Binary Recursive Partitioning Method for Modeling Hot-Stabilized Emissions From Motor Vehicles

SIMON WASHINGTON, JEAN WOLF, AND RANDALL GUENSLER

An alternative statistical modeling approach, hierarchical tree-based regression (HTBR), is presented for developing modal correction factors for hydrocarbon (HC) emissions from motor vehicles. The term *modal* refers to operating modes of vehicle activity such as cruise, idle, deceleration, and acceleration. Explanation of the statistical theory is provided, followed by a presentation of specific modeling results for HCs. The modeling results are based on 4,800 vehicle emissions tests representing 29 laboratory testing cycles. HTBR methods are indicated to overcome statistical difficulties that are problematic for classical ordinary least-squares (OLS) regression, a commonly applied statistical technique for analyzing emissions data. HTBR methods are more adept at treating interactions and monotonic transformations on independent variables, better at handling categorical independent variables with more than two levels, not adversely affected by multicollinearity, and good at capturing nonadditive behavior across the range of independent variables. Unfortunately, HTBR theory is less well developed than OLS regression theory, and statistical parameter properties, such as efficiency, unbiasedness, and consistency, need further development. The HTBR modeling results for HCs are insightful. Hydrocarbon emissions from normal-emitting motor vehicles are most sensitive to changes in power (instantaneous speed^2 · acceleration) requirements of a given driving sequence, while high-emitting vehicles are sensitive to both the amount of idle activity and positive kinetic energy (instantaneous speed · acceleration) in a given driving sequence. Vehicle model year, engine size (cubic centimeters of displacement), curbside weight, and fuel delivery type (fuel injected, throttle body injected, carbureted), also were indicated to influence emission rates. Finally, high- and normal-emitting vehicles are sensitive to different operational and vehicle specific factors.

Through efforts to better understand emissions generated by motor vehicles and to develop mathematical models based on modes of vehicle operation rather than average operating speeds, researchers have identified shortcomings of existing mathematical models for accurately predicting emissions from motor vehicles (*1–8*). The focus of this paper is the way in which vehicle modes of operation are related to emissions and how these relationships are mathematically modeled. Specifically, this paper focuses on the problem of forecasting hydrocarbon (HC) exhaust emissions from a fleet of automobiles (excluding grade-induced enrichment events) operating under any set of modal operations. Introduced are HC modal correction factors that adjust baseline or fundamental HC emission rates by specific factors, depending on vehicle-specific attributes and modes of operation encountered during a typical driving sequence. The result is a forecasted hot-stabilized HC emission rate in grams per second based on user-provided fleet characteristics and known modes of operation.

School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Ga. 30332-0355.

The modeling method presented here provides a foundation for predicting emissions rates for nongrade-induced emissions, while a companion effort ongoing at Georgia Tech explicitly considers the impact of grades (and other loads) on emission rates of HC, oxides of nitrogen (NO_X), and carbon monoxide (CO). These two methods are being integrated to enable to forecasting of all hot-stabilized emissions (HC, NO_X, and CO) from a fleet of motor vehicles. Future research will focus on forecasting deterioration rates of motor vehicles, start-related emissions increments, and modes of vehicle operation necessary as input to modal-based models.

STATISTICAL MODELING APPROACH AND JUSTIFICATION

For any given mode of operation, motor vehicles emissions (in grams per second) may differ among vehicles by one or more orders of magnitude. A relatively small fraction of the vehicle fleet is responsible for a disproportionately large share of total emissions due to engine or vehicle defects that produce high emissions (*9*). In addition, a small fraction of a vehicle's observed activity can exhibit disproportionately high HC emissions when the engine computer calls for decreased air-to-fuel ratios to provide power during hard accelerations at low speeds, moderate accelerations at high speeds, or accelerations against gravity (i.e., grade effects) (*10*). From a modeling viewpoint, it is vital to accurately predict the number of high emitters in the fleet (older vehicle technologies and poorly maintained or tampered-with vehicles), as well as the fraction of activities that yield high emissions for otherwise normal-emitting vehicles, since emission rates (in grams per second) from high emitters and high-emissions activities can be orders of magnitude higher than ordinary. From an emissions inventory perspective, emissions differences among cleanly running, well-maintained vehicles are relatively unimportant compared with elevated emissions from either high emitters or high-emitting activities. This modal modeling effort focuses on developing improved modeling capabilities in these two areas. Statistical analyses are focused on estimating high-emitting vehicle emission rates, emission rates for high-emitting activities, and differences between important vehicle technology classes. As older vehicles in the current fleet are replaced by modern-technology vehicles and as certification standards continue to be tightened, the emissions differences among "broken" vehicles (high emitters) and activity extremes are likely to remain a top concern. A likely future challenge will be to accurately model vehicle control system failures or malfunctions and to accurately model vehicle replacement and deterioration rates.

Description of Data Set

This section describes the creation of the data set and variables used in the statistical model estimations presented later in this paper. Similar data are being used for estimation of modal models for CO and NO_X in simultaneous research efforts.

Hot-Stabilized Driving Cycles

Emissions collected in a single bag over a driving cycle are commonly referred to as *bag emissions*. Bag emissions data from various test cycles were obtained by the authors from the Environmental Protection Agency (EPA) and the California Air Resources Board (CARB). A composite data set was created and includes only those bag test results obtained from vehicles operating in hot-stabilized conditions (e.g., data including engine operations affected by cold or hot starts were excluded from the analysis). Each record in the data set contains vehicle and test-cycle characteristics as well as bag-analyzed emission results from both the test cycle and the hot-stabilized portion of the federal test procedure (FTP) Bag 2 test cycle (*11*), which is used as the base test cycle to compute emissions factors [similar to the way in which speed correction factor algorithms currently function within MOBILE (*12*)]. Test results from vehicles that did not include a corresponding FTP Bag 2 emission rate also were excluded from the data set.

