High-Emitting Vehicle Characterization Using Regression Tree Analysis

JEAN WOLF, RANDALL GUENSLER, SIMON WASHINGTON, AND WILLIAM BACHMAN

A small fraction of motor vehicles on the roadway emit a disproportionate fraction of pollutant emissions, especially for carbon monoxide and hydrocarbons. Generally, these "high emitters" or "super emitters" exhibit higher emissions rates under all operating conditions than do "normal emitters." Since the instantaneous emissions response between normal- and high-emitting vehicles can differ by one or more orders of magnitude, so do their average emissions over a "typical" trip. Identifying the proportion of normal- and high-emitting vehicles in an urban area and quantifying their emissions is vital for accurate emission inventory accounting. A methodology by which high and normal emitters can be classified is presented. Unlike previous emitter classification approaches, the approach is data driven and relies entirely on hot-stabilized emissions results. A statistical classification scheme, better known as hierarchical tree based regression, is used to separate vehicles into homogenous emitter categories. The approach is shown to have a number of advantages. First, it is flexible with respect to both the number of classes and types of variables used to identify classes. Second, it considers the influence of a large number of vehicle and technology attributes on emitter status. Third, it ensures that the highest emitters can be isolated from the normal emitters, so that separate emission rate models can be developed for these vehicles. Finally, the approach does not combine the effects of starts and hot-stabilized operations within the definition of high emitter, leading to a classification scheme whereby vehicles with poor start emissions characteristics will not be incorrectly classified as vehicles with poor hot-stabilized emission characteristics.

A small fraction of motor vehicles on the roadway are responsible for a large fraction of pollutant emissions. These "high emitters" or "super emitters" are typically vehicles that exhibit high emissions rates under all operating conditions. The research literature provides wide ranges of estimated contributions from these high emitters. Study claims range from 20 percent of the vehicles being responsible for 50 percent of the emissions, to 10 percent of the vehicles being responsible for 60 percent of the emissions, to even 5 percent of the vehicles being responsible for 80 percent of the emissions. The differences in these estimates stem from differences in the definitions of "high emitters" and the differences in test methods used to estimate the activity and emissions rates for these vehicle groups. Nevertheless, it is clear that a small fraction of the fleet is responsible for a significant fraction of fleet emissions. Thus, tracking the activity of high-emitting vehicles and establishing accurate emission rates for the vehicles is a first-order modeling problem. Identifying these vehicles in an urban area (both spatially and temporally) and quantifying their emissions will lead to significantly improved emissions prediction and will provide a basis for improved emission reduction policies that target subsets of the vehicle fleet.

J. Wolf, R. Guensler, and S. Washington, School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332. W. Bachman, GIS Center, Georgia Institute of Technology, Atlanta, GA 30332.

Evidence in laboratory testing data indicates that high-emitting vehicles respond differently to changes in operating conditions than do normal-emitting vehicles. Many vehicles in the fleet usually undergo "enrichment" under certain operating conditions, creating higher emissions rates under heavy engine loads (such as hard accelerations or high operating speeds). However, high emitters do not respond similarly to changes in engine load conditions. Furthermore, the relative change in emissions rates under enrichment conditions for normal emitters is significant, whereas the relative change for high emitters is much less so. Hence, both the average emissions rate and emissions response to changes in operating mode appear to be significantly different for normal- and high-emitting vehicles. This means that activity tracking and the development of hot-stabilized running exhaust emissions algorithms need to proceed separately for these two emitter groups.

High emitters (i.e., vehicles that are malfunctioning or have been tampered with) are usually defined relative to the emissions of other vehicles in their technology class (representing model year groups with emissions control technology combinations that behave similarly with respect to emissions production). Thus, when a new vehicle and an old vehicle both exhibit a large emissions rate in grams per mile, the new vehicle might be considered a high emitter whereas the older vehicle might be considered a normal emitter. After undertaking a literature review on high emitter definitions, the research team concluded that previous studies did not determine whether the emitter groups behaved similarly in terms of emissions production. That is, engineering judgment and assumptions were used in all studies to split vehicles into technology classes and to define the high-emitter cutpoints (typically separated into groups either by multiples of certification standards or by position in the cumulative emissions rate distribution).

The contribution of this research is the development of an alternative to existing methods for high-emitter identification; this research presents a different methodology that separates vehicles into classes that behave similarly, exhibit similar technology characteristics, and exhibit similar mean emissions rates under standardized test conditions. This classification can be performed through regression tree analysis. Then, by selecting the highest-emitting vehicles within each class on the basis of targeted percentiles of a normal distribution, high emitters can logically be identified within each technology classification.

This paper describes the technique used by the Georgia Tech research partnership to develop emitter groups for incorporation into a geographic information system–based (GIS-based) modal emissions model. The laboratory testing database assembled from government and industry sources is described, including the variables available for emitter classification purposes. The analytical methods are outlined, and detailed references describing the modeling tech-

nique are cited. The model results are presented, and the method for separating test results into high-emitter and normal-emitter groups for hot-stabilized emissions rate algorithm development are provided. Finally, the programming methods used to incorporate the model algorithms into the GIS-based model are described.

