Statistical Validation of Traffic Simulation Models

Tomer Toledo and Haris N. Koutsopoulos

Traffic simulation models support detailed analysis of the dynamics of traffic phenomena and are important tools for analysis of transportation systems. In order to evaluate correctly the impact of different traffic management schemes, simulation models must be able to replicate reality adequately. Model validation (i.e., the process of checking to what extent the model replicates reality) is discussed. The role of validation is defined within the scope of model development and calibration, and the framework for performing the validation is discussed. A hierarchy of statistical methods to validate different types of simulation outputs against observed data is examined. Also, a validation method is proposed on the basis of statistical tests on metamodels fitted to the observed and simulated data. A case study illustrates the applicability of the various methods.

Intelligent transportation systems (ITS) applications, such as traffic controls and route guidance, have emerged in recent years as tools for traffic management. Traffic management methods and algorithms need to be calibrated and tested. In most cases only limited, if any, field tests are feasible because of prohibitively high costs and lack of public acceptance. Furthermore, the usefulness of such field studies is deterred by the inability to control fully the conditions under which they are performed. Hence tools to perform such evaluations in a laboratory are needed. Traffic simulation models support detailed analysis of the dynamics of traffic phenomena and are important tools for analysis of transportation systems, especially in the presence of ITS technologies. In order to evaluate correctly the impact of different traffic management schemes, simulation models must be able to replicate reality adequately.

Model validation (i.e., the process of checking to what extent the model replicates reality) is an essential step in the development and application of any traffic model. In this paper, the role of validation is defined within the scope of model development and application, and the framework for performing the validation is discussed. Statistical methods to model validation based on different types of observed data and simulation outputs are reviewed. Although some of these methods have been previously used in traffic simulation studies, others are adapted to the application domain from the broader simulation literature. In addition, a validation method is proposed, based on performing statistical tests on metamodels fitted to the observed and simulated data. Application of these methods is demonstrated with a case study. While the applications presented in the paper center around microscopic modeling, the methods are applicable to other types of traffic simulation models as well.

PROBLEM DESCRIPTION

Validation and calibration of simulation models are related tasks and ideally should take place before each new application. Calibration and validation of traffic simulation models consist of two steps (*1*). Initially, the individual models that comprise the simulator (e.g., driving behavior and route choice models) are estimated using disaggregate data. Disaggregate data include detailed driver behavior information such as vehicle trajectories. These individual models may be tested independently, for example, by using a holdout sample. The disaggregate analysis is performed within statistical software and does not involve the use of a simulation model. The level of effort required to collect and analyze trajectory data and the limited access to modify the models implemented within traffic simulators dictate that this step is most often only performed by the model developers. In the second step, the simulation model as a whole is calibrated and then validated using aggregate data (e.g., flows, speeds, occupancies, time headways, travel times, queue lengths). Aggregate calibration and validation are important both in the model development and in its application. In model development they serve to ensure that the interactions between the individual models within the simulator are captured correctly. In an application they are used to refine previously estimated parameter values for the specific site being studied.

Despite the increasing popularity of traffic simulation models, little attention has been given in the literature to model validation. Two types of validation approaches may be performed: visual and statistical (*2*). In visual validation, graphical representations of the outputs from the real and the simulated systems are displayed side by side to determine whether or not they can be differentiated. The visualization may be based on the animation modules available in most traffic simulation models. Alternatively, plots of different outputs (e.g., flows, speeds) may be generated. Turing tests (*3, 4*) may also be used. These tests involve presenting experts with two sets of outputs: observed and simulated. The test result depends on whether these experts are able to identify correctly the two sets apart. In any case, the process remains an inherently subjective and heuristic exercise. Statistical validation applies goodness-of-fit measures, confidence intervals, and statistical tests to quantify the similarity between the real and simulated systems.

Many published validation studies are based on visual comparison of outputs from the real and simulated systems or on comparison of simple descriptive statistics. For example, Abdulhai et al. (*5*) plot the observed and simulated headway distributions, lane-usage

T. Toledo, Center for Transportation and Logistics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, NE20-208, Cambridge, MA 02139. H. N. Koutsopoulos, Department of Civil and Environmental Engineering, Northeastern University, 437 Snell Engineering Center, Boston, MA 02115.

Transportation Research Record: Journal of the Transportation Research Board, No. 1876, TRB, National Research Council, Washington, D.C., 2004, pp. 142–150.

breakdown, and flow-density curves and calculate the mean relative percent error of total demands, link flows, and lane-usage breakdown. Other examples are found in the literature (*6–11*). Furthermore, in many cases the validation is limited to an isolated road section or a traffic corridor, thus avoiding the more complex behaviors and interactions associated with network applications. Jayakrishnan et al. (*12*) note that networkwide validation is necessary to ensure that interactions between different models within the simulation framework are captured correctly. Also, the validation is often oriented at a specific model (often the acceleration model). For example, Benekohal (*13*) focuses on the car-following model, and Fellendorf and Vortisch (*10*) separately and independently validate car-following and lane-changing models. Rao et al. (*14*) propose a multilevel approach, which consists of conceptual and operational validation. Conceptual validation focuses on the consistency of the simulation results with the theoretical foundations. For operational validation, a two-level evaluation of the simulated data against real-world observations using statistical tests is proposed: comparison of the respective means and comparison of the distributions. The application of the proposed methods is demonstrated using platoon data collected from video recordings.

VALIDATION FRAMEWORK

The main steps in developing an appropriate validation approach are:

- Generation of inputs for the simulation model,
- Choice of measures of performance (MOPs), and
- Choice of appropriate statistical tests for comparison of simulated and observed MOPs.

