Transportation (2008) 35:347–362
DOI 10.1007/s11116-007-9156-2

The principles of calibrating traffic microsimulation models

Yaron Hollander · Ronghui Liu

Published online: 15 January 2008

Abstract Traffic microsimulation models normally include a large number of parameters that must be calibrated before the model can be used as a tool for prediction. A wave of methodologies for calibrating such models has been recently proposed in the literature, but there have been no attempts to identify general calibration principles based on their collective experience. The current paper attempts to guide traffic analysts through the basic requirements of the calibration of microsimulation models. Among the issues discussed here are underlying assumptions of the calibration process, the scope of the calibration problem, formulation and automation, measuring goodness-of-fit, and the need for repeated model runs.

Keywords Calibration · Microsimulation · Traffic models

Introduction

Traffic microsimulation models (hereafter TMMs) are used by researchers and practitioners for a detailed analysis of the performance of transport systems. Estimates generated by a TMM are based on explicit representation of various aspects of individual behaviour. These aspects range from the driver's choice of route and departure time (Hu and Mahmassani 1997; Dia 2002; Liu et al. 2006; Zhang et al. 2006) to various features of driving behaviour, such as car following and lane changing (Toledo et al. 2005; Zhang and Kim 2005; Laval and Daganzo 2006; Ossen et al. 2006). Most TMMs include a large number of parameters that represent various characteristics of the travellers, the vehicles and the road system. These parameters must be calibrated before the TMM is used as a tool for prediction.

Y. Hollander (✉)
Steer Davies Gleave, 28-32 Upper Ground, London SE1 9PD, UK
e-mail: yarhol@gmail.com

R. Liu
Institute for Transport Studies, University of Leeds, 38 University Road, Leeds LS2 9JT, UK
e-mail: rliu@its.leeds.ac.uk

In the last few years there has been a wave of valuable research work that discussed procedures for TMM calibration, but there have been no attempts to identify general calibration principles based on their collective experience. The current paper concentrates on several related issues that are repeatedly brought up, based on around twenty different calibration methodologies. We hope our discussion can guide traffic analysts through some of the basic requirements of TMM calibration.

The paper is organised as follows. The following section discusses some basic concepts that the reviewed methodologies are based on. Next is a section that compares the different approaches to TMM calibration with respect to their scope. The subsequent section looks at the formulation of different calibration problems and techniques used to solve them. This is followed by a section that focuses on the different mathematical expressions used to minimise the discrepancies between simulation outputs and field data. Another section examines the number of times the TMM needs to be run at every stage of the calibration process. The subsequent section discusses the validation of calibration results. We conclude with some comments and practical recommendations.

Calibration conventions and underlying assumptions

A TMM typically consists of several sub-models, each of which tries to reproduce the mechanism of a single decision made by an individual traveller, such as the decision to change lane or to use a gap in the opposing traffic in order to enter an intersection. Each sub-model includes several parameters, and a complete TMM sometimes includes many dozens of parameters. Direct measurement of these parameters is very complicated, either because many of them represent subtle features that are hard to isolate, or because it requires extensive data collection. Works that directly study the value of a TMM parameter do exist, but we are not aware of any work where it was possible to do so for all parameters of a TMM.

In the calibration process, the parameters are adjusted so that the model outputs are similar to observed data. Due to the abovementioned difficulties, all the studies reviewed here do this using aggregate data, which do not describe the behaviour of individual drivers or vehicles. This type of data includes such measures as travel times, flows, speeds or queue lengths. When a model is calibrated using aggregate data, there is a risk that the result has limited behavioural power. The main justification for such calibration is the idea that the TMM is built of sub-models which are based on well-founded behavioural theories, and that the user of the full TMM only needs to verify the model works well for the situation of current interest. Nevertheless, we discuss later additional measures that should be taken to assure that the behavioural aspects of calibrated model are well-established.

It is worth noticing that some difference lies between the calibration of a TMM and calibration of other network models. The likely flaws in the forecasts made by an improperly-calibrated assignment model are of a local nature (e.g. erroneous flows at a specific location), while a TMM not adequately calibrated is prone to fail both locally and globally. There are other tools that go through a calibration process, such as volume-delay functions, which are similar to TMMs in that failure to calibrate them appropriately leads to wrong forecasts throughout the modelled network. However, such functions are aggregate tools by definition, and the parameters determined during their calibration are not said to represent behavioural features of individual drivers or vehicles.

TMMs, similar to other models, are not free from simplification; for instance, they often take limited account of the effects of roadside activity or road incidents. But when we

compare observed and simulated data during the calibration process, we unavoidably assume that the model includes all factors that exist in the actual network. This is a source of error in the calibrated parameters that we have no means to tackle; it should remind us that there is a need to constantly seek ways to improve the behavioural explanatory power of the TMM itself, independently of the calibration methodology.

The number of parameters we would ideally like to calibrate is high, but this is seldom possible because of the computational effort involved and limited data availability. All calibration methodologies reviewed here concentrate on a relatively small subset of parameters, but we found no study where this subset is chosen systematically. Analysts should remember that putting much effort in a powerful calibration methodology can bring little gain if some parameters that strongly influence the traffic measure of interest have not been included in the calibration subset.

