How to reduce the gambling element of some transport planning decisions

***Emily Seaman**, SIAS Limited*

Some practitioners remain wary of microsimulation models because they perceive a difficulty in drawing firm conclusions from their outputs. While these can be remarkably accurate if inputs are properly described, gut feeling derived from graphical output is not enough, and it is the lack of the simple number answer that causes some modellers to seek comfort in deterministic processes.

Modellers generally adopt microsimulation because they consider its representation of traffic flow and congestion to be closer to their expectations of real world conditions, and are not surprised when successive runs of the same model produce a range of results. Of course, deterministic models can produce a range too, but, unlike in the real world, their outputs are precisely determined by their inputs, which have difficulty in encompassing detailed descriptions of some common features such as UTC, bus priority systems and pedestrian controlled crossings. Microsimulation models can include these influences, and represent conditions experienced by real transport users, whose journey might not be entirely predictable. Transport users understand the concept of fuzzy modelling, and while some modellers tend not to appreciate this state of affairs, they may be more accepting if in possession of some simple statistical procedures to help interpret the results of models which in most respects deliver what they expect to see.

The answer lies in multiple simulation model runs. Deterministic models tend to produce single number answers based on convergent iterative assignment procedures embodied within single model runs. Because it does not exist in reality, the concept of assignment convergence does not exist with microsimulation. Instead, the model must be run enough times to produce a statistically sound conclusion in which there can be confidence in the model's validation.

Stephen Druitt, SIAS

Emily Seaman MSc is a contributor to *The Microsimulation Consultancy Good Practice Guide*, available from SIAS Transport Planning Consultants.

I'm a transport modeller. I've just built a microsimulation model of a local traffic network. I had a large amount of reliable input data. I've calibrated and validated my model and have fulfilled the validation criteria of the Design Manual for Roads and Bridges without problems. I have a final validation report ready for the client. It's going to be plain sailing from here, because the model outputs will speak for themselves. Am I right? Probably not.

Whilst the use of microsimulation modelling for traffic assessment is becoming widely accepted, many practitioners do not fully appreciate how the simulation process translates into validated model outputs. Microsimulation modellers tend to concentrate on the inputs and the visuals on their way to creating the best model to assess the consequences of future traffic growth and development scenarios. But how can they properly report that one scheme is better than another?

Microsimulation systems directly model the contributory factors to traffic flow in real time, and are our best attempt to date at representing reality. Reality tends to contain a high element of random chance, and a traffic model worth its true salt should reflect this. Random numbers govern many aspects of microsimulation modelling, such as the allocation of a certain type of driver behaviour to a particular vehicle (eg gap acceptance) and the exact time vehicles enter the road network. Every time the model is run, a unique stream of random numbers is used to govern events in that run. Each different random number seed will cause a different set of outputs to be produced, so which simulation run should be reported on? Although this may appear to be a problem, having more than one run to report on is in fact a real benefit.

In reality traffic flow is not exactly the same at the same time every day. Even the most predictable of events, the time we leave home for work, can at best be stated to most likely fall within a ten minute window. A microsimulation model can accommodate this for every journey, and, along with many of the events which might happen on that journey, provide the ability to model a range of scenarios that could happen in reality. The variability across multiple model runs is an approximation to the variability that exists in the real world, but the question is how to accommodate this within an evaluation framework.

It might reasonably be assumed that the 'most representative', 'average' or 'worst case' scenario should be reported on, but there is no way of defining these positions, and in reality such situations do not exist.

In theory it would be possible to run a model a very large number of times, and base our conclusions on every possible scenario with a different combination of release times and model-wide vehicle interactions.

But 'very large' equates to near infinity, a clearly hopeless situation. In fact we don't need to do this, because some simple statistical measures can provide an acceptable level of confidence in a sample of results from a limited number of simulation runs, such that our reported figures are close to the actual average.

Microsimulation systems reflect the fact that real traffic flow changes from day to day on the same journey and can report on the effects of a scheme taking into account these fluctuating conditions. In our examples which follow, we use the example of a Mr Jones, who leaves his house to drive to work at 08:34:56 one Monday. On any other day that year it is unlikely that he will leave at exactly the same time, although he'll probably leave between 08:30:00 and 08:45:00. Mr Jones knows his journey takes about 15 minutes but it really all depends on the traffic that day and a lucky run through the MOVA-controlled traffic signals. His journey time would be better represented as 14:50 ±1:05 minutes, an interval either side of the average calculated from a set of representative days.