Generation of Inputs

The purpose of aggregate validation is to determine the extent to which the simulation model replicates the real system, with data readily available from loop detectors and other sources. Ideally, the validation should be based on comparing outputs that were generated by feeding the real and simulated systems with identical inputs. Therefore, the real system should be observed not only for its outputs but also for its input variables, which are then used in the simulation study. This practice reduces the variance of the differences between observed and simulated outputs and therefore increases the efficiency of the comparison. Travel demand, most commonly given in the form of dynamic origin–destination (O-D) matrices, is an important input in traffic simulation models. However, in most applications the O-D matrix is not observed, and so O-D flows must be estimated. The resulting aggregate validation process is shown in Figure 1. For the purposes of validation there is interest in O-D reconstruction (i.e., the demand realization that is most likely to have generated the observed traffic counts). See Hazelton (*15*) for a discussion of the differences between O-D reconstruction and estimation of mean O-D flows.

In many cases multiple sets of traffic data are available (e.g., data from several days) and the question of using all of them arises. In congested networks small changes in the demand may have a large impact on the simulation output. Hence separate O-D matrices must be estimated for each data set. Each one of these O-D matrices is then applied to the simulation model, and the resulting outputs are compared with the observed data that were used to generate the particular matrix. In contrast, using mean traffic counts to estimate mean O-D flows may lead to biased estimation of the model performance, because both the O-D estimator and the simulation model itself are nonlinear functions. The magnitude of the bias depends on the extent of nonlinearities and on the day-to-day variability in demand.

Choice of Measures of Performance

Measures of performance are statistics produced as outputs (or post-processed from the outputs) of the real and simulated systems. The model validation is based on the similarity between the simulated and real MOPs. The following criteria can assist in selecting MOPs for validation.

Context of Application

MOPs should be statistics that are important in the intended study. For example, point-to-point travel times are useful MOPs for validation when a traveler information system is to be evaluated on the basis of travel time savings. However, if a sensor-based incident detection system is studied, MOPs extracted from the sensors (e.g., occupancies, flows, speeds) may be more useful.

Independence

MOPs used for validation should be independent of any measurements used for calibration or to estimate inputs to the simulated

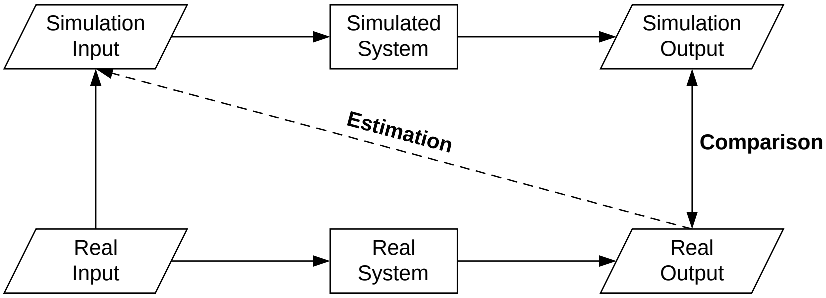


FIGURE 1 Aggregate validation process.
]

system. O-D flows are commonly estimated by minimizing a measure of the discrepancy between observed and simulated traffic counts. Therefore, validation of the simulation model (only) against traffic counts may lead to overestimating the realism of the model. For example, Rakha et al. (*6*) observed that INTEGRATION simulated flows match observations more closely than simulated speeds.

Error Sources

In traffic analysis the discrepancy between observed and simulated outputs can be explained by the following sources of error (*16*):

- Travel demand (O-D flows),
- Route choice,
- Driving behavior, and
- Measurement errors in the observed outputs.

The first three sources contribute errors to the simulated output. The last one represents errors in the observed output relative to the "true" output. In most cases, the contributions of the three simulation error sources are confounded and cannot be isolated in the validation. The locations and types of MOPs to be collected should be chosen to reflect errors from all these sources and reduce the effect of measurement errors as much as possible. Measurement locations should be chosen to provide spatial coverage of all parts of the network. Moreover, measurements close to the network entry points will mostly reveal errors in the O-D flows with little effect of the route choice and driving behavior models. As many measurement points as possible should be used in order to reduce the effect of measurement errors, assuming that they are independent for different locations.

Traffic Dynamics

MOPs and the level of temporal aggregation at which they are calculated (e.g., 15 min, 30 min) should be chosen such that they facilitate testing whether or not the model correctly captures the traffic dynamics. This is especially true in network applications in which both the temporal and the spatial aspects of traffic are important.

Level of Effort Required for Data Collection

In many cases this is the most constraining factor in practice. Point measurements (e.g., flows, speeds, and occupancies) are often readily and cheaply available from the surveillance system. Other types of measurements (e.g., travel times, queue lengths, and delays) are more expensive to collect. It is also important to note that data definitions and processing are not standardized. For example, statistics such as queue lengths may be defined in different ways, and surveillance systems may apply various time-smoothing techniques. It is therefore necessary to ensure that the simulated data are defined and processed the same way as the observed data.

Most traffic simulation models are stochastic (Monte Carlo) simulations. Hence MOPs should be calculated from a number of independent replications. There are mainly two approaches to determine the number of replications: sequential and two-step (*17*). In the sequential approach, one replication at a time is run until a suitable stopping criterion is met. Assuming that the outputs, Y_i, from different simulation runs are normally distributed, Fishman (*18*) suggested the following criterion:

$$R \geq R_i = \max\left[2, \left(\frac{s_R(Y_i)t_{\alpha/2}}{d_i}\right)^2\right] \tag{1}$$

where

R = number of replications performed,
R_i = minimum number of replications required to estimate the mean of Y_i with tolerance d_i,
$s_R(Y_i)$ = sample standard deviation of Y_i based on R replications, and
$t_{\alpha/2}$ = critical value of the t-distribution at significance level α.