EMISSIONS DATABASE DEVELOPMENT

To develop accurate mobile-source emissions models, a comprehensive database of laboratory testing had to be assembled. In January 1997, the entire emissions database collected and maintained by the Environmental Protection Agency's (EPA's) Air Quality Office of Mobile Sources (OMS) was obtained. The database contains emissions test information (e.g., vehicle, engine, and test characteristics) and results from 48 separate testing programs initiated by OMS. The tests were conducted on laboratory dynamometers using standard test conditions outlined in the Code of Federal Regulations (40CFR86). The test data were supplemented by a variety of industry and other state programs.

A three-step process was defined to transform the existing OMS database and supplements into a standard database on which emissions modeling could be performed:

Step 1. Data conversion (transforming data from existing sources to target structure),

Step 2. Data cleansing (addressing text-coded fields, missing values, etc.), and

Step 3. Data screening (eliminating tests on the basis of vehicle type, fuel type, test type, etc.).

Data Conversion

Preliminary analysis of the contents of the OMS directories and files indicated which testing programs contained either the minimum criteria of vehicle information and Federal Test Procedure (FTP) test results (necessary for the development of a hot- and cold-start database) or the minimum criteria of vehicle information, FTP test results, and the results of at least one other hot-stabilized testing cycle (necessary for the development of a database to predict hot-stabilized emission rates). All programs not meeting these criteria were eliminated from further analysis.

Next, the data dictionaries of the OMS files were reviewed for parameter content. A standard file structure was then designed to accommodate the available information; the resulting record format contained 114 fields for storing vehicle and engine characteristics, test characteristics, and test results. Vehicle variables include model year, vehicle identification number, engine identification number, engine displacement and number of cylinders, fuel injection type, catalyst type, and so forth. Information about the test program, the specific test cycle, test date, odometer readings, dynamometer settings, and inertial weight are also stored. Finally, the test results in grams per mile are recorded for each pollutant. An additional 32 fields per record were defined and generated by Georgia Tech researchers to provide information on the modal characteristics of the hot-stabilized test cycles.

Patterns of file organization within the OMS database were recorded and a conversion plan was developed to transform the data from the original format into the standard file structure as defined above. Each of the OMS testing programs yielded multiple database files (vehicle files, test files, engine files, and FTP files). Because data were collected in programs that were conducted intermittently, separately, and sometimes concurrently over a 20-year period, there were significant discrepancies in file formats, specific variables collected, and data dictionaries used to code data. Database reconciliation and assembly was a daunting task, taking more than six graduate-student-months of labor to complete.

The development of the emissions database produced two separate data sets—one with FTP test results only and one with all hot-stabilized test results. These data sets contain 30,834 FTP test results (for 19,092 vehicles) and 17,417 test results (for 8,171 vehicles) on alternative hot-stabilized testing cycles, respectively.

Data Cleansing and Screening

Once data conversion and data dictionary reconciliation were complete, the new data sets were cleansed for completeness and consistency in modeling. Text-based variable values were converted to standard codes, missing values were appropriately tagged, and true "0" values in the emission rate variables were identified and set to minimum default values for modeling purposes.

Next, in preparation for the high-emitting vehicle (HEV) analysis, the FTP test data set was screened on the basis of vehicle type and test sequence. Only tests on light-duty automobiles performed with the test sequence code "As Received" [which means it was the original test using 9.0 RVP Indolene fuel at 75°F (23.9°C)] were retained in an HEV test file. This reduced FTP file contained 17,420 records representing 17,249 vehicles. (Of the 17,249 vehicles in this base file, only 78 had more than one FTP test "As Received"—56 of which had the same test date, making it impossible to determine which test was performed first. Therefore, it was decided to leave these few multiple test records in the data set to gain the additional test results.)

Because the vehicle data were collected over an extended period, it is not expected that the model year and technology distributions assembled in the database will be reflective of the in-use fleet. Older model year vehicles will have been tested many times under different testing programs, whereas data from newer model year vehicles will only have recently begun to be collected. Figure 1 provides the model year distributions in the testing database. The current Atlanta fleet distribution by model year is presented in the same figure.

Figure 1 shows that the latest model year vehicles are underrepresented in the database. Hence, analytical results are more certain for older than for newer model year vehicles. Although additional in-use testing of these newer vehicles should help improve the accuracy of emissions modeling efforts, the overall extent to which the emission rates will be improved through such testing requires further investigation given that emission rates for the more recent model year fleet are significantly lower than for older vehicles.

HEV CLASSIFICATION AND IDENTIFICATION

The next sections focus on the methodologies used to perform the HEV classification and identification. A general discussion of the analytical methods is followed by the modeling results.

