Are Microsimulation Models Random Enough?
A Comparison of Modeled and Observed Stochasticity

John W. Shaw*
Researcher
University of Wisconsin Traffic Operations & Safety Laboratory
B243 Engineering Hall, 1415 Engineering Drive, Madison, WI 53706
Phone: (414) 263-9049
Email: jwshaw@wisc.edu

David A. Noyce
Professor
University of Wisconsin - Madison
1204 Engineering Hall, 1415 Engineering Drive, Madison, WI 53706
Phone: (608) 265-1882, Fax: (608) 262-5199
Email: noyce@engr.wisc.edu

Word Count (3114 + 5 Figures × 250 + 3 Tables × 250) = 5114 words
Revised from the Original Submittal
* Corresponding Author
ABSTRACT

Traffic microsimulation models contain many parameters which are randomized during the simulation process. The randomized parameters influence trip generation, vehicle characteristics, gap acceptance, and other aspects of simulated driver behavior, which ultimately affect output Measures of Effectiveness (MOEs) such as modeled traffic volumes and speeds. As part of a microsimulation model review process, it became necessary to investigate whether the statistical distribution of a specific model’s outputs matched the observed distributions for the study area. The sponsoring agency’s expectations (prior to the research) were that modeled and observed stochasticities were broadly similar, and that individual model runs could be likened to traffic conditions on “different days.” Comparison of the statistical distributions of volume and speed data showed that field and model stochasticities differed substantially: statistical dispersions in the microsimulation model were much narrower than the field observations. The modeled data roughly followed a normal distribution, but the field volume data was skew-distributed and the field speed data had a statistically multi-modal (double-hump) distribution. Traditional modeling practices often assume that the field data is normally distributed, which was certainly not the case in the congested urban freeway corridor used for this evaluation. The findings have implications for the model calibration process, the determination of the minimum number of simulation runs, and the interpretation of what comprises a “typical” set of traffic conditions in the model.

INTRODUCTION

Traffic microsimulation models such as Aimsun, Corsim, Quadstone Paramics (Q-Paramics), SAIS Paramics (S-Paramics), and Vissim contain many parameters which are randomized during the simulation process. Depending on the software, these parameters influence the number of vehicles generated by each zone-to-zone origin-destination pair, the type of vehicle that is generated in the simulation (e.g. small car, large car, light-duty truck, heavy-duty truck, bus, etc.), and driver-related parameters such as willingness to accept small gaps when changing lanes, perception-reaction time, and willingness to use second-best routes. (Q-Paramics and S-Paramics are different software packages derived from a common predecessor).

The intent of the randomization process is to assure that each model reflects a fairly wide range of the possible combinations of drivers, vehicles, and route choices that exist in the real roadway system that is being studied. Real-world driver behavior reflects individual psychological and physiological differences, and the randomization process is designed to reflect these differences, as well as the physical variations in the size, weight, and performance of vehicles using the roadway system. A guideline produced by the publisher of S-Paramics (1) explains the software design as follows:

In conjunction with release rates, S-Paramics uses random number generators in its simulations. This means every [model] run will vary slightly. Hence, performing multiple runs and combining the data is statistically more robust than relying on a single run.

A single run of the model may produce a random event that increases delays in a certain area of the model, leading to re-routing or increased queues, delays, and journey times. An example would be the simultaneous arrival of a number of HGVs [heavy trucks] at a junction [intersection]. A single run is therefore not necessarily representative of typical traffic conditions.

If multiple runs are performed using different seeds the stochastic nature of the release of vehicles that means that unusual conditions (e.g. large numbers of simultaneous HGV arrivals) will not skew
the results of the model. This enables users to derive statistically relevant average traffic conditions. [emphasis original].

 Undertake an appropriate number of runs to ensure that your results are statistically robust.

Similarly, the user guide for Q-Paramics Version 3 (2) read as follows:

[The software] has a choice of two random number generators, either Marsaglia or Merseene Twister... Users are recommended to run the model with different seed values to test the sensitivity of the model. A table showing the fluctuation in [vehicle] release rates can indicate the stability of the model results for a particular run... It is advisable for the user to assess these variations [between runs] and decide if they lie within reasonable limits for each particular model.

By default [Q-Paramics] produces a random probability of a vehicle release [from a traffic analysis zone] which may be seen as representative of daily or hourly changes in traffic flow... It is advisable to test the sensitivity of model results using a range of [random number generator] seed values. [emphasis added].