Table 1 summarizes the test cycles contained in the data set. It indicates the testing program or data source, the cycle name, and the number of tests conducted on a particular cycle. Again, it is important to note that each test record also includes the FTP Bag 2 test results. There are no records that contain only FTP Bag 2 results (these records would have an emissions factor of one). The total number of vehicle test results contained in the data set is 4,800, which represents 798 vehicles tested on various subsets of the 29 hot-stabilized test cycles. Thus, each vehicle represented in the data set was tested on six different test cycles on average. Each record in the data set contains vehicle and test cycle characteristics as well as emissions test results under the noted cycle and FTP Bag 2 cycle. Test cycle variables represent modal activity contained in the driving cycle. Because the goal of the statistical analyses is to quantify the influence of vehicle modes of operation on HC emission rates, a key element of this research is the creation of modal activity variables associated with the emissions testing cycles. Because vehicle speed/time trajectories of driving cycles are approximately known (drivers in the laboratory must follow a prescribed speed/time trace within a specified speed tolerance of about 4.8 kph), they can be disaggregated into constituent modal activity components: cruise, idle, acceleration, deceleration, and interactions of these variables (such as positive kinetic energy defined as instantaneous acceleration times speed). In total, 29 modal variables were created to describe the vehicle modes of operation contained in the various driving cycles. These variables are shown in Table 2.

Table 2 values are interpreted in the following fashion. Six modal variables were created to quantify acceleration activity in any given cycle. The first variable is defined as the percentage of cycle time spent with instantaneous (1 sec) acceleration rates greater than 9.7 kph/sec, the second variable as the percentage of cycle time spent with instantaneous acceleration rates greater than 8.0 kph/sec (which includes all seconds above 9.7 kph/sec by definition), and the final variable as the percentage of cycle time spent with instantaneous acceleration rates greater than 1.61 kph/sec. Thus, each acceleration variable defines a threshold of acceleration exceeded during a given driving cycle.

Similar interpretations should be made for other modal variables: deceleration, positive kinetic energy, power, and cruise. Table 2 also presents a sample of modal variable values for the 29 hot-stabilized driving cycles where cycle conditions exceed variable thresholds. For example, for the Unified Cycle—Bag 2, 4 percent of the cycle (expressed in time) consists of acceleration rates greater than 4.8 kph/sec; 9 percent of the cycle consists of deceleration rates greater than 4.8 kph/sec; 5 percent of the cycle consists of positive kinetic energy (PKE) values greater than 194 kph^2/sec; 16 percent of the cycle consists of power values greater than 5001 kph^3/sec; 6 percent of the cycle consists of constant cruise speeds greater than 96.6 kph (*cruise* is defined as constant speed from one second to the next); and 13 percent of the cycle consists of idling (speed = 0 kph). The table clearly indicates the variability in modes of operation across driving cycles contained in the data set.

A summary of all variables contained in each record of the data set is given in Table 3. The table presents the four general categories of variable: emission test details, emission test results, vehicle characteristics, and cycle characteristics. After creation of the

TABLE 1 Summary of Vehicle Tests, Test Programs, Test Cycles, and Number of Tests in Hot-Stabilized Data Set

CARB Freeway and Arterial Test Cycles		EPA and/or CARB Hot-Stabilized Test Cycles		EPA -- FTP Revision Test Cycles	
# Tests	Test Cycle Name	# Tests	Test Cycle Name	# Tests	Test Cycle Name
136	Arterial 1	25	High Speed 1	46	REP05
139	Arterial 2	25	High Speed 2	45	ARB02
139	Arterial 3	69	High Speed 3	41	HL07
136	Freeway 1	69	High Speed 4	2	HL04
138	Freeway 2	237	Low Speed 1		
137	Freeway 3	236	Low Speed 2		
138	Freeway 4	237	Low Speed 3		
136	Freeway 5	188	Unified Cycle - Bag 2		
138	Freeway 6	464	New York City Cycle		
137	Freeway 7	533	Highway Fuel Economy		
51	5 Freeway	517	Speed Correction Factor, 12		
50	65 Freeway	542	Speed Correction Factor, 36		
49	70 Freeway				

TABLE 2 Modal Activity Variable Definitions (as percentage of Cycle Time Spent in Specified Operating Condition) and Sample of Modal Activity Variable Values Calculated for Hot-Stabilized Cycles

Variable Name (abbreviation) and [Variable Threshold Values]

acceleration, kph/sec (ACC); [> 9.66, >8.04, >6.44, >4.83, >3.22, >1.61]
deceleration, kph/sec (DEC); [>4.83, >3.22, >1.61]
positive kinetic energy (acc • speed), kph^2/sec (PKE); [>311, >272, >233, >194, >155, >116, > 77.7]
power (acc • $speed^2$), kph^3/sec (POW); [>13340, >11670, >10000, >8335, >6668, >5001, > 3334]
cruise, kph (CRZ); [>113, >96.6, >80.5, >64.4, >48.3]
idle, percent (IDLE); N/A

Hot-stabilized Cycle Name	DEC > 4.83	PKE> 194	CRZ > 96.6	IDLE %
Arterial 1	7	1	0	0
Arterial 2	8	2	0	0
Arterial 3	5	3	0	0
Freeway 1	0	2	5	0
Freeway 2	0	3	1	0
Freeway 3	1	3	2	0
Freeway 4	2	1	0	0
Freeway 5	3	1	0	0
Freeway 6	2	0	0	0
Freeway 7	3	0	0	0
5 Freeway	1	0	0	0
65 Freeway	0	2	5	0
70 Freeway	0	1	8	0
High Speed 1	1	0	0	1
High Speed 2	1	0	0	1
High Speed 3	1	1	16	1
High Speed 4	1	3	18	1
Low Speed 1	0	0	0	34
Low Speed 2	0	0	0	36
Low Speed 3	0	0	0	45
Unified Cycle - Bag 2	9	5	6	13
New York City Cycle	4	1	0	32
Highway Fuel Economy	1	0	0	1
Speed Correction Factor, 12	6	0	0	26
Speed Correction Factor, 36	2	1	0	6
REP05	5	6	2	0
ARB02	8	9	4	0
HL07	4	11	48	0
HL04	4	10	44	0

data set, the contents were closely examined for completeness and accuracy. Because the bag data were collected at different time periods in different laboratories and sometimes for alternative purposes, some data were missing in the source data sets. For example, in the hot-stabilized data set of 4,800 records, 1,264 records were missing the odometer reading and 3,148 were missing the vehicle identification number (VIN). All missing values were identified and replaced with an asterisk. In addition, all emission rates of 0.00 grams per kilometer (g/km) were given default values of 0.0155 g/km for CO and 0.0016 g/km for HCs and NO_X due to measurement device tolerances (*5*).