Regression Tree Analysis

Hierarchical tree based regression (HTBR) can be thought of as a forward stepwise variable selection method, akin to forward step-

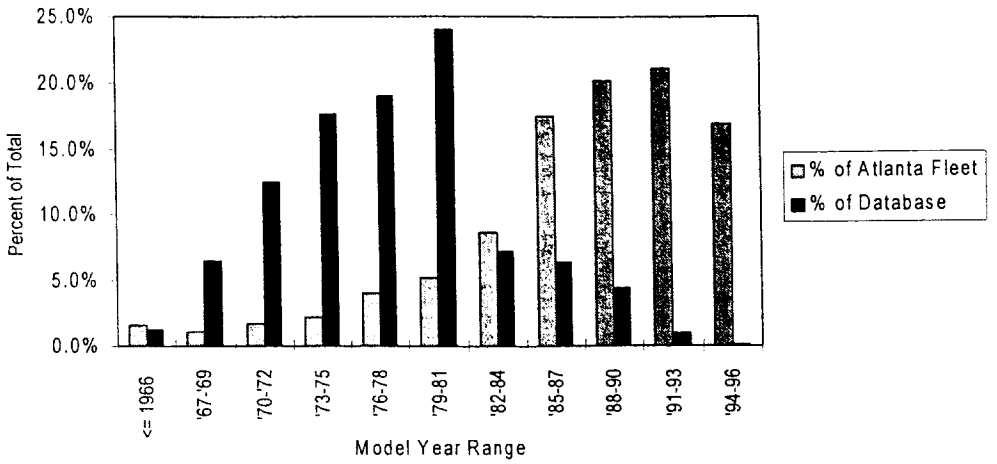

FIGURE 1 Model year distributions in 1996 Atlanta fleet and emissions testing database.

wise regression. The methods used to estimate regression trees have been around since the early 1960s and are sometimes referred to as CARTs (for classification and regression trees). The method proceeds by iteratively asking (and answering) the following two questions: (*a*) Which variable of all independent variables "offered" in the model should be selected to produce the maximum reduction in variability of the dependent response variable? (*b*) Which value of the selected variable (discrete or interval) 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.

To begin, consider the consistent terminology that will be used throughout the following discussion. The response variable, $\mathbf{Y}_n$, is a column vector of n random variables, and $\mathbf{X}_{n,p}$ is a matrix of $p - 1$ random independent variables measured for n cases. Specific to the mathematics employed in HTBR, the deviance D is defined as follows:

$$D = \sum_{l=1}^{L} (Y_l - \mu)^2 \tag{1}$$

$$\mu = \frac{1}{L}\sum_{l=1}^{L} Y_l = \text{arithmetic mean of } \mathbf{Y} \tag{2}$$

where

D = total deviance of $\mathbf{Y}$, or the sum of squared errors (SSE);
Y_l = lth observation in column vector $\mathbf{Y}$; and
L = sample size over which D is calculated ($L = n$ for total sample).

The observations in $\mathbf{Y}$ are partitioned on an independent variable X_1 that results in two subsamples, say samples b and c, each containing M and N of the original L observations ($M + N = L$). If the overall sample deviance is D_a, then the deviance reduction function is

$$\Delta = D_a - D_b - D_c \tag{3}$$

where Δ is the deviance reduction when sample a is partitioned on X_1 to obtain subsamples b and c,

$$D_a = \sum_{l=1}^{L} [Y_{(a)l} - \mu_{(a)}]^2 = \text{total deviance in sample (node) } a \tag{4}$$

$$D_b = \sum_{m=1}^{M} [Y_{(b)l} - \mu_{(b)}]^2 = \text{total deviance in sample (node) } b \tag{5}$$

$$D_c = \sum_{n=1}^{N} [Y_{(c)l} - \mu_{(c)}]^2 = \text{total deviance in sample (node) } c \tag{6}$$

$$\mu_{(b)} = \frac{1}{M}\sum_{m=1}^{M} Y_m = \text{mean of subsample (node) } b \tag{7}$$

$$\mu_{(c)} = \frac{1}{N}\sum_{n=1}^{N} Y_n = \text{mean of subsample (node) } c \tag{8}$$

In the preceding equations, M is the sample size of subsample (node) b, and N is the sample size of subsample (node) c.

A variable X_i taken from $\mathbf{X}_{n,p}$ is sought to partition the column vector $\mathbf{Y}$ such that the deviance reduction function is maximized, or, more formally, when

$$\Delta = \sum_{l=1}^{L} [Y_{(a)l} - \mu_{(a)}]^2 - \sum_{m=1}^{M} [Y_{(b)l} - \mu_{(b)}]^2 - \sum_{n=1}^{N} [Y_{(c)l} - \mu_{(c)}]^2 = \text{maximum} \tag{9}$$

While searching the matrix $\mathbf{X}_{n,p}$, two items must be sought to maximize Equation 9—the variable X_i and the numerical value on which the corresponding partition of $\mathbf{Y}$ will produce the maximum reduction of the deviance reduction function. When this maximal partition is found, the original data in node a are partitioned into two subsamples b and c having minimal combined deviance compared with all possible subsamples. Thus, the reduction in node a deviance is greatest when the deviances at nodes b and c are smallest. Numerical search procedures are used to maximize Equation 9.

Note that partitions on independent variables in $\mathbf{X}_{n,p}$ can be done on interval, ordered factor, and unordered factor variables. An optimal split t on an interval variable is found such that responses are grouped according to $X_i < t$ and $X_i > t$. Similarly, for ordered factors with J levels there are $J - 1$ possible splits. For unordered factors with J levels there are 2^J possible splits, but if empty splits are disallowed and order is ignored, there are $2^{J-1} - 1$ possible splits (1).

