Statistical Analysis and Model Validation of Automobile Emissions

DANIEL SCHULZ

THEODORE YOUNGLOVE

MATTHEW BARTH

University of California, Riverside

ABSTRACT

A comprehensive modal emissions model has been developed and is currently being integrated with a variety of transportation models as part of National Cooperative Highway Research Program Project 25–11. Second-by-second engine-out and tailpipe emissions data were collected on 340 light-duty vehicles, tested under "as is" conditions. Variability in emissions of CO_2, CO, HC, and NO_x were observed both between and within groups over various driving modes.

This paper summarizes initial statistical analysis and model validation using bootstrap validation methods. The bootstrap methodology was shown to be a valuable tool during model development. A significant positive bias (overprediction) in NO_x during higher speed driving was identified in CMEM v1.0 and eliminated in CMEM v1.2.

INTRODUCTION

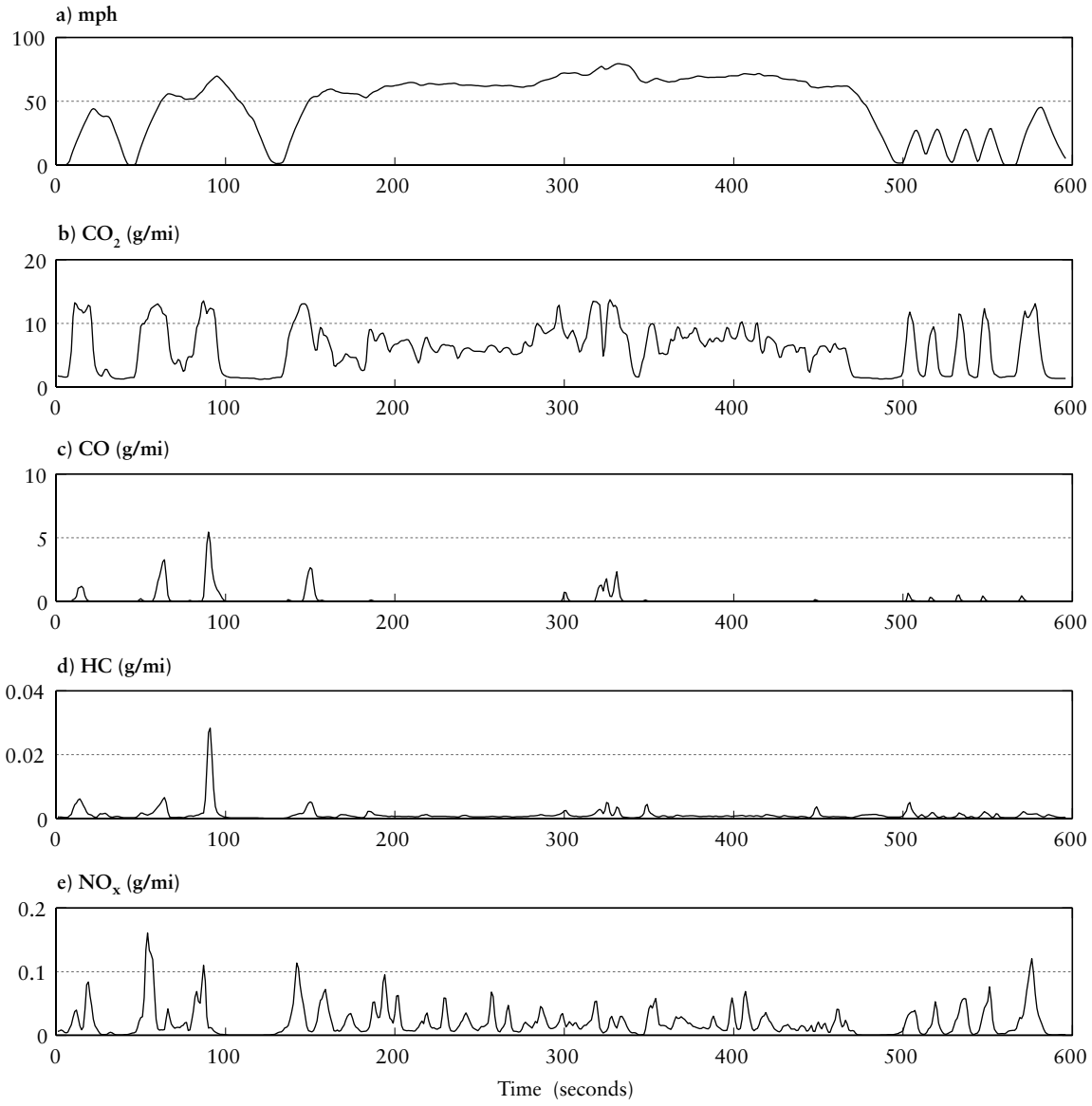
Measurements of automobile tailpipe emissions at second-by-second time resolution provide a statistically challenging data set for modeling and analysis. Emissions can vary by an order of magnitude within the space of a few seconds, with the response frequently nonlinear, due to enrichment

Theodore Younglove, Bourns College of Engineering, Center for Environmental Research and Technology, University of California, Riverside, CA 92521. Email: tyoung@cert.ucr.edu.

or enleanment of the air-fuel mixture. Figure 1 presents an emission trace from a representative, normally operating vehicle to illustrate the large differences in magnitude of tailpipe emissions over the driving schedule.