High-Emitting Vehicle Identification

To separate normal- from high-emitting vehicles, the hot-stabilized data set was reduced to a vehicle data set in which there was only one record per vehicle tested (for a total of 798 vehicles). Next,

TABLE 3 Summary of Variables Contained in Data Set

Emission Test Details	Emission Test Results	Vehicle Characteristics	Cycle Characteristics
Test Cycle, Test Site, Test Date.	CO, HC, NO_x, CO_2 (in g/m) for both the test cycle and for the FTP Bag 2 test.	VIN, Engine Family, Model Year, Odometer Reading, Inertial weight, Dynamometer HP, Number of Cylinders, Cubic Inch Displacement, Type of Transmission, Fuel Injection, Catalytic Converter, and Supplemental Air Recirculation.	Percent of Cycle Exceeding Various Thresholds of Positive Kinetic Energy, Power, Acceleration, Deceleration, and Cruise Speeds. Also includes Cycle Duration, Distance, and Average Speed.

regression tree models using FTP Bag 2 as the dependent variable and vehicle characteristics as independent variables were developed to identify technology traits that significantly influenced emission rates (the regression tree modeling process used is described in detail in the next two sections). The analysis resulted in four mutually exclusive HC technology groups: Model Year < 1980.5; 1980.5 < Model Year < 1987.5 and Cubic Centimeter Displacement > 3097; 1980.5 < Model Year < 1987.5 and Cubic Centimeter Displacement < 3097; and Model Year > 1987.5.

After technology groups were identified, FTP Bag 2 emission rates were analyzed for the vehicles in each technology group. Those vehicles with emission rates greater than the group's average rate plus two standard deviations (which represents the upper 2.27 percent of a normal distribution) were labeled high emitters. The remaining vehicles were labeled normal emitters. These labels were then applied to the 4,800 records in the hot-stabilized data set, and the data set was separated into two distinct data sets for HC modeling—a normal-emitter data set (4,510 records) and a high-emitter data set (182 records). An additional 108 test records were excluded from further analysis due to lack of information necessary to determine emitter status.

Hierarchical Tree-Based Regression Modeling Methodology

Binary recursive partitioning, commonly known as hierarchical tree-based regression (HTBR), can be thought of as a forward stepwise variable selection method, akin to forward stepwise regression. The methods used to estimate regression trees have been around since the early 1960s and are sometimes referred to as classification and regression trees (*13*). The method proceeds by iteratively asking (and answering) the following two questions: Which variable of all of the variables offered in the model should be selected to produce the maximum reduction in variability of the response? and Which value of the selected variable (discrete or continuous) results in the maximum reduction in variability of the response? The method continually asks and answers these questions (through numerical search procedures) until a desirable end condition is met, at which time the tree model is estimated. Tree terminology is similar to that of a real tree; there are branches, branch splits or internal nodes, and leaves or terminal nodes.

In mathematical terms, the deviance D is defined as

$$D_a = \sum_{l=1 \text{ to } L} (y_{la} - \mu_a)^2 \quad (1)$$

where

D_a = total deviance at node a, or the usual sum of squared error at the node;
y_{la} = lth observation on dependent variable y in Node a; and
μ_a = mean of L observations in node a.

Next is found a split of the observations at node a on a value of an independent variable X_i that results in two branches and corresponding Nodes b and c, each containing M and N of the original L observations ($M + N = L$). The deviance reduction function evaluated over all possible Xs then can be defined:

$$\Delta_{\text{all } X_i} = D_a - D_b - D_c \quad (2)$$

where

$\Delta_{\text{all } X_i}$ = the total deviance reduction function evaluated over the domain of all Xs;
$D_b = \sum_{m=1 \text{ to } M} (y_{mb} - \mu_b)^2$;
$D_c = \sum_{n=1 \text{ to } N} (y_{nc} - \mu_c)^2$;
D_b = total deviance in node b;
D_c = total deviance in node c;
y_{mb} = mth observation on dependent variable y in node b;
y_{nc} = nth observation on dependent variable y in node c;
μ_b = mean of population (estimated using M observations) in node b; and
μ_c = mean of population (estimated using N observations) in node c.

The variable X_i and its optimum split i is sought so that the reduction in deviance is maximized, or more formally when

$$\Delta_{\text{all } X_i} = \sum_{l=1 \text{ to } L} (y_{la} - \mu_a)^2 - \sum_{m=1 \text{ to } M} (y_{mb} - \mu_b)^2 - \sum_{n=1 \text{ to } N} (y_{nc} - \mu_c)^2 = \text{maximum} \quad (3)$$

The maximum reduction occurs at some $X_{i(i)}$; the independent variable X_i at value $X_i = i$. When the data are split at this X, the remaining samples have a much smaller variance than the original data set. Thus, the reduction in node a deviance is greatest when the deviances at nodes b and c are smallest. Numerical search procedures are employed to maximize Equation 3.

This iterative partitioning process is continued at each node until one of the following conditions is met: (a) the node of a tree has met minimum population criteria (i.e., based on statistical sampling theory), or (b) minimum deviance criteria at a node have been met. Nodes that are split are internal nodes or branch splits, whereas nodes that are not split (because of criteria provided earlier) are terminal nodes or leaves. The S-Plus software program by StatSci allows the user to select either of the foregoing criteria to control "growth" of a tree (*14*). When growing is terminated, the mean of the remaining sample in all terminal nodes is provided.

