Calibration and validation

Typical Steps of Simulation Studies

Traffic simulation models are typically used for

1. Assessment and planning of (road) infrastructures;
2. Evaluation of advanced traffic management and information strategies; and
3. Testing technologies and systems intended to increase the safety, capacity, and environmental efficiency of vehicles and roads.

A simulation study usually involves a comparison of a current and a future situation, with possible modifications in the network (infrastructure), traffic management, information strategies, and relevant technologies or systems. In addition, the effects of external conditions such as increased demand or changes in traffic composition are investigated. The simulation task involves multiple steps, depending on the specific study questions. Most authors' descriptions of the successive steps of a typical calibration study converge (see Dowling et al., 2004, for an example). A typical task list is:

1. Define the objectives of the study and the alternative scenarios to be tested.
2. Define the measures of performance that will be used to compare the current situation with the alternative.
3. Define the network to simulate by:
    a. Characterizing links: number of lanes, lengths, desired speeds, and upstream and downstream nodes.
    b. Characterizing nodes: allowed turning movements, upstream and downstream links, signalization (traffic lights with fixed phases, adaptive controllers, roundabouts, priority rules).
4. Define demand; this can be done from an additional model (usually static) or through the use of the existing data (in this case, measured flows on several links are needed), or by a combination of both.
5. Run the simulation and check whether the model performs as expected (verification).
6. Collect the data for calibration and validation by
    a. Collecting the data set that allows the definition of simulation entry variables (both static and dynamic).
    b. Collect the measures of performance that will allow comparison of simulation results with the observed current reality.
7. Calibrate and validate the traffic simulation tool for the specific site and the reference scenario.
8. Simulate the alternative scenarios; based on specific cases, describe at least one of the following:
   (a) new infrastructure; (b) new regulations; (c) new demand.
9. Analyze the impact of the scenario on the simulation results by carefully scrutinizing the impact of the evolution of the scenarios on the chosen measure of performance.
10. Write the report.

When building the reference and alternative scenarios, one must define which traffic phenomena or behaviors should be included in the study and should be accurately described by the simulation model. For example, a tool does not necessarily have to describe the queue formations upstream of roundabouts if the researcher wants to evaluate the impact of ramp metering installations on a highway, but an accurate lane-changing process is essential. Disaggregate data analysis is necessary for each key behavior included in a simulation scenario. Optimally, this detailed data analysis is performed and reported by the developers of the simulation tool.

Along with a detailed evaluation of the predictive capacity of the tool for each key behavior implied in the scenario to be simulated, aggregate calibration and validation must be completed for the application scenario. The calibration and validation of the model should focus on the specific site and traffic situations to be covered in the simulation study. Depending on the site chosen for the simulation, several variables should be considered:

- Type of network: size (number of links and nodes); urban or interurban routing, with and without traffic signals, roundabouts, curves, ramps, and combinations,
- Conditions of use: morning and evening peak hours, weekends, and holidays; weather conditions, traffic composition (percentages of trucks and passenger vehicles), evacuation needs,
- Traffic management system: adaptive control for traffic signals, driver information collected by means of GPS or by onboard devices, for individual information anti-collision, and other advanced driver assistance systems (ADAS).

With respect to traffic management systems, this component of calibration and validation focuses more on the settings of the systems since the disaggregate calibration and validation already demonstrated that the simulation model can reproduce these technologies and systems in general.

Generic Procedure for Calibration and Validation

Globally, calibration and validation of a simulation tool (with a set of parameters) is a process of comparison on an appropriate scale of the simulation results for chosen variables with a set of observations of the variables. (See Figure 1.) During calibration, the differences between the simulated and observed variables are minimized by finding an optimum set of model parameters. This is often formulated as minimizing an error measure e(p) (Vaze et al., 2009; Ciuffo and Punzo, 2010a):

$$e(p) = f(M(p) - d) \quad [1]$$

d is the observed measurement.