The conventions and assumptions discussed in the previous paragraphs are common to all the calibration methodologies reviewed here. In the subsequent sections we discuss issues where major differences exist between the different procedures. A systematic comparison between the reviewed methods and case studies is presented in Table 1.

Scope of the calibration problem

All the studies summarised in Table 1 deal with TMM calibration (or validation), but in fact there are considerable differences between the problems they discuss. A first major difference lies in the definition of the problem itself: while some studies concentrate on the calibration of driving behaviour parameters only (e.g. Jayakrishnan et al. 2001; Ma and Abdulhai 2002; Hourdakis et al. 2003; Kim and Rilett 2003, 2004), some others (e.g. Toledo et al. 2003; Ben Akiva et al. 2004; Chu et al. 2004; Dowling et al. 2004; Oketch and Carrick 2005) incorporate this in a broader problem, where a route choice model and/or an origin-destination matrix are calibrated too. The authors who propose the broader problems present evidence that procedures which simultaneously tackle multiple problems result in stronger models, and that solving the sub-problems separately might lead to biased estimates.

The various sub-problems might differ from each other in their data requirements. For example, to calibrate driving behaviour parameters it is important to use data from a range of traffic settings (e.g. both arteries and minor streets), while for estimating the demand matrix it is more important that they are collected in a large number of locations (independent of the road type). Still, even if data availability is limited, solving a reduced problem does not in itself reduce the risk of bias. If the analyst intends to use an available set of data to solve various problems, doing this simultaneously is methodologically more correct.

Among the case studies that accompany the calibration methodologies there is substantial variation in the number of parameters being calibrated (from 3 to 19 parameters). The advantage in focusing on a smaller number of parameters is that it enables paying more attention to each parameter when its value is modified; in some cases this is done through a manual procedure (see more on this issue later in the paper). Bigger parameter subsets are normally calibrated using automated algorithms, and hence get more efficiently closer to an optimal solution, but also make it harder to follow changes in the value of each parameter. Overall, when an analyst chooses a set of calibration parameters, the ambitious task is to choose a number of parameters that is big enough to cover the various behavioural elements in the model, but small enough to enable paying individual attention to the

Table 1 Summary of the reviewed studies

Source	The problem	Formulation	Compared measures	No. of measurement sites	No. of calib. Parameters	No. of runs per evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Jayakrishnan et al. (2001)	Calibration	Verbal	N/A	N/A	N/A	N/A	None	PARAMICS	No
Ma and Abdulhai (2002)	Calibration	Genetic algorithm	Turning flow	20	5	1	470-node urban network	PARAMICS	Yes
Hourdakis et al. (2003)	Calibration	Verbal	Flow, speed	21	12	N/A	One freeway (20-km section)	AIMSUM	No
Park and Schneeberger (2003)	Calibration & validation	Verbal	Travel time, queue length	12	7	5	Urban arterial (12 intersections)	VISSIM	No
Toledo et al. (2003)	Calibration inc. route choice & OD matrix, & validation	Math. program	Speed, flow, travel time, queue length	N/A	N/A	1	Mixed freeway & urban network (size not specified)	MITSIMlab	Yes (not full)
Kim and Rilett (2003)	Calibration	Math. program (Simplex algorithm)	Flow	5	3–19	1	One freeway (23-km sections)	TRANSIMS and CORSIM	Yes
Barcelo and Casas (2004)	Calibration & validation	Verbal	N/A	Up to 700	N/A	N/A	Interurban network (totally 1800 road-km)	AIMSUM	No
Ben-Akiva et al. (2004)	Calibration inc. route choice & OD matrix	Math. program (Box's Complex algorithm)	Speed, density, flow	Up to 68	3–6	N/A	Freeway & 2 arterials/298-node urban network	MITSIMlab	Yes
Chu et al. (2004)	Calibration inc. route choice & OD matrix	Verbal (quantitative objective)	Flow, travel time	52	N/A	N/A	Mixed freeway (5–10 km sections)	PARAMICS	No

Table 1 continued

Source	The problem	Formulation	Compared measures	No. of measurement sites	No. of calib. Parameters	No. of runs per evaluation	Type of transport system in case study	TMM used	Solution algorithm?
Dowling et al. (2004)	Calibration inc. route choice	Verbal	Capacity, flow, travel time, queue length, density.	2	N/A	N/A	Freeway & urban (inc. arterial with 6 intersections)	N/A	No
Kim and Rilett (2004)	Calibration	Genetic algorithm	Flow	11	3–19	1	Two freeways (23-km sections)	TRANSIMS & CORSIM	Yes
Merritt (2004)	Calibration & validation	Verbal	Queue time, delay time, queue length	8	4	N/A	Inter-urban arterial (5km section)	CORSIM	No
Toledo and Koutsopoulos (2004)	Validation	None	Speed	4	None	10	One freeway (4-km section)	N/A	No
Kim et al. (2005)	Calibration	Genetic algorithm	Distribution of travel time	3	6	1	Urban arterial (1-km section)	VISSIM	Yes
Shaaban and Radwan (2005)	Calibration & validation	Verbal	Queue length, total travelled distance	2	4	20	Single segment with 2 signalised intersections	SimTraffic	No
Park and Qi (2005)	Calibration	Genetic algorithm	Travel time	1	8	5	One signalised intersection	VISSIM	Yes
Oketch and Carrick (2005)	Calibration inc. OD matrix & validation	Verbal	Link flow, turning flow, travel time, queue length	55	N/A	N/A	Urban network (8 km^2, 16 intersections)	PARAMICS	No
Brockfeld et al. (2005)	Calibration	Verbal (Simplex algorithm)	Speed	3	4–15	N/A	One freeway (1-km section)	10 different models	Yes (not presented)

value of each parameter, and also small enough to make the procedure computationally feasible. We recommend tackling this dilemma by undertaking an initial review of all the parameters of the TMM that is being used, and dividing them into the following groups:

