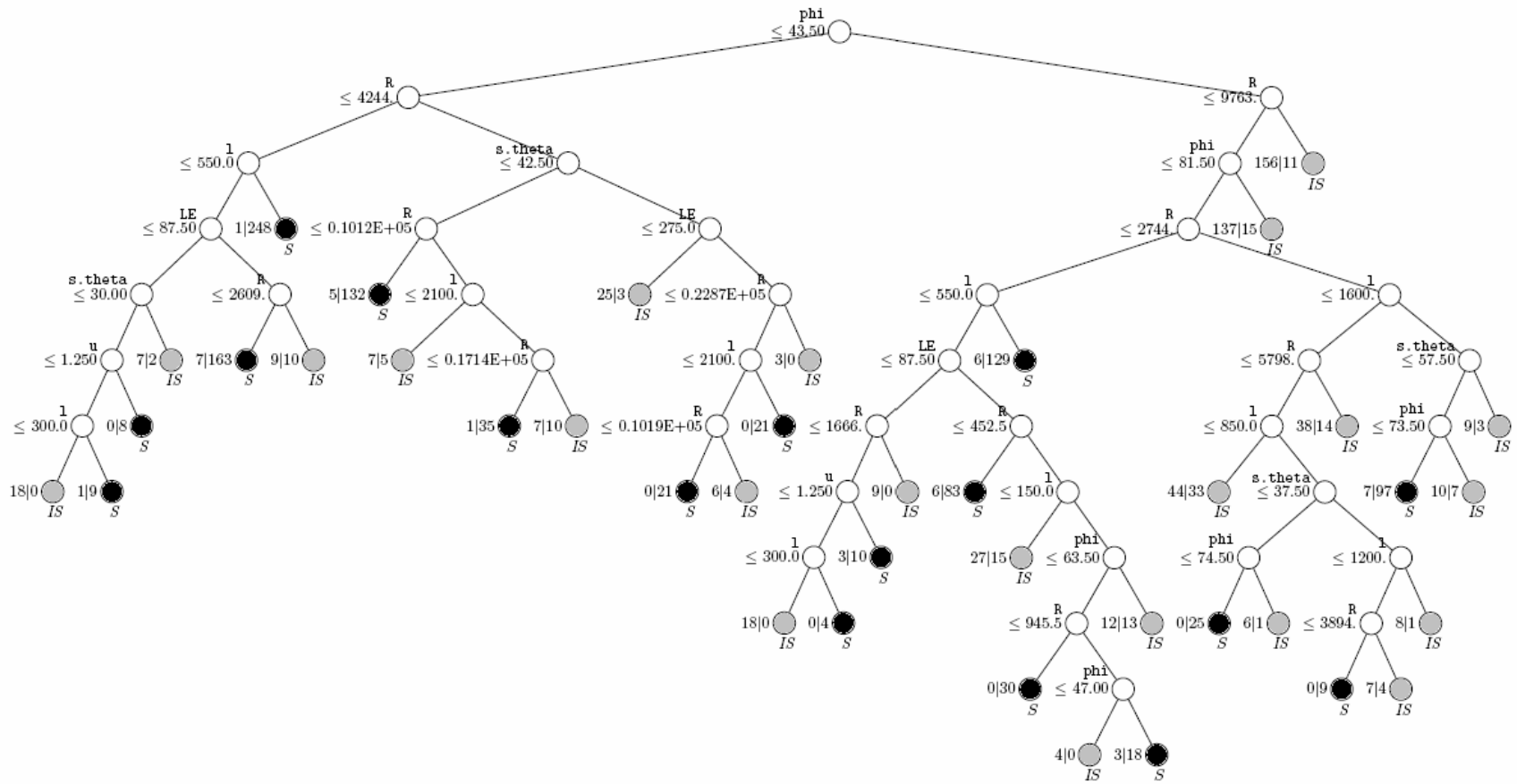


Figure 8b - QUEST Tree for Particulates. Sold gray represents a predicted insignificant and solid black represents a predicted significant terminal node. The list of numbers beside a terminal node gives the number of learning samples for each class in the node. Their class labels, from left to right are IS, S. Splitting rule for each intermediate node is given beside the node. Variables are R & phi = receptor polar coordinates, LE = linear emissions, u = wind speed, s.theta = wind directional variability and l = link length. Every terminal node represents a logical if-then rule. For instance the right most terminal node corresponds to if phi > 43.5° and if R > 9763m then link is insignificant.

