Research Needs for Determining Spatially Resolved Subfleet Characteristics

WILLIAM BACHMAN, JESSICA GRANELL, RANDALL GUENSLER, AND JOHN LEONARD

Future emission models will need spatially resolved subfleet characteristics to determine mobile emission inventories. The use of geographic information systems and regional registration data for developing location-specific vehicle characteristics that can feed future models is addressed. Issues regarding data availability and quality are explored to define gaps in the research that may prevent development of comprehensive and accurate estimates. As a component of a larger research project studying vehicle emission modeling, a six-step process was developed and implemented for a 100 km^2 area in Atlanta. Vehicles were geocoded by using registration addresses, and vehicle characteristics were determined through a series of computer programs, commercial software, and related datasets. During the process, many research issues were identified that prevent a comprehensive assessment of spatially resolved fleet characteristics. The data and research needed to further improve the capability to generate spatially resolved subfleet characteristics were identified.

Within the next few years, new emission rate models will be available that will improve the ability of transportation and environmental planners to measure the emission impact of transportation change. Although the modeling approach still is being debated, all research efforts indicate that an improved capability to identify the emission significant components of the operating fleet will be crucial to model accuracy (*1*). Currently, emission models only allow users to vary model year distributions in describing their fleet (other characteristics are used in determining base emission rates but are imbedded in the model code). Many other variable characteristics (and their interactions) can be identified by users and hold significant explanatory power for predicting emission rates (*2,3*). Further, spatially resolved subfleet characterization is needed because spatially variant mobile emission estimates are used in subsequent models that combine mobile emission estimates with stationary sources and climatic conditions to predict ambient air quality (*4*). Therefore, there is a need for identifying procedures that can accurately predict spatially resolved vehicle characteristics for urban areas.

This report describes an experimental procedure developed at the Georgia Institute of Technology (Georgia Tech) that attempts to develop vehicle characteristic distributions for an area in Atlanta. The procedure was developed as part of an emission modeling research effort funded by the U.S. Environmental Protection Agency and the Federal Highway Administration. Several gaps in the knowledge and data availability prevent comprehensive and accurate subfleet estimates. However, significant strides have been made.

The key components of the subfleet characterization effort are a geographic information system (GIS), a regional vehicle registration dataset for Atlanta, and a series of programs that decipher vehicle identification numbers (VINs). GIS provides the ability to geographically allocate, store, and manipulate vehicle information. The regional registration database for Atlanta consists of 2.2 million individual vehicles represented by VINs and other owner information. Vehicle identification number decoding programs and related information provide a means for determining a wide range of vehicle characteristics for a fleet of vehicles. In this study, a 10 km^2 area in Atlanta was used (Figure 1).

Center for Geographic Information Systems, School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

DATA DEVELOPMENT PROCESS

The sample vehicle data was developed as part of an effort by Georgia Tech to create a research-grade, GIS-based modal emission model (*4*). The model was intended to explore transportation and air quality relationships and provide guidance to future software development efforts. The model predicts emission-specific automobile characteristics, automobile activity, and emissions, aggregating the estimates to blocks (polygons, bounded by roads) and road segments (lines, bounded by intersections). Vehicle characteristics are determined through regional registration data. Vehicle activity is predicted through a combination of travel demand forecasting model results and observed speed and acceleration data. The pollutants estimated in the model framework are carbon monoxide, hydrocarbons, and oxides of nitrogen.

One of the objectives for the research model was to better characterize spatial variability in emissions due to the spatial variability of vehicle characteristics. Emission-specific vehicle characteristics were identified in a regression tree analysis of emission test data (*5*). The specific characteristics are model year, engine size (cubic inch displacement), emission control type (oxidation, three-way catalyst, both, or none), fuel injection type (port, carburetor, or throttle body), and dynamometer test weight. Vehicles were divided into high and normal emitters for each of the three pollutants of concern (carbon monoxide, hydrocarbons, and oxides of nitrogen). Each emitter group was subdivided further into technology groups based on vehicle characteristics that statistically explained variability found in observed emission rates. The sample site was used as a model development area because it contains a variety of land uses, road classifications, and a major highway interchange.

The conceptual process for developing the subfleet location and characteristics is shown in Figures 2 and 3. In the figures, divided boxes represent entities, or specific files. Undivided boxes represent a process or series of processes. Shaded boxes indicate that the entity is a GIS dataset (one that has topology).