1. Parameters whose values are relatively easy to measure directly (e.g. average car length or average bus boarding time per passenger).
2. Parameters whose values can be taken from previous studies that are applicable to the time and place being currently modelled.
3. Parameters whose influence on the outputs can be considered negligible. To check if a parameter belongs to this group, the TMM should be run several times with a range of values of the parameter, without varying the other parameters.
4. Parameters which are inappropriate to adjust because of the nature of the input data. For example, if the data are all taken from a motorway that is not used by buses, it is improper to include bus acceleration as a calibration parameter.
5. All the parameters that remain after omitting the above. Modellers should attempt to include all the parameters in this group in the calibration set.

There are also significant differences between the various calibration studies in terms of their geographical scale. Such differences exist both in the size of the simulation network and in the spread and density of data sources over this network. In terms of network size, the studies vary from a single intersection (e.g. at Ma and Abdulhai 2002) to an extensive metropolitan area (e.g. at Park and Qi 2005). The dispersion of sources of input data is sometimes as limited as two observation points in a medium-sized network (Dowling et al. 2004) or, in contrast, dozens of points in a network that is not much bigger (Chu et al. 2004; Oketch and Carrick 2005). In principle, many calibration methodologies can be implemented in various networks. But most methodologies are at least partially tailored to the scale in which they are later implemented: automated calibration is preferred if data is available from many measurement points (e.g. Ma and Abdulhai 2002 or Ben Akiva et al. 2004); comparison of multiple traffic measures is used in cases where there is much data but only from a small number of locations (e.g. Dowling et al. 2004; Merrit 2004). We discuss these issues further in subsequent sections, but would stress at this point that if the modeller wishes to use the TMM as a general tool for multiple purposes, this must be reflected in the geographical and typological scope of the calibration inputs.

The scope of the calibration problem also has to do with the choice of traffic measures used to compare observed data to the simulation outputs. Some of the proposed procedures use a single measure; for instance, Ma and Abdulahi (2002) and Kim and Rilett (2003, 2004) compare only flows. Some others use more than one measure, normally by performing a sequence of calibration sub-processes, each one of which uses a different traffic measure to calibrate a separate group of parameters. In the procedure proposed by Dowling et al. (2004) simulated and observed capacities are compared in the first stage to calibrate driving behaviour parameters, then flows are compared to calibrate route choice parameters, and finally all parameters are fine-tuned by comparing travel times and queue lengths. Hourdakis et al. (2003) start with calibrating global parameters (such as maximum acceleration and other vehicle characteristics) by comparing flows; then they calibrate local parameters (such as speed limits) by comparing speeds; an optional third calibration stage is suggested, where any measure chosen by the user can be compared. A similar multi-stage concept is also proposed by Chu et al. (2004).

Decomposing the main calibration problem into sub-problems is tempting since these can be solved more efficiently. If some parameters are of a local nature, or if different traffic measures seem more appropriate for calibrating different parameters (e.g. flows for

route choice parameters and speeds for car following parameters), then it is indeed sensible to assign different parameters to different sub-problems. But as mentioned previously, simultaneous optimisation reduces the risk that improved fit is achieved by adjusting the wrong parameters. Therefore, decomposition of the calibration problem should *replace* the simultaneous procedure only if the sub-problems are independent from each other; this is seldom the case. In most cases we would advise that the decomposition should *precede* the simultaneous calibration (as done, for instance, by Dowling et al. 2004).

Formulation and automation of the calibration process

Many of the discussions of TMM calibration (Jayakrishnan et al. 2001; Hourdakis et al. 2003; Park and Schneeberger 2003; Barcelo and Casas 2004; Dowling et al. 2004; Merrit 2004; Chu et al. 2004; Shaaban and Radwan 2005; Oketch and Carrick 2005) stress the need for consistent judgement but do not include an explicit formulation of a calibration problem. Thus they form an intermediate stage in the evolution of more cohesive concepts of calibration.

When an explicit calibration procedure is presented, it often has the form of an optimisation problem. It is sometimes presented as a mathematical program and in other cases only described verbally, but in most cases, at least an objective function is introduced. Systematic calibration procedures must use a solution algorithm which is normally an iterative process, and is often described as a flow chart.