Where the model M can, of course, depend on much more than parameters p. Thus, calibration becomes an optimization problem. The validation process estimates the difference between the simulation variables using the parameter set, resulting from calibration and an independent set of observed variables. In the following, a variable is referred to as a measure of performance (MoP). The comparison scale is the goodness-of-fit measure (GoF). The practical specification of calibration and validation relies on several factors:

- Error function, generally GoF
- Traffic measurement, usually called the MoP of the simulation model
- Optimization method for calibration

- Traffic simulation model used
- Transportation system (traffic scenario) to be simulated
- Demand pattern used and accuracy of its estimation
- Parameters to calibrate
- Quality (in terms of error presence) of observed measurement
- Possible constraints

These factors make calibration and validation more complicated because it is difficult to define the general methodologies to be followed in all the cases. Depending on the particular specification of the calibration or validation, the strategy to be adopted in practice may be different. However, the calibration and validation framework shown in Figure 1 can still be maintained.

For a specific traffic scenario, after defining the transportation network and corresponding demand pattern, the sensitive parameters of the model can be identified by means of a sensitivity analysis then, decisions should be made on the MoP, GoF, and optimization algorithm variables.
Figure 1 depicts the comparison between the measurement of real values of the MoP and the simulated values with the help of a GoF measurement, which is the heart of both calibration and validation processes. The MoP choice must be made during the second step of the simulation study, linked strongly with the study objectives and in agreement with the operational aspects of the study. The choice of the GoF is technical and impacts the simulation results.

During calibration, the parameter optimization loop permits a progressive definition of an optimal set of parameters corresponding to the subsample of the data chosen for calibration (Set 1 in the figure). The second subsample (Set 2) is used for validation. To minimize the impact of the choice, the subsample for calibration must be as representative as possible of the various observed situations of the studied transportation system. One can ensure this independence by duplicating the process: use Set 2 for calibration, reset the parameter optimization process, compare the optimal parameter set with those obtained from the first calibration procedure, and compare the GoF values resulting from validation with Set 1.
Let us designate the parameter set obtained after successful calibration as pˆ. Calibration is incomplete without validation. Validation asks whether based on the model M above and the parameters pˆ estimated based on observed measurement d, how well another data set d′ approximated by the model. To answer this, once more e (p) must be computed, but now without the minimization of e (p) done for calibration.

$$e_v = e(.,\hat{p}) \qquad 2$$

In conclusion, the limited capacity of our models to depict reality makes it necessary, for their correct use, to follow a long iterative process of continuous verification of the steps.