Enrichment occurs in modern computer-controlled vehicles based on proprietary engine control strategies. The computer enriches the air-fuel mixture at high power to protect the catalytic converter from heat damage, resulting in short-term spikes in emissions. The size and timing of the increases in emissions vary from vehicle to vehicle, even for identical models. Enleanment occurs in some modern computer-controlled vehicles during coastdown and braking events. In normal powered driving, the amount of condensed fuel on the walls of the intake manifold is in rough equilibrium with the addition of fresh condensate from fuel injection and with the loss by evaporation into the air moving into the cylinders. The amount of fuel on the walls depends to some extent on the recent history of fuel injection, that is, the recent power level. When engine power is negative, there is still significant air-flow but little or no fuel injection. The condensed fuel will be removed by evaporation over a period of seconds and will pass through

FIGURE 1 Second-by-Second Data for a 1986 Buick for a) Speed; b) CO_2; c) CO; d) HC; and e) NO_x



the cylinders. The critical fact is that during these events the fuel-air ratio is typically very lean, so lean that there is little or no combustion. In this case, hydrocarbon (HC) emissions can become quite high relative to normal operation. Second-by-second changes in emissions can occur during constant speed cruising, due in part to small changes in throttle position that, in turn, affect manifold air pressure without affecting vehicle speed.

In addition to these large differences in emissions for individual vehicles during driving, there are large differences in emissions from vehicle to vehicle. Changes in emissions behavior under different driving conditions occur because of changes in vehicle-emissions control technology. Large reductions in the emission of carbon dioxide (CO_2), carbon monoxide (CO), hydrocarbon (HC), and nitrogen oxides (NO_x) have been achieved over the past 25 years, resulting in great differences in emission rates between vehicle/technology groups (Calvert et al. 1993).

In late 1995, the Bourns College of Engineering, Center for Environmental Research and Technology (CE-CERT) at the University of California, Riverside undertook a cooperative investigation with the University of Michigan and Lawrence Berkeley National Laboratory in order to develop a comprehensive modal emissions model (CMEM). The overall objective of this research project was to develop and verify a modal emissions model that accurately reflects emissions from light-duty vehicle (LDV), cars and small trucks, produced as a function of the vehicle's operating mode. The model is comprehensive in the sense that it will be able to predict emissions for a wide variety of LDVs in various conditions (e.g., properly functioning, deteriorated, malfunctioning). The model is capable of predicting second-by-second tailpipe and engine-out emissions and fuel consumption for a wide range of vehicle/technology categories. The principal sponsor of this project is the National Cooperative Highway Research Program, NCHRP, Project 25–11 (see An et al. 1997). CMEM is a physical, parameter-based model requiring parameterization of many processes involving the vehicle, engine, emissions control system, and catalytic converter, and affecting how the vehicle is driven. Many of the relationships must be approximated within the model, and the parameters themselves are estimated from measurement data subject to error. This model differs from other conventional emissions models in that it is modal in nature: it predicts emissions for a wide variety of light-duty vehicles over a wide variety of driving modes, such as acceleration, deceleration, and steady-state cruise. The two primary models currently in use are MOBILE, developed by the U.S. Environmental Protection Agency, and EMFAC, developed by the California Air Resources Board. Both MOBILE and EMFAC predict vehicle emissions based in part on average trip speeds and depend on regression coefficients derived from a large number of trip average emission measurements for a driving schedule representative of "typical" driving. For more detail, see Barth et al. (1996), Barth et al. (1997), and An et al. (1997). Only emissions from light-duty vehicles are considered in this paper.

For model validation, the key question to answer is whether the model predicts emissions with reasonable accuracy and precision. Bornstein and Anderson (1979) have pointed out the need for communication between modelers and statisticians in air pollution research. Since then, Hanna has done considerable research into the development of statistical methods for air quality investigations (Hanna and Heinold 1985; Hanna 1988 and 1989). Of particular interest is his use of the normalized mean square error (NMSE) methods for estimating bias based on a percentile of observed and predicted differences, as well as his application of Efron's bootstrap resampling methods to compare different air pollution models (Efron 1982; Efron and Tibshirani 1986). Bootstrap bias plots, shape statistic plots, histograms of bias values, bootstrap confidence interval length plots, and maximum and minimum bias plots have also recently been used in the context of validating a complex modal emission model (Schulz et al. 1999).