Some optimisation approaches are repeatedly used for different TMM calibration studies. Several studies conduct the search for the best parameter set using a Genetic Algorithm (Ma and Abdulhai 2002; Kim and Rilett 2004; Kim et al. 2005; Park and Qi 2005). Various other studies use the Downhill Simplex Method (e.g. Kim and Rilett 2003) or the similar Box's Complex Algorithm (Ben Akiva et al. 2004). The choice of these concepts illustrates some features that calibration of a TMM commonly requires:

1. The optimisation technique should not restrict the number of variables.
2. A solution technique that requires fewer evaluations of the objective function during the process is preferable. This is due to the fact that the TMM often needs to be run more than once even for a single evaluation of the objective (we discuss this later), hence each evaluation is time-consuming.
3. The technique must not use derivatives of the objective function, because the objective of a calibration problem is not explicitly a function of the optimised parameters. Calculating derivatives in this case would require, again, a large number of TMM runs.

There is often a trade-off between the run time per iteration and the number of iterations required to reach a satisfactory solution. The Simplex Method and Genetic Algorithms represent two extreme cases in this respect. The former approach only requires a single evaluation of the objective function in most iterations, but improvement between iterations is slow. In contrast, the latter examines many candidate solutions concurrently but generally requires fewer iterations. The practice of using both methods suggests that one additional iteration of a Genetic Algorithm, in which the objective value is calculated for K new candidates, is not likely to improve the solution as much as multiple additional iterations of the Simplex Method, with K evaluations in total. Observe that almost all calibration methodologies (see Table 1) that use genetic algorithms only use one run of the TMM for each evaluation of the objective function; we find this inappropriate, as we discuss later. Park and Qi (2005) use a genetic algorithm and base each evaluation of the

objective function on five runs, but they only model a single intersection, and therefore do not face considerate run time problems. Overall, of the automated solution techniques used in the studies we reviewed, we find the Simplex Method the most appropriate, because it is more likely to minimise the total number of objective function evaluations. Nevertheless, note that innovative optimisation methods are intensively discussed in the literature in Mathematics and in Operational Research; we advise traffic analysts to constantly review recent developments in this field and seek improved solution methods for the TMM calibration problem.

As mentioned earlier, some calibration methods are partially or entirely manual. Merritt (2004) and Shaaban and Radwan (2005) predetermine several discrete values of each parameter and then check each feasible combination to find which one gives the best fit. Hourdakis et al. (2003) and Oketch and Carrick (2005) use automated search but calibration is only performed at one location at a time. Generally, manual calibration should be considered only if the expected application of the TMM is of a very limited scale. As discussed earlier, the risk that an automatic procedure might not be sensitive enough to the behavioural foundations of each parameter provides strong motivation for undertaking manual calibration. But in most cases we find that only an automated approach is practical.

Even when the calibration problem is formulated as an optimisation problem, it is unlikely to lead to a global optimum, due to the multidimensionality of the solution search space and the tendency of the observed data to exhibit various inconsistencies. Therefore, we urge modellers not to belittle the importance of the stages that precede or follow the solution procedure: proper definition of the likely range of values for each parameter (e.g. based on other studies), a clear validation methodology (as we discuss later), and a search for irregularities in the performance of the calibrated model by scrutinising the graphical display as it runs.

Measuring goodness of fit

At the heart of any calibration technique is a comparison between simulation outputs and observed measurements of various traffic measures. The measures of fit used for this purpose by the different calibration methodologies are summarised in Table 2. The notation used is explained below the table.

It is important to note that:

1. Most of the measures will identify poor fit between the central tendencies of the compared samples, while only few measures (especially Theil's indicators) are sensitive to the variance and covariance. The latter should be used when the calibrated model is to be used for analysis of variation.
2. Some measures (PE, ME, MNE) let errors with a similar size but a different sign balance each other. Such measures are useful for detecting systematic bias, but they should generally be avoided in calibration procedures.
3. Some measures (MAE, MANE) use the absolute value of the difference between the observed and simulated measurements; thus they give equal weights to all errors. Other measures (SE, RMSE, RMSNE) depend on the squared difference, and hence place a higher penalty on large errors. In the context of stochastic traffic modelling, penalising small errors is wrong; it might lead to an over-specified model, because minor fluctuations around the mean are in the nature of traffic phenomena. Using the squared error is more appropriate, and it is in fact surprising that none of the reviewed