In HTBR, the deviance reduction function is used iteratively to partition a multivariate data space into mutually exclusive and collectively exhaustive subsamples, or nodes. The partitioning process is generally continued until (*a*) a subsample has met minimum population criteria (i.e., on the basis of statistical sampling theory) or (*b*) minimum deviance criterion at a subsample has been met. Nodes that are split are internal nodes or branch splits, whereas nodes that are not split (because of criteria provided above) are called terminal nodes or leaves. The S-Plus software program by StatSci allows the user to select either of the above criteria to control "growth" of a tree (*2*). When tree growth is terminated, the mean of the remaining sample at each terminal node is estimated as provided in Equations 7 and 8.

In the analyses that follow, the regression tree approach is used to identify vehicle characteristics and technology variables that can be used to create technology classes that exhibit emissions responses (in grams per mile) similar to the testing conditions encountered in the Federal Test Procedure. Those vehicles with similar technology traits that exhibit similar mean emission rates will be grouped together by the model, because these groupings will reduce the overall variability in emission rate response of the overall tested fleet.

Regression Tree Model Definition

Regression tree analysis, using S-Plus HTBR modeling tools, was performed on the FTP Bag 2 emission rates for each pollutant of interest. The objective of this analysis was to identify and isolate technology traits that significantly influenced the emission levels. The FTP Bag 2 emission rate was chosen as the rate for analysis for the following reasons: (*a*) the FTP test is the baseline test performed on every vehicle within a given testing program; (*b*) Bag 2 of the FTP contains hot-stabilized activities only—no engine starts are included; and (*c*) Bag 2 contains little to no enrichment events compared with other hot-stabilized test cycles.

The data in the OMS database were incomplete for a number of vehicle technology variables; consequently, the analyses only included variables that were present in the vast majority of the 17,420 test results (see Table 1). Model specifications included the minimum number of observations (MINSIZE) used to determine node splits, which was set equal to 150, and the missing data flag, which was set to OMIT, which automatically excludes all test records that are missing any of the variables specified above from the model. Consequently, the data set of 17,420 FTP Bag 2 test results was reduced to 15,061 tests.

Allowing each pollutant model to run with these initial tree specifications generated regression trees with 118, 97, and 93 terminal nodes, respectively, for carbon monoxide (CO), hydrocarbons (HC), and oxides of nitrogen (NOx). These initial trees were reduced to four terminal nodes to reduce the complexity and detail of the resultant trees (for implementation purposes). Model year and catalytic converter type were the two primary variables in these reduced trees. Since the database contains a large proportion of older model vehicles, the reduced trees were "grown" to gain additional branches (and thus produce better resolution) for more recent model year vehicles.

Model Results

Table 2 summarizes the technology classes developed by the regression tree analysis for each pollutant. Technology Class 1 represents the highest emitter class, Technology Class 2 represents the next highest emitter class, and so on.

The technology classifications identified through regression analysis are reasonable from a theoretical standpoint, in that there are logical physical reasons why certain technologies are expected to behave differently. In examining the hydrocarbon tree, the model year breaks are consistent with the implementation of new motor vehicle emissions standards for 1975 [1.5 g/mi (0.9 g/km)] and 1980 [0.41 g/mi (0.25 g/km)], assuming manufacturers were about 1 year ahead of the curve on technology implementation. The breaks associated with catalytic converter technology and model year also appear consistent with catalyst performance. Post-1979 model year vehicles using three-way catalysts emit less than their model year counterparts, and post-1988 model year vehicles with three-way catalysts pollute even less.

The regression tree for carbon monoxide is much more complicated, involving model year breakpoints at 1975 [15-g/mi (9.3-g/km) standards], 1979 [7-g/mi (4-g/km) standards came on line in 1980], 1981 [3.4-g/mi (2.1-g/km) standards], and 1990 (perhaps associated with general advances in computer control of air/fuel ratios). Smaller engines [less than 141 in.^{3} (2310 cm^3)] appear to behave significantly differently from their larger counterparts. These 3,000+ small engine

TABLE 1 Technology Variables Used in HEV Analysis

Variable Name	Description
MY	Last two digits of the vehicle's model year
INERTIA	Curb weight of the vehicle loaded with fuel and oil in kilograms
DYNOHP	Dynamometer horsepower setting, compensating for such factors as drag, coast down, friction, etc
CID	Cubic inch displacement of the vehicle's engine
TRAN	Transmission type (1 = automatic, 2 = semiautomatic, 3 = 3 speed manual, 4 = 4 speed manual, 5 = 5 speed manual)
FINJ	Fuel delivery technology (1 = port injected, 2 = carburetor, 3 = throttle body injected, 4 = pre 1981, fuel injected, but of unknown type)
CATA	Catalytic converter type (1 = none, 2 = oxidation only, 3 = 3-way catalyst, 4 = oxidation plus 3-way catalyst)

TABLE 2 CO, HC, and NOx Technology Classes and Distributions Developed Through Regression Tree Analysis

