# Analysis and Comparison of Microscopic Traffic Flow Models with Real Traffic Microscopic Data

Vincenzo Punzo and Fulvio Simonelli

**The evermore widespread use of microscopic traffic simulation in the analysis of road systems has refocused attention on submodels, including car-following models. The difficulties of microscopic-level simulation models in the accurate reproduction of real traffic phenomena stem not only from the complexity of calibration and validation operations but also from the structural inadequacies of the submodels themselves. Both of these drawbacks originate from the scant information available on real phenomena because of the difficulty with the gathering of accurate field data. In this study, the use of kinematic differential Global Positioning System instruments allowed the trajectories of four vehicles in a platoon to be accurately monitored under real traffic conditions on both urban and extraurban roads. Some of these data were used to analyze the behaviors of four microscopic traffic flow models that differed greatly in both approach and complexity. The effect of the choice of performance measures on the model calibration results was first investigated, and intervehicle spacing was shown to be the most reliable measure. Model calibrations showed results similar to those obtained in other studies that used test track data. Instead, validations resulted in higher deviations compared with those from previous studies (with peaks in cross validations between urban and extraurban experiments). This confirms the need for real traffic data. On comparison of the models, all models showed similar performances (i.e., similar deviations in validation). Surprisingly, however, the simplest model performed on average better than the others, but the most complex one was the most robust, never reaching particularly high deviations.**

The calibration and validation of models with field data are not only necessary for their correct use in simulating real systems but also allow them to be powerful tools for the investigation of real phenomena. Analysis of the performances of models whose modeling approaches differ significantly can give a much deeper insight into the behavior of a real system than the mere analysis of observed measurements can, and such analyses can substantially improve the models that are developed and make them more reliable. As the challenging task of assessing the potential effects of new technologies applied to road systems (e.g., intelligent transportation systems) has shown that the available simulation tools are generally inadequate, calibration and validation of microscopic traffic flow models are even more important.

In traffic flow modeling, many microscopic laws and models that attempt to capture longitudinal interactions among vehicles have been proposed [a comprehensive review is provided elsewhere (*1*)]. However, not as many studies have been carried out to calibrate and validate them, probably because of the difficulties and costs involved with both gathering and processing field data as well as setting up calibration studies. Moreover, the findings from such studies have often been contradictory because of the bias of the data as well as the use of inappropriate data collection schemes (*1, 2*).

Thus, as a result of the motivations described above, renewed efforts are being directed to the testing of models against real traffic data. A recent study compared different microscopic traffic flow models by using the travel times of single vehicles recorded at eight fixed locations on a rural road in the United States (*3*).

Improved accuracy in data collection allows the comparison of models from a microscopic perspective as well. For instance, in the field of instrumented vehicles, on the one hand, investment in driver assistance systems has made available accurate automotive distance sensors (*4, 5*); on the other hand, advances in Global Positioning System (GPS) technology allow vehicle positions to be recorded with an accuracy up to 1 cm (*6*). Thus, by using time series data for car-following variables gathered along United Kingdom motorways with an instrumented vehicle, Wu et al. validated a fuzzy logic-based model (*7*). The ability of the model to fit the experimental data was compared with those of other models. However, the latter had not previously been calibrated with field data, but parameter values from the literature were used.

Brockfeld et al. (*8*) and Ranjitkar et al. (*9*) used time series data recorded for nine vehicles that formed a single platoon on a test track in Japan (*6*). In the work of Brockfeld et al., 10 car-following models were calibrated by using distances as the error measurements and the models were validated with different data sets (*8*). None of the models appeared to be significantly better than any other. In the work of Ranjitkar et al., six models were calibrated by using instead both speeds and distances as error measurements (*9*). The results of the two models with the two error measurements chosen differed significantly, and the cause was attributed to data errors. Both studies showed that interpersonal variations were greater than intermodel variations in most cases, highlighting the influence of individual drivers on car-following behavior.

Recent studies have validated two models (the Newell and CELLSIM models, respectively) by using vehicle trajectory data sets (*10, 11*). The first of those studies used data on the motions of queued vehicles discharging into signalized intersections, while the second of those studies used data from a highway (*12*) and a five-lane freeway.

Despite the efforts mentioned above, several issues related to microscopic traffic flow model calibration still need to be investigated. The first regards data accuracy: the way in which they influence model calibration results needs to be investigated; and, consequently, the

Department of Transportation Engineering, University of Napoli "Federico II," Via Claudio, 21, 80125 Naples, Italy.

*Transportation Research Record: Journal of the Transportation Research Board, No. 1934*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 53–63.

minimal requirements of data for such studies have not been sufficiently stressed in the literature. The choice of the most suitable field data collection scheme or at least the best compromise between data accuracy and realism is another problem whose solution has not yet been consolidated. Moreover, methodological issues concern the choice of appropriate measures of performance and calibration methods as well as of suitable validation protocols.

In this context, this paper sets out not only to investigate methodological issues of model calibration and validation but also to produce preliminary results from a comparison models with real traffic microscopic data. Initial problems regarding the collection of field data are extensively discussed in another work (*13*), which presents a detailed description of the estimation process and the data accuracy used here.

The next section describes the models tested. A brief description of the data collection and the estimation process is then provided. A discussion of the methodology adopted throughout this study and issues of model calibration and validation follows. Finally, analysis and comparison of model performances complete the paper.

## MODELS TESTED

The study focused on the analysis of four well-known models, which were chosen because of the different approaches that they use, as well as the different standards of complexity and the different numbers of parameters that need to be calibrated.

The following models were analyzed and are ordered as follows by increasing numbers of parameters (and complexity):

1. A trajectory translation model, Newell's model (*14*);
2. A safety distance model, Gipps model (*15*);
3. A continuous response model, the intelligent driver model (IDM) (*16*); and
4. A stimulus–response model, MITSIM (*17, 18*).

The first model is the so-called low-order model by Newell. It was used as a reference because of its simplicity and the minimum number of parameters (only two) to be calibrated. It simply states that if a driver is impeded from traveling at his or her desired speed, the driver follows the leader at the same speed and observes a desired intervehicle spacing, *d*, that varies linearly with the speed. If the leader changes speed, the driver ($n$) following imitates that speed after a time interval, $\tau_n$, that is required to reach the new desired spacing. This is the same as saying that the follower's trajectory is simply translated by a time $\tau$ and a space $d$ with respect to those of the leader. In mathematical terms,

$$x_n(t + \tau_n) = x_{n-1}(t) - d_n$$

where $x_n(t+\tau_n)$ and $x_{n-1}(t)$ represent the positions of the follower and the leader, respectively. The model is able to describe the behaviors only of following vehicles, so that a leader trajectory must be provided (i.e., boundary conditions must be defined).