In the two-step approach, first an estimate of the standard deviation of Y_i is obtained by performing R_0 replications. Assuming that this estimate does not change significantly as the number of replications increases, the minimum number of replications required to achieve the allowable error d_i is given by

$$R_i = \left[\frac{s_{R_0}(Y_i)t_{\alpha/2}}{d_i}\right]^2 \tag{2}$$

The required number of replications is calculated for all measures of performance of interest. The most critical (highest) value of R_i determines the number of replications required.

STATISTICAL VALIDATION

The general simulation literature includes a large number of approaches for the statistical validation of simulation models. These approaches include goodness-of-fit measures, confidence intervals, and statistical tests of the underlying distributions and processes. In many cases though, they may not be applicable because both the real and the simulated traffic processes of interest are nonstationary and autocorrelated. The choice of the appropriate methods and their application to the validation of traffic simulation models depends on the nature of the output data. The following cases are considered:

- Single-valued MOPs (e.g., average delay, total throughput), and
- Multivariate MOPs (e.g., time-dependent flow or speed measurements at different locations, travel times on different sections).

Single-valued MOPs are appropriate for small-scale applications in which one statistic may summarize the performance of the system. Multivariate MOPs capture the temporal and/or spatial distribution of traffic characteristics and are therefore useful to describe the dynamics at the network level. It may also be useful to examine the joint distribution of two MOPs (e.g., flow and headway), as this provides more information regarding the interrelationships among MOPs. The types of statistical approaches that are discussed include:

- Goodness-of-fit measures,
- Hypothesis testing and confidence intervals, and
- Test of underlying structure.

Goodness-of-Fit Measures

A number of goodness-of-fit measures can be used to evaluate the overall performance of simulation models. Popular among them are the root-mean-square error (*RMSE*), the root-mean-square percent error (*RMSPE*), the mean error (*ME*), and the mean percent error (*MPE*) statistics. These statistics quantify the overall error of the simulator. Percent error measures provide information on the magnitude of the errors relative to the average measurement directly. *RMSE* and *RMSPE* penalize large errors at a higher rate relative to small errors. The two measures are given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}} - Y_n^{\text{obs}}\right)^2} \tag{3}$$

$$RMSPE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{Y_n^{\text{sim}} - Y_n^{\text{obs}}}{Y_n^{\text{obs}}}\right)^2} \tag{4}$$

where Y_n^{obs} and Y_n^{sim} are the averages of observed and simulated measurements at space–time point n, respectively calculated from all available data (i.e., several days of observations and multiple simulation runs).

ME and *MPE* indicate the existence of systematic under- or overprediction in the simulated measurements. These measures are given by

$$ME = \frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}} - Y_n^{\text{obs}}\right) \tag{5}$$

$$MPE = \frac{1}{N}\sum_{n=1}^{N}\frac{Y_n^{\text{sim}} - Y_n^{\text{obs}}}{Y_n^{\text{obs}}} \tag{6}$$

These two statistics are most useful when applied separately to measurements at each time–space point rather than to all measurements jointly. This way they provide insight to the spatial and temporal distribution of errors and help identify deficiencies in the model.

Another measure that provides information on the relative error is Theil's inequality coefficient, U (*19*)

$$U = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}} - Y_n^{\text{obs}}\right)^2}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}}\right)^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{obs}}\right)^2}} \tag{7}$$

where U is bounded, $0 \leq U \leq 1$. $U = 0$ implies perfect fit between the observed and simulated measurements. $U = 1$ implies the worst possible fit. Theil's inequality coefficient may be decomposed to three proportions of inequality: the bias (U^M), variance (U^S), and covariance (U^C) proportions, which are, respectively, given by

$$U^M = \frac{\left(\bar{Y}^{\text{sim}} - \bar{Y}^{\text{obs}}\right)^2}{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}} - Y_n^{\text{obs}}\right)^2} \tag{8}$$

$$U^S = \frac{\left(s^{\text{sim}} - s^{\text{obs}}\right)^2}{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}} - Y_n^{\text{obs}}\right)^2} \tag{9}$$

$$U^C = \frac{2(1-\rho)s^{\text{sim}}s^{\text{obs}}}{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n^{\text{sim}} - Y_n^{\text{obs}}\right)^2} \tag{10}$$

where $\bar{Y}^{\text{obs}}$, $\bar{Y}^{\text{sim}}$, s^{obs}, and s^{sim} are the sample means and standard deviations of the average observed and simulated measurements, respectively, and ρ is the correlation between the two sets of measurements.

The bias proportion reflects the systematic error. The variance proportion indicates how well the simulation model replicates the variability in observed data. These two proportions should be as small as possible. The covariance proportion measures the remaining error and therefore should be close to one. If the different measurements are taken from nonstationary processes, the proportions can be viewed only as indicators of the sources of error.

Hypothesis Testing and Confidence Intervals

Classic hypothesis tests (e.g., two-sample *t*-test, Mann-Whitney test, and two-sample Kolmogorov–Smirnov test) and confidence intervals may also be used. Law and Kelton (*20*) suggest the use of confidence intervals, which provide richer information compared to statistical tests, for the validation of complex simulation systems.