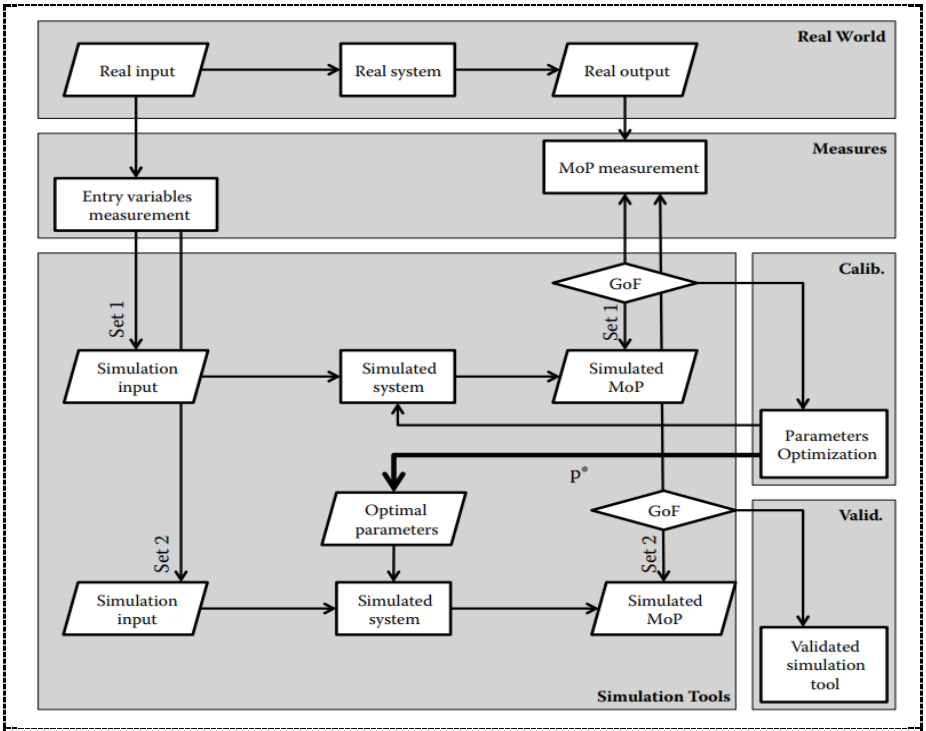

Fig. 1 Depiction of the global framework of calibration and validation. The first data set (Set 1) is used for the calibration procedure. The successive steps of the parameter optimization procedure are represented with bold arrows. The second subset is used for validation (Set 2). The set of optimal parameters (p^), resulting from calibration serves as model parameters during the validation. When comparing the measured and simulated measure of performance (MoP) during the validation process, one must determine the distance between them measured with the help of the same goodness of fit (GoF) of the same amplitude at the end of the calibration process. (Source: Ciuffo, B. et al., 2012. The Calibration of Traffic Simulation Models: Report on the Assessment of Different Goodness-of-Fit Measures and Optimization Algorithms. MULTITUDE Project and JRC Scientific and Technical Reports. With permission.)

Defining measure of performance (MoP)

The MoP must be defined in strong interaction with the application of the simulation tool in mind and, most importantly, it is objective. For example, if the objective of the network modification between reference and future scenarios is to improve the mode share, one has to use the mode percentage as the MoP instead of, for example, vehicle queue length. Also, the MoP must be observable in reality with the available measurement devices and must be easily calculable from the simulation outputs.

Defining and collecting a data set

After choosing a MoP adapted to the objective of the work, the next steps are designing the experiment and collecting the data. We can distinguish between single-valued MoP (characterizing the global behavior of the simulation scenario with a single value) and multivalued MoP reflecting the evolution of the system during the simulation duration.

When the latter MoP is chosen, care must be taken in defining aggregation periods: overly long aggregation periods may average out specific characteristics of the traffic system; aggregation periods that are too short may include too much noise. The question of the coherence of the definitions of the observed and simulated MoPs should be carefully addressed, because the definitions may be the sources of multiple errors.
A classic example is when an observed variable is defined as the arithmetic temporal mean speed and the output of the simulation is computed as the harmonic temporal mean speed. Specifically, if extreme single-valued MoPs are used (such as maximum waiting time), the question of the sample representativeness is crucial and the MoP comparison must be made with an appropriate number of simulation runs.
Table 1 proposes a list of MoPs, appropriate data collection procedures, and descriptions of their drawbacks and advantages.

Measures of performance (MoPs)
Typical MoPs used in the literature
The measure of the performance of a system can be defined as a collection of variables necessary to describe the status of the system (Law, 2007). Depending on a system's complexity, several MoPs may exist. However, since a system is usually observed for a specific purpose, the MoPs that best describe its status depend also on the analyses to be carried out. This concept also applies to transportation systems. Their specific characteristics thus influence the calibration and validation of a traffic simulation model.
Common MoPs for the calibration and validation of a traffic simulation model are time series of speeds and counts collected on a road section (possibly differentiated per lane) aggregated over a certain time interval (Hourdakis et al., 2003; Toledo et al., 2003; Kim and Rilett, 2003; Toledo et al., 2004; Chu et al., 2004; Dowling et al., 2004; Kim and Rilett, 2004; Brockfeld et al., 2005; Schultz and Rilett, 2005; Balakhrishna et al., 2007; Ma et al., 2007; Ciuffo et al., 2008; Lee and Ozbay, 2008; Menneni et al., 2008; Vaze et al., 2009; Punzo and Ciuffo, 2009; Ciuffo and Punzo, 2010b). The other fundamental variable of the traffic, namely, density, is used less frequently for calibration and validation purposes (Dowling et al., 2004; Ma et al., 2007) because it is more difficult to observe. Usually, density data are derived from the occupancy data from a single detector, which is not really accurate.
Other measures that are frequently used in more specific studies are queue lengths and turning flows at intersections (Ma and Abdulhai, 2002; Park and Schneeberger, 2003; Toledo et al., 2003; Dowling et al., 2004; Merritt, 2004; Shaaban and Radwan, 2005; Oketch and Carrick, 2005). Finally, more recently, based on the availability of more detailed information, point-to-point network travel times have been studied, both as aggregated measures and as distributions (Park and Schneeberger, 2003; Toledo et al., 2003; Chu et al., 2004; Dowling et al., 2004; Kim et al., 2005; Park and Qi, 2005; Oketch and Carrick, 2005; Hollander and Liu, 2008b; Vaze et al., 2009).
Vehicle trajectory data may also be used for the calibration of traffic simulation models. Because trajectory data are difficult to collect, they are rarely used. However, since data from the NGSIM project (NGSIM, 2011) have been made available, new applications are possible (Chiu et al., 2010)