Table 2 Measures of goodness-of-fit

Name	Measure	Used by	Comments
Percent error (*PE*)	$\frac{x_i - y_i}{y_i}$	Shaaban and Radwan (2005), Park and Qi (2005), Merritt (2004)	Applied either to a single pair of observed-simulated measurements or to aggregate networkwide measures
Squared error (*SE*)	$\sum_{i=1}^{N} (x_i - y_i)^2$	Ben-Akiva et al. (2004), Chu et al. (2004)	
Mean error (*ME*)	$\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)$	Toledo and Koutsopoulos (2004)	Indicates the existence of systematic bias. Useful when applied separately to measurements at each location
Mean normalized error (*MNE*)	$\frac{1}{N}\sum_{i=1}^{N} \frac{x_i - y_i}{y_i}$	Toledo et al. (2003), Toledo and Koutsopoulos (2004), Chu et al. (2004)	Indicates the existence of systematic bias. Useful when applied separately to measurements at each location
Mean absolute error (*MAE*)	$\frac{1}{N}\sum_{i-1}^{N} \lvert x_i - y_i \rvert$	Ma and Abdulhai (2002)	Not particularly sensitive to large errors
Mean absolute normalized error (*MANE*)	$\frac{1}{N}\sum_{i=1}^{N} \frac{\lvert x_i - y_i \rvert}{y_i}$	Ma and Abdulhai (2002), Kim and Rilett (2003), Merritt (2004), Kim et al. (2005)	Not particularly sensitive to large errors
Exponential mean absolute normalized error (*EMANE*)	$A \cdot \exp(-B \cdot \text{MANE})$ (A, B are parameters)	Kim and Rilett (2004)	Used as a fitness function in a genetic algorithm
Root mean squared error (*RMSE*)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2}$	Toledo and Koutsopoulos (2004), Dowling et al. (2004)	Large errors are heavily penalised. Sometimes appears as mean squared error, without the root sign
Route mean squared normalized error (*RMSNE*)	$\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{x_i - y_i}{y_i}\right)^2}$	Hourdakis et al. (2003), Toledo et al. (2003), Toledo and Koutsopoulos (2004), Ma and Abdulhai (2002)	Large errors are heavily penalised
GEH statistic	$\sqrt{\frac{2(x_i - y_i)^2}{x_i + y_i}}$	Barcelo and Casas (2004), Chu et al. (2004), Oketch and Carrick (2005)	Applied to a single pair of observed-simulated measurements. $GEH < 5$ indicates a good fit
Correlation coefficient (*r*)	$\frac{1}{N-1} \cdot \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$	Hourdakis et al. (2003)	
Theil's bias proportion (*Um*)	$\frac{N(\bar{y} - \bar{x})^2}{\sum_{i=1}^{N} (y_i - x_i)^2}$	Hourdakis et al. (2003), Barcelo and Casas (2004), Brockfeld et al. (2005)	A high value implies the existence of systematic bias. $Um = 0$ indicates a perfect fit, $Um = 1$ indicates the worst fit

Table 2 continued

Name	Measure	Used by	Comments
Theil's variance proportion (*Us*)	$\frac{N(\sigma_y - \sigma_x)^2}{\sum_{i=1}^{N}(y_i - x_i)^2}$	Hourdakis et al. (2003), Barcelo and Casas (2004), Brockfeld et al. (2005)	A high value implies that the distribution of simulated measurements is significantly different from that of the observed data. $Us = 0$ indicates a perfect fit, $Us = 1$ indicates the worst fit
Theil's covariance proportion (*Uc*)	$\frac{2(1-r)\cdot N\cdot \sigma_x\sigma_y}{\sum_{i=1}^{N}(y_i - x_i)^2}$	Hourdakis et al. (2003), Barcelo and Casas (2004)	A low value implies the existence of unsystematic error. $Uc = 1$ indicates a perfect fit, $Uc = 0$ indicates the worst fit. $\boldsymbol{r}$ is the correlation coefficient
Theil's inequality coefficient (*U*)	$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}y_i^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}}$	Ma and Abdulhai (2002), Hourdakis et al. (2003), Toledo and Koutsopoulos (2004), Barcelo and Casas (2004), Brockfeld et al. (2005)	Combines effects of all 3 Theil's error proportions (*Um, Us, Uc*). $U = 0$ indicates a perfect fit, $U = 1$ indicates the worst fit
Kolmogorov-Smirnov test	$\max(\lvert F_x - F_y \rvert)$	Kim et al. (2005)	F is the cumulative probability density function of x or y
Moses' test and Wilcoxon test	The detailed procedure is described by Kim et al. 2005		

x_i, simulated measurement; y_i, observed measurement; N, number of measurements; $\bar{x}$, $\bar{y}$, sample average; σ_x, σ_y, sample standard deviation

measures uses a power higher than 2. Alternatively, avoiding the effect of small errors is also possible by examining the probability density function (as in the Kolmogorov-Smirnov (aka K-S) test) rather than directly examining each individual observation.

Most measures involve summation of errors over a series of pairs of simulated and observed values. It is not always obvious how to create these pairs; each pattern of pairing might lead to a different level of fit, but unfortunately, none of the reviewed methods elaborates on this issue. If a test such as K-S is used, this problem is avoided, since individual observations are not examined explicitly.

The reviewed methodologies tend to consider the space of simulation outputs as one-dimensional, as only one index (denoted *i*) is used for the series of measurements in all the measures of fit in Table 2. But in fact the outputs form a multi-dimensional space; in different studies, the index *i* is used in different dimensions. The most common dimension is time (namely, each measurement is taken at a different time interval), as used by Toledo et al. (2003), Chu et al. (2004), Hourdakis et al. (2003), Kim et al. (2005) and others. But sometimes the series of measurements consists of values from different locations in the study network, and in other cases, each element in the series corresponds to a different vehicle. The fact that in each dimension the measures of fit have a different meaning is most apparent when examining variations between the measurements. For instance, calibrating a TMM by focusing on estimates of variation of the travel speed over different time

periods will probably lead to different results from calibration that focuses on speed variation between vehicles. In addition, it is clearly a mistake to compare observed speed variation between days to modelled speed variation between vehicles, even if all values refer to a single location. It is therefore important to choose not only a measure of fit that is suitable for the particular needs of every application, but also to use it in the appropriate dimension.