Two-sample tests assume that both sets of outputs are independent draws from identical distributions (IID). Therefore, these tests should be performed separately for each time–space measurement point. If the number of observations at each time–space point is not sufficient to obtain significant results, observations from appropriate time intervals (such that the IID assumption holds, at least approximately) may be grouped together. Furthermore, the standard two-sample *t*-test also assumes that the two distributions (observed and simulated) are normal and share a common variance. These assumptions, in particular the variance equality, may be unrealistic in the context of traffic simulation. Law and Kelton propose an approximate *t*-solution procedure according to Scheffe (*21*), which relaxes the variance equality assumption. To test for the equality of the mean of observed and simulated measurements, $H_0: Y_n^{\text{sim}} = Y_n^{\text{obs}}$ against $H_1: Y_n^{\text{sim}} \neq Y_n^{\text{obs}}$ at the α significance level, reject H_0 if

$$\frac{\left|Y_n^{\text{sim}} - Y_n^{\text{obs}}\right|}{\sqrt{\frac{\left(s_n^{\text{sim}}\right)^2}{m_n^{\text{sim}}} + \frac{\left(s_n^{\text{obs}}\right)^2}{m_n^{\text{obs}}}}} \geq t_{\alpha/2,\,\hat{f}} \tag{11}$$

where s_n^{obs} and s_n^{sim} are the sample standard deviations of the observed and simulated measurements at time–space point n, respectively; m_n^{obs} and m_n^{sim} are the corresponding sample sizes; and $\hat{f}$ is the modified number of degrees of freedom given by

$$\hat{f} = \frac{\left(\frac{s_n^{\text{sim}}}{m_n^{\text{sim}}} + \frac{s_n^{\text{obs}}}{m_n^{\text{obs}}}\right)^2}{\frac{\left(s_n^{\text{sim}}\right)^4}{\left(m_n^{\text{sim}}\right)^2\left(m_n^{\text{sim}} - 1\right)} + \frac{\left(s_n^{\text{obs}}\right)^4}{\left(m_n^{\text{obs}}\right)^2\left(m_n^{\text{obs}} - 1\right)}} \tag{12}$$

The corresponding $(1 - \alpha)$ confidence interval is given by

$$Y_n^{\text{sim}} - Y_n^{\text{obs}} \pm t_{\alpha/2,\,\hat{f}} \sqrt{\frac{\left(s_n^{\text{sim}}\right)^2}{m_n^{\text{sim}}} + \frac{\left(s_n^{\text{obs}}\right)^2}{m_n^{\text{obs}}}} \tag{13}$$

The above approaches can be used for the analysis of individual time–space points. However, the behavior of traffic networks in many applications is autocorrelated and nonstationary as a result of time-varying travel demands and traffic dynamics (i.e., congestion buildup and dissipation). Therefore, measurements at different space–time points cannot be considered as independent draws from identical distributions. In this case, the above methods cannot be used to test the overall validity of the simulation, and joint hypothesis tests, which better reflect the dynamics of the system, are more appropriate. Kleijnen (*22*) recommends the application of Bonferroni's inequality to test multiple hypotheses jointly at a prescribed significance level α. Let α_n be the significance level at each individual time–space point *n*. An upper bound for the simultaneous significance level α at the network level is given by

$$\alpha \leq \sum_{n=1}^{N} \alpha_n \tag{14}$$

where *N* is the number of measurement points (over time and space).

Equation 14 holds under very general conditions. In practice the significance levels at each time–space point *n* are usually set to $\alpha_n = \alpha/N$. α_n is then used as the level of significance independently at each time–space location to perform hypothesis testing or develop confidence intervals. Bonferroni's inequality can similarly be used to create composite tests and joint confidence intervals for multiple single-valued MOPs.

In practice, this technique may only be applied to a small number of measurement points *N*. For large *N* the corresponding significance levels α_n become small and the confidence intervals increase to the point at which it is difficult to reject any model as invalid. Another limitation of Bonferroni's inequality is that it is too conservative, especially for highly correlated test statistics, hence resulting in a high probability of Type II errors (i.e., failure to reject false null hypotheses). Holm's test (*23*) is a more powerful version of Bonferroni's test and works better when the tests are correlated and when several of the null hypotheses are false. Other methods that improve the power of the test have also been proposed (*24*).

An alternative sequential test, adapted from Rao et al. (*14*), involves a two-sample Kolmogorov–Smirnov (K-S) test for each one of the measurements (i.e., time–space points or different MOPs) and recording the corresponding *p*-values. A one-sided *t*-test is then conducted to test whether the mean of these *p*-values is smaller than the desired significance level (i.e., the model is invalid).

As previously discussed, two-sample tests, such as the K-S test, can be used to test whether simulated and observed measurements are drawn from the same distribution, and so, to examine the validity of the simulation. The application of the K-S test is straightforward for the case of a single MOP. Rao et al. (*14*) suggest the use of a two-dimensional two-sample K-S test to validate a pair of MOPs jointly. This test is useful, since MOPs are usually correlated (for example, flow and headways). Although the test involves several steps, it is simple to implement and apply. Details can be found in Fasano and Franceschini (*25*) and Press et al. (*26*), which also provides an approximate function for the test *p*-values.