Step 1—Address Geocoding

Vehicle registration data is compiled by most state or county governments annually. The forms submitted by vehicle owners identify the registered mailing address, the VIN, the make, the model, the

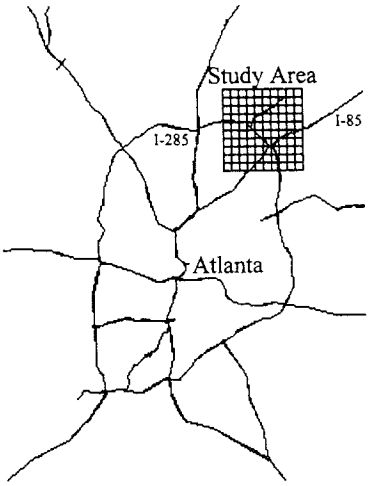

FIGURE 1 Study area in Atlanta.

mileage, and the model year. Under the assumption that the mailing address provides an accurate representation of the home location for the vehicle, the address can be used as the location parameter, allowing spatial aggregation. GIS software and a comprehensive road database can be used to determine geographic coordinates for the street address. Errors associated with the road dataset or the mailing address can result in incomplete or incorrect matching. Generally, address matching rates of 60 to 80 percent success are considered fairly good. Unmatched vehicles default to the registered zip code as their location parameter.

The raw Atlanta regional registration database of 2.2 million records was searched for vehicles registered in zip codes found in the 100 km^2 study area. The resulting vehicles (163,848 records) were geocoded by using three road datasets: the Atlanta Metropolitan Planning Organization database, a commercially available database, and 1994 TIGER (a U.S. Bureau of the Census product). The three individual databases were used because the road datasets were either not comprehensive or not current. The rejects from one geocoding process were fed to the next dataset, resulting in additional matches. At the end of the process two files existed—a successfully geocoded file (134,725 vehicles) and an unsuccessfully geocoded file (29,123 vehicles). The match rate varied from zip code to zip code but averaged 82.2 percent. Although correctly geocoded within their zip codes, many of the matched vehicles were removed because their registered location was outside the 10 km^2 limits of the study area (zip code boundaries extended outside the study limits). The resulting geocoded files consisted of 56,611 vehicles suspected of residing at specific points in the study area and 29,123 vehicles suspected of residing somewhere within their zipcode.

Sources of Error

Studies have shown that about 80 percent of trips generated in an urban area start or end at home. For this research effort, it was assumed that the registration site is the residence of the owner and, for modeling purposes, the origin of the trips. Sometimes, however, vehicles are registered at one location, but the owner lives elsewhere. Although unvalidated, it was assumed for the research that vehicle registration data is a good indicator of the origins of the trips.

The process of developing state registration datasets is hampered by the transcription process. In Georgia, registration forms are scanned, resulting in data loss and data error. The extent of the errors is unknown.

Geocoding failures can occur for several reasons, but most failures are the result of incomplete road datasets or addresses with additional or nonstandard character strings (i.e., apartment numbers, suite numbers, P.O. boxes). Most road databases tend to have problems with accurately matching new addresses because it takes a continuous effort to keep the road datasets up to date. Multifamily housing districts are underrepresented because the apartment numbers can cause the address to be misinterpreted.

Research and Data Needs

Research is needed that studies the relationship between registered addresses and actual residences. Improvements are needed in the data communication between vehicle owners, counties, and state agencies. Comprehensive, accurate, well-maintained road datasets also are needed.

Step 2—Aggregating to Zones

The geocoded vehicles were aggregated to U.S. Census block zones (925 in the study area) because of privacy concerns for individual vehicle owners. The zones generally were areas bounded by roads. For zones with geocoded vehicles, the average number of vehicles per zone was 39.9. Given the scale of ozone formation and the spatial quality of other important factors, the aggregation to small zones is not expected to reduce the overall spatial resolution of the final emissions model.

Sources of Error

There are potential problems with the relational accuracy of the zones and the road datasets used in geocoding. Because Census block boundaries may represent roads, and vehicles are assigned to roads based on address, the zone boundaries and roads must be similar in relative positional accuracy. Two solutions to this problem are possible: offset the addresses from the road to be sure they fall within the zones, or define the zones to be large enough to lessen the impact of the edge addresses. The first solution was used in the Georgia Tech model. The second option will reduce spatial resolution (larger zonal aggregations); however, the optimum zone size has not yet been defined in research.