Table 1 Data collection techniques associated with measures of performance (MoPs).

MoP type	*Name*	*Data collection device*	*Comment*
Multivalued collective variables	Flow	Any detector described in Section 2.2.2; typically a loop detector	Error can be as high as 20% on total flow; take care with aggregation period definition; if aim is to validate network-wide simulation (OD estimation), link coverage must be defined carefully
	Density	Any detector described in Section 2.2.2; typically a loop detector	Low-quality measurements; take care with aggregation period definition
	Speed	Any detector described in Section 2.2.2; typically a loop detector	Low-quality measurements, especially for low speeds; take care with aggregation period definition
	Queue length	Visual observation or set of loop detectors with speed observations	Human errors generally reduce data accuracy with visual observation; for loops, the quality of queue length measurement relates directly to spatial densities of detectors
	Travel times	Floating car data (FCD) Mean loop measured speeds	With FCD, size and representativeness of the sample must be verified carefully; quality of travel times data relates directly to spatial densities of loops
Multivalued individual variables	Trajectories	Camera placed at elevated position with automatic numeration	Numerical treatment can lead to errors that must be corrected
	Individual travel times	FCD or individual vehicle identification	Representativeness of observation sample (coverage, duration, number of days) is crucial
	Last decile waiting time	FCD or mean speed	Representativeness of observation sample (duration, number of days) is crucial
Single-valued	Last decile queue length	Manual observation or automatic loop detector	Human errors generally reduce data accuracy with visual observation; representativeness of observation sample (duration, number of days) is crucial
	Mode share	Manual observation	

(*continued*)

Criteria for MoPs selection

The criteria discussed below are useful for selecting MoPs for calibration and validation.

Context of the application: MoP statistics should be 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 sensors (occupancies, flows, speeds) may be more useful.

Independence: MoPs used for validation should be independent of any measurements used for calibration or estimating inputs to a simulated system. Origin–destination (OD) 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.

Error sources: In traffic analysis, a discrepancy between observed and simulated outputs can be explained by the following sources of error (Doan et al., 1999):

- Travel demands (OD flows)
- Route choices
- Driving behaviors
- Measurement errors in observed outputs

The first three sources contribute to errors in the simulated output. The last source 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 a validation. The locations and types of MoPs to be collected should be chosen to reflect errors from all these sources and reduce the effects of measurement errors as much as possible. Measurement locations should provide spatial coverage of all parts of a network.
Moreover, measurements near the network entry points will usually reveal errors in the OD flows with little effect from route choice and driving behavior models. As many measurement points as possible should be used to reduce the effects of measurement errors, assuming that the measurement errors are independent for different locations.

Traffic dynamics: MoPs and the level of temporal aggregation at which they are calculated (15 minutes, 30 minutes) should be chosen to facilitate testing whether or not the model correctly captures the traffic dynamics. This is especially true in network applications where both the temporal and spatial aspects of traffic are important.

Level of effort required for data collection: In practice, this is often the most constraining factor. Point measurements (flows, speeds, and occupancies) are often readily available from an existing surveillance system. Other types of measurements (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 various ways and surveillance systems may apply a number of time-smoothing techniques. It is therefore necessary to ensure that the simulated data are defined and processed the same way as the observed data.