In the preceding discussion no assumption was made regarding the nature of the input data. If the input data is known (trace-driven simulation), Kleijnen (*22, 27*) proposes a regression procedure to validate the simulation model using an *F*-test of the joint equality of the means and variances of the real and simulated measurements. Let us assume that there are *N* different input data sets (common to the simulated and true system). For pairs of observations $(y_n^{\text{obs}}, y_n^{\text{sim}})$, $n = 1, \ldots, N$ the following regression is performed:

$$\left(y_n^{\text{obs}} - y_n^{\text{sim}}\right) = \beta_0 + \beta_1\left(y_n^{\text{obs}} + y_n^{\text{sim}}\right) + \epsilon_n \tag{15}$$

where β_0 and β_1 are parameters and ϵ_n is a random error term. The hypothesis that the observed and simulated outputs are drawn from identical distributions is tested with the null $H_0 : \beta_0 = 0$ and $\beta_1 = 0$.

Test of Underlying Structure

We now propose another approach, particularly suited for the validation of traffic simulation models with limited data. The method is based on developing metamodels that capture the underlying relations among important traffic variables of interest (e.g., speed–flow relationships), time evolution of flows and statistically testing the hypothesis that the metamodel parameters are equal in the simulated and observed data.

The validation proceeds by using the outputs from the real and simulated systems to estimate two separate metamodels, which describe the structure of these outputs. The choice of the appropriate metamodel depends on the nature of the application and relationships among variables established by the traffic flow theory. Statistical tests for the equality of coefficients across the two metamodels are then used to validate the simulation model. The equality of the models is then tested with the null hypothesis $H_0 : \beta^{\text{obs}} = \beta^{\text{sim}}$ against $H_1 : \beta^{\text{obs}} \neq \beta^{\text{sim}}$ using an *F*-test. The test uses two models: restricted and unrestricted. The restricted model, which forces the equality of parameters of the two metamodels, is estimated with the combined data set (both real and simulated observations). The unrestricted model is the combination of two separate models: one estimated with the real data and the other with the simulated data. The test statistic is calculated:

$$F_{K,\,N^{\text{sim}}+N^{\text{obs}}-2K} = \frac{\left(ESS^{R} - ESS^{UR}\right)\big/K}{ESS^{UR}\big/\left(N^{\text{sim}} + N^{\text{obs}} - 2K\right)} \tag{16}$$

where

ESS^R and ESS^{UR} = sums of squared residuals of the restricted and the unrestricted models, respectively, calculated as $ESS^R = ESS^{\text{com}}$ and $ESS^{UR} = ESS^{\text{obs}} + ESS^{\text{sim}}$;

ESS^{com}, ESS^{obs}, and ESS^{sim} = statistics calculated from the models estimated with the combined, the real, and the simulated data, respectively;

N^{obs} and N^{sim} = numbers of observations in the real data and simulated data, respectively; and

K = number of parameters in the model.

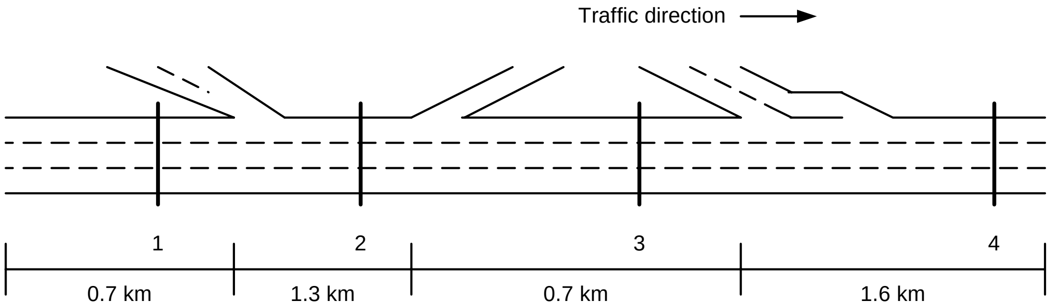


FIGURE 2 M27 freeway section in Southampton, England.

This method provides great flexibility in validating simulation models and, in many cases, overcomes the difficulty of limited data. In many traffic studies measurements may be available for only a few days. This poses a problem for most of the methods discussed above since limited data curbs the development of statistically significant test statistics and goodness-of-fit measures. Moreover, disaggregate data (e.g., observations of individual vehicles, 1-min sensor reading) can be used when fitting the metamodel without having to aggregate observations (e.g., to 5-min intervals) as would be required if real and simulated outputs were directly compared. Another advantage is the flexibility in choosing the functional form of the metamodel to accommodate different modeling needs. For example, a metamodel describing an MOP as a function of time (e.g., time-dependent speeds) can be used to test whether or not the formation and dissipation of congestion are captured correctly. Another metamodel can be formulated to test the realism of the underlying fundamental diagram.

CASE STUDY

The application of the methods discussed above are demonstrated in the validation of a microscopic simulation model by using 3-h a.m. peak sensor data from a section of the M27 freeway in Southampton England, shown in Figure 2. The observed data include counts and speeds for 5 days at 1-min intervals at the locations indicated in the figure. The simulation outputs include similar measurements from 10 runs. For the purpose of this paper the measurements at Sensor 3 are considered.

First goodness-of-fit statistics for the sensor speed are calculated. They are based on measurements at 1-min intervals. The results are summarized in Table 1. The results indicate a reasonable overall fit, with a small bias and good ability to replicate the variability in the data.