The Gipps model is a safety-based model. It provides two different transfer functions according to the two different driving regimes assumed. In the free-flow regime, the speed planned for the following instant is obtained from an inequality of the experimental origin that joins two conditions: (*a*) the speed does not exceed the desired speed of the driver, and (*b*) free acceleration decreases on an increase in the speed until it becomes null once the desired speed has been reached. In the car-following regime, instead, the driver adopts such a speed as to safely stop the vehicle in case of sudden braking of the leading vehicle. Then, the speed planned by follower $n$ ($v_n$) is

$$v_n(t + \tau) = \min[v_{a,n}(t + \tau), v_{b,n}(t + \tau)]$$

where

$$v_{a,n}(t + \tau) = v_n(t) + 2.5 \cdot a_n \cdot \tau \cdot \left[1 - \frac{v_n(t)}{V_n}\right] \cdot \sqrt{0.025 + \frac{v_n(t)}{V_n}}$$

$$v_{b,n}(t + \tau) = b_n\tau + \sqrt{b_n^2\tau^2 - b_n \cdot 2[x_{n-1}(t) - s_{n-1} - x_n(t)] - v_n(t) \cdot \tau - \frac{v_{n-1}^2(t)}{\hat{b}_{n-1}}}$$

$x$ = vehicle position;
$\tau$ = follower's reaction time (that is, the simulation step);
$a_n$ = follower's maximum desired acceleration;
$V_n$ = follower's desired speed;
$b_n$ = follower's maximum deceleration;
$s_{n-1}$ = effective size of the leader (that is, physical length plus a margin into which the following vehicle is not willing to intrude); and
$\hat{b}_{n-1}$ = follower's estimate of maximum deceleration that the leader intends to adopt.

IDM can be seen to be a continuous response model. The model does not consider a reaction time, which is the same as saying that its expression is an ordinary differential equation. It assumes that the follower's acceleration is a continuous function of the follower's speed ($v_n$), the spacing ($d_n$) from the leader, and the speed difference from the leading vehicle ($\Delta v_n$). In particular, the follower's acceleration ($\dot{v}_n$) is calculated as

$$\dot{v}_n = a^{(n)}\left\{1 - \left(\frac{v_n}{V_0^{(n)}}\right)^{\delta} - \left[\frac{d(v_n, \Delta v_n)}{d_n}\right]^2\right\}$$

which mediates the tendency to accelerate to reaching the desired speed, $V_0^{(n)}$, and the tendency to decelerate when the spacing is less than the desired spacing, $d(v_n, \Delta v_n)$. The expression of the desired spacing as a function of $v$ and $\Delta v$ is

$$d(v, \Delta v) = d_0^{(n)} + d_1^{(n)}\sqrt{\frac{v}{V_0^{(n)}}} + T^{(n)}v + \frac{v\Delta v}{2\sqrt{a^{(n)}b^{(n)}}}$$

where

$d_0^{(n)}$ = desired spacing with zero speed of driver $n$,
$V_0^{(n)}$ = desired speed, and
$a^{(n)}$, $\delta$, $d_1^{(n)}$, $T^{(n)}$, and $b^{(n)}$ = nonphysical parameters of driver $n$.

The acceleration model of Ahmed (*17*), implemented in MITSIMLab, provides two main driving regimes. The first is valid for car-

following conditions and is a classic model of the type response = sensitivity × stimulus, derived from the work by Gazis et al. (*19*). The major differences of the acceleration model from the model of Gazis et al. are as follows: (*a*) the sensitivity term changes for acceleration or deceleration maneuvers; (*b*) parameters vary among drivers; (*c*) a term that takes into account the density of the segment (i.e., that accounts for a look-ahead capability) is considered; and (*d*) a stochastic residual term is added to acceleration–deceleration. Points *c* and *d* of the model are not considered here because of the lack of density measurements and for the sake of simplicity (with regard to taking into account the stochastic term). Thus, the mathematical form of the car-following model tested here was

$$a_n^{cf,i}(t) = \alpha^i \frac{V_n(t-\tau_n)^{\beta^i}}{\Delta X_n(t-\tau_n)^{\gamma^i}} |\Delta V_n(t-\tau_n)|^{\lambda^i} \quad (1)$$

where

$a_n^{cf,i}(t)$ = follower's acceleration–deceleration in a car-following situation,
$V_n(t-\tau_n)$ = follower's speed at time $(t-\tau_n)$,
$\tau_n$ = follower's reaction time,
$\Delta X_n(t-\tau_n)$ = distance from bumper to bumper between the leader and the follower at time $(t-\tau_n)$,
$\Delta V_n(t-\tau_n)$ = speed difference between the leader and the follower at time $(t-\tau_n)$,
$i$ = acceleration or deceleration based on the value of $\Delta V_n(t-\tau_n)$ [if $\Delta V_n(t-\tau_n)$ is greater than 0, there is acceleration; otherwise, there is deceleration], and
$\alpha^i$, $\beta^i$, $\gamma^i$, and $\lambda^i$ = eight parameters to be calibrated for the car-following regime (with *i* values for acceleration and deceleration).

The second model is a free-flow regime (in which the vehicle accelerates at a normal acceleration rate to attain its desired speed). An emergency regime for the avoidance of vehicle collision is provided as well. A change from the car-following regime to the free-flow regime is conditioned by thresholds on time headway between the leader and the follower (*h_upper* and *h_lower*, respectively, in Table 1).

**TABLE 1 Average Values of Parameters After Calibrations**

| | | Mean | VAR | Cv |
|---|---|---|---|---|
| Newell | $\tau_n$ | 1.027 | 0.074 | 0.264 |
| | $d_n$ | 0.370 | 0.228 | 0.109 |
| Gipps | $a_n$ | 3.331 | 4.189 | 0.614 |
| | $b_n$ | –3.801 | 5.949 | 0.642 |
| | $\hat{b}_{n-1}$ | –4.783 | 10.613 | 0.681 |
| | $V_n$ | 16.152 | 12.280 | 0.217 |
| | $\tau_n$ | 0.567 | 0.024 | 0.272 |
| IDM | $a$ | 2.568 | 0.619 | 0.306 |
| | $b$ | 1.694 | 0.493 | 0.415 |
| | $V_0$ | 28.362 | 203.987 | 0.504 |
| | $T$ | 0.690 | 0.046 | 0.312 |
| | $\delta$ | 2.836 | 3.499 | 0.660 |
| | $d_0$ | 0.743 | 0.130 | 0.484 |
| | $d_1$ | 0.557 | 1.637 | 2.299 |
| MITSIM | $\alpha$_acc | 2.512 | 1.563 | 0.498 |
| | $\beta$_acc | 0.150 | 0.099 | 2.102 |
| | $\gamma$_acc | 0.509 | 0.324 | 1.120 |
| | $\lambda$_acc | 1.073 | 0.539 | 0.684 |
| | $\alpha$_dec | –2.328 | 2.545 | 0.685 |
| | $\beta$_dec | 0.861 | 0.485 | 0.809 |
| | $\gamma$_dec | 1.116 | 0.389 | 0.559 |
| | $\lambda$_dec | 1.293 | 0.338 | 0.449 |
| | *h*_upper | 2.044 | 0.285 | 0.261 |
| | *h*_lower | 0.289 | 0.014 | 0.404 |
| | $\tau_n$ | 0.580 | 0.093 | 0.526 |

VAR = variance; Cv = covariance.

## EXPERIMENTAL DATA

More details on the data used here (for which only brief information is provided here) and a comprehensive discussion about the issues regarding experimental data collection are provided elsewhere (*13*).