Of the misclassified classes, it is worthwhile to analyze the false negatives i.e. those significant links that have been classified as insignificant. Identification of this group is important because if the misclassified significant links are such that they are having a major impact on the receptors and are still misclassified as insignificant, that can cause serious doubts about the model reliability.

For CO, there are 121 false negative cases in the learning sample and 43 in the test sample. Figure 9a shows the frequency of observations of misclassified significant links for different contributed CO concentrations. In the learning dataset for CO, 51 out of 121 misclassified links were contributing only 0.1 ppm to receptors while only 25 of the misclassified links were making a contribution of more than 0.5 ppm. In the test sample 25 of the 45 misclassified significant links were contributing 0.1ppm, while only 4 links were contributing more than 0.5 ppm. Similarly for particulate matter, 151 cases in learning sample and 32 in test sample are false negatives. The frequencies of PM concentrations for misclassified significant links are shown in Figure 9b. Out of 151 misclassified significant links in learning sample, 65 links were contributing $0.1 \mu\text{g}/\text{m}^3$ and only 26 were contributing more than $0.5 \mu\text{g}/\text{m}^3$. Corresponding number of false negative links for test sample with a total of 32 such cases are 13 and 4 respectively. Therefore it would be reasonable to say that even the incorrectly classified insignificant cases share the characteristics of insignificant ones and developed screening rules are applicable in conformity analyses.

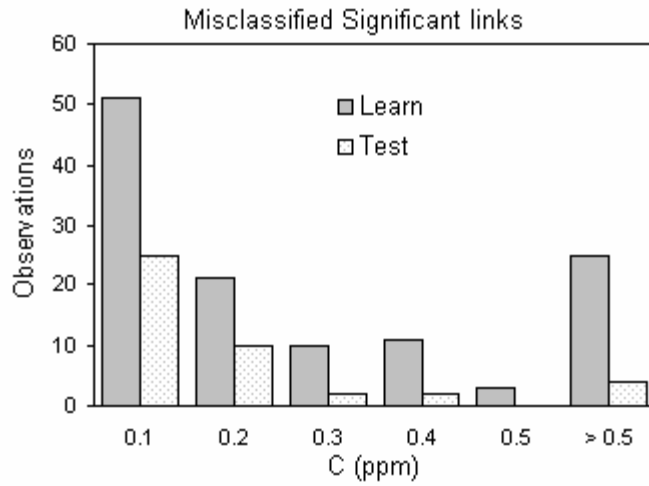


Figure 9a: Carbon Monoxide

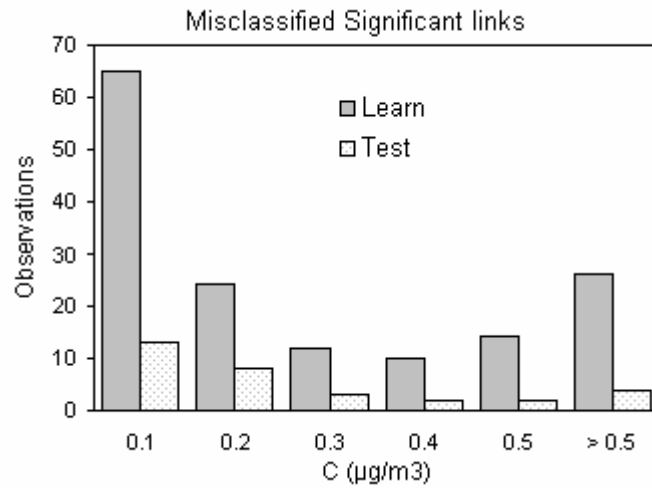


Figure 9b: Particulate Matter

Figure 9: Frequency Distribution of Contributed Concentrations of Misclassified Significant Links or False Negative Links.

Conclusions and Recommendations

Undertaking impact analyses of transportation projects on air quality to determine potential pollutant hotspots is an elaborate and complex process. Exhaustive impact studies, which need to take into account each roadway link, add to the complexity of the project increasing the required computational resources and computational time. Elimination of the roadway links, whose dispersion zones do not impact receptor under consideration significantly, will result in substantial reduction in the overall processing time of air quality modeling in major infrastructure projects. The statistical screening rules that have been developed can be used in the initial stages of hotspot modeling for any microscale region. The rules are conservative in that they provide for individual assessment of a link on an individual receptor for worst-case scenario.