TABLE 1 Goodness-of-Fit Statistics for Southampton Network

Statistic	Value
RMSPE (%)	5.08
RMSE (km/hr)	5.09
MPE (%)	-0.66
ME (km/hr)	-0.83
U (Theil's inequality coefficient)	0.024
U^M (Bias proportion)	0.135
U^S (Variance proportion)	0.023
U^c (Covariance proportion)	0.842

The limitation of these statistics is that they do not evaluate the ability of the simulation model to represent the dynamics of traffic behavior. Next, two-sample tests focused on a specific time interval are performed. Measurements were grouped in 15-min intervals in order to have enough observations for the statistical tests. The assumption is that within each interval observations are independent draws from the same distribution. The tests were performed separately for each interval. For example, Table 2 summarizes the test results for the third time period. The null hypothesis $H_0 : Y^{\text{sim}} = Y^{\text{obs}}$ is rejected at 95% confidence. The corresponding confidence interval is $-4.72 \leq Y^{\text{sim}} - Y^{\text{obs}} \leq -1.86$. While the goodness-of-fit statistics show good overall fit, the focused analysis of specific time intervals reveals weaknesses of the simulation model in capturing traffic dynamics. This result illustrates the need to use statistical methods that are as detailed as possible.

Individual tests over all time periods may yield conflicting results, and application of Bonferonni's inequality may be too conservative. Therefore, the validity of the model is tested with the application of hypothesis testing on metamodels. First metamodels that describe the fundamental diagrams underlying the two sets of data are developed. The functional form of the Pipes–Munjal model (*28*) was selected.

$$V(t) = V_f\left[1 - \left(\frac{\rho(t)}{\rho_{\text{jam}}}\right)^{\beta}\right] + \epsilon(t) \tag{17}$$

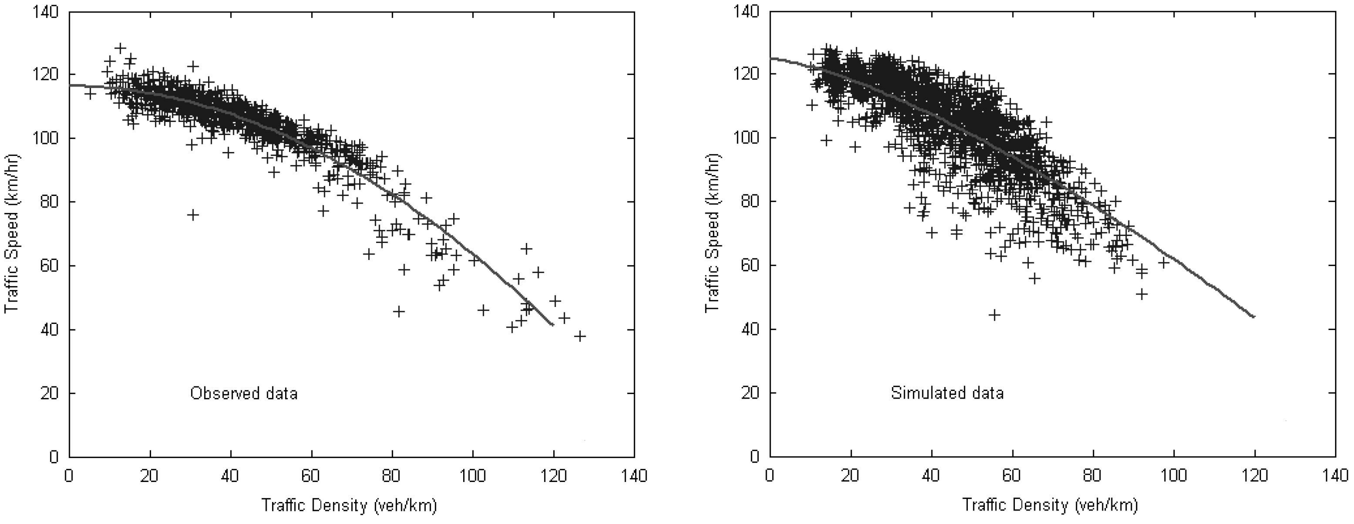
where

$V(t)$ and $\rho(t)$ = traffic speed and density at time t, respectively;
V_f = free-flow speed;
ρ_{jam} = jam density; and
$\epsilon(t)$ = an error term.

V_f, ρ_{jam}, and β are the underlying parameters of the model.

TABLE 2 Two-Sample *t*-Test for Third Time Interval

	Observed Data	Simulated Data
Mean	108.77	105.49
Variance	11.53	56.09
Observations	75	150
t-statistic	4.52	
T critical value	1.97 (95% confidence)	



(a)

(b)

FIGURE 3 Observed and simulated speed–density data and fitted metamodels.

The observed data and the simulated data used for estimation and the corresponding estimated metamodel regression lines are shown in Figure 3. In addition, for the purpose of statistical testing a third metamodel was estimated with the combined data set including both observed and simulated data. Estimation results for the three models are presented in Table 3. The *F*-test described above may be used to test the validity of the simulation output.

$$F_{3,2694} = \frac{(52034.3 - 47698.4)/3}{47698.4/(1800 + 900 - 6)} = 81.63 \qquad (18)$$

The critical value of the *F*-distribution with (3, 2694) df at the 95% confidence level is 8.53, and so the hypothesis that the parameters of the observed and simulated metamodels are equal can be rejected.