To facilitate statistical inference, we generally make assumptions about the probability distribution of the response variable and, subsequently, the resulting error terms [as in normal-theory ordinary least-squares (OLS) regression]. In a previous OLS regression effort (*2*), the dependent variable HC in grams per kilometer was transformed to Log_{10}(HC), where HC is in grams per second, to obtain a normal distribution. In HTBR, however, a normally distributed response variable results in a chi-square distributed variable at terminal nodes. This can be observed from the following. Assume that $Y = \text{Log}_{10}(\text{HC})$ is approximately normally distributed, such that $Y \approx N(\mu, \sigma^2)$. It follows that a new variable $z_i = (y_i - \mu)/\sigma \approx N(0, 1)$, where the y_is are random variables drawn from Y. Then, since the deviance at any node j is given by $D_j = \sum_n (y_i - \mu_j)^2$, then $\sqrt{D}/\sigma \approx N(0, 1)$. By definition, a chi-square distribution results when we divide a sum of squares from a normal distribution by its variance, so we get $D/\sigma^2 \approx \chi^2_{(n-1)}$. Thus, the deviance divided by the sample variance is approximately chi-square distributed with n-1 degrees of freedom. As a result, a normally distributed response variable results in a chi-square distributed sample at the terminal nodes. Statistical inferences based on this model were reserved for later research but inferences are possible without having to use Monte Carlo techniques; however, transformations of the sample at terminal nodes may be necessary.

Model Selection Criteria

As in OLS regression, there are formal methods for selecting an appropriate HTBR model. The general procedure is to estimate a full

tree model that includes all of the variables of interest and then trim the tree with one of two methods. In reality, a fully specified tree model can predict every observation (each terminal node has only one observation); however, this is generally neither a good practical starting point nor desirable from a statistical standpoint. The tradeoffs involved in tree model selection are analogous to statistical models in general. The more complex the tree, the better the tree's ability to make forecasts and to describe the data on which it is estimated; whereas, an "overfitted" tree model results in overstated confidence in predictions and inclusion of insignificant variables into the model.

The two methods used to trim trees are called *pruning* and *shrinking*. For brevity, only the first of these methods is discussed here. Each method uses criteria about model complexity to trim the full tree model to a smaller and more manageable or practical tree size. In S-Plus, pruning successively snips off the least important splits using the following cost-complexity measure:

$$D_k(T') = D(T') + k \cdot size(T') \quad (4)$$

where

$D_k(T')$ = deviance of subtree T' with cost-complexity parameter k,
k = cost-complexity parameter,
$size(T')$ = the number of terminal nodes (leaves) of T'.

When pruning a tree, the general objective is to reduce the complexity of the tree by comparing the relative benefit (reduction in deviance) of overall tree splits. By applying Equation 4, then, cost-complexity pruning determines the subtree T' that minimizes the equation over all subtrees (*14*). When the cost-complexity parameter k is large, the number of terminal nodes in T' significantly raises the value of $D_k(T')$, so that a minimum to Equation 4 results in a subtree where its size is small. In contrast, when k is zero, there is no cost for having a subtree with many terminal nodes, and therefore a minimum to Equation 3 occurs with the original tree.

HTBR Results for HCs

Before conducting the modeling exercise, several working hypotheses were developed. These hypotheses served to guide the model development and implementation process. The working hypotheses for the analyses presented here are discussed in turn.

1. **A statistical approach is adequate for forecasting the influence of vehicle and activity factors on emission ratios.** The HTBR method will be used to model relative changes in emissions for certain vehicle classes and activities instead of absolute magnitudes of emissions. Hence, the approach taken in the analyses that follow is analogous to the approach incorporated in both the EMFAC and MOBILE models. The absolute values of emissions will be obtained from recent testing results, while model correction factors (based on the data discussed here) will be used to forecast how certain vehicles and operating modes affect emissions relative to other vehicles and modes of operation.

2. **The second bag of the federal test procedure (FTP Bag 2) is a useful and meaningful test cycle for use as a baseline emission rate in modal correction factors or ratios.** The FTP Bag 2 cycle will be incorporated into the development and implementation of modal correction factors. The emissions ratio (more commonly known as a correction factor) is the value of emissions (grams/sec) under a test condition relative to the emissions (grams/sec) under a control condition. In the MOBILE model, the FTP Bag 2 is the control condition. This is a convenient test to use, as most of the bag-testing programs have subjected vehicles to multiple test cycles and have almost always included the FTP Bag 2 tests in the testing sequence.

3. **Various vehicle technology classes will respond differently to modal activity.** Vehicle classes or technology groups, once identified, can be classified into groups of vehicles that behave relatively homogeneously with respect to HC emissions. Across these technology groups, however, emissions responses to modes of operation will be heterogeneous.

4. **A subfleet of vehicles exists in the overall vehicle fleet that exhibits fundamentally different emissions response behavior than other vehicles in their technology class.** These high-emitting vehicles will require their own emissions model, and the emissions from them need to be forecasted separately from normal-emitting vehicles because their emissions rates are so much greater than the emissions rates of normally functioning vehicles.

Normal-Emitter Model Selection and Results

Before estimation of the normal- and high-emitter regression trees, the dependent variable, hydrocarbon emissions in grams per kilometer, was transformed into grams per second units. The corresponding FTP Bag 2 HC emission rates similarly were transformed. Next, the test cycle's HC emission rate was divided by the FTP Bag 2 HC emission rate to establish an emission factor. Finally, the HC g/sec factor was transformed to be approximately normally distributed by taking the Log_{10}(factor). This transformation is not necessary to estimate a regression tree; however, it will be used later to enable statistical inference and will also enable comparison with previous models developed using the same response variable.

S-PLUS provides three methods for limiting the size of regression trees. Within the initial tree specification, the *minsize* command can be used to limit the node size at which the last split is performed. Then, pruning or shrinking can be applied to an existing tree to further reduce its size. When growing the initial tree, as defined in the S-PLUS statistical analysis manual (*14*), the binary partitioning algorithm recursively splits the data in each node until either the node is homogeneous or contains too few observations. Node homogeneity is defined by deviance—a deviance of zero would indicate a perfectly homogeneous node. For the purposes of this research effort, initial cutoffs for "too few observations" were set at 50 samples (as specified by the minsize command) for the normal-emitter data set, which has 4,510 test records, and at 30 samples for the high-emitter data set, which has only 182 test records.