In developing CMEM, several validation techniques were used: 1) validation of intermediate variables, such as modeled engine RPM against observed RPM, 2) composite vehicle schedule validation, and 3) second-by-second individual vehicle validation. Validation was undertaken on a sec-

ond-by-second basis for individual vehicles to provide a robust data set on which to test the model and to ensure that a sufficient number of vehicles would be available for the bootstrap analysis. It should be noted that although this validation is accomplished at a second-by-second basis, the model was intended for use on driving modes lasting ten or more seconds. This difference was necessary for model development because of the need to identify situations in which problems were occurring. Practically speaking, many of the errors will "average out" over a driving schedule.

The focus of this paper is the validation methods employed on a second-by-second basis for use by the modeling team in model diagnostics and model improvements. The statistics used for model evaluation on a second-by-second basis must be valid under many possible distributions of emissions but must also be easily understood by nonstatisticians. In addition, while the initial validation presented in this paper was conducted on two large groups of vehicles, the methodology employed also needed to be valid for analysis of model performance on the individual vehicle/technology groups with 10 to 25 vehicles in each group. For these reasons, second-by-second validation methods inspired by Hanna's work are described and applied to two versions of the model.

METHODOLOGY

Vehicle Recruitment and Testing

The gasoline powered light-duty fleet was divided into 24 categories for vehicle recruitment, with divisions based on vehicle type (car or truck), emissions status (normal or high emitter), fuel control technology, emission control technology, power-to-weight ratio, and accumulated mileage. High-emitting vehicles were defined as those having CO, HC, or NO_x emissions 1.5 or more times higher than the certification standard for the vehicle. Vehicles ranged in age from a 1965 Ford Mustang to a 1997 Dodge Ram pickup and represented all major foreign and domestic auto manufacturers. The vehicle/technology groups were chosen to cover the range of vehicle technology types within the gasoline powered light-duty vehicle fleet. A total of 340 in-use vehicles were recruited and tested, primarily from the South Coast Air Basin, with a small subset brought in from other states. Particular care was given to target forty-nine state-certified vehicles, as well as California-certified vehicles, to ensure the model was representative of the national LDV population. Vehicles were selected at random from the Department of Motor Vehicles registration list for Southern California. Recruitment was conducted through a mailing to vehicle owners within the 24 categories, but category sample sizes were selected by model development needs rather than population proportions. Once recruited, the vehicles were tested on CE-CERT's forty-eight-inch electric chassis dynamometer using three driving schedules: the Federal Test Procedure (FTP), which the federal government uses to represent normal in-use driving; the US06 driving schedule, which the federal government uses to represent in-use hard driving; and the Modal Emission Cycle (MEC), developed as part of NCHRP Project 25–11 to measure emissions during specific driving modes (Barth et al. 1996). It should be noted that the third driving segment of the FTP driving schedule and the US06 driving schedule were not used in model development. For this reason, they were used as independent validation schedules. During testing, emissions of CO_2, CO, HC, and NO_x were measured on a second-by-second basis.

Time-Alignment of Data

To perform a meaningful second-by-second validation, the emissions test results first had to be time-aligned. The time delay between the start of data recording and the start of the vehicle is not automated and can vary by several seconds from one vehicle to the next. Prior to the application of bootstrap analysis to the vehicle data, all values were time-aligned to reflect acceleration from a common starting point. Small differences between the driving trace and the schedule speed trace are inevitable during the test schedules, so the time alignment is not perfect. Although some error can arise when time-aligning the files to the nearest second, it should be negligible when compared with the deviations from the driving trace resulting from human error.

Validation Statistics

A measure of closeness, called model bias but not the same as the statistical definition of bias (a property of an estimator of an unknown population parameter) is given by

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n} i^{\text{th}} \text{ modeled value} - i^{\text{th}} \text{ observed value} =$$

$$\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i) = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i - \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{\hat{y}}_i - \bar{y}_i, \qquad (1)$$

where y_i is the i^{th} observed emission value, $\hat{y}_i$ is the corresponding i^{th} value predicted by the model, and there are n observations in the sample. This is consistent with the standard definition of bias historically used in environmental pollution studies (Zannetti 1990), but in the language of statistics it is referred to as mean prediction error. If this bias value is larger (smaller) than an acceptable predetermined cutoff value, then the model significantly overpredicts (underpredicts).

A point estimate of bias is useful, but statistics are random quantities that vary from sample to sample. Confidence intervals provide a better description of a reasonable range of values for the bias statistic. If the confidence interval (95% confidence intervals are used in this paper) contains the bias value of 0, then the model bias is not significantly different from 0, and the model is performing well. If the interval does not contain the bias value of zero, the model may have some prediction problems, thereby warranting further investigation.