Constant meteorological parameters have been used in this research. Future improvements in the screening model can involve incorporating meteorological parameters as variable in the model and new rules can be developed. The model can be expanded to include other pollutants which are important in the context of ambient air quality. Model can also be extended in scale so that to cover an entire metropolitan region. Finally, a user interface can be developed that has the codified screening rules which can be used in gridded pollutant concentration predictions.

APPENDIX A

Programming Rules for CO

RULE 1: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE \leq 87500$, $u \leq 2.25$, $l \leq 300$, $u \leq 1.25$))
Then “Insignificant”

RULE 2: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE \leq 87500$, $u \leq 2.25$, $l \leq 300$, $u > 1.25$))
Then “Significant”

RULE 3: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE \leq 87500$, $u \leq 2.25$, $l > 300$))
Then “Significant”

RULE 4: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE \leq 87500$, $u > 2.25$, $s.\theta \leq 25$))
Then “Significant”

RULE 5: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE \leq 87500$, $u > 2.25$, $s.\theta > 25$))
Then “Insignificant”

RULE 6: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE > 87500$, $R \leq 5154$))
Then “Significant”

RULE 7: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE > 87500$, $R > 5154$, $l \leq 1600$, $s.\theta \leq 37.5$,
 $R \leq 10080$))
Then “Significant”

RULE 8: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE > 87500$, $R > 5154$, $l \leq 1600$, $s.\theta \leq 37.5$,
 $R > 10080$))
Then “Insignificant”

RULE 9: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE > 87500$, $R > 5154$, $l \leq 1600$, $s.\theta > 37.5$))
Then “Insignificant”

RULE 10: IF (AND ($\phi \leq 43.5$, $R \leq 14110$, $LE > 87500$, $R > 5154$, $l > 1600$))
Then “Significant”

RULE 11: IF (AND ($\phi \leq 43.5$, $R > 14110$, $LE \leq 275000$))
Then “Insignificant”

RULE 12: IF (AND ($\phi \leq 43.5$, $R > 14110$, $LE > 275000$, $R \leq 22450$))
Then “Significant”

RULE 13: IF (AND ($\phi \leq 43.5$, $R > 14110$, $LE > 275000$, $R > 22450$))
Then “Insignificant”

RULE 14: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE \leq 125000$, $s.\theta \leq 32.5$, $u \leq 1.25$))
Then “Insignificant”

RULE 15: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE \leq 125000$, $s.\theta \leq 32.5$, $u > 1.25$, $\phi \leq 47.5$))
Then “Insignificant”

RULE 16: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE \leq 125000$, $s.\theta \leq 32.5$, $u > 1.25$, $\phi > 47.5$))
Then “Significant”

RULE 17: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE \leq 125000$, $s.\theta > 32.5$))
Then “Insignificant”

RULE 18: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE > 125000$, $R \leq 539$))
Then “Significant”

RULE 19: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE > 125000$, $R > 539$, $l \leq 150$))
Then “Insignificant”

RULE 20: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE > 125000$, $R > 539$, $l > 150$, $\phi \leq 63.5$))
Then “Significant”

RULE 21: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l \leq 550$, $LE > 125000$, $R > 539$, $l > 150$, $\phi > 63.5$))
Then “Insignificant”

RULE 22: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R \leq 2704$, $l > 550$))
Then “Significant”

RULE 23: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l \leq 1600$, $s.\theta \leq 37.5$, $R \leq 5124$, $l \leq 850$))
Then “Insignificant”

RULE 24: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l \leq 1600$, $s.\theta \leq 37.5$, $R \leq 5124$, $l > 850$, $\phi \leq 74.5$))
Then “Significant”

RULE 25: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l \leq 1600$, $s.\theta \leq 37.5$, $R \leq 5124$, $l > 850$, $\phi > 74.5$))
Then “Insignificant”

RULE 26: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l \leq 1600$, $s.\theta \leq 37.5$, $R > 5124$))
Then “Insignificant”