Volume III of the FHWA Traffic Analysis Toolbox (TAT-III) provides guidelines for applying traffic microsimulation modeling software. (3) Appendix B of the current edition (written in 2004 and now under revision) explains the need for stochasticity as follows:

Multiple repetitions of the same model are required because microsimulation results will vary depending on the random number seed used in each run. The random number seed is used to select a sequence of random numbers that are used to make numerous decisions throughout the simulation run (Should a vehicle be loaded on the network now or later? Should the driver be aggressive or timid? Which vehicle type should it be? Will the driver chooses the shortest path or a slightly longer one instead?). The outcomes of all of these decisions will affect the simulation results. The results of each run will usually be close to the average of all of the runs, however, each run will be different from the other.

TAT-III also states that “microsimulation model results can vary by 25% between model runs.” The document goes on to provide guidance on how to calculate the minimum number of model runs required to achieve a desired level of statistical confidence for Measures of Effectiveness (MOEs) generated by microsimulation models. Three important assumptions are implicit in the guidance:

• The randomness in the models is fundamentally similar to the randomness of field conditions.
• MOEs are normally distributed (i.e. statistical measures such as the average, standard deviation, and Student’s T-test are relevant when computing confidence intervals).
• Observed values are independent.

A recent microsimulation model review for a freeway design project in Milwaukee, Wisconsin provoked questioning of some assumptions which often underlie efforts to calibrate and validate simulation models:

• Are model outputs (such link traffic volumes and travel times) normally distributed?
• Are the same parameters normally distributed in the field?
Do the outputs from microsimulation models contain as much variation as day-to-day field conditions? Should they?

Although most modeling practitioners currently make model calibration adjustments by hand, in recent years a number of research advances have been made toward automating the microsimulation model calibration process, e.g. (4), (5), though questions of uniqueness and over-fitting remain. Typically, these efforts focus on adjusting model parameters systematically to close the gap between simulated and field values for key MOEs. But in most commercial microsimulation packages, only the coefficients for model parameters can be specified by the user. The underlying parameter distribution is hard-coded in the software: if a parameter is modeled as normally-distributed but the actual distribution in the field is a skew distribution, the model user can adjust settings in an effort to minimize differences, but an exact match generally cannot be achieved. The hard-coded nature of the distributions serves as a counterpoint to Hollander and Liu’s observation that, "We...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.”(6)

METHODOLOGY

We used a Q-Paramics microsimulation model of the I-94 corridor in Milwaukee, Wisconsin as a case study for comparing the actual stochasticity of speeds and traffic volumes with field data from reliable traffic detector stations at selected locations along the same corridor. The project limits included about 4 miles of I-94 from approximately 27th Street to 85th Street and about 2 miles of US 41/341 (Miller Park Way) adjoining the Stadium Interchange. The corridor currently experiences significant congestion during the AM and PM peak hours and is under consideration for capacity expansion.

The Paramics model for the I-94 corridor is intended to represent “typical” peak-hour traffic conditions in the corridor. More specifically, the model builders established target traffic volumes based on the annual average hourly weekday volumes for 2009 (the model was developed in 2013-14, but newer data was not representative due to construction). At the request of the client’s traffic forecasting unit, the model builders applied small uniproportional adjustments to these volumes to assure that the peak directional volumes at selected locations matched the 250th Highest Hourly Directional Volumes (K250). Twenty model runs were completed to establish the distribution of traffic volumes and link speeds, which were considered the most important MOEs for the project.

The researchers compared the traffic volume distributions of these MOEs with the observed hourly field volumes for all weekdays (Monday-Thursday) in 2009. Traffic speed distributions from the model were compared with 5-minute average speeds spanning 32 weekdays (Monday-Thursday) in May and October 2009. May and October are considered operationally “typical” for the corridor: volumes are close to the annual average and speeds are usually unaffected by severe weather. May 25, 2009 was excluded from the speed data set because it was a holiday. The volume and speed data sets originated separately: the volume data set is “Planning” data from Automatic Traffic Recorder (ATR) stations (archived at hourly intervals), and the speed data set is “Operations” (freeway traffic management system) data from side-fired radar detectors (archived at 5 minute intervals).
RESULTS

Although there are a number of detector stations along the test corridor, for brevity a small number of representative locations are discussed below.

Traffic Volume. As shown in the Figure 1 histogram, the modeled hourly traffic volumes are tightly clustered around the target volume set by the modeling team, and roughly follow a normal distribution. Figure 2 compares the modeled and observed volumes for the same location and time period (AM peak west of 37th Street). It is evident that the spread of the observed values (900 to 7138 veh/hr) is far larger than the spread of the modeled volumes (5721 to 5991 veh/hr). The observed volumes follow a skew distribution, with a long left tail that is likely influenced by incidents and a short right tail that probably reflects periods of high demand (e.g. professional sports events) and capacity constraints. Additional statistical comparisons of the two data sets can be found in the boxes on Figure 2; notably the skewness of the modeled volumes (a measure the symmetry of the distribution) was +0.62, while the skewness of the observed field volumes was –2.69 (a standard normal distribution has zero skewness; negative values indicate left skew and positive values indicate right skew).