Data Needs

Improved absolute and relative positional accuracy among datasets is needed, as is research into the spatial resolution needs of mobile emission estimates.

Step 3—Decoding VINs

The VIN, assigned to each vehicle during assembly, is a code that can be deciphered to identify a variety of vehicle characteristics. The information obtainable from the VIN has changed from year to year. Before the end of the 1970s, there were no rules regarding the length

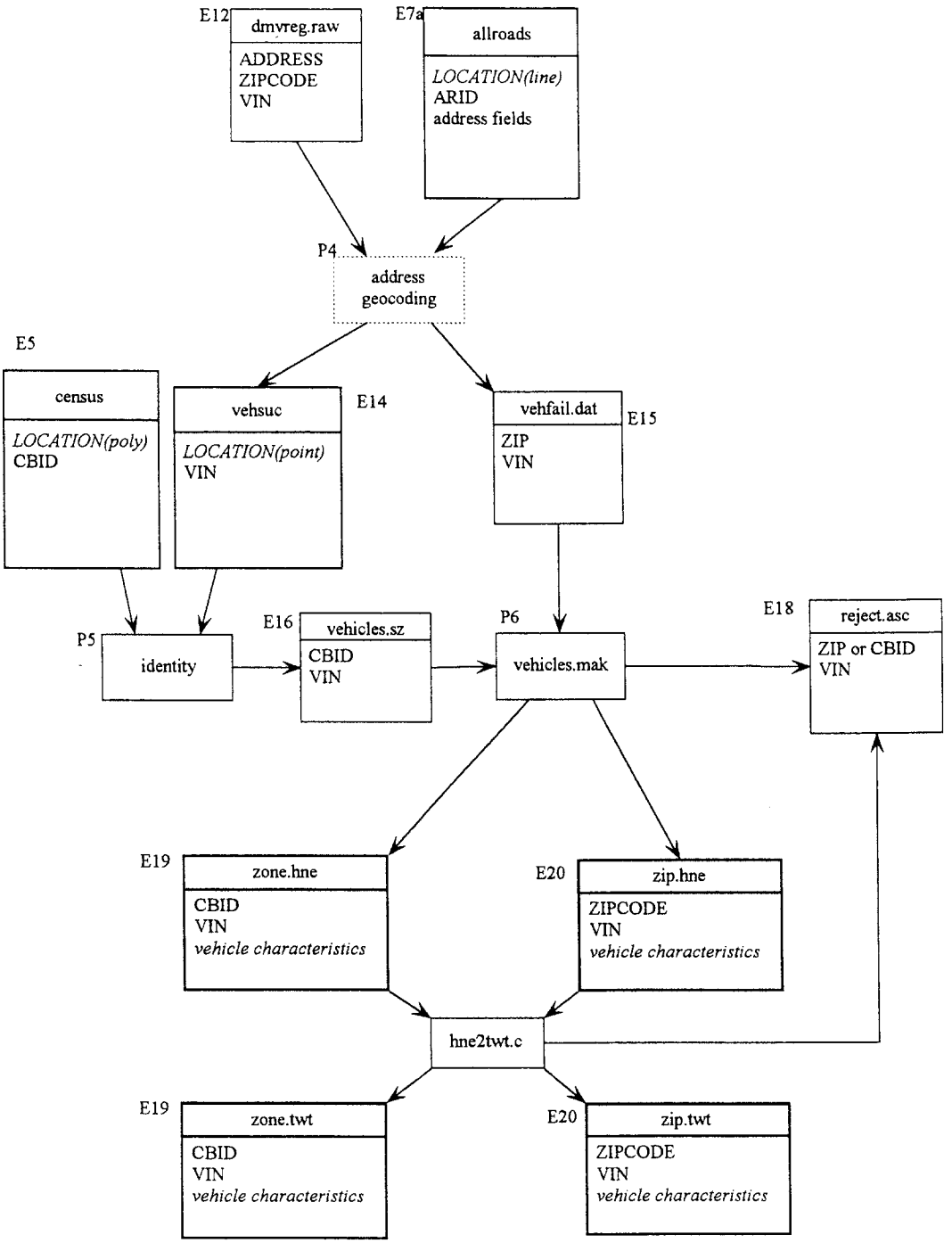

FIGURE 2 Data development process, Part 1.

or content of the VIN, resulting in duplication and lack of uniformity. Manufacturers finally agreed on a standard VIN format resulting in the 17-digit number. For model years after 1980, the VIN of any vehicle is alphanumeric and consists of 17 characters as follows:

- The first three digits uniquely identify the manufacturer, make, and class of vehicle, providing information about the nation of origin. This is an international identifier because all world manufacturers are assigned an identifier number within their country.
- Characters four through eight uniquely identify decodable information for the vehicle, as described in Table 1.
- The ninth character is a check digit, used to verify the sequence and format accuracy of the VIN.
- The tenth character is the vehicle's model year.
- The eleventh character is the assembly plant.
- The last portion of the VIN is a six-digit number that constitutes the production number and is sequentially assigned by the manufacturer.