In standard parametric statistical theory, confidence intervals are constructed assuming the statistic of interest follows a known distribution. The assumed distribution is frequently a normal distribution. These distributional assumptions are valid for simple statistics like the mean and variance. Here, for bias, a mean is calculated, but it is not the usual sample mean. Averaging involves emission values predicted from the model, which could have a strange, underlying distributional form. Therefore, it is undesirable to assume that bias follows a normal distribution since its true form is unknown. Also, there is no obvious calculation to estimate the standard error of the bias. For these reasons, the method of choice is the bootstrap method to determine confidence intervals (Efron and Tibshirani 1993).

The bootstrap algorithm can now be described in detail in this context. The bootstrap sampling is conducted at each time point in the driving schedule with new sequencing of the bootstrapped samples. First, assume a sample of n paired observations drawn from the population of interest. The first value in each pair is the observed emissions value, and the second value is the corresponding predicted emissions value. To construct the first bootstrap sample, a sample pair is chosen at random from the original sample. Its values are recorded, and the selected pair is returned to the original sample. A second pair is chosen at random from the original sample, its values recorded, and is then returned to the original sample. This is the second pair of values in the first bootstrap sample. Pairs of values are chosen from the original sample until the first bootstrap sample contains n pairs and thus is the same size as the original sample. In this fashion, a random group of vehicles the same size as the actual group is created. The first value of the bias statistic can be calculated from these paired values.

The second bootstrap sample is calculated in a similar way to the first with a new randomization of pairs chosen with replacement until there are n pairs in the bootstrap sample. The second value of the bias statistic is then computed. This procedure is repeated until B bootstrap samples, each of size n, have been drawn, and B bias estimates have been calculated. B must be quite large in order to obtain reasonably accurate results. For the present study, B = 1,000 is used. Of the 1,000 bias estimates calculated, the 25th smallest bias estimate, or 2.5 percentile, is determined, as well as the 25th largest bias estimate, or 97.5 percentile. The difference between these two numbers is an approximate 95% bootstrap confidence interval on the bias.

While there are other bootstrap methods for establishing confidence intervals (Efron and Tibshirani 1993), the percentile method is preferred for the present study due to its simplicity and because the intervals can be asymmetric, unlike traditional confidence intervals. Concerns about potential accuracy and underprediction are offset in this study by the number of vehicles considered, 340, as well as the number or replications

of the procedure, 1,000. Consequently, for a given constituent emitted on a specific driving schedule, the 95% bootstrap confidence interval is calculated based on 1,000 replications for each second in time over the length of the driving schedule. The US06 driving schedule is about 589 seconds long, resulting in 589 intervals with different random sequences of vehicles. Formally speaking, these intervals are not to be used for strict statistical hypothesis testing. To do so could lead to overstated, erroneous conclusions. Informally speaking, the plots are quite useful for summarizing the available information in the data and for observing underlying patterns and trends through time. Plots of the length of the confidence intervals over time are used as a measure of variability of the bias statistic. Wider intervals indicate more variability. Narrower intervals indicate less variability.

In addition to the plots of bootstrap confidence intervals, called bias bootstrap plots, other potentially informative plots over time, such as plots of the shape statistic, can be constructed (Efron and Tibshirani 1993). The shape statistic is a measure of skewness, which numerically describes the shape of the distribution of the statistic of interest.

RESULTS

Due to the large differences in emissions and the possible differences in emissions behavior over driving modes, the normal-emitting (emissions less than 150% of the vehicle's certification standard) and high-emitting (emissions greater than or equal to 150% of the emissions standard) vehicles were analyzed separately. Second-by-second bias plots with bootstrap confidence limits were constructed for CO_2, CO, HC, and NO_x after calculation of model results. The US06 NO_x results are presented for CMEM v1.0 and CMEM v1.2. Differences in CMEM v1.0 and CMEM v1.2 are described below. The bias plots for CO_2, CO, and HC followed the same general pattern as those of NO_x but did not show large changes from CMEM v1.0 to CMEM v1.2 and are not presented here. NO_x results for CMEM v1.0 normal-emitting vehicles and high-emitting vehicles are shown in figures 2 and 3, respectively.