### Experimental Setup and Data Collection

The data used in this study were obtained from a series of experiments carried out along roads in areas surrounding Naples, Italy, under real traffic conditions between October 2002 and July 2003. Experiments were performed by driving four vehicles in a platoon along urban and extraurban roads under different traffic conditions. All vehicles were equipped with kinematic differential GPS receivers [dual-frequency, GPS + Global Navigation Satellite System (GLONASS) receivers] that recorded the position of each vehicle at 0.1-s intervals.

As the aim of the experimental surveys was to collect data for comparison of car-following models, careful attention was devoted to the choice of routes and the roads to be used for data collection in this study. Roads with only one lane per direction were considered so that car-following behavior was unaffected by other behaviors, like lane changing. Moreover, roads along a route had to differ by type, level of congestion, etc., to capture after a while the behaviors of the same drivers (i.e., drivers in the same psychophysical condition) coping with different environments.

The trajectory data used here were collected along the same route on 2 days: October 30, 2002, and February 25, 2003. In all, five trajectory data sets were extracted from data collected along the route: three data sets from the October trial and two sets from the February trial. These are named 30A, 30B, 30C, 25B, and 25C, respectively. Sets 30A and 30C are for one-lane urban roads and are 3.3 and 6 min long, respectively, while Set 30B is for a two-lane extraurban highway that bypasses the historical center of Pozzuoli (a town near Naples) and is 4.2 min long. Sets from February 25 were gathered on the same urban roads used for data collection in October and are 5.3 and 5 min long, respectively.

Careful attention was paid to the setup of the experimental protocol. The leader of the platoon was one of the authors. The following drivers were informed of the path to be taken and were familiar with it, but they were unaware of the aim of the experiment. The leader took care to prevent intrusions into the platoon by giving way to extraneous vehicles at intersections. When intrusions occurred, the corresponding data were discarded. The number of instrumented vehicles in the platoon was limited to four because of the experimental difficulties mentioned above as well as because of budget limitations.

Figure 1 shows as an example the speeds of the platoon leaders from Experiments 30A, 30B, and 30C. The average speed for the extraurban data set (Set 30B) was 47 km/h, while that for the urban data sets, Sets 30A and 30C, were 26 and 28 km/h, respectively. The driving patterns were also qualitatively different, with the urban patterns showing repeated vehicle stops but with the other patterns being more regular.

### Data Estimation

GPS technology is known to allow the positions of the receivers to be estimated in a common space–time reference system. From these positional data, time series of intervehicle spacings is immediately available, whereas vehicle speeds and accelerations must be calculated through successive derivations of the space traveled. Despite the expected precisions of differential GPS positional measurements (approximately 10 mm), the data needed to be filtered, in view of the high levels of measurement noise because of the urban environment on the one hand (e.g., the multipath effect) and because of the stringent requirements of data for car-following studies on the other.

The core of the problem was to filter noisy trajectory data for each vehicle without altering the platoon data consistency; i.e., the speeds and the accelerations of the following vehicles had to be estimated so that the intervehicle spacings calculated from them were equal to the real ones. Otherwise, for example, even slight differences between the estimated and the actual speeds of a vehicle could easily have entailed negative spacings in case of a stop. This was accomplished by use of a nonstationary Kalman filter, which considers following vehicles as a sole dynamic system and which allows one consistent estimation problem instead of several independent (and inconsistent) ones to be solved.

Thus, accurate and consistent time series data for intervehicle spacings, speeds, and accelerations were obtained. Identification and adoption of intervehicle spacings as reference measurements [an explanation is provided elsewhere (*13*)] also allowed quantitative evaluation of the estimation accuracy. The values of the estimation errors for the trajectories of the five data sets used here are provided elsewhere (*13*).

## METHODOLOGY

As mentioned above, the aim of the study was to investigate methodological issues concerning comparison of microscopic traffic flow models as well as comparison of some different well-known models. Thus, the first step was the calibration of the models. The calibration results gave a measure of the models' ability to fit the experimental data but did not necessarily represent the ability of the models to reproduce real phenomena, i.e., to capture real system dynamics. Thus, validations were performed to address this issue.

All the calibrations and validations of the models were carried out for one driver at a time. In particular, the models simulated the trajectory of each vehicle being fed the experimental trajectory of its leader.

### Calibration

#### *Problem Formulation and Solution*

Calibration of the model of a real system by indirect techniques (i.e., by techniques based on the use of the model itself to estimate its parameters) starts by comparison of the model outputs with those of the real system fed the same inputs. It is equivalent to the solution of a constrained minimization problem in which the objective function expresses the deviation of the simulated output measurements from those observed. Among the estimators commonly used there is the generalized least-squares estimator, according to which the problem formulation is set as follows:

$$\min_{a} \gamma = (\boldsymbol{Y}^{\text{obs}} - \boldsymbol{Y}^{\text{sim}})^T \boldsymbol{P}^{-1} (\boldsymbol{Y}^{\text{obs}} - \boldsymbol{Y}^{\text{sim}})$$

$$\boldsymbol{Y}^{\text{sim}} = S(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{a})$$

$$g_i(\boldsymbol{x}, \boldsymbol{a}) \geq 0 \qquad i = 1, \ldots, n_d$$

$$h_j(\boldsymbol{x}, \boldsymbol{a}) = 0 \qquad j = 1, \ldots, n_e$$

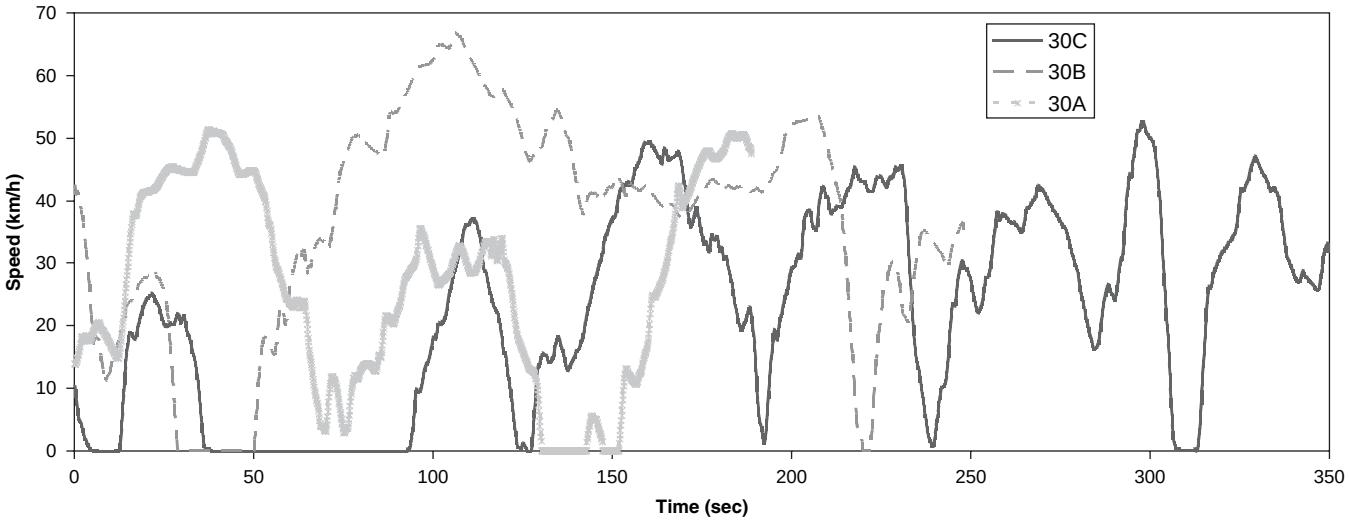


FIGURE 1 Speed profiles of platoon leaders from Experiments 30A (urban), 30B (extraurban), and 30C (urban).