It can now be revealed that Mr Jones is our client. In common with many clients he knows his own patch, and is concerned that the simulation model must accurately reflect the travel time along his morning route. If we get that right he'll be happy with the rest of the model, but he's adamant that his journey time must be estimated to within a 95% confidence interval of no greater than 10% of the sample average. This means that we must gather enough data from the model such that he is 95% confident that his true average journey time lies between the calculated sample average journey time plus or minus 10%. We refer to these numbers below as the 'convergence criteria', but be aware that the use of the word 'convergence' here does not imply any form of assignment iteration, a process that does not occur with microsimulation.

We have arranged for Mr Jones's home to fall within Zone A of the microsimulation model, and his workplace to be within Zone B. Initially only one run is performed and all the journeys made from Zone A to Zone B between 08:30am and 08:45am are considered and averaged. The average journey time from Zone A to Zone B is calculated from the model as 12:40 minutes. From only one average result it's not possible to calculate a standard deviation figure or a confidence interval. We can't claim that one model run represents all the possible model scenarios that might occur, so we run the model another couple of times and take an average of the three individual run averages. This statistic is acceptable because the times we are averaging are 'independent identically distributed random variables' from individual simulation runs. These runs are independent of all others and identical in their input data.

The average of the three averages has increased the journey time by over a minute and a half over the average from the single run. Considering the three separate averages in Table 1, it seems that the first run underestimated the travel time from Zone A to Zone B. This does not make Run 1 any less valid as a possible scenario, but simply means that it has experienced less congestion than the other scenarios due to reasons with which we are all familiar. The standard deviation of the a verages (note: not the standard deviation of journey times within a single model run) and 95% confidence intervals are produced (top tip: these are standard MS-Excel functions). This is summarised in Table 2.

Were the three runs enough to satisfy the convergence criteria? For this to be satisfied, the 95% confidence interval of the three runs must fall within the average of the three run sample plus or minus 10%. Considering Table 2, in this case the 95% confidence interval for the three runs is much wider than the 10% interval and so we must keep going. A similar conclusion is reached after performing 5 runs in total, for which the average journey time has altered by another minute.

After ten runs have been performed we appear to be getting somewhere. The 95% confidence interval gives limits of between 13:45 minutes and 15:55 minutes whereas the 10% limits are between 13:21 minutes and 16:19 minutes. Notice that, as the number of runs in the sample increases, the 95% confidence interval width, shown in Table 2, reduces, as more information is gathered about the underlying population average value (the average of all possible model runs). The journey time from Zone A to Zone B could now be reported as 14:50 ±1:05 minutes. This gives an average figure based on a model that has satisfied its principal convergence criteria, and provides a useful indication to the client of the spread in times that his journey might take.

Table 1: Single Run Journey Time Results

Run	Zone A to Zone B Average Journey Time (mins)
1	12:40
2	15:16
3	14:59
4	17:05
5	16:46
6	15:24
7	14:23
8	12:47
9	15:32
10	13:30
11	14:30
12	13:45
13	16:01
14	12:45
15	14:15

Table 2: Multiple Run Confidence Intervals

Number of Runs Considered	Zone A to Zone B Average Journey Time (mins)	95% Confidence Interval		10% of Average		Confidence Interval Width
		Minus Limit	Plus Limit	Minus Limit	Plus Limit	
1	12:40	-	-	11:24	13:56	-
3	14:18	10:46	17:51	12:53	15:44	07:05
5	15:21	13:10	17:32	13:49	16:53	04:22
10	14:50	13:45	15:55	13:21	16:19	02:10
15	14:39	13:52	15:25	13:11	16:06	01:33

(The confidence interval has been calculated based on the assumption that across an infinite number of uniquely seeded model runs the journey time of Mr Jones would be approximately normally distributed. The 'students t-test' can therefore been used to calculate the confidence interval. A 95% confidence interval equates to two standard deviations in the sample.)