Figure 2 shows that the model overpredicts NO_x emissions to a small degree in normally operating vehicles during the high-speed cruise section of the US06 driving schedule. Figure 3 indicates that for the high-emitting vehicles there is no model over-

FIGURE 2 US06 Normal Emitter Second-by-Second Average Bias

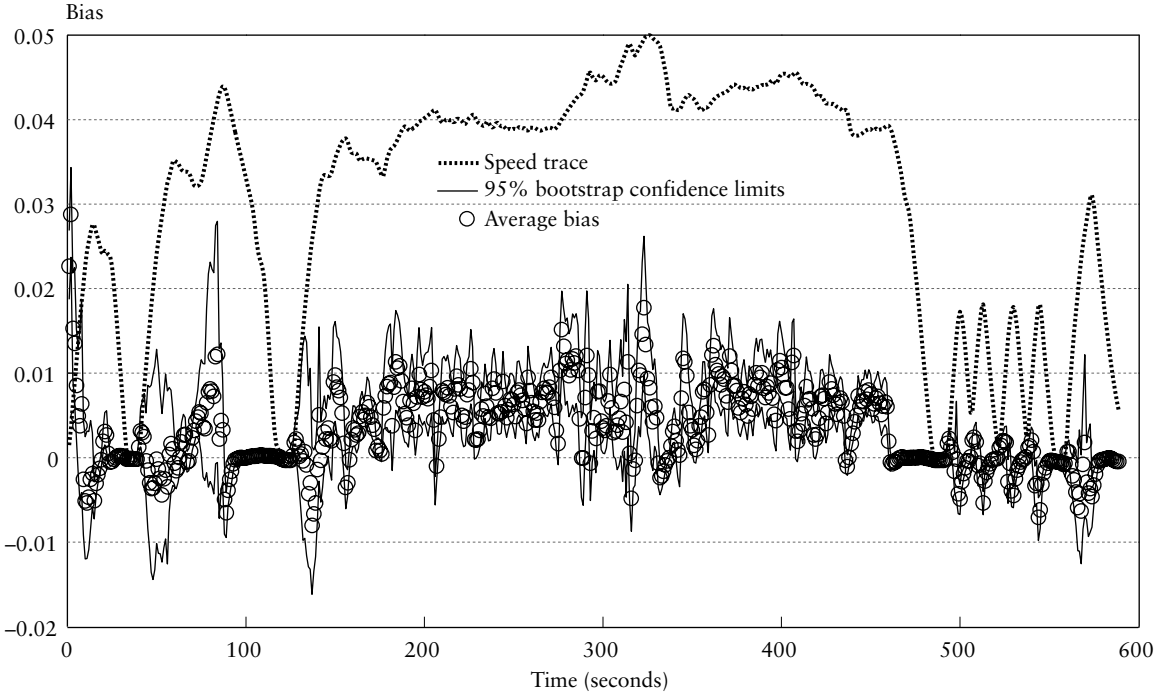
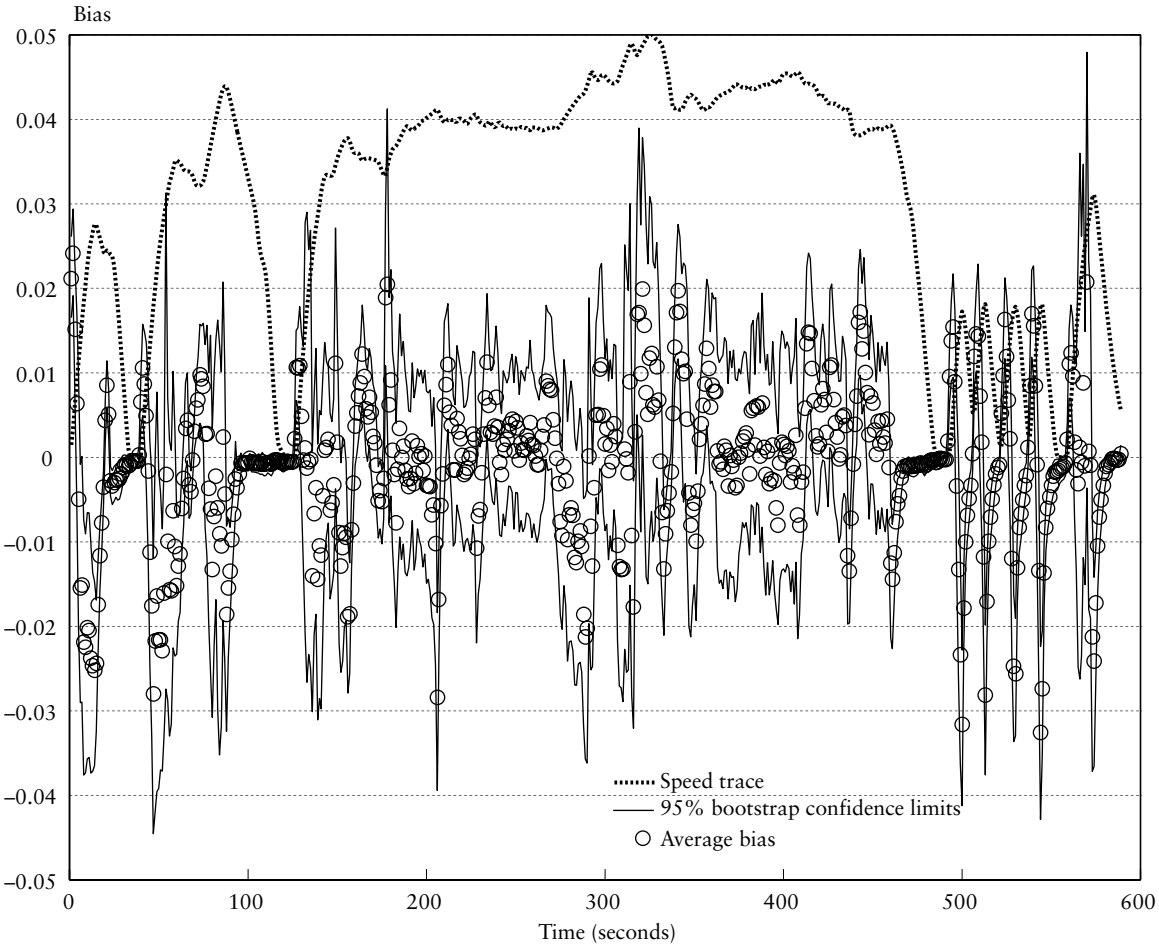