Unlike classical OLS regression, variables offered in the tree model estimation process may not end up in the estimated model, and variables can be selected multiple times (i.e., as interactions with other variables). The 37 independent variables offered in the initial regression are shown in Table 4 (recall that there are multiple modal variables). The dependent variable is listed as *logHC-fact*. The normal emitter regression tree model was estimated and, of the 37 independent variables originally offered for model estimation, 21 variables were selected for inclusion in the model. The resulting regression tree contained 102 terminal nodes, the magnitude of which reflects the size of the data set. Both the number of

TABLE 4 Variables Offered in Tree-Based Model

Variable Name	Variable Description
logHCfact	The log 10 transformation of the HC emission factor (in grams / second) logHCfact = Log_{10}(HCgps / Bag2 HCgps)
MY	The last two digits of the model year of the vehicle tested
INERTIA	The curb weight of the vehicle loaded with fuel and oil (kilograms)
DYNOHP	The dynamometer-measured horse power (including drag, friction, and frontal area factors)
CID	The displacement of the tested vehicle's engine (cubic centimeters)
TRAN	The transmission type of the vehicle (1 = automatic, 3 = 3 speed manual, 4 = 4 speed manual, 5 = 5 speed manual)
FINJ	The fuel injection type (1 = port, 2 = carburetor, 3 = throttle body)
CATA	The catalytic converter type (1 = none, 2 = oxidation only, 3 = 3-way catalyst, 4 = oxidation and 3-way catalyst)
AVGSPD	The average speed of the test cycle (kph)
PKE.x	The percent of the cycle time with positive kinetic energy (accel*speed) greater than x (kph^2/second)
POW.x	The percent of the cycle time with power (acc*$speed^2$) greater than x (kph^3/second)
ACC.x	The percent of the cycle time with acceleration rates greater than x (kph/second)
DEC.x	The percent of the cycle time with deceleration rates greater than x (kph/second)
CRZ.x	The percent of the cycle time with cruising speeds greater than x (kph)
IDLE	The percent of the cycle time at idle

variables (21) and the number of terminal nodes (102) in the initial model are likely to overstate the confidence in modeled factors and are likely to be cumbersome to apply in practice. The prune function was subsequently used to reduce the size of the initial tree to a manageable and justifiable size and depth.

Examination of the pruned tree graph (Figure 1) indicates significant thresholds near k-values of 1.7 and 2.8. Given the immediate objective of limiting the terminal nodes to a manageable number, the pruned tree created with a k-value of 2.8 and with 14 terminal nodes was selected for presentation in this paper (there were 28 terminal nodes when the k-value was set at 1.7). Figure 1 presents a graph of the trade-off between the cost-complexity parameter k, the number of terminal nodes, and the mean residual deviance summed over all terminal nodes. Examination of the pruned tree graph indicates that, as the cost-complexity factor increases, the number of terminal nodes decreases and the mean residual deviance increases. The trade-offs depicted enable the analyst to make decisions about the necessary level of detail in the tree model. Abrupt vertical changes in the curve indicate that there are large changes in mean residual deviance given small changes in the number of terminal nodes.

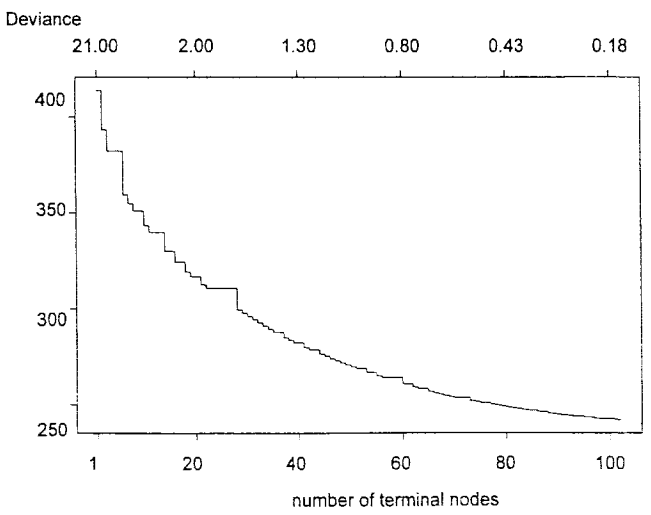

FIGURE 1 Pruned tree graph for model with *minsize* = 50 (cost-complexity K-values appear at top of graph).

Interpreting the Normal-Emitter Regression Tree

The pruned tree with 14 terminal nodes is presented in Figure 2. To interpret the graphical representation of the tree, one should start at the top node of the tree (the root) and move downward through the tree to a specific terminal node. Along that path, each binary split of the tree is labeled with a decision rule that determines the correct path to take. The terminal nodes of the tree are labeled with values that represent the expected value of the dependent variable [in this case, Log_{10}(HC/Bag 2 HC)]. For example, the decision rules and resulting path to reach the node designated with the arrow (shown in the figure with value 1.310) can be defined. Begin at the root and move downward through the tree until a terminal node is reached. The decision rules to reach the terminal node indicated are given by the following path (starting at the root): If MY > 84.5, go right; otherwise go left. If POW.20 > 5.975, go right; otherwise go left. If MY > 89.5, go right; otherwise go left. If CID > 186, go right; otherwise go left. If INERTIA > 3562.5, go right; otherwise go left. The 1.310 value at the terminal node is the average Log_{10}(HC/Bag 2 HC) for normal-emitting post-1989 model year vehicles with engines smaller than 3048 cm^3, curbside weight greater than 1616 kgs, and operating conditions in which the percentage of driving time with $speed^2$ · acceleration greater than 8335 kph^3/sec is less than 5.975 percent.

Rules encountered during a specific path traversal that uses the same variable can be combined into a single rule. For example, MY > 84.5 and MY > 89.5 can be reduced to a single rule of MY > 89.5. The rules applicable for the specific tree traversal (as listed earlier) can be combined and potentially coded as such—if MY > 89.5 and if POW.20 > 5.975 and if CID < 186 and if INERTIA > 3562.5, then Log_{10}(HC/Bag 2 HC) = 1.310.