where

$\gamma$ = objective function of the optimization problem, which measures the overall performance of the model;
$\mathbf{Y}^{obs}$ and $\mathbf{Y}^{sim}$ = vectors of observed and simulated measures of performances (MOPs) obtained from the outputs of the model $S$, respectively;
$\mathbf{P}$, $\mathbf{u}$, $\mathbf{x}$, and $\mathbf{a}$ = vectors of weights, inputs, state variables, and parameters, respectively;
$g_i$ = $i$th inequality constraint;
$h_j$ = $j$th inequality constraint; and
$n_e$ and $n_d$ = numbers of equality and inequality constraints, respectively.

When $S$ is a simulation model, to calculate the value of the objective function at every step of the algorithm searching for the minimum, one or more simulations are performed whether or not the model is stochastic. In this work, the optimization software LINDO API (*20*) was used to solve the minimization problem presented above. The software uses a multipoint nonlinear optimization algorithm, which starts by searching for the minimum from different points to circumvent local minima.

### Choice of Performance Measures

A fundamental aspect of the problem is the choice of more adequate MOPs to represent system and model output measurements. If the model were capable of reproducing the dynamics of the real phenomenon exactly and, thus, all the system output measurements coincided with the model output measurements, the objective functions obtainable with any MOP would be equally null in terms of their global minima and the corresponding sets of optimum parameters would coincide. As models are more or less accurate approximations of reality, the choice of the functional form of the objective function, as well as the choice of MOPs, influences the results of calibrations. The form of response surfaces of a model obtained by calibrating the model vis-à-vis different MOPs, for example, may prove different, as the same may happen with the minima of objective functions and the corresponding sets of optimum parameters.

In the case of calibration of car-following models, the MOPs used must capture the dynamics of the phenomenon as it develops. These are derived directly from disaggregated traffic surveys and consist of time series of vehicle speeds or of their intervehicle spacings or time headways. Once simulations have been performed, to measure overall model performance and to check whether the simulated measurements really match the observed ones, error tests are usually adopted. In fact, most of the common statistical tests cannot be used in this case, as the measurements concerned are not stationary and self-correlated. Also, as mentioned above, the inputs used for calibration consist of the trajectory of the leader.

Error tests of common use are the root mean square error (RMSe), the root mean square percentage error (RMSPe), or Theil's inequality coefficient ($U$):

$$\text{RMSe} = \sqrt{\frac{1}{N}\sum_i \left(Y_i^{obs} - Y_i^{sim}\right)^2}$$

$$\text{RMSPe} = \sqrt{\frac{1}{N}\sum_i \left(\frac{Y_i^{obs} - Y_i^{sim}}{Y_i^{obs}}\right)^2}$$

$$U = \frac{\sqrt{\frac{1}{N}\sum_i \left(Y_i^{obs} - Y_i^{sim}\right)^2}}{\sqrt{\frac{1}{N}\sum_i \left(Y_i^{obs}\right)^2} + \sqrt{\frac{1}{N}\sum_i \left(Y_i^{sim}\right)^2}}$$

The square of RMSe can be decomposed into the following terms (*21*):

$$U^M = \frac{(\mu_{sim} - \mu_{obs})^2}{(1/N)\sum_i \left(Y_i^{obs} - Y_i^{sim}\right)^2}$$

$$U^S = \frac{(\sigma_{sim} - \sigma_{obs})^2}{(1/N)\sum_i \left(Y_i^{obs} - Y_i^{sim}\right)^2}$$

$$U^C = \frac{2(1-\rho)\sigma_{sim}\sigma_{obs}}{(1/N)\sum_i \left(Y_i^{obs} - Y_i^{sim}\right)^2}$$

where

$U^M$, $U^S$, and $U^C$ = bias, variance, and covariance proportions of $U$, respectively (these are useful as a means of understanding the sources of the simulation error);
$\mu_{sim}$ and $\mu_{obs}$ = means of simulated and observed values, respectively;
$\sigma_{sim}$ and $\sigma_{obs}$ = standard deviations of simulated and observed values, respectively;
$\rho$ = correlation coefficient; and
$Y_i^{obs}$ and $Y_i^{sim}$ = $i$th observed and simulated variables, respectively, with $i$ ranging from 1 to $N$.

As pointed out above, in the case of calibration of car-following models, the choice of MOP in the objective function is expected to condition the results. First, the choice of time headway as the MOP, especially with nonlinear objective functions, may provide nonoptimal results. Indeed, as higher values of time headways are obtained as speeds become closer to zero, observations that fall in this range of speeds might have an excessive weight in the calibration of the model, which is especially the case for urban data sets, in which low speeds are more frequent. In this case a simple remedy may be to eliminate observations concerning speeds close to zero when the model is calibrated (a threshold of 1 m/s has been adopted here).

Some further considerations may therefore arise from the observation of Figure 2, in which the results of calibration of the three models carried out with three different objective functions (i.e., different MOPs) are reported for data from one experiment. Mean errors and RMSPe values are given. Once the models have been calibrated for experimental time headways, the error statistics for speeds and intervehicle spacings are also calculated. The same was done when the models were calibrated on the basis of speeds and intervehicle spacings. Thus, each graph represents the values of error statistics for the three MOPs, obtained by calibrating the models three times: on the basis of time headways, speeds, and intervehicle spacings. Thus, for example, the "headway ObjF" label reports the values of the test errors of the three MOPs calculated once the model has been calibrated on the basis of time headways.