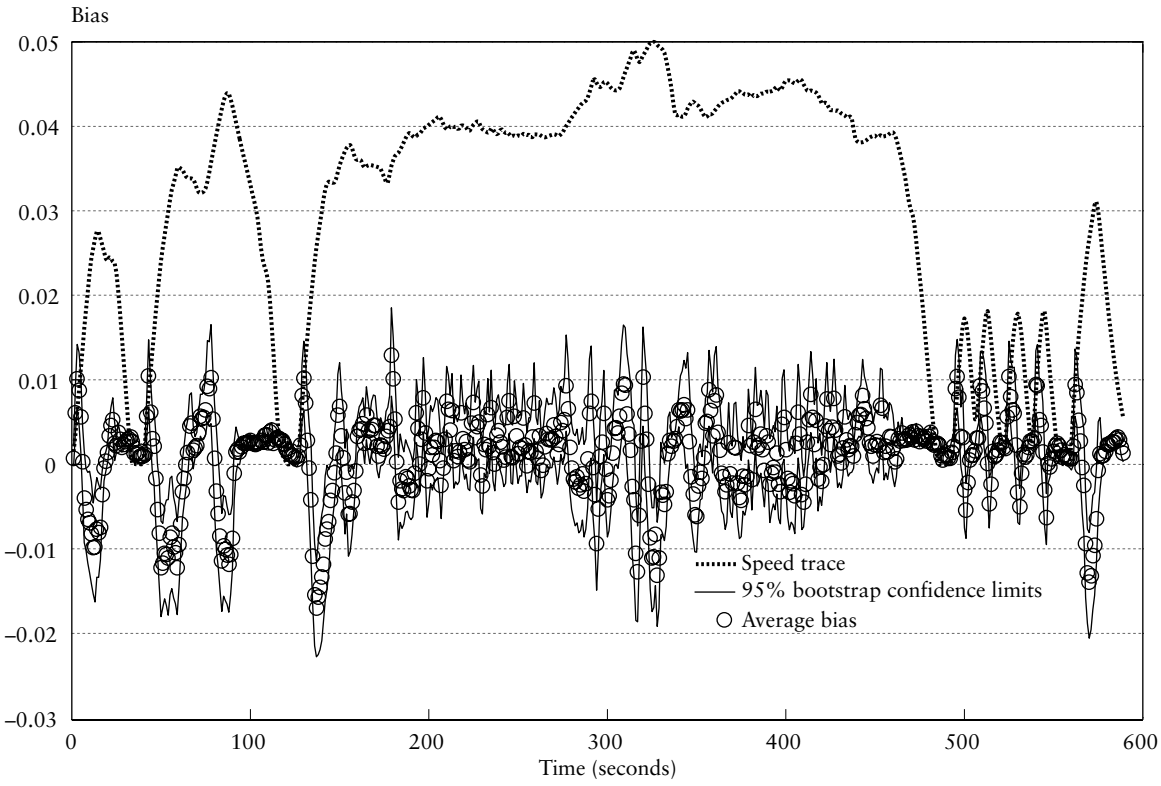
FIGURE 3 US06 High Emitter Second-by-Second Average Bias



prediction for the high-speed cruise section. Comparison of figure 2 and figure 3 also shows that bias is more variable for the high-emitting vehicles, apparent from the wider confidence limits and greater range of average bias values from second to second. This can be explained at least in part by the higher levels of emissions for the high-emitting vehicles and the higher variability in emissions of high-emitting vehicles. Additionally, both figures 2 and 3 suggest that the model overpredicts emissions at the start of an acceleration event and underpredicts them at the end of the acceleration event. Thus, the observed pattern in bias indicates that this version of the model may be inadequate for detailed second-by-second analysis while still appropriately capturing the intended range of emissions on the total driving trace and for driving modes. Driving modes are considered as individual events such as acceleration, deceleration, and steady-state cruising. For example, users of the model would be interested in the total emissions contribution of a vehicle accelerating onto the freeway and not in emissions at the start and end of the acceleration separately.