Table 3: 95% Confidence Interval calculations for five runs

	Average Journey Time (minutes)	95% Lower Confidence Bound	95% Upper Confidence Bound
Base	15:21	13:10	17:32
Option 1	09:32	07:39	11:25
Option 2	12:29	10:48	14:10

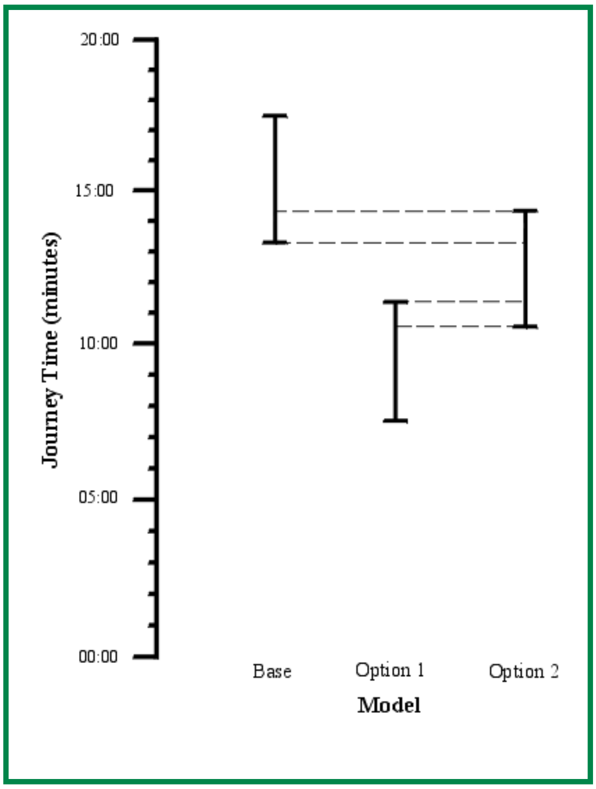


Figure 1: Graphical representation of 95% Confidence Interval calculations for five runs. The black vertical bars represent the 95% confidence interval width and upper and lower confidence bounds. The dotted lines show the overlapping confidence intervals.

The single run results are shown below in Table 1. Table 2 shows the calculated confidence limits from the Table 1 results. It is interesting to note that, had the average journey time been reported from just the one initial run, it would have been underestimated by over two minutes, with clear implications for erroneous decision making. It would be possible to perform more runs and either increase the confidence interval (to say 99%) or keep the confidence interval at 95% and reduce the average plus or minus criteria to 5%. With an infinite number of runs we can reduce the confidence interval to 0% as we are no longer sampling the model but measuring all its combinations. This is of course impractical. However, given that Mr Jones has only measured his journey time accurately in the last few days since instigating the project, he also has only a sample of the data and similarly has a mean, and a set of confidence intervals about that mean. Therefore we must beware of false accuracy given that our validation data is also merely a sample of reality.

The above discussion has focused on a single base model. When it comes to comparing design schemes the underlying variability within models should be analysed and reported on. Mr Jones didn't have us build that model for nothing, and he now wants us to undertake an operational comparison of two proposed schemes to speed up his journey to work. The first involves an additional lane for all traffic on part of the route, and the second sees this lane being for buses only. The two schemes are coded separately into two copies of the validated base model. Initially five runs are performed of each, and 95% confidence intervals calculated, shown in Table 3 and Figure 1.

The 95% confidence interval means that there is 95% probability that the underlying population average (the average if all possible simulations were performed) is within the interval stated. Considering Figure 1, even though it seems Option 1 out-performs Option 2 (a faster journey time by 02:56 minutes), the 95% confidence intervals overlap. It might be the case that the true Option 1 journey time is 11:25 minutes (the Option 1 upper bound) and the Option 2 journey time is in fact 10:48 (the Option 2 lower bound). In this case, Option 2 would then be the better choice. More model runs are needed in order to tighten the confidence intervals around the averages to be sure that Option 1 is better.

Similarly, there is no clear way of deciding whether Option 2 is better than the Base as their confidence intervals also overlap. All that can be said after five runs is that Option 1 is better than the Base. Five more runs are now performed, so a total of ten model runs are available for analysis. Table 4 and Figure 2 show the results tabulated and graphically, respectively.