An initial consideration is the fact that all the models are better at reproducing speeds than at reproducing spacings or headways. The deviations between the simulated and the observed speeds, which

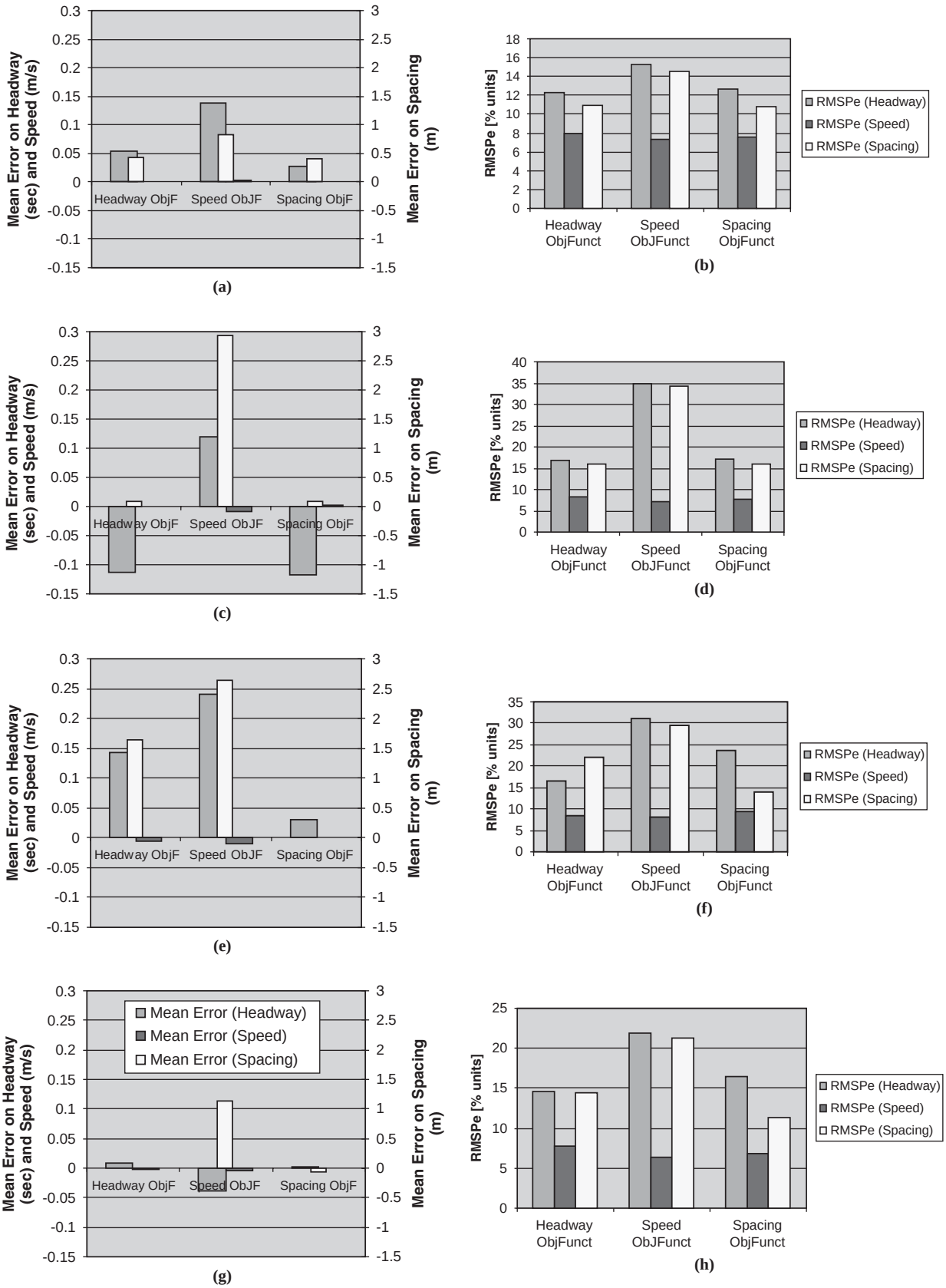


FIGURE 2 Mean errors and RMSPe values of the MOPs (headway, speed, and spacing) for each kind of calibration [calibrations differ for the choice of MOPs in the objective function (ObjF or ObjFunct)] (Experiment 30C): (*a–b*), IDM, (*c–d*) Gipps, (*e–f*) Newell, and (*g–h*) MITSIM.

were always lower than 10%, were always lower than the deviations obtained for the other two MOPs. This result was already found by Ranjitkar et al., who justified it by stating that it could be ascribed to data errors (*9*). Instead, the explanation is that speed deviations and spacing deviations from the observed data do not have the same meaning. For example, when a model is calibrated on the basis of speeds, an error made by the model in calculating the speed between instants $t_{k-1}$ and $t_k$ entails an error in the space traveled in the same interval (i.e., an error in the spacing from the leader). The latter, however, is kept equal for all the following instants; i.e., in all the following instants the result for the space traveled will be increased or decreased by this amount of error. Therefore, it is easier to fit models on the basis of speed measurements than on the basis of spacing measurements, but this definitely does not imply a better reproduction of real dynamics.

Another aspect of this difference is that by calibration of a model on the basis of speeds, the values of the error tests calculated for the other two measures are, sensibly, higher than the optimum ones. In other words, they are higher than the values obtained by calibrating the model directly on the basis of headways and spacings. For example, on the right side of Figure 2 it is shown that the Gipps model calibrated on the basis of speeds presents errors for headways and spacings equal to 35% and 34%, respectively, while when it is directly calibrated, headways and spacings present values of 17% and 16%, respectively. The mean errors on the left side of Figure 2, which provide information on the bias of the models, again show that calibration of the models on the basis of speeds implies nonnegligible errors for headways and spacings.

This proves that intervehicle spacing is the most reliable measure of performance for the calibration of car-following models. As a consequence, all subsequent calibrations were performed by using intervehicle spacing as the MOP.

### Validation

Unlike calibrations, validations consist of a simple simulation in which the model seeks to reproduce a trajectory from Data Set X by using parameters calibrated on the basis of another data set, Data Set Y.

As mentioned above, data for the data sets used in this study were collected on 2 different days. For each day, different data sets are nothing but the trajectories of the same drivers traveling in the same order along different stretches of roads belonging to the route covered. It is straightforward to verify whether the models are able to reproduce the behavior of the same driver along different parts of the route with the parameters calibrated on the basis of data for another part of it. This is an interesting point, because the roads from which the data for the different data sets along the route were extracted differ in their types and levels of congestion.

Hence, cross validations were accomplished by comparing for one driver at a time the observed trajectory from Data Set X with that simulated by using the parameters calibrated for the same driver in a different data set, Data Set Y, and vice versa. The error tests used for evaluation of the performances were the same as those used for the calibrations.

### Calibration and Validation Setup

As the trajectory of the preceding vehicle was unknown for the leader of the platoon, calibrations and validations could have been accomplished only for the three following drivers, which are referred to as Driver 2, Driver 3, and Driver 4.

First, model parameters were calibrated for each observed trajectory (i.e., 3 drivers × 5 data sets = 15 calibrations per model). Then, for each driver six cross validations were carried out between the three data sets from October 30 and two cross validations were carried out between the two data sets for February 25 (i.e., 3 drivers × 8 pairs = 24 validations per model). There were thus eight cross validations (per driver per model), referred to as 30AB, 30AC, 30BA, 30BC, 30CA, 30CB, 25BC, and 25CB, where 30AB, for example, means that each driver in Experiment 30A was simulated by using the optimal parameters for the driver calibrated from Data Set 30B.

All the models were simulated by adopting a simulation step of 0.1 s, consistent with the available field data. In other words, for models whose simulation step was a parameter to be calibrated, when this proved to be greater than 0.1 s, the values of the output variables at every 0.1 s were also calculated.

## CALIBRATION AND VALIDATION RESULTS

### Calibration Results

A few interesting remarks can be made on the basis of the results from the calibrations (Table 2 and Figure 3). It is surprising that the RMSPe values resulting from the experiments carried out under real traffic conditions in this study are mainly consistent with those found in the literature for experiments conducted on test tracks (*8, 9*).