Due to the validation results discussed above, modifications were made to the NO_x components of the CMEM model, leading to the establishment of CMEM v1.2. NO_x emissions predictions for normal-emitting vehicles on the US06 using CMEM v1.2 are presented in figure 4. Similar results for the high-emitting vehicles are presented in figure 5. The bootstrap results show the resulting changes in the model bias. Note that the overprediction of NO_x in normal-operating vehicles at the high-speed portion of the US06 has been eliminated (figure 4). However, the deceleration events for which CMEM v1.0 exhibited no under- or overprediction now do exhibit overprediction of emissions, as seen in the positive values and narrow confidence bands around times 100 and 475.

FIGURE 4 US06 Normal Emitter Second-by-Second Average Bias



This indicates that CMEM v1.2 overpredicts NO_x on long deceleration modes for normal-operating vehicles. These changes, while not perfect, represent a substantial improvement in the model prediction accuracy for normal-operating vehicles because the high levels of NO_x in the high-speed portions are much more important than the low NO_x levels produced in the deceleration events.

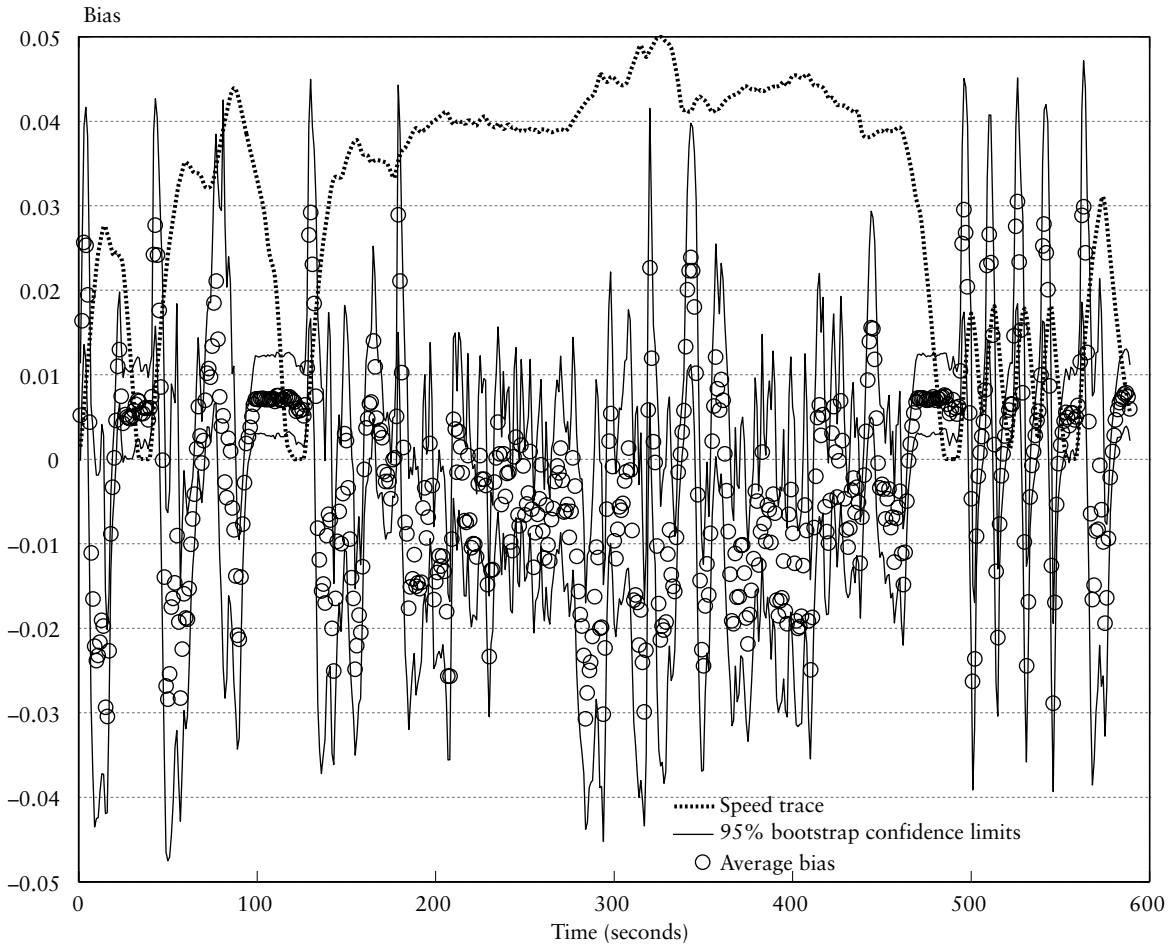
For the high-emitting vehicles, figure 5 suggests that the changes to the model have affected predictions of emissions at the high-speed portion of the schedule. The overprediction in the high-speed portions of the driving schedule is slightly lower for CMEM v1.2 than for CMEM v1.0 (figure 3 versus figure 5), with CMEM v1.2 tending towards underprediction of NO_x on the high-speed driving section. For CMEM v1.0, the confidence limits include zero indicating no under- or overprediction during the high-speed driving section, but for CMEM v1.2 some parts of the high-speed section do not include zero. The overprediction in NO_x for long deceleration events is also clearly visible on the high-emitting vehicles around times 100 and 475 (figure 5).