A number of runs: Most traffic simulation models are stochastic (Monte Carlo) simulations. Hence, MoPs should is calculated from a number of independent replications. The two main

approaches to determining the number of replications are sequential and two-step (Alexopoulos and Seila, 1998). 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 (1978) suggested the following criterion:

$$R \geq R_i = max\left(2, \left(\frac{s_R(Y_i)t_{\alpha/2}}{d_i}\right)^2\right) \quad [3]$$

Here, R is the number of replications performed and R_i represents the minimum number of replications required to estimate the mean of Y_i with tolerance d_i. $s_R(Y_i)$ is the sample standard deviation of Y_i based on R replications and $t_{\alpha/2}$ is the 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:

$$R_i = \left(\frac{s_{R_0}Y_i t_{\alpha/2}}{d_i}\right)^2 \quad [4]$$

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.

Choice of appropriate statistical tests for comparing simulated and observed MoPs

The general simulation literature includes several 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, however, they may not be applicable because both the real and the simulated traffic processes of interest are nonstationary and autocorrelated. The choices of the appropriate methods and their application to the validation of traffic simulation models depend on the nature of the output data. The following methods and their outputs are considered:

- Single-valued MoPs (e.g., average delay, total throughput).
- 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 a system. Multivariate MoPs capture the temporal and/or spatial distribution of traffic characteristics and thus are useful to describe the dynamics at the network level. It may also be useful to examine the joint distribution of two MoPs (e.g., flows and travel times) to gain more information regarding the interrelationships of MoPs. The next section describes the statistical tests needed to calibrate and validate a simulation tool.

The goodness of Fit (GoF)

A number of goodness-of-fit measures can be used to evaluate the overall performance of a simulation model. This section reports on the studies of Hollander and Liu (2008a) and Ciuffo and Punzo (2010a). Note that GoF methods can be used both for calibration and validation, and are reported here for both applications. Table 2 lists GoF measures used for the calibration of traffic flow models and indicates the works in which they have been used. Ciuffo and Punzo (2010a)

analyzed 16 GoF measures, in particular, using response surface techniques, and their suitability for use as error functions in Equation (1) was investigated. The authors derived the following findings:

- Response surfaces confirm, as argued in the introduction, the complexity of the calibration problem and the need to use global optimization. Most of the response surfaces showed several local minima as well as wide areas with approximately constant values. Clearly, different choices in setting up a calibration problem generate different response surfaces.
- U_m, U_s, $-U_c$, and $-r$ (correlation coefficient) proved less suitable than other factors to be used in the objective function of the calibration. In particular, U_s, U_c, and $-r$ was always more irregular, showing several different minima in all plots.
- The values of 3 and 5 as thresholds in GEH3 and GEH5 evaluations proved to be very high, and, consequently, a wide area in all their plots generated a constant value of the objective function. This suggests that, at least in the transportation field, 5 is probably a too high threshold to assess whether two series of data show a good fit, as proposed by the Highway Agency (1996).
- All other GoFs showed similar behaviors on the whole, even if SE appeared to be the least sensitive GoF (the widest deep area around the "true" solution) but also the most regular around the minimum value.
- RMSE seemed to offer higher irregularity than the other GoFs around the optimum value.
- GEH1 probably showed the best capacity in highlighting the position of the minimum (in this case the threshold used seems to have a good impact).
- MAE and MANE show the highest flatness of the objective function around the global solution. This could represent a problem in identifying an effective stopping rule for any optimization algorithm.
- All these GoFs proved robust with respect to the introduction of noise to the data even with large errors.

Table 2 Measures of goodness of fit.