Each terminal node value, which is calculated as Log_{10}(HC/Bag 2 HC) for all observations falling into each node, is transformed into an emissions rate by taking the Log_{10} inverse (e.g., $10^{1.310}$) to get the factor HC/Bag2HC. The Bag2HC rate is then calculated as the

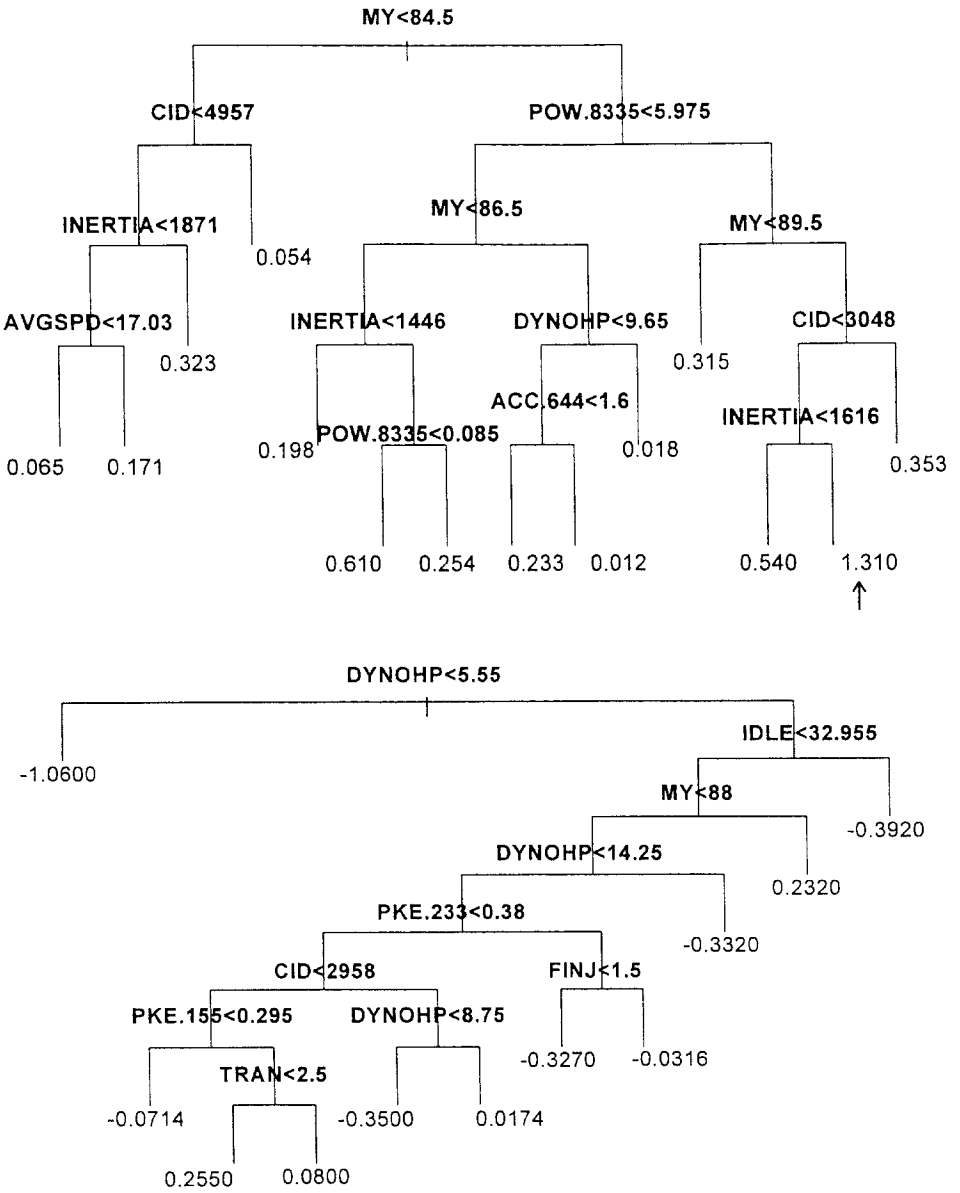

FIGURE 2 Plot of normal-emitter HC tree with *minsize* set to 50 and cost-complexity factor *K* set to 2.8 (top) and plot of high-emitter HC tree with *minsize* set to 30 (bottom).

average FTP Bag 2 HC rate for all observations falling into that node. This average FTP Bag 2 HC rate is multiplied by the emissions factor HC/Bag2HC to get the predicted HC emissions rate (in grams per second) for each node for all vehicles with the specified vehicle, engine, and modal activity characteristics.

It should be noted that the model output should be applied on a trip-based forecast, since the emissions data were derived from trip-based cycles. In other words, it may be inappropriate to forecast emissions from daily or weekly driving using the regression tree. It also is inappropriate to forecast instantaneous emissions with the regression tree.

High-Emitter Model Results

The high-emitter regression tree model is also presented in Figure 2. Recall that the high-emitter model is based on fewer observations than is the normal-emitter model—182 to be exact. The premise of the high-emitter model is that these vehicles behave differently from normal emitters. A quick comparison of the normal- and high-emitter models can illustrate this effect. Consider the same vehicle as described when illustrating the normal emitter. Recall that the vehicle's path rules were determined by: if MY > 89.5 and if POW.20 > 5.975 and if CID < 186 and if INERTIA > 3562.5, then Log_{10}(HC/Bag 2 HC) = 1.310. To trace the path for a similar, but high-emitting vehicle, we would need to first know the vehicle's dynamometer horsepower drag rating (including factors such as drag, coast-down friction, and frontal area) and the amount of idle activity before an emissions forecast could be made. Note that *POW.20*, *CID*, or *INERTIA* are not needed for these vehicles. From these results it can be concluded that the responses of these two classes of vehicle are indeed different, and that perhaps these high-emitting vehicle emissions are a function of fundamentally different mechanisms.

Interpretation of Regression Trees

The HTBR modeling results for HC are insightful. With respect to vehicle modes of operation, HC emissions from normal-emitting motor vehicles are most sensitive to changes in power (instantaneous speed^2 · acceleration) requirements of a given driving sequence, whereas high-emitting vehicles are sensitive to both the amount of idle activity and positive kinetic energy (instantaneous speed · acceleration) in a given driving sequence. It is not surprising that operational loads have great influence on HC emissions. This finding is consistent with ongoing research focused on the development of theoretical models of emissions from automobiles.

Model year, engine size (cubic centimeters of displacement), curbside weight, and fuel delivery type (fuel-injected, throttle body injected, carbureted) are factors that have significant influence on emission rates. These results also are not surprising. Model year is a surrogate for penetration of specific vehicle technologies, which tend to have specific model year groupings that are relatively homogeneous with respect to HC emissions. Curbside weight is a load-related variable, and thus is likely accounting for the relationship between vehicle weight, or mass, and emissions.