CONCLUSIONS

The bootstrap technique has been proven to be a useful method for graphically validating the predictions of CMEM on a second-by-second basis. This paper has also shown the bootstrap bias plot to be a useful tool for modelers during the model development process. It provides both detailed and summary information about the model's accuracy to facilitate model refinement. Using bootstrap bias plots, it can be determined if the model is predicting as well as desired, and if not, the bias plots identify where the bias is occurring in the driving schedule. Overall, the effects of model improvements can be observed directly in the plots, leading, in the particular case described, to the elimination of overprediction in NO_x under high-speed driving conditions for normal-operating vehicles. In the case presented here, the bias plots also identified unintended changes in model behavior resulting from the changes to the model.

The technique described here has been used on 340 vehicles split into 2 groups: normal emitters and high emitters. Differences in model bias were observed between the two groups. Further comparisons of these vehicles on the basis of the other

FIGURE 5 US06 High Emitter Second-by-Second Average Bias



classification criteria, such as carburetor versus fuel injection, could provide more valuable information for improving model bias. In addition, further research should be conducted to determine whether other statistics or other bootstrap methods of determining confidence intervals on emissions model predictions are more appropriate.

Finally, current efforts are focused on other ways to compare different versions of emissions models. Validation studies are targeting methods used to compare model results on the basis of an overall driving schedule in much the same way that the vehicles are expected to be used in practice, rather than on a second-by-second basis.

REFERENCES

An, F., M. Barth, M. Ross, and J. Norbeck. 1997. The Development of a Comprehensive Modal Emissions Model: Operating Under Hot-Stabilized Conditions. *Transportation Research Record* 1587:52–62.

Barth, M., F. An, J. Norbeck, and M. Ross. 1996. Modal Emissions Modeling: A Physical Approach. *Transportation Research Record* 1520:81–8.

Barth, M.J., T. Younglove, T. Wenzel, G. Scora, F. An, M. Ross, and J. Norbeck. 1997. Analysis of Modal Emissions from a Diverse In-Use Vehicle Fleet. *Transportation Research Record* 1587:73–84.

Bornstein, R.D. and S.F. Anderson. 1979. A Survey of Statistical Techniques Used in Validation Studies of Air Pollution Prediction Models. Technical Report 23, Stanford University. Stanford, CA.

Calvert, J.G., J.B. Heywood, R.F. Sawyer, and J.H. Seinfeld. 1993. Achieving Acceptable Air Quality: Some Reflections on Controlling Vehicle Emissions. *Science* 261:37–45.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society of Industrial Applied Mathematics CBMS–National Science Foundation Monograph 38.

Efron, B. and R. Tibshirani. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* 1, no. 1:54–77.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. New York, NY: Chapman & Hall.

Hanna, S.R. and D.W. Heinold. 1985. *Development and Application of a Simple Method for Evaluating Air Quality Models*. Publication 4409. Washington, DC: American Petroleum Institute.

Hanna, S.R. 1988. Air Quality Model Evaluation and Uncertainty. *Journal of the Air Pollution Control* Association 38:406–12.

_____. 1989. Confidence Limits for Air Quality Model Evaluations, as Estimated by Bootstrap and Jackknife Resampling Methods. *Atmospheric Environment* 23, no. 6:1385–9.

Schulz, D., T. Younglove, M. Barth, C. Levine, and G. Scora. 1999. Development of a Model Validation Procedure for Use in Evaluating Changes in a Complex Modal Emissions Model, in *Proceedings of the Ninth CRC On-Road Vehicle Emissions Workshop*. Atlanta, GA: Coordinating Research Council, Inc.

Zannetti, P. 1990. *Air Pollution Modeling: Theories, Computational Methods and Available Software*. Boston, MA: Computational Mechanics Publications.