CO Technology Classes	Observations	% of Total
1) CATA < 1.5 and MY < 74.5	3070	20.4
2) CATA < 1.5 and MY > 74.5	826	5.5
3) CATA > 1.5 and MY < 78.5	4012	26.6
4) CATA > 1.5 and 78.5 < MY < 89.5 and CID < 140.5 and FINJ = 3	473	3.1
5) CATA = 2 or 3 and MY > 80.5 and CID > 140.5	1214	8.1
6) CATA = 2 or 3 and MY = 79 and CID > 140.5	876	5.8
7) CATA > 1.5 and 78.5 < MY < 89.5 and CID < 140.5 and FINJ <> 3	2471	16.4
8) CATA = 3 and MY = 80 and CID > 140.5	213	1.4
9) CATA > 1.5 and MY > 89.5 and CID < 140.5	217	1.4
10) CATA = 4 and MY > 78.5 and CID > 140.5	1222	8.1
11) CATA = 2 and MY = 80 and CID > 140.5	467	3.1

HC Technology Classes	Observations	% of Total
1) MY < 74.5	3073	20.4
2) 74.5 < MY < 78.5	4700	31.2
3) MY = 79 and CATA <> 3	1019	6.8
4) MY > 79.5 and CATA <> 3	3202	21.3
5) 78.5 < MY < 88.5 and CATA = 3	2511	16.7
6) MY > 88.5 and CATA = 3	556	3.7

NOx Technology Classes	Observations	% of Total
1) CATA < 2.5 and MY < 76.5	5824	38.7
2) CATA < 2.5 and 76.5 < MY < 80.5	3604	23.9
3) CATA > 2.5 and MY < 87.5 and INERTIA > 1787.6	521	3.5
4) CATA < 2.5 and MY > 80.5	313	2.1
5) CATA > 2.5 and MY < 87.5 and INERTIA < 1787.6	3849	25.6
6) CATA > 2.5 and MY > 87.5	950	6.3

vehicles were split into three separate technology group combinations. It is unclear why the 1980 model year vehicles in the emission database are such low CO emitters relative to the rest of the tested fleet. Additional analyses are being undertaken to identify probable causes associated with general sample selection or the distribution of 1980 model year vehicle age at time of testing.

The NOx technology classes split first on catalytic converter type. Three-way catalysts are clearly the most important technology variable affecting NOx emissions rates, and model year splits are secondary, apparently associated with standards changes in 1975, 1980, and 1981. Again, the 1988 model year is identified as a breakpoint, perhaps associated with improved fuel injection and computerized control systems. The inertial weight of vehicles appears as a factor in the NOx tree, probably associated with its effect on engine load for those model years.

The technology classes developed through the HTBR analyses are different from those used by EPA in developing MOBILE5a. The 13 technology classes used by EPA in developing speed correction factors and other algorithms did use technology groupings based on model year, fuel delivery, and catalytic converter type (*3*). These groups were based on engineering judgment associated with the phase-in of new emissions control standards and expected effects of specific technologies. The technology groups based on HTBR analyses indicated some similarities. As discussed, the phase-in of standards and technologies logically explain why groups of vehicles fall together. However, none of the EPA technology classes map exactly into the HTBR combinations. Different interactions are identified through the statistical analysis, and new variables such as cubic inch displacement and vehicle weight enter into the HTBR-derived relationships.

Distribution Analysis

The next step in HEV identification was to apply the technology class rules to the original HEV data file and to develop emissions distributions within each technology class. A separate distribution analysis was performed for each pollutant within each technology class. The grams per mile emission rates were converted first to grams per second and then to their base 10 logarithm equivalent so that the distributions would be normally distributed and standard normality assumptions could be used. Given these normality assumptions, the mean plus two standard deviations was established as the cutpoint for normal- versus high-emitter status within each technology class (where 97.73 percent of the population should fall into the normal-emitter status). The strength of this approach is that it provides a statistical basis for HEV identification. In addition, the actual HEV cutpoints can be easily modified by changing the targeted HEV percentage (which is set at 2.27 percent in the case above). The HEV cutpoint analsis is summarized in Table 3.

Comparison of this Procedure with EPA Method for HEV Classification

In developing emission rates for MOBILE5a, analyses undertaken by EPA defined high emitters as vehicles that produced composite emissions at a rate greater than five times the applicable certification standard for the pollutant and model year in question. Table 4 contains the applicable emissions standards by model year and the FTP composite emissions rates that would be defined by EPA staff as the cutpoint for high-emitting vehicle identification.