Further insight into the performance of the simulation model can be derived by developing a piecewise linear metamodel describing the temporal change in traffic speeds.

$$V(t) = V_0 + \alpha_1 t + \alpha_2(t - \beta_1)x_1(t) + \cdots + \alpha_i(t - \beta_{i-1})x_{i-1}(t) + \cdots + \alpha_N(t - \beta_{N-1})x_{N-1}(t) + \epsilon(t) \qquad (19)$$

where

$V(t)$ = traffic speed at time *t;*
V_0, α, and β = parameters of the model;
α_i = change in the slope of section *i* from the previous section;
β_i = boundary point between sections *i* and *i* + 1;
$\epsilon(t)$ = an error term; and

$$x_i(t) = \begin{cases} 1 & \text{if } t \geq \beta_1 \\ 0 & \text{otherwise.} \end{cases} \qquad i = 1, \ldots, N - 1$$

TABLE 3 Fundamental Diagram Metamodel Estimation Results

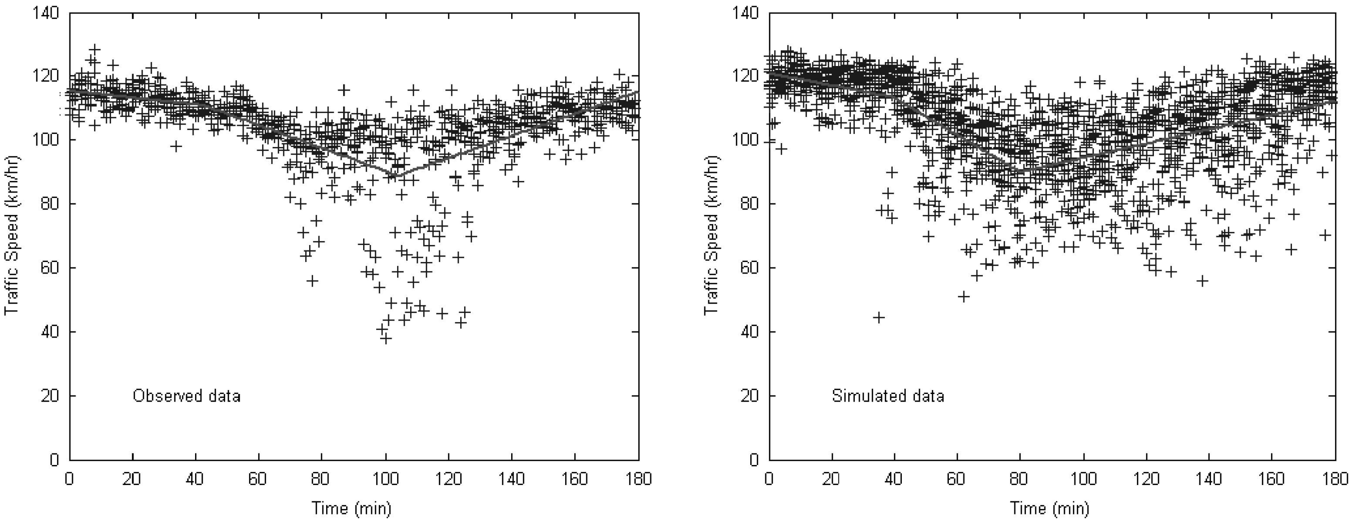
Parameter	Observed Data	Simulated Data	Combined Data
V_f	116.400	124.813	121.609
ρ_{jam}	149.797	162.915	158.065
β	1.964	1.406	1.571
R^2	0.867	0.882	0.866
ESS	19367.7	28330.7	52034.3
Observations	900	1800	2700

Metamodels of this form were estimated for the real data, simulated data, and the combined data set using the Gallant–Goebel estimator for nonlinear least squares problems with time-series data (*29*). The number of linear segments was set at three by maximizing $\bar{R}^2$ in the observed data nonparametrically. Estimation results are shown in Table 4 and the regression lines in Figure 4.

For this representation of the data, an *F*-test for the equality of coefficients in the data yields the statistic $F_{6,2688} = 18.31$. The critical value of the *F*-distribution at the 95% confidence level is 3.67, and so the null hypothesis can be rejected (i.e., with respect to the time–speed space, the simulation model is not a valid representation of the real system). Similar tests assuming partial equality of the parameters may also be conducted to better understand the sources of the differences between the two data sets. For example, a test for

TABLE 4 Time–Speed Metamodel Estimation Results

Parameter	Observed Data	Simulated Data	Combined Data
V_0	115.313	120.859	118.372
α_1	-0.106	-0.190	-0.105
α_2	-0.253	-0.394	-0.265
α_3	0.706	0.807	0.656
β_1	42.860	40.001	26.857
β_2	104.011	79.976	96.121
R^2	0.404	0.411	0.384
ESS	87024.4	99827.3	194490.8
Observations	900	1800	2700



(a)

(b)

FIGURE 4 Observed and simulated time–speed data and fitted metamodels.

the equality of the boundaries of the linear pieces ($H_0 : \beta_1^{\text{obs}} = \beta_1^{\text{sim}}$ and $\beta_2^{\text{obs}} = \beta_2^{\text{sim}}$) will reveal whether or not the buildup and dissipation of congestion occur at the same times in the two data sets.

CONCLUSION

This paper discusses issues in validation of microsimulation models and proposes relevant statistical tests of various types of MOPs and underlying model structure. The tests cover a wide range of data requirements and range from overall goodness-of-fit statistics to hypothesis testing at various levels of detail. The authors propose a method for validation by fitting metamodels to observed and simulated outputs and conducting statistical tests on the equality of the parameters across these metamodels. This approach does not require grouping of the data (e.g., by time interval) and therefore may yield significant results even if only limited data are available as shown in the case study. The case study results highlight an important aspect of validation: results from different validation methods and across different MOPs may be conflicting. Therefore, comprehensive validation using multiple MOPs and their theoretically expected relationships should be used.

ACKNOWLEDGMENT

This material is based on work supported by the National Science Foundation.