The results also indicated that high- and normal-emitting vehicles are sensitive to different operational and vehicle-specific factors. Again, recent studies have demonstrated that emissions from "dirty" vehicles can have instantaneous emission rates that are several orders of magnitude higher than a similar cleanly running vehicle. Thus, the factors that influence the HC emissions characteristics of these vehicles would be expected to be fundamentally different from the factors that influence HC emissions from clean vehicles.

Application of Regression Tree Model Results

To apply the results of the foregoing models in practice, one must accomplish several tasks. It is assumed that the intention is to model emissions from a fleet of motor vehicles from a large metropolitan region. A brief discussion of the necessary steps follows.

1. The proportion of high emitters by model year and vehicle technology could be estimated by surveying FTP Bag 2 test results from the state and comparing the threshold criteria with those used in developing the high emitters as part of this study (thresholds not provided here). From this survey, estimates of the proportion of vehicles by vehicle technology and model year that are high emitters could be determined. An alternate method could be devised using remote sensing devices in conjunction with FTP Bag 2 results to correlate HC concentrations with gram-per-second emissions rates and "critical" thresholds.
2. Normal emitters in the vehicle fleet would then be surveyed to determine the number and proportion of vehicles by model year and technology. Again, a sample of FTP Bag 2 test results from the region or state would aid in characterizing the normal-emitting fleet. For example, for accurate forecasting of HC emissions, we need the vehicle fleet distribution by model year, curbside weight, dynamometer horsepower rating, and average FTP Bag 2 emission rates. These data could be coupled with regional vehicle registration data to arrive at an estimate of the regional vehicle fleet distribution.
3. With these data in conjunction with forecasted modes of vehicle operation (research currently under way at Georgia Tech and the University of California, Davis), the regression tree can be used to forecast emissions for any given network link and vehicle distribution. Average gram-per-second emission rates would need to be weighted by vehicles in a specific class and summed over all terminal nodes of the trees (normal and high emitters) to obtain an estimate of total HC emissions.

All three steps currently are being studied as part of the ongoing modeling effort under way at Georgia Tech. Of particular difficulty is the ability to forecast modes of vehicle operation, but much progress is being made on this effort. In addition, algorithms are being developed to forecast an emissions increment induced by grade and accessory loads.

DISCUSSION OF RESULTS

The model presented here demonstrates the power of HTBR methods. In general, HTBR is a competing method to OLS regression, which was used in a previous modeling effort by the authors and by others. Readers interested in comprehensive treatments of OLS and tree-based regression theory, respectively, should consult Neter et al. (*15*) and Breiman et al. (*13*). Examples of the method applied in practice can also be found (*16–18*).

Recall that an OLS regression model takes the form, $Y = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon$, where Y is the response variable, alpha is the intercept term, the betas are estimated regression coefficients, and the Xs are independent variables. One of the drawbacks of OLS regression models, in contrast to HTBR models, is that all regression coefficients affect each fitted value—a constraint brought about by the additive form of the OLS regression model. HTBR is tolerant of nonadditive behavior between the response and predictor variables, which allows an appropriate subset of all variables to influence specific fitted values. For example, the model in Figure 2 indicates that the emission response for some pre-1985 model year vehicles is significantly influenced by average travel speed, whereas 1985 and later model year vehicles are not influenced by this variable.

Discrete variables with many more than two responses are often cumbersome to include in an OLS regression model. HTBR is adept at modeling multivalued discrete variables. Again referring to Figure 2, model year is treated as a discrete variable with various overlapping and nonadditive ranges of model years: pre-1985, 1985 and 1986, 1987 to current, 1985 to 1989, and 1990 to current. As expected, unit changes in MY should not result in constant and incremental changes to predicted emissions across the range of model years—as would be predicted by an OLS regression model. In other words, MY is a discrete variable that has heterogeneous effects on emissions across model years. The HTBR method is adept at modeling heterogeneity associated with multivalued discrete variables.

Missing data often pose a problem for both model estimation and prediction. Although missing data cause similar problems in estimation of both OLS regression and HTBR models, prediction with missing data is handled much better using HTBR models. To illustrate, consider the following scenario. Recall that at each node (both internal and terminal) of the HTBR, there is a chi-square probability distribution with n-1 degrees of freedom. In the case of missing data, one cannot continue all the way down the branches of a particular tree, but a prediction can be made at an internal node. Assume a 1977 normal-emitting vehicle with unknown engine size. The node in the regression tree marked "CID < 4957" denotes the farthest node that can be traveled given the existing information on the vehicle (Figure 2). Fortunately, the internal node carries with it

a specific mean μ that can be used to make a prediction for the emission factor for this vehicle with unknown engine size. Although the outputs indicated here do not illustrate these values, S-PLUS can easily be commanded to provide internal node means and deviances.

Outlier identification and analysis are an important aspect of any statistical model-building process. In OLS regression, outliers can pose significant problems, as partial slope coefficients can be unduly influenced by an outlying observation—outlying with respect to x, outlying with respect to y, or outlying with respect to both—thus causing the fit of the line to be nonrepresentative of the population. There are various techniques available to remedy problems in OLS regression when there are outliers, including absolute deviation regression, m-estimates regression, and generalized least-squares estimation techniques. In HTBR, outliers are fairly robust with respect to measurement variables but less so with respect to the response variable (*13*). HTBR methods tend to help identify the outlying Y-observations, and also help to lessen their impact on the resulting model. Heteroscedastic errors (constant errors across all values of the Xs), however, result in biased low estimates of standard errors of the terminal nodes samples (*13*). It turns out that heteroscedasticity cannot only affect the standard error of the node samples, but also y_{ave} may be a poor estimate of μ at the terminal nodes. As a result, it often is advisable to select tree depths such that the model favors larger samples at the terminal nodes to minimize the effect of biased estimates of μ and σ (*13*).

It is difficult to arrive at objective criteria for comparing the models in a quantitative fashion. The choice of model estimation method should rely on many factors, including their respective performance on a validation data set, their agreement with rational expectation, the intended end use of the model, and quantitative measures, such as mean residual deviance. Further development needs to be done, however, to formulate relevant statistical tests for differences in HTBR model formulations.