In the calibration phase, MITSIM is capable of reproducing the experimental data better than the other models are. Indeed, the average error is about 12% for MITSIM, whereas the average errors are about 16% for the IDM model and approximately 17% for the models of Newell and Gipps. The values of the statistical indexes $U^M$, $U^S$, and $U^C$, which provide information on the nature of errors, are close to the optimal configuration ($U^M = 0$, $U^S = 0$, $U^C = 1$) for all the models except Newell's, which, even though it does not introduce systematic errors ($U^M$ is always close to 0), does not seem to reproduce correctly the fluctuations of the experimental data, i.e., measured spacings ($U^S$ is about 0.17). This result was actually expected because of the simplicity of Newell's model, in which the spacing varies linearly with speed and oscillations in the distance-keeping behavior with the leading vehicle are not allowed (in fact, the error $U^S$ presented in Table 2 is almost always due to an underestimation of the variance of the real data).

The worst values of RMSPe in the calibration phase were attained with all models for Driver 25-3: in Experiment 25B for the Newell model (22.45%), the Gipps model (23.02%), and IDM (23.75%) and in Experiment 25C for MITSIM (19.09%). It is therefore reasonable to suppose that the behavior of Driver 25-3 is not easily reproducible by models. The fact that there are drivers whose behaviors are more easily reproduced is confirmed by the trend of errors of the models. Indeed, the calibration results among the different models for the same driver did not generally differ significantly, while the performances of models dealing with different drivers vary consistently (Figure 3 and Table 2), as highlighted in previous work as well (*8, 9*). It can thus be argued that traditional models fail to capture some aspects of driving behavior.

In Table 1, the average values of the mean, the variance, and the covariance of the calibrated model parameters are reported. It is worth noting that the exponents $\lambda^i$ (with $i$ values for acceleration and deceleration) of the $\Delta V_n$ term in Equation 1 are nearly equal to 1, as found

**TABLE 2 Calibration and Validation Results**

| | | RMSPe | $U^M$ | $U^S$ | $U^C$ |
|---|---|---|---|---|---|
| Calibration | | | | | |
| Newell | Mean | 16.9% | 0.054 | 0.173 | 0.773 |
| | Max | 22.5% | 0.148 | 0.424 | 0.958 |
| | Min | 11.3% | 0.000 | 0.003 | 0.427 |
| | Amplitude (max–min) | 11.1% | 0.148 | 0.421 | 0.531 |
| Gipps | Mean | 17.2% | 0.038 | 0.052 | 0.910 |
| | Max | 23.0% | 0.146 | 0.241 | 0.988 |
| | Min | 12.2% | 0.000 | 0.000 | 0.680 |
| | Amplitude (max–min) | 10.8% | 0.145 | 0.241 | 0.308 |
| IDM | Mean | 15.6% | 0.040 | 0.066 | 0.894 |
| | Max | 23.8% | 0.122 | 0.188 | 0.989 |
| | Min | 10.6% | 0.000 | 0.001 | 0.729 |
| | Amplitude (max–min) | 13.1% | 0.122 | 0.187 | 0.259 |
| MITSIM | Mean | 12.4% | 0.025 | 0.052 | 0.923 |
| | Max | 19.1% | 0.105 | 0.124 | 0.980 |
| | Min | 7.3% | 0.000 | 0.002 | 0.795 |
| | Amplitude (max–min) | 11.8% | 0.105 | 0.122 | 0.184 |
| Validation | | | | | |
| Newell | Mean | 22.5% | 0.149 | 0.172 | 0.678 |
| | Max | 41.4% | 0.444 | 0.369 | 0.958 |
| | Min | 13.6% | 0.010 | 0.003 | 0.454 |
| | Amplitude (max–min) | 27.7% | 0.435 | 0.366 | 0.504 |
| Gipps | Mean | 24.2% | 0.130 | 0.094 | 0.776 |
| | Max | 45.4% | 0.343 | 0.256 | 0.975 |
| | Min | 17.1% | 0.006 | 0.004 | 0.434 |
| | Amplitude (max–min) | 28.3% | 0.338 | 0.252 | 0.541 |
| IDM | Mean | 23.5% | 0.387 | 0.124 | 0.490 |
| | Max | 44.0% | 0.659 | 0.389 | 0.951 |
| | Min | 13.8% | 0.003 | 0.002 | 0.187 |
| | Amplitude (max–min) | 30.2% | 0.656 | 0.387 | 0.765 |
| MITSIM | Mean | 22.9% | 0.210 | 0.097 | 0.692 |
| | Max | 29.1% | 0.696 | 0.259 | 0.979 |
| | Min | 15.5% | 0.000 | 0.001 | 0.283 |
| | Amplitude (max–min) | 13.6% | 0.696 | 0.257 | 0.695 |

Max = maximum; Min = minimum.

in previous studies (*1*). Another remark concerns the Gipps model, which has a low average reaction time, confirming the stringent car-following behavior from the available experimental data.

## Validation Results

In the validation phase, the different models were essentially equivalent, on average, with all models giving RMSPe values between 22.5% and 24.2% (Table 2 and Figure 4). The Newell model exhibited the smallest increment of error between the calibration and the validation phases (5.6%), while MITSIM had the highest increment (about 10.6%, on average) (Table 2). This could be confirmation of a tendency for MITSIM to overfit the experimental data. The better performance of MITSIM in the calibration phase may well be due to the larger number of parameters and, therefore, the larger number of degrees of freedom compared with those in the other models. Except for the two cases discussed below, MITSIM almost always had worse RMSPe values than the other models. Nevertheless, it showed the most robust behavior, as the validation results never reach particularly high values.

By looking at the validations for each driver, all models had similar responses except for those for Driver 30-4 for IDM and the Newell model (Validations 30BA and 30CA) and except for all drivers for Validation 30BC for the Gipps model.

In the first case, it appears that the two models overfit the calibration data for Driver 30-4 in Experiment 30A. Indeed, the parameters calibrated for Experiment 30A differed greatly from the optimum ones obtained for Experiments 30B and 30C. This can be interpreted as anomalous behavior for that driver.

The second major case regards the most extreme validation. Indeed, it is the validation between the urban and the extraurban trials (Validation 30BC). In this case, the Gipps model failed to simulate correctly the extraurban (Experiment 30B) trajectories of all the vehicles with parameters calibrated on the basis of the urban data (Experiment 30C). If one looks at the optimum parameters from the two calibrations (Experiments 30B and 30C), it can be noted that the values of maximum deceleration of the drivers varied by approximately 600% between the two trials. Unlike the case of Driver 30-4, in which anomalous behavior of the driver occurred, here the Gipps model seems to be sensitive to the calibration context (urban versus extraurban), failing to reproduce the results for all drivers. In general, the performance of Validation 30BC, unlike those of the other validations, differed significantly among the different models, confirming this difficulty with urban and extraurban cross validation.