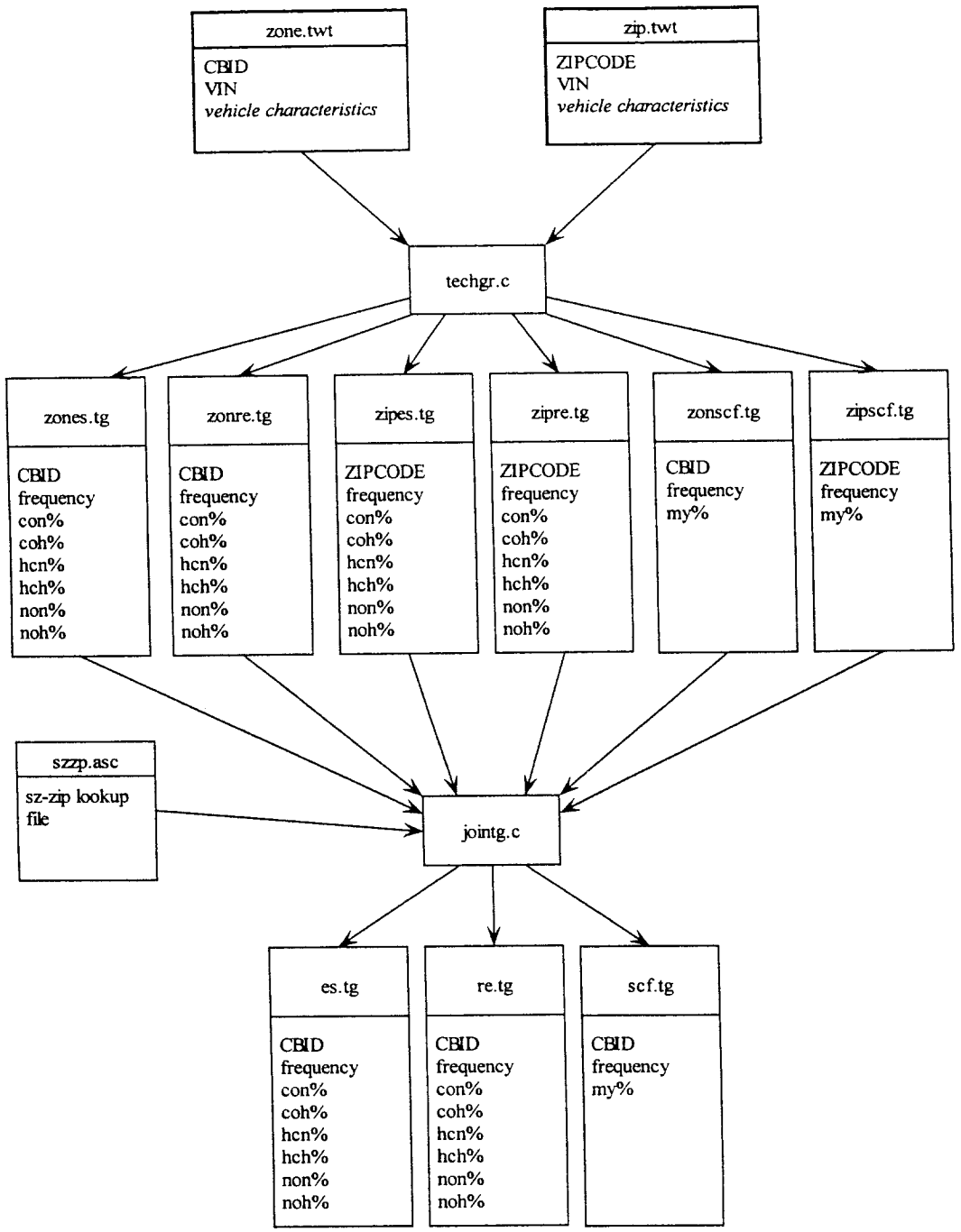

FIGURE 3 Data development process, Part 2.