TABLE 3 CO, HC, and NOx Technology Class Distribution Analysis

CO Technology Classes	Count	Mean (g/sec)	Mean + 2SD (Cutpoint)	# Observations > Cutpoint	% High Emitter
1	3070	0.211	1.105	26	0.8%
2	826	0.058	0.412	15	1.8%
3	4012	0.029	2.084	0	0.0%
4	473	0.021	0.393	25	5.3%
5	1214	0.012	0.213	47	3.9%
6	876	0.011	0.994	0	0.0%
7	2471	0.011	0.347	62	2.5%
8	213	0.009	0.178	13	6.1%
9	217	0.004	0.069	5	2.3%
10	1222	0.003	0.275	55	4.5%
11	467	0.002	0.334	6	1.3%

HC Technology Classes	Count	Mean (g/sec)	Mean + 2SD (Cutpoint)	# Observations > Cutpoint	% High Emitter
1	3073	0.017	0.058	117	3.8%
2	4700	0.004	0.039	49	1.0%
3	1019	0.002	0.032	11	1.1%
4	3202	0.001	0.011	160	5.0%
5	2511	0.001	0.014	116	4.6%
6	556	0.000	0.004	26	4.7%

NOx Technology Classes	Count	Mean (g/sec)	Mean + 2SD (Cutpoint)	# Observations > Cutpoint	% High Emitter
1	5824	0.009	0.025	99	1.7%
2	3604	0.006	0.020	116	3.2%
3	521	0.003	0.016	19	3.6%
4	313	0.002	0.008	6	1.9%
5	3849	0.002	0.012	83	2.2%
6	950	0.002	0.010	16	1.7%

TABLE 4 CO, HC, and NOx Certification Standards (*4*) by Model Year and EPA HEV Cutpoints

	CO		HC		NOx	
Model Year	Standard g/km (g/mi)	HEV Cutpoint (5*Standard)	Standard g/km (g/mi)	HEV Cutpoint (5*Standard)	Standard g/km (g/mi)	HEV Cutpoint (5*Standard)
pre-control	135.2 (84)	676.2 (420)	17.1 (10.6)	85.3 (53)	6.6 (4.1)	33.0 (20.5)
1968	82.1 (51)	410.6 (255)	10.1 (6.3)	50.7 (31.5)	--	--
1970	54.7 (34)	273.7 (170)	6.6 (4.1)	33.0 (20.5)	--	--
1972	62.8 (39)	314 (195)	5.5 (3.4)	27.4 (17)	--	--
1973	--	--	--	--	4.8 (3)	24.2 (15)
1975	24.2 (15)	120.8 (75)	2.4 (1.5)	12.1 (7.5)	5.0 (3.1)	25.0 (15.5)
1980	11.3 (7)	56.4 (35)	0.66 (0.41)	3.30 (2.05)	3.2 (2)	16.1 (10)
1981	5.5 (3.4)	27.4 (17)	--	--	1.6 (1)	8.1 (5)
1994	--	--	0.40 (0.25)	2.01 (1.25)	0.6 (0.4)	3.2 (2)

-- not applicable; no applicable standards for this pollutant-model year combination

Rather than presume that vehicles with the same certification standard should be grouped together, the regression tree approach allows examination of whether specific technologies explain more variability than the groups created by the certification standards. Because the regression tree approach uses the emissions value located at two standard deviations above the mean of the normalized emissions distribution for each technology class to distinguish between normal and high emitters, the regression tree method will consistently predict that approximately 2.27 percent of the vehicles in each technology class are high emitters, where each technology class is a collection of vehicles that belong together with respect to emissions magnitude. The EPA method is not similarly constrained. Table 5 gives a comparison of the number of vehicles defined as high emitters by the regression tree and EPA approaches. The number of vehicles identified as high emitters by both approaches is in the first row and the number of different vehicles identified by each approach is in the second row.

The FTP composite emissions rate is calculated as a weighted average of the emissions rates noted under FTP Bag 1, Bag 2, and Bag 3 tests. The Bag 1 test is conducted immediately after an overnight soak, so bag emissions contain incremental emissions associated with cold-engine startup. The Bag 3 test also contains some incremental emissions associated with a hot-engine start (after a short 10-min soak). The contribution of the cold- and hot-start emissions to the FTP composite score makes comparisons between the EPA definition and the regression tree definition of high emitters difficult to undertake. Under the EPA definition, vehicles can fall into the high-emitter category when engine start emissions are very high, even if hot-stabilized emissions are low. As mentioned previously, the regresssion tree approach is based on hot-stabilized emissions only (in part because one of the modeling goals is to separate engine start emissions behavior from hot-stabilized behavior).

REGRESSION TREE ANALYSIS TO GENERATE EMISSION FACTORS

Having identified HEVs, each record (i.e., test result) in the master FTP data set was tagged with its appropriate technology class and HEV status (normal or high, based on its FTP Bag 2 emissions rate). Then, separate data sets were created for normal and high emitters for each pollutant of interest. This process generated a total of six data sets for subsequent modeling efforts. By separating high and normal emitters, the effect of changes in the operating environment can be modeled separately for each group. Each data set is used in the next stage of HTBR modeling to determine final technology group emission rates for hot-stabilized operations.

Emission rate models were developed using emission test results across a wide variety of testing cycles with different modal operating characteristics. Although these analyses also group vehicles by mean emissions, this step allows emissions to be defined as a function of both vehicle technology and test cycle (i.e., modal) characteristics. Details of the regression tree methods used to develop models herein are described in detail by Washington et al. (*5*). Research at Georgia Tech is under way to integrate OLS regression and HTBR modeling methods.