Name	*Measure*	*Comments*
Percent error *(PE)*	$\frac{x_i - y_i}{y_i}$	Applied to single pair of observed and simulated measurements or to aggregate network-wide measures
Squared error *(SE)*	$\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$	Most used GoF; low values show good fit; strongly penalizes large errors; serves as basis of least squares method which, according to Gauss-Markov theorem (Plackett, 1950), provides best parameter estimation for linear models with zero-mean, unbiased and uncorrelated errors
Mean error *(ME)*	$\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)$	Indicates systematic bias; useful when applied separately to measurements at each location; cannot be used in calibration because low values prevent good fit (same high errors with opposite sign will result in zero *ME*)
Mean normalized error *(MNE)* or mean percent error *(MPE)*	$\frac{1}{N}\sum_{i=1}^{N}\frac{x_i - y_i}{y_i}$	Indicates systematic bias; useful when applied separately to measurements at each location; cannot be used in calibration because low values prevent good fit (same high errors with opposite sign will result in zero *ME*)
Mean absolute error *(MAE)*	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i - y_i\rvert$	Not particularly sensitive to large errors
Mean absolute normalized error (*MANE*) or mean absolute error ratio (*MAER*)	$\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert x_i - y_i\rvert}{y_i}$	Not particularly sensitive to large errors; using absolute values would result in using same weights for all errors (it is preferable to assign more importance to high errors); gradient of absolute value analytical function has discontinuity point in zero; second most used GoF
Exponential mean absolute normalized error (*EMANE*)	$A\ \exp(-B\ MANE(x, y))$	Used as fitness function in genetic algorithm; *A*, *B* are parameters
Root mean squared error (*RMSE*)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$	Large errors heavily penalized; may appear as mean squared errors without root sign

Root mean squared normalized error (*RMSNE*) or root mean squared percent error (*RMSPE*)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - y_i}{y_i}\right)^2}$	Large errors heavily penalized; normalized measures (also *MANE*) are attractive GoFs since they allow model calibration using different measures of performance (only relative error is considered); instabilities due to low values among measurements in fraction denominator may affect their use
GEH statistic	$\sqrt{2\frac{(x_i - y_i)^2}{x_i + y_i}}$	Applied to single pair of observed and simulated measurements, not over data series; *GEH* < 5 indicates good fit; looks suspiciously like one term of χ^2 sum; when applied to time series, *GEH* < 5 for 75% of observed and simulated measurements indicates good fit (Highway Agency, 1996); *GEH* can also be used by counting the number of times its value is under a certain threshold; this avoids taking very small errors into account; threshold considered also defines error neglected); to remove dependence on number of available observations, number of times *GEH* is under threshold can be divided by total number of observations; in this way, indicator would be bounded between 1 (perfect fit) and 0 (worst fit)
Correlation coefficient (r)	$\frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$	
Theil's bias proportion (U_m)	$\frac{N(\bar{y} - \bar{x})^2}{\sum_{i=1}^{N}(y_i - x_i)^2}$	High value implies systematic bias; U_m = 0 indicates perfect fit; U_m = 1 indicates worst fit
Theil's variance proportion (U_s)	$\frac{N(\sigma_y - \sigma_x)^2}{\sum_{i=1}^{N}(y_i - x_i)^2}$	High value implies that distribution of simulated measurements is significantly different from that of observed data; U_s = 0 indicates perfect fit; U_s = 1 indicates worst fit
Theil's covariance proportion (U_c)	$\frac{2N(1-r)\sigma_x \sigma_y}{\sum_{i=1}^{N}(y_i - x_i)^2}$	Low value implies unsystematic error; U_c = 1 indicates perfect fit; U_c = 0 indicates worst fit; r is correlation coefficient

Theil's inequality coefficient (U)	$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i)^2}}$	Combines effects of all Theil's error proportions (U_m,U_s,U_c); $U = 0$ indicates perfect fit; $U = 1$ indicates worst fit
Kolmogorov-Smirnov test	$\max(\lvert F_x - F_y \rvert)$	F is cumulative probability density function of x or y; requires more detailed traffic measurements; when comparing time series, two series may show same distribution of values and thus good KS-test result, but may be completely different
Speed flow graph	$Y - (Y \cap X)$	Parameter combination allows simulated and observed speed flow diagrams to overlap
Moses and Wilcoxon tests		Detailed procedure described by Kim et al. (2005)

x_i = simulated measurements.

y_i = observed measurements.

N = number of measurements.

$\overline{x}$, $\overline{y}$ = sample average.

σ_x, σ_y = sample standard deviation.

X,Y area of speed flow diagram covered by simulated and observed measurements.

Sources: Adapted from Hollander, Y. and Liu, R. 2008a. *Transportation Research Part C,* 16(2), 212–231; Ciuffo, B. and Punzo, V. 2010a. Proceedings of 84th Annual Meeting. Washington, D.C.: Transportation Research Board. With permission.