Both geocoded and ungeocoded files of vehicles (zones and zip codes, respectively) were decoded by using Radian International's VIN decoder software (*6*). This program decodes each VIN by cross-checking it with a group of files that contain descriptive information (i.e., vehicle manufacturer, make, model,). After decoding, the software stores the results in two files—an output file with the decoding results, and a file containing the VINs that were not decoded.

The VIN decoder parameters that are useful for fleet characterization are as follows:

- Model year;
- Vehicle class: car, truck, multipurpose vehicle, van, bus, and incomplete vehicle chassis;
- Engine displacement: minicompact, subcompact, compact, midsize, and large vehicles;
- Number of cylinders;
- Fuel delivery system: one-, two-, three-, and four-barrel carburetor systems, fuel injection system, diesel fuel, electronic fuel injection, multiple fuel injection system, multipoint injection system,

TABLE 1 VIN Decoded Vehicle Attributes

Vehicle Class	Decoded information
Passenger Car	Line, series, body type, engine type, and restraint system
Multipurpose	Line, series, body type, engine.
Truck	Model or line, series, chassis, cab type, engine type, brake system, and gross vehicle weight rating.
Bus	Model or line, series, body type, engine type, and brake system.
Trailer	Type of trailer, body type, length and axle configuration.

throttle body injection system, central point injection system, diesel fuel injection system, sequential fuel injection system, and liquefied petroleum gas; and

- Emission control system: air injection emissions control system, evaporative emission control system, three-way catalyst control system, oxidizing catalyst control system, exhaust gas recirculation system, open-loop mode code control, and closed-loop control system.

Many of these fields are not identified explicitly in the VIN. At least two important variables (emission technology and fuel delivery system) are identified by using a lookup routine in the decoder. This initial assessment of the decoder did not account for measuring the accuracy of the actual output, only if the output was possible. A FORTRAN routine was written by Georgia Tech to prepare the vehicle files (consisting of a location identification and the vehicle's VIN), run the decoder, and pull out information important to the project. For the zone file, 7.8 percent (4,450) of the vehicles could not be decoded. Of the remaining 52,161 vehicles, 37,371 were automobiles, the vehicle type of concern to the study. For the zip code files, 7.6 percent (2,226) of the vehicles could not be decoded. Of the remaining 29,124 vehicles, 26,897 were automobiles.

Sources of Error

Because there were no rules governing VIN assignment before the 1970s, the VIN decoder does not decode vehicles built before 1971.

To undertake the assessment of the accuracy of the VIN decoder, a random sample of 100 VINs was decoded manually by using guides for the identification of domestic and imported vehicles (*7,8*). This series of eight books contains VIN information for passenger cars, light trucks, and vehicles not destined for use in the United States, from 1938 to 1995. The results showed that 71 percent of the sample corresponded to the decoder results. Another 17 percent of the sample was decoded but was unidentifiable in the guides. The remaining 12 percent was decoded incorrectly by the decoder.

A large portion of the incorrectly decoded vehicles were built before the 17-digit standard was implemented. Figures 4 and 5 show that the 1973 autos and 1978 nonautos had significantly high errors. A further analysis of the 1973 vehicles revealed that the VIN decoder incorrectly identified some pre-1972 vehicles as 1973 Volvos and BMWs.

Data Needs

A comprehensive, maintained VIN decoder is needed.

Step 4—Determining High and Normal Emitter Groups

There are large differences in emissions between vehicles with functional control systems (catalytic converters, oxidation, etc.), and those with malfunctioning, deteriorated, altered, or nonexistent control systems. The latter group can have significantly higher emissions and thus have been termed high emitters. From a fleet or subfleet perspective, it is difficult to identify which vehicles are high and which are normal emitters. Furthermore, a vehicle can