In general, the results of the validations from this study show errors larger than those reported in similar studies (*8*). In addition, the distribution of the error among its components deviates from the optimal, as shown in Table 2. The fact that there are significant increments of error compared with those from the calibration results suggests that the behavior of the same driver may differ in different contexts. It is therefore not advisable to calibrate car-following models on the basis

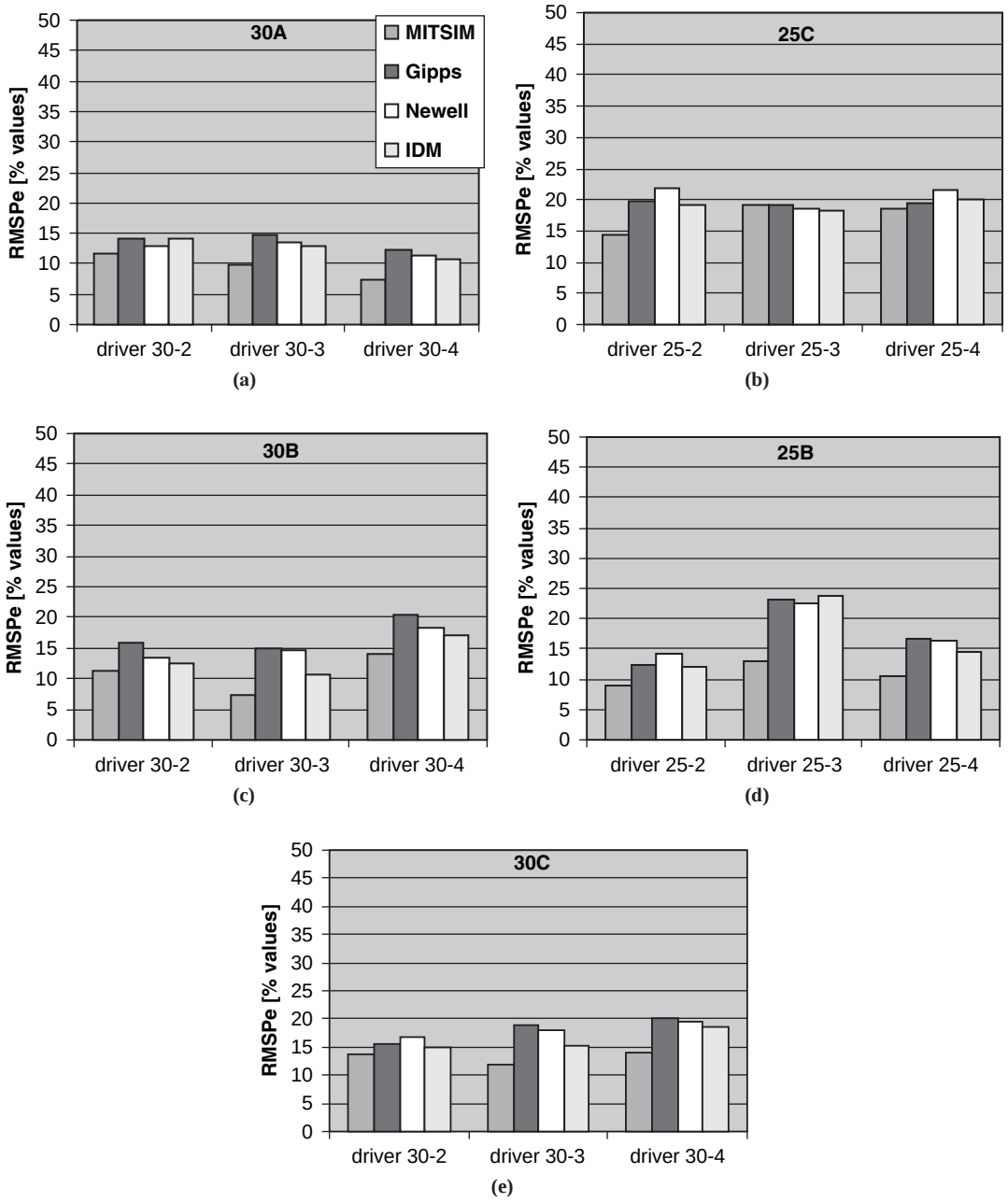


FIGURE 3 Model performances in calibrations (RMSPe values).

of experimental data pertinent to contexts limited in space and time (i.e., data from a single video camera), but it is necessary to detect the behavior of a user for a fairly long period, if possible, who is coping with different types of road and traffic characteristics.

## CONCLUSIONS

This paper sought not only to investigate methodological issues of model calibration and validation but also to produce preliminary results of a comparison of models on the basis of real microscopic traffic data.

First, the influence of the choice of performance measurement used for model calibration was examined. Intervehicle spacing was found to be the most reliable measure, and a physical interpretation was provided. Numerical evidence of the need to perform model validations was provided. The same drivers showed different optimal calibration parameters when data were calibrated on the basis of data from data sets collected for different parts of a route over a short distance in time. Moreover, the calibration results appeared to be surprisingly similar to those of previous work performed with data from test tracks [on average, 15.50% RMSPe versus 15.51% RMSPe from a previous study (*8*) for the same models], but this did not hold for cross validations that performed worse [on average, 22.31% RMSPe versus 17.37% RMSPe from a previous study (*8*)].

It is worth noting, on the one hand, that this difference in validation performance was obtained by considering only the data that were not overfitted and that it can be explained by the qualitative difference