REFERENCES

1. Toledo, T., H. N. Koutsopoulos, A. Davol, M. E. Ben-Akiva, W. Burghout, I. Andréasson, T. Johansson, and C. Lundin. Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1831,* TRB, National Research Council, Washington, D.C., 2003, pp. 65–75.
2. Rao, L., and L. Owen. Validation of High-Fidelity Traffic Simulation Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1710,* TRB, National Research Council, Washington, D.C., 2000, pp. 69–78.
3. Turing, A. M. Computing Machinery and Intelligence. *Mind,* Vol. 59, 1950, pp. 433–460.
4. Schruben, L. Confidence Interval Estimation using Standardized Time Series. *Operations Research,* Vol. 31, 1980, pp. 1090–1108.
5. Abdulhai, B., J. B. Sheu, and W. Recker. *Simulation of ITS on the Irvine FOT Area Using "Paramics 1.5" Scalable Microscopic Traffic Simulator: Phase I: Model Calibration and Validation.* California PATH Research Report UCB-ITS-PRR-99-12. University of California, Berkeley, 1999.
6. Rakha, H., B. Hellinga, M. Van Aerde, and W. A. Perez. Systematic Verification, Validation, and Calibration of Traffic Simulation Models. Presented at 75th Annual Meeting of the Transportation Research Board, Washington, D.C., 1996.
7. Bloomberg, L., and J. Dale. Comparison of VISSIM and CORSIM Traffic Simulation Models on a Congested Network. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1727,* TRB, National Research Council, Washington, D.C., 2000, pp. 52–60.
8. Hall, F. L., L. Bloomberg, N. M. Rouphail, B. Eads, and A. D. May. Validation Results for Four Models of Oversaturated Freeway Facilities. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1710,* TRB, National Research Council, Washington, D.C., 2000, pp. 161–170.
9. Rilett, L. R., K.-O. Kim, and B. Raney. Comparison of Low-Fidelity TRANSIMS and High-Fidelity CORSIM Highway Simulation Models with Intelligent Transportation System Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1739,* TRB, National Research Council, Washington, D.C., 2000, pp. 1–8.
10. Fellendorf, M., and P. Vortisch. Validation of Microscopic Traffic Flow Model VISSIM in Different Real-World Situations. Presented at 80th Annual Meeting of the Transportation Research Board, Washington, D.C., 2001.
11. Lee, D.-H., X. Yang, and P. Chandrasekar. Parameter Calibration for PARAMICS Using Genetic Algorithm. Presented at 80th Annual Meeting of the Transportation Research Board, Washington, D.C., 2001.
12. Jayakrishnan, R., J.-S. Oh, and A.-E.-K. Sahraoui. Calibration and Path Dynamics Issues in Microscopic Simulation for Advanced Traffic Management and Information Systems. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1771,* TRB, National Research Council, Washington, D.C., 2001, pp. 9–17.

13. Benekohal, R. F. Procedure for Validation of Microscopic Traffic Flow Simulation Models. In *Transportation Research Record 1320,* TRB, National Research Council, Washington, D.C., 1991, pp. 190–202.
14. Rao, L., D. Goldsman, and L. Owen. Development and Application of a Validation Framework for Traffic Simulation Models. *Proc., 1998 Winter Simulation Conference* (D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, eds.), Washington, D.C., 1998.
15. Hazelton, M. L. Inference for Origin–Destination Matrices: Estimation, Prediction and Reconstruction. *Transportation Research B,* Vol. 35, 2001, pp. 667–676.
16. Doan, D. L., A. Ziliaskopoulos, and H. Mahmassani. On-Line Monitoring System for Real-Time Traffic Management Applications. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1678,* TRB, National Research Council, Washington, D.C., 1999, pp. 142–149.
17. Alexopoulos, C., and A. Seila. Output Data Analysis. In *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice* (J. Banks, ed.), John Wiley & Sons, New York, 1998.
18. Fishman, G. S. *Principles of Discrete Event Simulation.* John Wiley & Sons, New York, 1978.
19. Theil, H. *Economic Forecasts and Policy.* North-Holland, Amsterdam, Netherlands, 1961.
20. Law, A. M., and W. D. Kelton. *Simulation Modeling and Analysis,* 3rd ed. McGraw-Hill, New York, 2000.
21. Scheffe, H. Practical Solutions of the Behrens-Fisher Problem. *Journal of the American Statistical Association,* Vol. 65, 1970, pp. 1501–1508.
22. Kleijnen, J. P. C. Verification and Validation of Simulation Models. *European Journal of Operational Research,* Vol. 82, 1995, pp. 145–162.
23. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics,* Vol. 6, 1979, pp. 65–70.
24. Shaffer, J. P. Multiple Hypothesis Testing. *Annual Review of Psychology,* Vol. 46, 1995, pp. 561–584.
25. Fasano, G., and A. Franceschini. A Multidimensional Version of the Kolmogorov-Smirnov Test. *Monthly Notices of the Royal Astronomical Society,* Vol. 225, 1987, pp. 155–170.
26. Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C,* 2nd ed. Cambridge University Press, Cambridge, United Kingdom, 1992.
27. Kleijnen, J. P. C. Validation of Models: Statistical Techniques and Data Availability. Presented at 1999 Winter Simulation Conference, Phoenix, Ariz., 1999.
28. Pipes, L. A. Car-Following Models and the Fundamental Diagram of Road Traffic. *Transportation Research,* Vol. 1, 1967, pp. 21–29.
29. Gallant, A. R., and J. J. Goebel. Nonlinear Regression with Autocorrelated Errors. *Journal of the American Statistical Association,* Vol. 71, 1976, pp. 961–967.