The methodology described by Park and Schneeberger (2003) does not use any of the measures in Table 2 but proposes an alternative concept, which estimates the model parameters without explicit calculation of the goodness of fit. This is done by creating a regression model where the calibration parameters are used as the explanatory variables and a traffic measure is the dependent variable. Calibration of the TMM is performed through seeking the parameter values with which the regression-based value of the traffic measure is the closest possible to the observed value. The procedure presented by Kim et al. (2005) is the only one where the evaluation of fit uses the family of statistical techniques known as *two-sample tests*. These tests are more commonly used for validation of the calibration results. It should be stressed that two-sample tests are as suitable as the other measures mentioned above for examining model fit. We return to this issue later.

Repeated runs

Unless the user explicitly disables all randomness features, a TMM will generate different outputs in every run, and therefore it is insufficient to only examine the results of a single run. The different calibration methodologies are not equally rigorous in this respect, and the number of runs per one evaluation of the fit of a single candidate solution varies from 1 to 20. Some of the methodologies use the following formula to determine the required number of runs (Merritt 2004; Toledo and Koutsopoulos 2004; Chu et al. 2004; Shaaban and Radwan 2005):

$$R = \left(\frac{s \cdot t_{\alpha/2}}{\bar{x} \cdot \varepsilon} \right)^2$$

where R, required number of model runs; s, standard deviation of the examined traffic measure; $\bar{x}$, mean of the traffic measure; ε, the required accuracy, specified as a fraction of $\bar{x}$; $t\alpha/2$, critical value of Student's t-test at confidence level α.

When R is calculated with this formula, an estimate of s is necessary as an input; but s is unknown prior to running the model. This is commonly tackled by sequentially running the model and re-calculating s and R till the number of runs that has already been performed is found high enough. If more than one traffic measure is used, a sufficient number of runs should be verified for each measure separately.

Note that the abovementioned formula only determines the number of runs that is required to achieve a certain level of confidence about the *mean* value of the estimated traffic measure. We are not aware of studies that seek the required number of runs for estimating other statistics but the mean, such as the variance; we discuss this in greater detail in a separate article (see Hollander and Liu 2008). If the modeller wishes to analyse other statistics but the mean, we recommend preceding the main estimation with an experiment that checks how many runs are needed in practice for the estimate of interest to converge to a stable value, and then use the higher between this empirical value and the value based on the formula given above.

Validation of the chosen parameter set

The validation stage is meant to confirm the predictive power of the calibrated model, using an independent set of data. The idea that validation must follow the calibration process is agreed by all, but a variety of techniques are used to implement it:

1. Visual validation (mentioned by Park and Qi 2005; Oketch and Carrick 2005; Toledo and Koutsopoulos 2004; and many others). This is done by inspecting the graphical presentation of the modelled network as the model runs, trying to spot any unusual behaviour. Most authors agree that visualisation is an efficient way to detect significant errors but cannot replace a more quantitative validation.
2. Validation using measures of fit, like those presented in Table 2. Toledo and Koutsopoulos (2004) point out that these measures are sometimes used for validation, but in practice we found very few studies that do this.
3. Statistical validation by arranging the simulated and observed measurements as two time series and then comparing these two series (Barcelo and Casas 2004).
4. Statistical validation using two-sample tests (Toledo and Koutsopoulos 2004; Barcelo and Casas 2004; Park and Qi 2005; Park and Schneeberger 2004). These are tools that examine the level of confidence about the hypothesis that the simulated and observed data have the same statistical properties. The most popular is the two-sample *t*-test, but many others are available (see Maisel and Gnugnoli 1972; and Kleijnen 1995). Some tests are parametric, i.e. designed for cases where we know the distribution of the measurements in the compared datasets. Nonparameteric tests do not rely on such information but are less powerful, namely they require more data for a certain level of confidence. Although we normally do not know what distribution describes the TMM outputs best, Kleijnen (1995) and others point out that it is common to make some distributional assumption in order to be able to use a parametric test.
5. Indirect statistical validation, by testing whether some product of the simulation outputs resembles the respective product of the field data. Toledo and Koutsopoulos (2004) build meta-models that capture relations between various traffic measures, such as the speed-flow relationship or the time evolution of flows; meta-models are estimated independently based on the simulated and the observed measurements, and it is then tested whether the two models are identical. Earlier versions of this approach were proposed by Kleijnen (1995) and Rao et al. (1998). A key drawback of this approach is that the estimation of the meta-model is in itself a potential source of error or bias.

The review of measures of fit, earlier in the paper, shows that the different measures used during the calibration process do not use any uniform scale or a consistent criterion to indicate good fit. In contrast, in the validation stage most authors use more statistically rigorous tests, which state well-defined levels of confidence. We find this unnecessary, because the uncertainty about the input data (e.g. travel demand) is very high, and the level of accuracy of the outputs is unknown. Validation is neither more nor less rigorous than calibration, and every test used in the calibration process can be also used for validation (or vice versa). The strengths and weaknesses of the various tests apply similarly to validation and calibration.