This exploration of the volume distribution also suggests that the common practice of using average hourly volume as the target for microsimulation does not result in a robust indicator of the “typical” operating conditions for this corridor: the average is skewed by incidents on the low end, and capacity constraints on the high end. In this instance the median hourly volume would probably have been a more robust measure of the “typical” traffic volume for the corridor.

<table>
<thead>
<tr>
<th>Modeled Volume (Vehicles Per Hour)</th>
<th>Proportion of Model Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>10%</td>
</tr>
<tr>
<td>550</td>
<td>15%</td>
</tr>
<tr>
<td>600</td>
<td>20%</td>
</tr>
<tr>
<td>650</td>
<td>25%</td>
</tr>
<tr>
<td>700</td>
<td>30%</td>
</tr>
</tbody>
</table>

Figure 1. Histogram of modeled hourly traffic volumes on eastbound I-94 near 37th Street from 20 simulation runs of the AM Peak model.
Figure 2. Histogram comparing modeled (red) and observed (blue) hourly traffic volumes on I-94 near 37th Street for the AM Peak period.

Speed. Generating link speed distributions is not straightforward in Q-Paramics, so our analysis of the speed data compared model-wide speeds with the observed speeds at two representative sites. For simplicity, speeds during the 7:30-7:35 AM time interval (near the middle of the simulation period) are used here for comparative purposes. As shown in Figure 3, the modeled (red) traffic speeds fall into a much narrower range than the observed (blue) values. The observed values are seen to have a statistically multi-modal distribution, i.e. the most common speeds observed at the detector are approximately 35 mph and approximately 50 mph. Similarly, Table 1 indicates that the maximum, minimum, mean, and median speed values from the model fall into a narrow range, while the field observations from two sites are much more broadly distributed. Notably, the standard deviation of the field observations was much higher than the standard deviation of the modeled values.
Figure 3. Histogram of modeled (red) and observed (blue) 5-minute speed distributions for the I-94 model near the Hawley Road detector station.

Table 1. Comparison of modeled and observed 7:30-7:35 AM speeds in the I-94 test corridor.

<table>
<thead>
<tr>
<th>Site</th>
<th>I-94 at Hawley Rd (Observed)</th>
<th>I-94 at 35th Street (Observed)</th>
<th>Network-Wide (Modeled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>37.4 mph</td>
<td>42.1 mph</td>
<td>43.5 mph</td>
</tr>
<tr>
<td>Median</td>
<td>38.3</td>
<td>42.2</td>
<td>43.4</td>
</tr>
<tr>
<td>Standard Dev</td>
<td>9.8</td>
<td>7.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Min</td>
<td>15.0</td>
<td>24.4</td>
<td>41.5</td>
</tr>
<tr>
<td>Max</td>
<td>53.0</td>
<td>59.0</td>
<td>45.6</td>
</tr>
<tr>
<td>% Difference</td>
<td>254%</td>
<td>142%</td>
<td>10%</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.22</td>
<td>0.17</td>
<td>0.09</td>
</tr>
<tr>
<td>Excess Kurtosis</td>
<td>-0.90</td>
<td>0.02</td>
<td>-1.09</td>
</tr>
</tbody>
</table>

As noted in the introduction and illustrated in Figure 4, TAT-III states that “microsimulation model results can vary by 25% between model runs.” Figure 5 shows a similar graph generated using one-minute total volume data from the I-94 Paramics simulation; in contrast to the TAT-III information the speed variation between runs ranged from 8.5% to 10.2%. This may reflect differences between Q-Paramics and the unspecified software package that was used for the TAT-III example, or perhaps differences in the way the two models were calibrated.
Figure 4. Variation of vehicular speeds in a model discussed in FHWA Traffic Analysis Toolbox Volume III. (3)

Figure 5. Mean speeds for all modeled vehicles in the I-94 model (1 minute temporal resolution).
HOW MANY MODEL RUNS?

TAT-III provides advice on the minimum number of model runs to be used in a microsimulation modeling process. The current (2004) edition of the document bases this calculation on Student’s T-test, which implicitly assumes that the data set in question is normally distributed. (In this test corridor the distribution of volume and speed data from model runs generally was normally distributed, but the field data was not.)