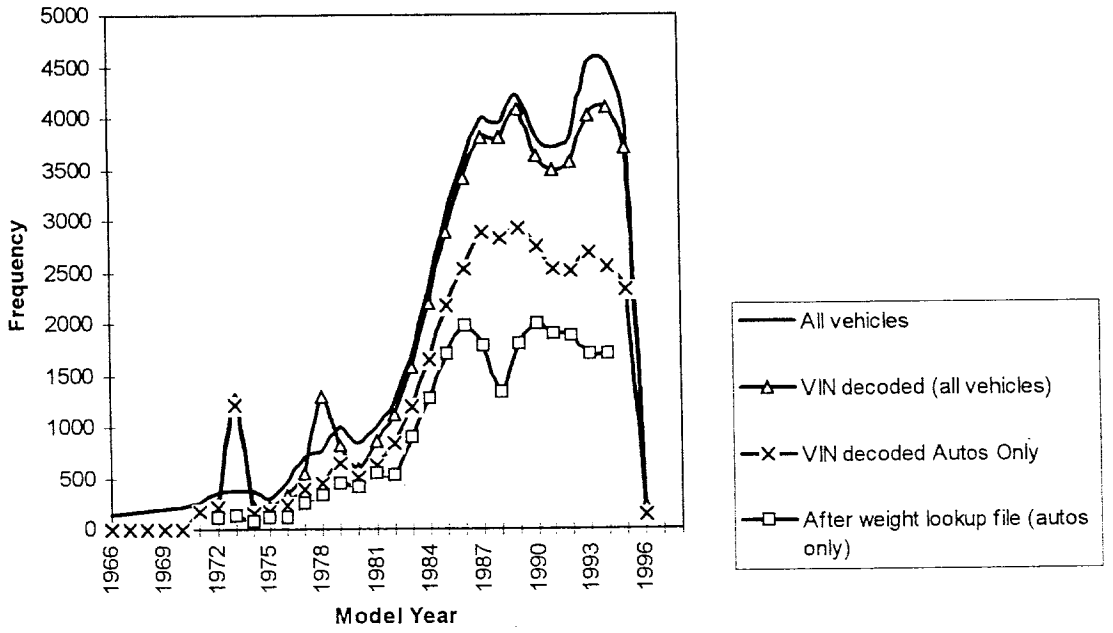

FIGURE 4 Model year frequency.

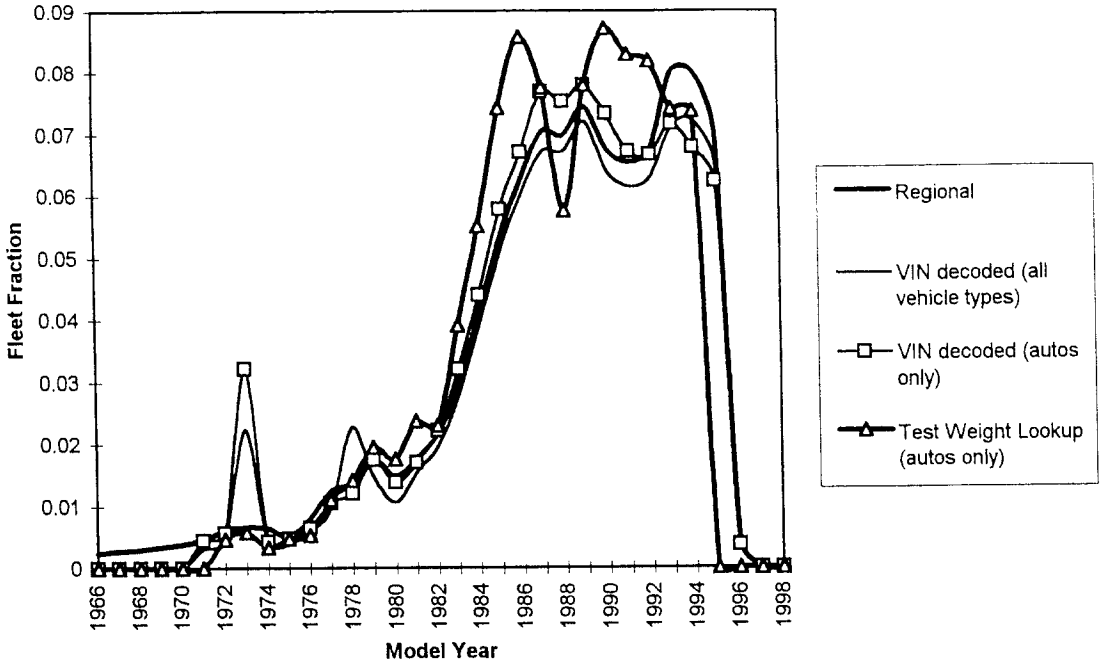

FIGURE 5 Model year fraction.

be high emitting for a single pollutant, thus complicating the process. To alleviate this problem, Georgia Tech defined groups of vehicle characteristics that result in a different likelihood of being high emitters. The fleet or subfleet is divided into the emitter groups and randomly assigned a pollutant-specific high emitter flag based on the basis of observed groups' specific likelihood fraction (1 to 3 percent). Thus, all the vehicles are assigned a normal or high flag for each pollutant. This process was described fully by Wolf et al. (*5*).