Nevertheless, it is important to ensure that the validation test does not simply repeat what has already been tested in the calibration process. The basic requirement, which every calibrated TMM must meet, is that it can be successfully validated with a new set of data of the same type. For example, a model that has been calibrated with queue length data from

Table 3 Guidelines for TMM calibration

Stage	Action
Pre-calibration	Consider improving the behavioural explanatory power of the TMM itself, independent of the calibration methodology.
Determining scope of problem	If the modeller intends to use the available dataset to solve various problems, doing this simultaneously is preferable.
	Decomposing the calibration problem into sub-problems is recommended if some parameters are of a local nature, or if different traffic measures are more appropriate for calibrating different parameters (the latter is very common but is subject to data availability).
	If the sub-problems are independent, the decomposed problem can replace the full calibration problem. Otherwise, it should be followed by simultaneous calibration.
Choosing calibration parameters	Do not calibrate:
	• parameters whose values can be determined directly by observation of measurement;
	• parameters whose values can be taken from previous studies that are applicable to the time and place being modelled;
	• parameters whose influence on TMM outputs can be considered negligible;
	• parameters that do not have effect on the observed measurements in the available dataset.
	Attempt to calibrate all remaining parameters.
Choosing measure of goodness-of-fit	Do not use measures that let errors with a similar size but a different sign to balance each other (e.g. PE, ME, MNE) unless there is only interest in detecting systematic bias.
	Prefer measures in which the simulation error is squared (e.g. SE, RMSE, RMSNE), or raised to a higher power, to measures that give small errors equal weights (e.g. MAE, MANE). Use a consistent method for pairing simulated and observed values.
	Alternatively, compare the distribution of measurements (e.g. K-S test or other two-sample tests) and thus avoid the need for pairing and the unwanted effect of small errors.
	Use measures with sensitivity to the distribution of model outputs if the model is to be used for analysis of variation.
	Make sure that the dimension in which the observed measurements form a series is the same dimension from which the simulated outputs are taken.
Specifying constraints	Define carefully the feasible range of each parameter
Specify calibration procedure	Manual calibration is advantageous for a small number of parameters (say, up to 5) but is not practical in other cases.
	If an automated process is undertaken, a preferable solution procedure should be:
	• suitable for a multidimensional problem;
	• one that does not use derivatives;
	• one that requires few evaluations of the objective function per iteration.
	Consider reviewing recent developments of optimisation methods.

Table 3 continued

Stage	Action
Determining required number of runs	To calculate mean values, use the formula given above. To analyse other statistics, either seek theoretical guidance on required number of runs, or find the number of runs that are needed for the statistic of interest to converge to a stable value empirically.
Statistical validation	Good fit must be shown when outputs of the calibrated model are compared to a new set of data of a similar type to the data used for calibration. A higher standard of validation is reached if it can be shown that good fit is found with other types of data, but this cannot be taken for granted.
Additional validation	Search for irregularities by inspecting the graphical display as the model runs. Compare parameter values to other sources.
Implementation	Use the calibrated model in similar settings to those used for calibration: geographical scope, road/intersection types included, traffic measures calculated, the level of sensitivity to other statistics apart from the mean, and the dimension in which variation is measured. Your calibrated model is not a credible tool in other settings.

one set of intersections must pass the validation test using queue length data from another set of intersections. A higher standard of validation is reached if it can be confirmed that the model calibrated with queue length data can also give good estimates of times or flows. But in practice the relations between the different dimensions in the TMM itself are not always reliable enough to achieve such standard. It should therefore be mainly stressed that if validation is undertaken in the same dimension that has been examined during calibration, the TMM can be later used reliably as a tool for prediction only in this dimension.

Conclusions

We have presented a review of methodologies for calibration of traffic microsimulation models, discussed similarities and differences between them, and made some recommendations. The reviewed methodologies differ from each other both in principle issues, such as their objective and their scope, and in technical issues, such as their formulation and solution approach.

We find that many calibration methodologies are not rigorous enough in terms of the number of repetitions of the model used throughout the calibration procedure. Despite their high time consumption, repeated runs are essential because the output of a single TMM run is a very small sample from an unknown distribution. We also find that most authors mainly use traffic microsimulation for estimating mean values of various traffic measures, despite the fact that the stochastic nature of microsimulation creates an excellent opportunity for examining their variation[1]. Another finding is that modellers tend to use a different type of statistical tools for calibration and validation, while in fact the tools commonly used for calibration can be efficiently used for validation and vice versa.

[1] We dedicate a separate discussion to the use of traffic microsimulation for the estimation of other statistics but the mean. See Hollander and Liu 2008

Our main recommendations for microsimulation modellers who embark on a calibration process are listed in Table 3. Note that the table does not attempt to prescribe a full calibration procedure, since as we have illustrated, various approaches are available, and which approach is used depends on the particular application.

Acknowledgements The authors are grateful to four anonymous reviewers for their valuable comments.