Table 2 shows the computation of the minimum number of model runs based on the recommendations in TAT-III using the traffic volume data. In this case, just 2 model runs would be required if the computation is based on the modeled volumes for the freeway mainline (the statistical dispersion for ramp volumes was broader). If the same computation is performed based on the observed stochasticity, 169 model runs would be recommended – though in practice it is unlikely they would yield a statistical spread much wider than what was achieved with 20 runs. Table 3 repeats this computation using the speed data. The results are similar, as might be expected since traffic speed and volume are related. Experienced microsimulation modelers will recognize that neither answer seems reasonable: 2 model runs is not enough to rule out the effects of model glitches (such as stuck vehicles) and for most projects 169 model runs is currently infeasible due to runtimes and associated budgetary and schedule constraints.

Table 2. Minimum number of model runs based on volume data for a 95% confidence interval of ±275 veh/hr.

<table>
<thead>
<tr>
<th>Standard Deviation</th>
<th>Model</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>73</td>
<td>902</td>
</tr>
<tr>
<td>Minimum Repetitions</td>
<td>2</td>
<td>169</td>
</tr>
</tbody>
</table>

Table 3. Minimum number of model runs based on speed data for a 95% confidence interval of ±5 mph.

<table>
<thead>
<tr>
<th>Standard Deviation</th>
<th>Model</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>1.35</td>
<td>8.75</td>
</tr>
<tr>
<td>Minimum Repetitions</td>
<td>2</td>
<td>48</td>
</tr>
</tbody>
</table>

DISCUSSION

Much of today’s “standard” modeling practice has been inherited from an earlier era when traffic data was scarce. As comprehensive year-round traffic data at fine temporal resolution becomes more widely available, modeling practitioners and researchers can re-examine and improve upon established procedures and assumptions. The following observations can be made for this test corridor:

- Field-observed speed and volume data were not normally-distributed
Field observations were not independent. Spatial and temporal autocorrelation was ubiquitous: when traffic volume is high at the west end of the model, it is also high at the east end. Speed disturbances propagate along the corridor.

Practitioners would have benefitted from enhanced guidance on data preparation and filtering methods that take the effects of incidents and special events into consideration when selecting target volume sets.

The results of this work suggest that (at least for congested corridors), more needs to be done to clarify what it means for a microsimulation model to represent “typical” conditions. In locations with congestion, the standard practice of using average volume and speed values to set modeling targets may cause the targets to be skewed by outliers (e.g. low flows caused by incidents or high flows caused by special events). Alternative statistical measures for setting targets, such as the use of median values, could overcome some of these potential concerns. (In Wisconsin’s case, computing median values would require software changes in the legacy systems used to manage traffic volume data).

In this case study, run-to-run variations in traffic volumes and link speeds fell into a much narrower range than the day-to-day variations observed in the field. Increasing the internal stochasticity of the modeling software is probably not the right solution: commercial software developers need to deliver products that produce reasonably consistent results. Therefore, replicating the full range of conditions that occur in the field would require much more than running the model with a larger number of random seeds – varying input parameters such as traffic demand levels would be necessary – and this in turn could greatly increase the level of effort required to complete a microsimulation study. The value of such processes requires further study, bearing in mind that in the vast majority of cases, models do not need to reproduce every field condition in order to be useful.

Some model users have fallen into the habit of thinking of individual simulation runs as representing “different traffic days.” This research suggests that is an oversimplification. Running multiple seeds is an essential part of the microsimulation modeling process: the software is stochastic by nature; multiple runs are necessary to obtain reliable results. Avoidance of results skewed by internal modeling anomalies (such as the simultaneous release of several heavy trucks at one location) is a sufficient reason for making multiple runs.

The research suggests that the application of simple statistical measures to compute the minimum number of runs also carries risks: the inner workings of commercial traffic microsimulation models are complicated; there are many internal parameters (not all of which are visible to the end user), and internal randomizations may be tempered by cut-offs that software developers have deemed necessary to stabilize the output. As a result, model outputs may not always follow distributions that are statistically normal. Hollander and Liu (7) discuss a number of statistical measures that can potentially be used for optimizing model calibration without imposing the assumption that the underlying data is normally distributed. Closer collaboration between software developers and agencies that have accumulated large traffic databases could contribute to long-term improvement in model fidelity by allowing developers to select statistical distributions that more closely represent the parameters of interest.
ACKNOWLEDGEMENTS

The authors wish to thank the Wisconsin Department of Transportation’s Bureau of Traffic Operations for support of this research, and the microsimulation modeling team of HNTB Corporation’s Wisconsin offices for providing Paramics models and model output data.

REFERENCES