The preliminary HTBR model presented here demonstrates the advantages and disadvantages compared with classical OLS regression. Many considerations suggest that the method is more appropriate than OLS regression for the HC emissions data, since there appears to be nonadditive behavior among variables and multivalued discrete variables.

The results provided here suggest that modeling of HCs can be performed for an entire fleet of motor vehicles given auxiliary information. For example, it requires link- or trip-based measures of modal activity consistent with the definitions used in the models. It also requires acquisition of detailed vehicle fleet information. Future work is planned to demonstrate the benefits and power of this approach through validation exercises, further model refinement, and inclusion of new model algorithms to forecast grade-induced emissions increments.

FUTURE RESEARCH

Much needs to be done to continue and improve the current modeling effort. The following related research needs currently are being addressed by researchers at Georgia Tech and by others.

1. Emission models that are sensitive to modal activity will require input about on-road modes of vehicle operation. Research currently under way is investigating the causal factors of various modes of vehicle activity, including underlying roadway and traffic conditions, vehicle fleet characteristics, and socioeconomic variables. Models are being developed to forecast the modes of operation under known traffic conditions, both for simulation and for travel demand modeling applications.

2. The HTBR method needs to be explored to better understand its limitations and strengths. Toward this goal, researchers are investigating the possibility of mixed models containing both classical OLS and HTBR components to capitalize on the strengths of both methods. Also being investigated are the properties of the split variables and split values to determine whether they are stable and robust with respect to random samples from the parent population.

3. The HTBR models need to be cross-validated with a data set not used for model estimation. For these purposes, we plan to employ the results of second-by-second emission test results and obtain independent estimates of μ and σ at terminal nodes. In this fashion, we can employ various goodness-of-fit measures to assess the accuracy of the models, such as mean squared prediction error, correlation coefficient, and Theil's U-statistic.

4. Once robust statistical model formulations are developed, similar models will be developed for forecasting emissions of CO and NO_X. As the data sets for these pollutants become ready for statistical analyses, these models will be refined and cross-validated in a similar fashion. The data necessary to fill in missing values in the data set will be collected so that all and any variables can be employed in the analyses.

5. Finally, the models must be refined to include the ability to forecast emissions induced by grade and accessory loads, such as air conditioning. In a parallel research effort, a load-based model is being integrated with the HTBR models to enable forecasting of the emission increment induced by grade and air-conditioning loads.

ACKNOWLEDGMENTS

The authors are grateful for the research sponsorship of the EPA's Office of Research and Development and the Federal Highway Administration. Particularly, the authors thank Ted Ripberger, Lois Platte, Chuck Mann, and Sue Kimbrough for their thoughtful reviews of this manuscript.

REFERENCES

1. *NCHRP Report, Project 25-7: Improving Transportation Data for Mobile Source Emissions Estimates*. TRB, National Research Council, Washington, D.C., July 1996.
2. Washington, S. *Estimation of a Motor Vehicle Emissions Model and Assessment of an Intelligent Transportation Technology*. Ph.D. dissertation. Institute of Transportation Studies, University of California at Davis, 1994.
3. Washington, S. *A Cursory Analysis of EMFAC7G: Reconciling Observed and Predicted Emissions*. Research Report prepared for the Union of Concerned Scientists. Institute of Transportation Studies, University of California at Davis, May 1994.
4. Croes, B. E., and E. M. Fujita. *California's Motor Vehicle Emission Inventory: Top Down and Bottom Up Assessments of Accuracy*. California Air Resources Board, Sacramento, 1993.
5. Guensler, R. *Vehicle Emission Rates and Average Vehicle Operating Speeds*. Ph.D. dissertation. Institute of Transportation Studies, University of California at Davis, 1993.
6. Guensler, R. Data Needs for Evolving Motor Vehicle Emission Modeling Approaches. In *Transportation Planning and Air Quality II* (Paul Benson, ed.), American Society of Civil Engineers, New York, 1993.
7. Purvis, C. Sensitivity of Transportation Model Results to Uncertainty in Input Data: Transportation Modeling Tips and Trip-Ups. *Proc., Air and*

Waste Management Association Specialty Conference. Pittsburgh, Pa., March 1992.
8. *Transportation Research Circular 389: Environmental Research Needs in Transportation*. TRB, National Research Council, Washington, D.C., 1992.
9. Pollack, A., J. Heiken, and R. A. Gorse. Comparison of Remote Sensing Data and Vehicle Emissions Models: The High Proportion of Emissions from High Emitting Vehicles. *Proc., 85th Annual Meeting of the Air and Waste Management Association*, Pittsburgh, Pa., June 1992.
10. Cicero-Fernandez, P., and J. R. Long. Modal Acceleration Testing on Current Technology Vehicles. *Proc., Emission Inventory: Perception and Reality*, Pasadena, Calif., Oct. 1993.
11. *Code of Federal Regulations*, Title 40, Part 86, Appendix I, July 1, 1990.
12. *User Guide to MOBILE5*. EPA-AA-AQAB-94-91. Air Quality Analysis Branch, EPA, Ann Arbor, Mich., 1994.
13. Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, Calif., 1984.
14. *S-Plus Guide to Statistical and Mathematical Analysis: Version 3.3 for Windows*. StatSci Division, MathSoft, Inc., Seattle, Wash., 1995.
15. Neter, J., M. Kutner, C. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, 4th ed. Irwin, Chicago, Ill., 1996.
16. Prager, R. CART/CMAC Hybrid: Regression Trees with Interpolation. *Proc., International Conference on Pattern Recognition 2*, Piscataway, N.J., IEEE, 1994.
17. Waterhouse, S. R., and A. Robinson. Pruning and Growing Hierarchical Mixtures of Experts. *Proc., Artificial Neural Networks*, Cambridge, U.K., IEE, 1995.
18. Grajski, K., L. Breiman, G. Di Prisco, and W. Freeman. Classification of EEG Spatial Patterns With a Tree-Structured Methodology: CART. *IEEE Transactions on Biomedical Engineering*, Vol. BME-33, No. 12, Dec. 1986.

Publication of this paper sponsored by Committee on Transportation and Air Quality.