References

Barcelo, J., Casas, J.: Methodological notes on the calibration and validation of microscopic traffic simulation models. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Ben-Akiva, M.E., Darda, D., Jha, M., Koutsopoulos, H.N., Toledo, T.: Calibration of microscopic traffic simulation models with aggregate data. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Brockfeld, E., Kühne, R.D., Wagner, P.: Calibration and validation of microscopic traffic flow models. In: Proceedings of the 84th TRB Annual Meeting, Washington, DC (2005)

Chu, L., Liu, H.X., Oh, J.S., Recker, W.: A calibration procedure for microscopic traffic simulation. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Dia, H.: Agent-based approach to modelling driver route choice behaviour under the influence of real-time information. Transp. Res. **10C**, 331–349 (2002)

Dowling, R., Skabardonis, A., Halkias, J., McHale, G., Zammit, G.: Guidelines for calibration of microsimulation models: framework and application. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Hollander, Y., Liu, R.: Estimation of the distribution of travel times by repeated simulation. Transp. Res. C, (2008, in press)

Hourdakis, J., Michalopoulos. P.G., Kottommannil, J.: Practical procedure for calibrating microscopic traffic simulation models. Transp. Res. Rec. **1852**, 130–139 (2003)

Hu, T.-Y., Mahmassani, H.S.: Day-to-day evolution of network flows under real-time information and reactive signal control. Transp. Res. **5C**, 51–69 (1997)

Jayakrishnan, R., Oh, J.S., Sahraoui, A.E.K.: Calibration and path dynamics issues in microscopic simulation for advanced traffic management and information systems. Transp. Res. Rec. **1771**, 9–17 (2001)

Kim, K.O., Rilett, L.R.: Simplex-based calibration of traffic microsimulation models with intelligent transportation systems data. Transp. Res. Rec. **1855**, 80–89 (2003)

Kim, K. O., Rilett, L. R.: A genetic algorithm based approach to traffic micro-simulation calibration using ITS data. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Kim, S. J., Kim, W., Rilett, L. R.: Calibration of micro-simulation models using non-parametric statistical techniques. In: Proceedings of the 84rd TRB Aannual Meeting, Washington, DC (2005)

Kleijnen, J.P.C.: Verification and validation of simulation models. Eur. J. Oper. Res. **82**, 145–162 (1995)

Laval, J.A., Daganzo, C.F.: Lane-changing in traffic streams. Transp. Res. **40B**, 251–264 (2006)

Liu, R., Van Vliet, D., Watling, D.: Microsimulation models incorporating both demand and supply dynamics. Transp. Res. **40A**, 125–150 (2006)

Ma, T., Abdulhai, B.: Genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters. Transp. Res. Rec. **1800**, 8–15 (2002)

Maisel H., Gnugnoli G.: Simulation of discrete stochastic systems. Science Research Associates Inc., Chicago (1972)

Merritt, E.: Calibration and validation of CORSIM for Swedish road traffic conditions. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Oketch, T., Carrick, M.: Calibration and validation of a micro-simulation model in network analysis. In: Proceedings of the 84rd TRB Annual Meeting, Washington, DC (2005)

Ossen, S., Hoogendoorn, S.P.,Gorte, B.G.H.: Interdriver differences in car-following: a vehicle trajectory-based study. Transp. Res. Rec. **1965**, 121–129 (2006)

Park, B., Qi, H.: Development and evaluation of simulation model calibration procedure. In: Proceedings of the 84rd TRB Annual Meeting, Washington, DC (2005)

Park, B., Schneeberger, J.D.: Microscopic simulation model calibration and validation—case study of VISSIM simulation model for a coordinated signal system. Transp. Res. Rec. **1856**, 185–192 (2003)

Rao, L., Owen, L., Goldsman, D.: Development and application of a validation framework for traffic simulation models. In: Proceedings of the Winter Simulation Conference, Washington, DC (1998)

Shaaban, K.S., Radwan, E.: A calibration and validation procedure for microscopic simulation model: a case study of sim traffic arterial streets. In: Proceedings of the 84rd TRB Annual Meeting, Washington, DC (2005)

Toledo, T., Choudhury, C.F., Ben-Akiva, M.E.: Lane-changing model with explicit target lane choice. Transp. Res. Rec. **1934**, 157–165 (2005)

Toledo, T., Koutsopoulos, H.N.: Statistical validation of traffic simulation models. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC (2004)

Toledo, T., Koutsopoulos, H.N., Davol, A., Ben-Akiva, M.E., Burghout, W., Andreasson, I., Johansson, T., Lundin, C.: Calibration and validation of microscopic traffic simulation tools – Stockholm case study. Transp. Res. Rec. **1831**, 65–75 (2003)

Zhang, H.M., Kim, T.: A car-following theory for multiphase vehicular traffic flow. Transp. Res. **39B**, 385–399 (2005)

Zhang, M.H., Ma, J., Dong, H.: Calibration of departure time and route choice parameters in microsimulation with macro measurements and genetic algorithm. In: Proceedings of the 85th TRB Annual Meeting, Washington, DC (2006)

Author Biographies

Yaron Hollander is a transport analyst, working for Steer Davies Gleave in London. His work combines advanced demand modelling, Stated Preference, appraisal, design of public transport systems, transport policy and network modelling. He completed his PhD at the Institute for Transport Studies in Leeds, and previously worked for the Technion – Israel Institute for Technology; for the Israeli Institute for Transportation Planning and Research; and for the public transport department at Ayalon Highways Co.

Ronghui Liu is a Senior Research Fellow at the Institute for Transport Studies, University of Leeds. Her main research interests are modelling of traffic and microsimulation of driver behaviour and dynamical systems. She develops and applies network microsimulation models to a wide range of areas from transport policy instruments such as road pricing, to public transport operations and traffic signal controls.