TABLE 5 Comparison of HEVs Identified by Regression Tree Method (Approximating 2.27 percent HEVs) Versus Standard Multiplier Method[a]

		CO		HC		NOx	
		Tree Method	FTP Method	Tree Method	FTP Method	Tree Method	FTP Method
High	# same	244	244	387	387	16	16
	# different	10	784	92	170	323	0
	% of total	1.7%	6.8%	3.2%	3.7%	2.3%	0.1%
Normal	#	14807	14033	14582	14504	14722	15045
	% of total	98.3%	93.2%	96.8%	96.3%	97.7%	99.9%

[a]based on the 15061 test records classified by regression tree method

GIS MODEL IMPLEMENTATION

The GIS-Based Modal Emission Model being developed at Georgia Tech uses the tree-generated emission rates in developing location- and time-specific emission estimates (*6,7*). The objective of the model is to explore transportation and air quality relationships and provide guidance to future software development efforts. The modeling process is based on a geographic information system, a computer modeling framework designed to store, organize, and manipulate spatial information. The model predicts emission-specific vehicle characteristics, vehicle activity, and emissions, aggregating the estimates for blocks (zonal, bounded by roads) and road segments (lineal, bounded by intersections). The model structure is illustrated in Figure 2. Vehicle characteristics are determined through regional registration data and on-road license plate studies. Vehicle activity is predicted through a combination of travel demand forecasting models and observed speed and acceleration data.

Coding of Emitter Groups

Vehicle characteristics in the model are based on the rules generated by the tree-based regression analysis. Technology classes are used to divide the fleet into pollutant-specific high- and normal-emitter groups. Each of the resulting classes of vehicles is passed through technology group rules to determine the specific emission rate for vehicles in that group. Development of location-specific fleet distributions relies on regional registration data. Regional vehicle registration data are geocoded to the study area using the registered owner's address and ZIP code. The vehicle identification number (VIN) of each record (2.2 million in Atlanta, Georgia) is decoded using software developed by Georgia Tech and a VIN decoder created by Radian International Corporation. The decoding process consists of three major steps:

1. Develop vehicle characteristics,
2. Determine technology class (to identify high versus normal emitters), and
3. Determine technology group (to assign appropriate emission rates).

Currently, the emission-specific vehicle characteristics include model year, engine size, emission control types, fuel delivery system, and dynamometer test weight. Dynamometer test weight is developed from a "lookup" file developed from insurance databases. Because of a lack of comprehensive decoding capability, approximately 8 percent of the vehicle registration data cannot be decoded (W. Bachman, J. Granell, R. Guensler, and J. Leonard, unpublished paper, 1997).

Developing On-Road Distributions (Tracking Emitter Group Activity)

Because registration databases can only provide location information at a zone level, on-road subfleet distributions need to be predicted. Currently, the Georgia Tech model uses default "local" and "regional" fleet contributions on each road segment. The local fleet is defined as all registered vehicles within a certain distance from the road (2 to 3 km). The regional fleet consists of all vehicles in the study area. The local and regional technology distributions are joined with the following conditions:

- Interstates: 60 percent regional, 40 percent local;
- Arterials: 50 percent regional, 50 percent local; and
- Other: 40 percent regional, 60 percent local.

It is not suggested that these splits be used universally. These contributions are based on preliminary work done by Tomeh using fleet technology distributions observed at five remote-sensing device

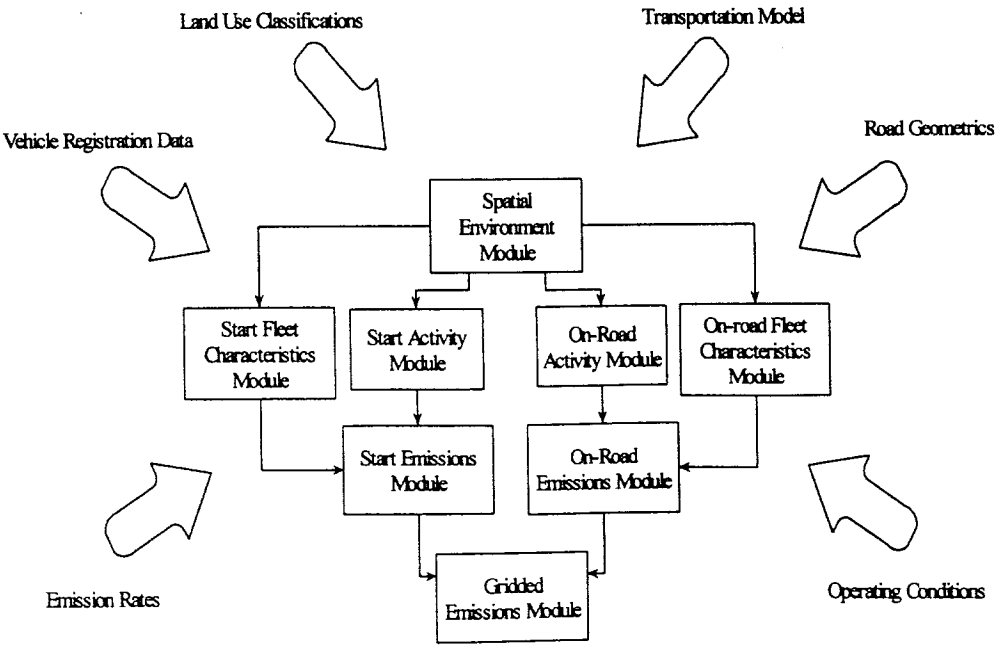

FIGURE 2 GIS-based modal model conceptual design.