RULE 27: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l \leq 1600$, $s.\theta > 37.5$))
Then “Insignificant”

RULE 28: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l > 1600$, $s.\theta \leq 57.5$, $\phi \leq 75.5$))
Then “Significant”

RULE 29: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l > 1600$, $s.\theta \leq 57.5$, $\phi > 75.5$))
Then “Insignificant”

RULE 30: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi \leq 81.5$, $R > 2704$, $l > 1600$, $s.\theta > 57.5$))
Then “Insignificant”

RULE 31: IF (AND ($\phi > 43.5$, $R \leq 9763$, $\phi > 81.5$))
Then “Insignificant”

RULE 32: IF (AND ($\phi > 43.5$, $R > 9763$))
Then “Insignificant”

APPENDIX B

Programming Rules for PM

RULE 1: IF (AND (phi <= 43.5, R <= 4244, l <= 550, LE <= 87.5, s.theta <= 30, u <= 1.25, l <= 300))

Then “Insignificant”

RULE 2: IF (AND (phi <= 43.5, R <= 4244, l <= 550, LE <= 87.5, s.theta <= 30, u <= 1.25, l > 300))

Then “Significant”

RULE 3: IF (AND (phi <= 43.5, R <= 4244, l <= 550, LE <= 87.5, s.theta <= 30, u > 1.25))

Then “Significant”

RULE 4: IF (AND (phi <= 43.5, R <= 4244, l <= 550, LE <= 87.5, s.theta > 30))

Then “Insignificant”

RULE 5: IF (AND (phi <= 43.5, R <= 4244, l <= 550, LE >87.5, R <= 2609))

Then “Significant”

RULE6: IF (AND (phi <= 43.5, R <= 4244, l <= 550, LE >87.5, R > 2609))

Then “Insignificant”

RULE 7: IF (AND (phi <= 43.5, R <= 4244, l > 550

Then “Significant”

RULE 8: IF (AND (phi <= 43.5, R > 4244, s.theta <= 42.5, R <= 10120))

Then “Significant”

RULE 9: IF (AND (phi <= 43.5, R > 4244, s.theta <= 42.5, R > 10120, l <= 2100))

Then “Insignificant”

RULE 10: IF (AND (phi <= 43.5, R > 4244, s.theta <= 42.5, R > 10120, l > 2100, R <= 17140))

Then “Significant”

RULE 11: IF (AND (phi <= 43.5, R > 4244, s.theta <= 42.5, R > 10120, l > 2100, R > 17140))

Then “Insignificant”

RULE 12: IF (AND (phi <= 43.5, R > 4244, s.theta > 42.5, LE <= 275))

Then “Insignificant”

RULE 13: IF (AND (phi <= 43.5, R > 4244, s.theta > 42.5, LE > 275, R <= 22870, l <= 2100, R <= 10190))
Then “Significant”

RULE 14: IF (AND (phi <= 43.5, R > 4244, s.theta > 42.5, LE > 275, R <= 22870, l <= 2100, R > 10190))
Then “Insignificant”

RULE 15: IF (AND (phi <= 43.5, R > 4244, s.theta > 42.5, LE > 275, R <= 22870, l > 2100))
Then “Significant”

RULE 16: IF (AND (phi <= 43.5, R > 4244, s.theta > 42.5, LE > 275, R > 22870))
Then “Insignificant”

RULE 17: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE <= 87.5, R <= 1666, u <= 1.25, l <= 300))
Then “Insignificant”

RULE 18: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE <= 87.5, R <= 1666, u <= 1.25, l > 300))
Then “Significant”

RULE 19: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE <= 87.5, R <= 1666, u > 1.25))
Then “Significant”

RULE 20: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE <= 87.5, R > 1666))
Then “Insignificant”

RULE 21: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE > 87.5, R <= 452.5))
Then “Significant”

RULE 22: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE > 87.5, R > 452.5, l <= 150))
Then “Insignificant”

RULE 23: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE > 87.5, R > 452.5, l > 150, phi <= 63.5, R <= 945.5))
Then “Significant”

RULE 24: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE > 87.5, R > 452.5, l > 150, phi <= 63.5, R > 945.5, phi <= 47))
Then “Insignificant”

RULE 25: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE > 87.5, R > 452.5, l > 150, phi <= 63.5, R > 945.5, phi > 47))