Remote sensing (not satellite remote sensing, but data collection by a device that is remote from the vehicle) refers to a process that can be used to collect a vehicle's tailpipe emissions as it is driving on the road. This is accomplished through a device that sits on the side of a road and monitors an infrared beam projected across the traffic (at tailpipe height), measuring the selective absorption of infrared radiation by carbon monoxide, carbon dioxide, and hydrocarbon gases (unfortunately, oxides of nitrogen cannot be measured with this device). Simultaneously, digital cameras capture the vehicle's license plate. The license plate can be used to identify the VIN in the registration database, thus providing a link to vehicle characteristics. Address-matching process can identify the registered location of a vehicle, and remote sensing can capture the vehicle's characteristics as it passes through a point on the network. The disadvantage is that the process is time consuming and is not comprehensive. There also is the difficult issue of reconciling remotely sensed measurements to the large numbers of laboratory tests.

Sources of Error

The success of this strategy is based on the ability to determine the appropriate percentages of high emitter vehicles by specific groups. Further, because distributions are being used, large samples are needed.

Data Needs

Accurate identification of the percentage of high emitters by a predefined class of vehicles will improve the ability to appropriate represent high emitting vehicles. Accurate nitrous oxide measurement from remote sensing devices is needed.

Step 5—Add Weight Field

The vehicle weight field was discovered to hold significant emission-rate explanatory power, but it was excluded from the list of characteristics identifiable in the VIN. Therefore, an outside source must be used to look up the appropriate weight given the make, model, and model year of the vehicle. Information about the established curb weight of the vehicles as provided by the manufacturers was obtained from the sources given in Table 2.

Little information was discovered for vehicles made before 1972 and after 1994. The data gaps result in a substantial loss of information. Of the 37,371 automobiles that were geocoded successfully and VIN decoded, the weight was successfully identified for only 23,010. Of the 19,531 automobiles based at zip codes, 11,836 were matched successfully with weights. The resulting overall data loss was approximately 39 percent.

TABLE 2 Curb Weight

Make Year	Source
1972- 1978	N.A.D.A. Classic, Collectible and Special Interest Car Appraisal Guide, 1997
1979- 1988	N.A.D.A. Appraisal Guide, 1996
1988- 1994	Consumer Reports Cars, The Essential Guide for Buyers and Owners, 1997

Sources of Error

Missing data is the major cause of error, resulting in a 39 percent data loss. Information about the established curb weight was obtained from various sources. In comparing the information provided by the sources, it was found that there was a wide variability in the levels of precision of the data. Errors also can occur in the matching process because of varying data make and model descriptions between the lookup tables and the registration.

Data Needs

A comprehensive source of vehicle weight data that will aid in the development of a lookup file is needed. Adding characteristics to the VIN decoder by incorporating lookup routines based on make, model, and model year will ease the data development process.

Step 6—Zonal Technology Group Fractions

When all the emission-specific vehicle characteristics were identified, vehicles were assigned a technology group. A technology group is an emission rate classification that combines vehicles with similar emission behavior into a single subfleet category. The rules for the groups were defined through a regression tree analysis of vehicle emission test data (*5*). Technology groups are defined for engine start and running exhaust modes, for normal and high emitters, and for each pollutant of concern. For example, the carbon monoxide normal emitter technology groups for engine start emissions were defined as follows:

- Model year < 1981 and weight < 3,250;
- Model year < 1980 and weight ≥ 3,250 and weight < 4,375;
- Model year < 1980 and weight ≥ 4,375 and cubic inch displacement (CID) < 351;
- Model year < 1980 and weight < 4,375 and CID ≥ 351;
- Model year ≥ 1980 and weight ≥ 3,250;
- Model year = 1981 and weight < 3,688 and CID < 131;
- Model year = 1981 and weight < 2,938 and CID ≥ 131;
- Model year = 1981 and weight ≥ 2,938 and weight < 3,688 and CID ≥ 131;
- Model year ≥ 1982 and model year < 1987 and weight < 3,688;
- Model year ≥ 1981 and model year < 1987 and weight ≥ 3,688; and
- Model year ≥ 1987.