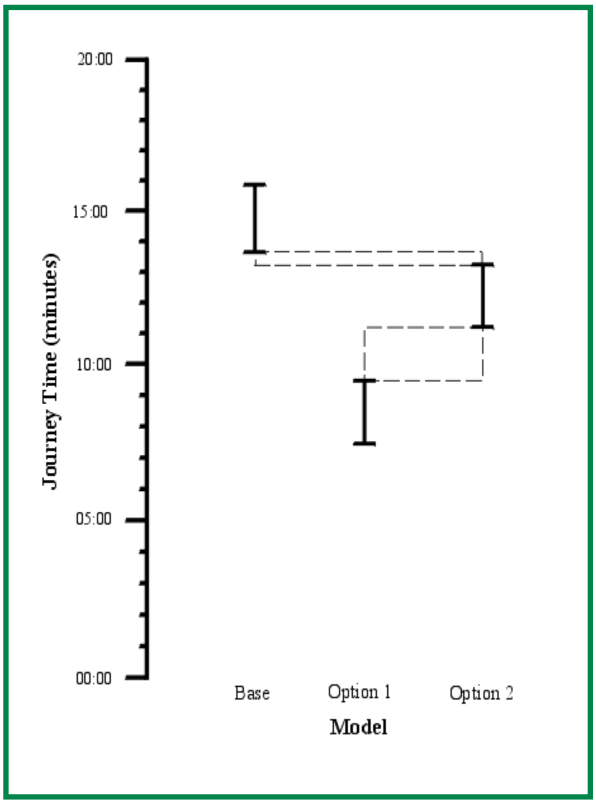
With none of the confidence intervals now overlapping, we can now conclude that Option 1 is the best, and Option 2 is better than the Base but not better than Option 1.

For heavily congested models, the variability across model runs may be high and a large number of runs may be required before the 95% confidence intervals between options do not overlap. It may also be the case that a choice must be made between options that, although not hugely variable, demonstrate very close average results. In this case, if performing a large number of runs does not eliminate the overlap, an ANOVA (Analysis of Variance) analysis could be performed to ascertain whether the options are truly different. (Once again, MS-Excel's add-in tools come to the rescue, but you might need to use the help key first.). Results from alternative options do not need to be different, it may simply be that neither option is better when compared under the convergence criteria.

To date little guidance documentation exists in the UK for microsimulation modelling methodologies, and the Design Manual for Roads and Bridges (DMRB) is geared towards deter-

Table 4: 95% Confidence Interval calculations for ten runs

	Average Journey Time (minutes)	95% Lower Confidence Bound	95% Upper Confidence Bound
Base	14:50	13:45	15:55
Option 1	08:35	07:45	09:24
Option 2	12:14	11:17	13:10



ministic modelling approaches. For this reason it is especially important to understand how to interpret microsimulation outputs properly. Some guidance is produced by the Federal Highways Authority in the US[1] although its practice is not mandatory there. The US guidance stresses the importance of multiple simulation model runs for proper model interpretation.

There is no fixed answer as to how to set the statistical convergence criteria for various outputs from microsimulation models. Microsimulation is an inherently stable methodology, especially when used in conjunction with a reporting tool which has the ability to analyse multiple runs. Analysis of output variability gives the modeller the opportunity to compare this with observed data variability, and it is sometimes necessary to be reminded that, as in all matters computing, the model can only be as good as the input data. The foregoing text refers to the accuracy of the model's ability to reflect the conditions engendered by the input data. In this respect the DMRB's advice is relevant to microsimulation modelling, and provides guidance on how to get the best from data sources.

Performing multiple model runs can be time consuming although tools exist to manage multiple runs and to process the resulting data. However, a balance needs to be struck between model robustness, suitability, and project time and budget constraints to ensure the integrity of a very powerful and realistic tool for assisting transport planning.

1. Traffic Analysis Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software. Published June 2004 by US Department for Transport (Federal Highways Administration). Section 6 and Appendix B of this guidance highlight the importance of performing multiple simulation runs and identifying statistical procedures for analysing them.

Figure 2: Graphical representation of 95% Confidence Interval calculations for ten runs. The dotted lines show the now non-overlapping confidence intervals