(RSD) sites in Atlanta (*8*). Modeled relationships between registration, distribution, and on-road distributions are still tenuous. Detailed analyses of 57 sets of data are under way to develop the best fit contribution from local and regional fleets to yield observed fleet technology characteristics, and a variety of spatial aggregation levels are being examined (*6*).

In the GIS model, the fraction of on-road vehicle activity in each technology class on each roadway segment is divided into a high-emitter fraction and a normal-emitter fraction. The fraction of vehicles predicted to be high emitters will eventually be a function of inspection and maintenance program characteristics and socioeconomic variables associated with vehicle ownership (*6*).

Emissions from more than 300,000 vehicles in Atlanta have been measured by remote sensing. The equipment simultaneously collects a passing vehicle's tailpipe emissions and license tag. The license tags are compared with the state registration database to add the VIN and other variables. Relationships between socioeconomic variables and emitter distributions will be developed through the Atlanta remote sensing databases (*6*). The effects of I/M programs will be gleaned from cross-city studies where controlled RSD experiments have been performed.

However, until the socioeconomic relationships are developed, the model assumes default on-road VMT fractions (typically 2 to 4 percent) of high emitters by technology class. Once the fractions of high-and normal-emitting vehicles are predicted for a roadway, the fraction of activity in each high-emitter and normal-emitter technology group is determined as a function of the on-road vehicle characteristics. Then, appropriate regression-tree generated hot-stabilized emission rates (grams per second) are applied to each technology group.

CURRENT FINDINGS AND NEXT STEPS

While the regression tree approach may appear arbitrary at first glance, there are several advantages. First, emitter classes are determined statistically, so that the within-class emission rate variability is minimized with respect to between-class emission rate variability. This ensures that vehicle classes are homogeneous with respect to emissions, avoiding unnecessary class distinctions. Second, the process is flexible, so that both the number of classes and the variables selected for classifications can be changed as vehicles age (say, for example, emitter definitions are updated every 2 to 3 years). Third, high emitters are identified and selected on the basis of their performance relative to other vehicles in their class. Unlike a fixed standard, this approach permits the emission rates of vehicles within a class to drift with time without changing their emitter status.

Finally, since the primary motivation of the method is to be able to capture differences between normal and high emitters, and since high emitters can represent a marginal increase of one or more orders of magnitude, the mean of the high-emitter vehicles is largely affected by the number of high emitters identified. Stated another way, a large number of high emitters will result in a lower mean emissions rate for them relative to normal emitters. This "averaging" will diminish the difference between normal and high emitters. Since the distinction between normal-and high-emitting vehicles is vital, the regression tree method, which separates the "lowest" from the "highest," ensures a good overall accounting of emissions differences across the fleet.

To address issues with model year representation within the data set, statistical weights will be incorporated into the analysis. The research team will also further analyze the cutpoint issue between normal and high emitters. Threshold or cutpoint analyses will be conducted to determine at which points vehicle emissions change significantly in response to operating conditions. Multiple cutpoints (e.g., normal-, high-, and super-emitter groupings) will be examined for potential model improvement. Supplemental studies will yield appropriate hot-stabilized emissions rates for high-emitting and normal-emitting vehicle technology groups within technology classes.

REFERENCES

1. Venables, W. N., and B. D. Ripley. *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York, 1994.
2. *S-Plus Guide to Statistical and Mathematical Analysis: Version 3.3 for Windows.* StatSci Division, MathSoft, Inc., Seattle, Wash., 1995.
3. Chatterjee, A., T. F. Wholley, Jr., R. Guensler, D. T. Hartgen, R. A. Margiotta, T. L. Miller, J. W. Philpot, and P. R. Stopher. *Improving Transportation Data for Mobile Source Emissions Estimates.* NCHRP Project 25-7. TRB, National Research Council, Washington, D.C., 1997.
4. *Emissions and Fuel Economy Regulations.* Chrysler Corporation, Auburn Hills, Mo., Dec. 1996.
5. Washington, S., J. Wolf, and R. Guensler. Binary Recursive Partitioning Method for Modeling Hot-Stabilized Emissions from Motor Vehicles. In *Transportation Research Record 1587,* TRB, National Research Council, Washington, D.C., 1997.
6. Guensler, R., M. O. Rodgers, S. Washington, and W. Bachman. Overview of the MEASURE GIS-Based Modal Emissions Model. In *Transportation Planning and Air Quality III* (T. Wholley, ed.), American Society of Civil Engineers, New York, 1997.
7. Bachman, W., W. Sarasua, and R. Guensler. Geographic Information System Framework for Modeling Mobile-Source Emissions. In *Transportation Research Record 1551,* TRB, National Research Council, Washington, D.C., 1996.
8. Tomeh, O. *Source Apportionment of the On-Road Fleet.* Dissertation. Georgia Institute of Technology, June 1996.

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