All the normal emitter flagged vehicles are aggregated into one of the technology groups based on these conditions. Each technology group later is assigned a gram-per-start emission rate. Technology group fractions are summarized by zone, thus defining a mode, emitter type, pollutant-specific subfleet.

Sources of Error

The main source of new error in this step is the accurate group definitions. Misrepresentation of vehicles is possible because of the limited nature and condition of the vehicle emission test datasets used by Wolf et al. (*5*) in developing the group rules.

Data Needs

Comprehensive vehicle emission test data will improve the ability to accurately characterize a vehicles emission rate.

Developing On-Road Subfleet Distributions

Currently, most agencies use regional distributions of vehicle model year distributions to account for technology impact. The improved modeling regime being researched by Georgia Tech needs to better characterize the fleet distribution on the road. After the origin zone subfleet characteristics are defined, the results are used to predict on-road distributions. Because registration databases can provide location information only at a zone level, on-road subfleet distributions need to be predicted. Currently, the Georgia Tech model predicts a local and a regional fleet distribution for 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 two 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.

Although the approach is very coarse, it represents an initial improvement over using a regional distribution. The resulting on-road distribution estimates for Atlanta are being validated through data from a remote-sensing effort conducted by Georgia Tech's Air Quality Laboratory. More than 50 road segments of various classifications and 200,000 vehicles were observed by the study. The observed fleet distribution will be compared to the predicted. New strategies may be developed as data collection reveals more information.

SUMMARY OF RESEARCH NEEDS

The process discussed in this report represents a first step toward developing spatially resolved subfleet characteristics needed for developing emission models. Significant gaps in the data and research prevent comprehensive and accurate spatially resolved subfleet characterization. Currently, the VIN can provide only a limited amount of information about a make, model, and model year vehicle. Additional lookup data are needed to develop a complete database. Other important vehicle characteristics, such as weight and frontal area influence, could be used to predict engine load, an important factor in determining emission rates. However, these obstacles can be overcome with effort placed into developing new, expanded VIN decoders.

Further, resolving issues regarding the level of spatial aggregation required to accurately characterize the fleet will help define the level of detail warranted by emission modeling. Data availability and quality will vary around the country, thereby requiring the development of specific spatial modeling guidelines.

The following research needs were identified:

1. Research is needed that studies the relationship between vehicle registered addresses and actual residences of owners.

2. Improvements are needed in data communication among vehicle owners, counties, and state agencies.

3. Comprehensive, accurate, well-maintained road datasets are needed.

4. Improved positional and relational accuracy among datasets is needed.

5. Research into the spatial resolution needs of mobile emission estimates is needed.

6. There is a need for a comprehensive, maintained VIN decoder.

7. Accurate identification of the percentage of high emitters by a predefined class of vehicles will improve the ability to appropriate represent high emitting vehicles.

8. A comprehensive source of vehicle weight data is needed that will aid in the development of a lookup file.

9. Adding characteristics to the VIN decoder by incorporating lookup routines based on make, model, and model year will ease the data development process.

10. Comprehensive vehicle emission test data will improve the ability to accurately characterize a vehicles emission rate.

REFERENCES

1. Siwek, S. J. *Summary of Proceedings*. EPA-FHWA Modeling Workshop, Ann Arbor, Mich., 1997.
2. Guensler, R. Data Needs for Evolving Motor Vehicle Emission Modeling Approaches. *Transportation and Air Quality II*. ASCE, 1994.
3. Barth, M., F. An, J. Norbeck, and M. Ross. Modal Emissions Modeling: Physical Approach. In *Transportation Research Record 1520*. TRB, National Research Council, Washington, D.C., 1996, pp. 81–88.
4. Bachman, W., W. A. 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, pp. 123–132.
5. Wolf, J., S. Washington, R. Guensler, and W. Bachman. High Emitting Vehicle Characteristics Using Regression Tree Analysis (in press). TRB, National Research Council, Washington D.C., 1998.
6. Vehicle Identification Number Decoder, Version 95, Radian Corporation, Austin, Tex., 1995.
7. *N.A.D.A. Classic, Collectible and Special Interest Car Appraisal Guide, 1997*. Lee Books, Washington, D.C., 1997.
8. *N.A.D.A. Appraisal Guide, 1996*. Lee Books, Washington, D.C. 1996.

Publication of this paper sponsored by Task Force on Geographic Information Systems for Transportation (GIS-T).