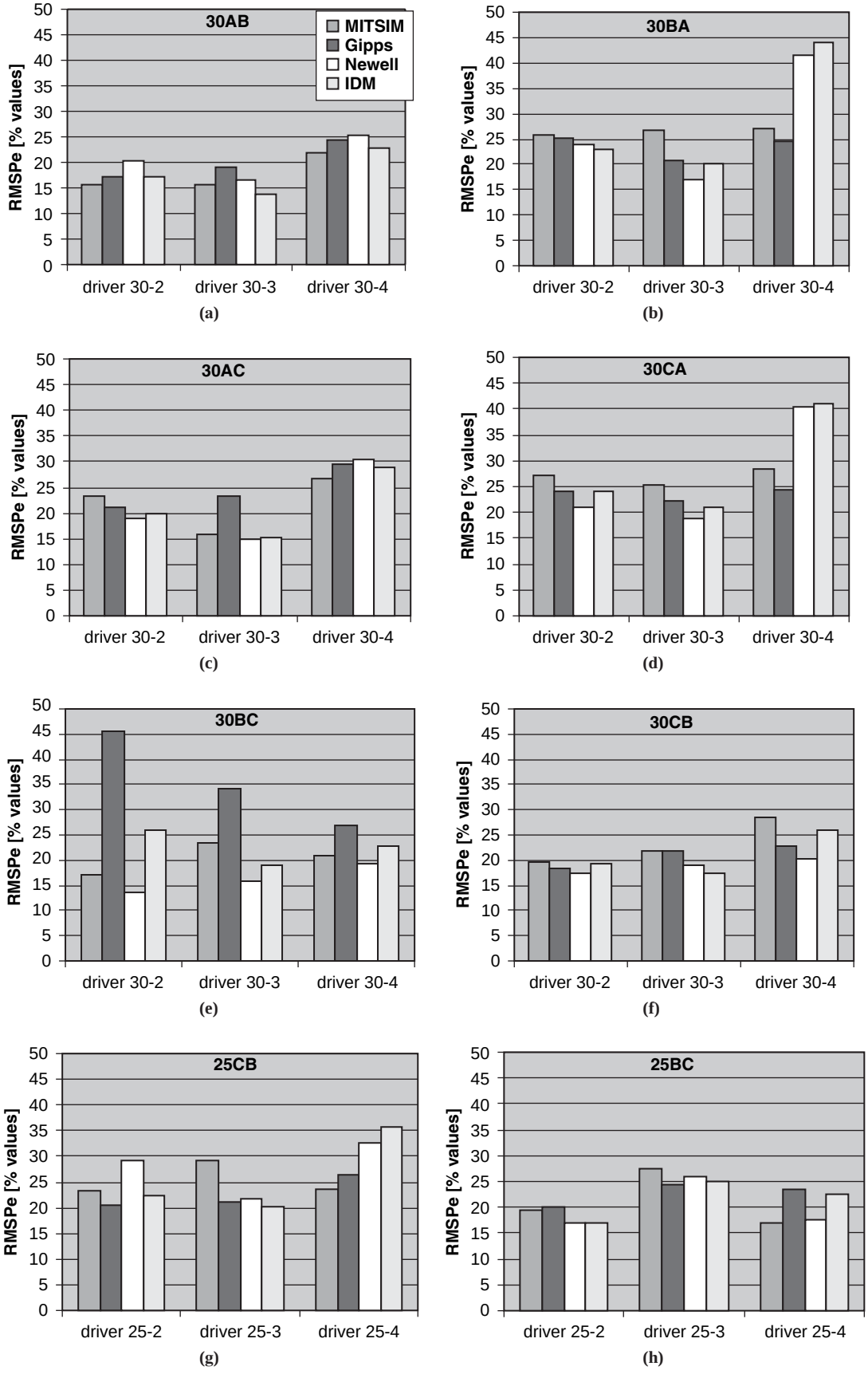


FIGURE 4 Model performances in cross validations (RMSPe values).

in the data sets used in the studies (data for real traffic versus data from a test track). On the other hand, this study overfit the data in the urban–extraurban cross validations. The first consideration highlights the importance of validations with real traffic data, while the second suggests that data collection schemes that allow the observation of drivers who are driving for long periods and who are coping with different types of road and traffic characteristics be adopted.

In a comparison of the models, the simplest model (the Newell model) performed the best, on average, while MITSIM showed a tendency to overfit the data on comparison of the calibration and validation results. A high degree of variability of parameters was observed not only among the different drivers but also for the same driver coping with different contexts. The results require confirmation in other studies and with other experimental data sets.

## ACKNOWLEDGMENTS

The authors thank the Massachusetts Institute of Technology for allowing them to use the MITSIMLab code and the TEST (Technology, Environment, Safety and Transport) Laboratory for its hardware and software support. This research was sponsored by the Italian Ministry of Research (MIUR) and the Regione Campania.

## REFERENCES

1. Brackstone, M., and M. McDonald. Car-Following: A Historical Review. *Transportation Research, Part F*, Vol. 2, 1999, pp. 181–196.
2. Kim, T., D. J. Lovell, and Y. Park. Limitations of Previous Models on Car-Following Behaviour and Research Needs. Presented at 82nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2003.
3. Brockfeld, E., R. D. Kühne, A. Skabardonis, and P. Wagner. Toward a Benchmarking of Microscopic Traffic Flow Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1852*, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 124–129.
4. Allen, R. W., R. E. Magdeleno, C. Serafin, S. Eckert, and T. Sieja. Driver Car Following Behaviour Under Test Track and Open Road Driving Conditions. SAE Paper 970170. SAE, Warrendale, Pa., 1997.
5. Brackstone, M., B. Sultan, and M. McDonald. Motorway Driver Behaviour: Studies on Car-Following. *Transportation Research, Part F*, Vol. 5, 2002, pp. 329–344.
6. Gurusinghe, G. S., T. Nakatsuji, Y. Azuta, P. Ranjitkar, and Y. Tanaboriboon. Multiple Car-Following Data Using Real-Time Kinematic Global Positioning System. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1802*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp.166–180.
7. Wu, J., M. Brackstone, and M. McDonald. The Validation of a Microscopic Simulation Model: A Methodological Case Study. *Transportation Research, Part C*, Vol. 11, 2003, pp. 463–479.
8. Brockfeld, E., R. D. Kühne, and P. Wagner. Calibration and Validation of Microscopic Traffic Flow Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1876*, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 62–70.
9. Ranjitkar, P., T. Nakatsuji, and M. Asano. Performance Evaluation of Microscopic Traffic Flow Models Using Test Track Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1876*, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp .90–100.
10. Ahn, S., M. J. Cassidy, and J. Laval. Verification of a Simplified Car-Following Theory. *Transportation Research, Part B*, Vol. 38, 2004, pp. 431–440.
11. Bham, G., and R. F. Benekohal. A High Fidelity Traffic Simulation Model Based on Cellular Automata and Car-Following Concepts. *Transportation Research, Part C*, Vol. 12, 2004, pp. 1–32.
12. Treiterer, J. *Investigation of Traffic Dynamics by Aerial Photogrammetry Techniques.* Final Report EES 278. Transportation Research Center, Department of Civil Engineering, Ohio State University, Columbus, 1975.
13. Punzo, V., D. J. Formisano, and V. Torrieri. Nonstationary Kalman Filter for Estimation of Accurate and Consistent Car-Following Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1934*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 3–13.
14. Newell, G. F. A Simplified Car-Following Theory—A Lower Order Model. *Transportation Research, Part B*, Vol. 36, 2002, pp. 195–205.
15. Gipps P. G. A Behavioural Car-Following Model for Computer Simulation. *Transportation Research, Part B*, Vol. 15B, No. 2, 1981, pp. 105–111.
16. Treiber, M., A. Hennecke, and D. Helbing. Congested Traffic States in Empirical Observation and Numerical Simulations. *Physical Review, Part E*, Vol. 62, 2000, pp. 1,805–1,824.
17. Ahmed, K. I. *Modeling Drivers' Acceleration and Lane-Changing Behavior*. Ph.D. thesis. Massachusetts Institute of Technology, Cambridge, 1999.
18. Yang, Q., and H. N. Koutsopoulos. A Microscopic Traffic Simulator for Evaluation of Dynamic Traffic Management Systems. *Transportation Research, Part C*, Vol. 4, No. 3, 1996, pp. 113–129.
19. Gazis, D. C., R. Herman, and R. W. Rothery. Non-Linear Follow-the-Leader Models of Traffic Flow. *Operations Research*, Vol. 9, No. 4, 1961, pp. 545–567.
20. Lindo Systems, Inc. LINDO API Manual. www.lindo.com. Accessed March 2004.
21. Pindyck, R. S., and D. L. Rubinfeld. *Econometric Models and Economic Forecasts*, 4th ed. McGraw-Hill, New York, 1998.

*The Traffic Flow Theory and Characteristics Committee sponsored publication of this paper.*