Then “Significant”

RULE 26: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l <= 550, LE > 87.5, R > 452.5, l > 150, phi > 63.5))

Then “Insignificant”

RULE 27: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R <= 2744, l > 550))

Then “Significant”

RULE 28: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R <= 5798, l <= 850))

Then “Insignificant”

RULE 29: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R <= 5798, l > 850, s.theta <= 37.5, phi <= 74.5))

Then “Significant”

RULE 30: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R <= 5798, l > 850, s.theta <= 37.5, phi > 74.5))

Then “Insignificant”

RULE 31: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R <= 5798, l > 850, s.theta > 37.5, l <= 1200, R <= 3894))

Then “Significant”

RULE 32: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R <= 5798, l > 850, s.theta > 37.5, l <= 1200, R > 3894))

Then “Insignificant”

RULE 33: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R <= 5798, l > 850, s.theta > 37.5, l > 1200))

Then “Insignificant”

RULE 34: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l <= 1600, R > 5798))

Then “Insignificant”

RULE 35: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l > 1600, s.theta <= 57.5, phi <= 73.5))

Then “Significant”

RULE 36: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l > 1600, s.theta <= 57.5, phi > 73.5))

Then “Insignificant”

RULE 37: IF (AND (phi > 43.5, R <= 9763, phi <= 81.5, R > 2744, l > 1600, s.theta > 57.5))
Then "Insignificant"

RULE 38: IF (AND (phi > 43.5, R <= 9763, phi > 81.5))
Then "Insignificant"

RULE 39: IF (AND (phi > 43.5, R > 9763))
Then "Insignificant"

REFERENCES

Anderson, D. A., and K. P. Burnham. (1998). “*Model selection and inference; a practical information-theoretic approach*”. Springer-verlag, New York, NY.

Benson P. E. (1984) “CALINE4 – A dispersion model for predicting air pollutant concentrations near roadways, FHWA-CA-TL-84-15”. Sensitivity Analysis. Chapter 7. (June 1989 update). California department of transportation, Sacramento, CA.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. stone. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, CA.

CFR Title 40, Protection of the Environment, Part 93: Determining Conformity of Federal Actions to State or Federal Implementation Plans. US Environmental Protection Agency; available at <http://www.epa.gov/lawsregs/search/40cfr.html>. Accessed September 2007

Chambers, J. M., and T. J. Hastie, eds. (1993). “Tree-based models”. Chapter 9, *Statistical models in S*. Chapman & Hall, New York, NY.

EPA-454/R-92-006: User’s Guide to CAL3QHC: A Modeling Methodology for Predicting Pollutant Concentrations Near Roadway Intersections, US Environmental Protection Agency; available at http://www.epa.gov/scram001/dispersion_prefrec.htm. Accessed September 2007.

EPA: Support Center for Regulatory Atmospheric Modeling, US Environmental Protection Agency; available at <http://www.epa.gov/scram001/dispersionindex.htm>. Accessed September 2007.

Guensler, R., M.O. Rodgers, W. Bachman, and J.D. Leonard II. (1999). “Microscale Carbon Monoxide Impact Assessment for the Atlantic Steel Development Project”; March. <http://transaq.ce.gatech.edu/guensler/publications/consulting/atlantic%20steel/as%20final%20report.pdf>

Loh, W.-Y., and Y.-S. Shih, (1997). “Split selection method for classification trees”. *Statistica Sinica*; 7, 815-840.

Quinlan, J. R., and R. M. Cameron-Jones, (1995). “Oversearching and layered search in empirical learning”; Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Vol. 2, Morgan Kaufman, pp. 1019-1024.

Ramaswami, A., J. B. Milford, M. J. Small (2005). *Integrated environmental modeling – Pollutant transport, fate, and risk in the environment*. John Wiley & Sons, Hoboken, NJ.

Washington, S. (2000). “Iteratively specified tree based regression – Theory and trip generation example”. Journal of transportation engineers; Nov/Dec 2000.

Zoumakis, N. M., A. G. Kelessis, M. Petrakakis, K. Nikolaou, and T. I. Kozyraki. (1994). “A microscale dispersion model for Carbon Monoxide emitted from motor vehicles”. The Science of the Total Environment; 146/147 (1994) 401-407.