Australasian Transport Research Forum 2011 Proceedings
28 - 30 September 2011, Adelaide, Australia
Publication website: http://www.patrec.org/atrf.aspx

DEVELOPMENT OF A STATISTICAL FRAMEWORK TO GUIDE TRAFFIC SIMULATION STUDIES

David Shteinman
Chief Operating Officer, Australian Centre for Commercial Mathematics (ACCM)
School of Mathematics and Statistics, University of New South Wales
Sydney NSW 2052
d.shteinman@unsw.edu.au

Christian Chong-White
Traffic Algorithm Development Manager, Roads and Traffic Authority of New South Wales
Australian Technology Park, Eveleigh, NSW 1430 Australia

Gary Millar
Software Support & Development Manager, Azalient (Australia) Pty Ltd
Suite 145 National Innovation Centre, 4 Cornwallis Street, Eveleigh NSW 2015 Australia

ABSTRACT

This paper describes a statistical framework to be used in a set of simulation and modelling guidelines. The framework was developed out of the necessity for defensible study results by the application of rigorous statistics. It is the result of a collaborative RTA-ACCM research project and provides a framework for analyzing simulation outputs, and also informs the design stage of a simulation study. The project was based on an analysis of thirty-six PARAMICS models supplied by the RTA. The guidelines apply Exploratory Data Analysis techniques (EDA) to the design and analysis of traffic micro-simulations, and include the graphing of output distributions to expose salient features and rigorous methods to detect and handle outliers in output data.

The framework includes methods to quantify and correct biases that result from the phenomenon of unreleased vehicles or incomplete trips.

Diagnostic tests are described for discriminating between running error, model error or extreme sensitivity to congested conditions

A model to predict the run cost of a simulation as a function of critical network features is also described. A regression model was built, based on an index of model complexity and the combination of critical input factors. The regression model was limited by the small sample size of models used (36). Further research is continuing on cost prediction, using a larger set of simulation outputs and model types.

The guidelines developed have provided value to RTA traffic control modelling practice and can be used by simulation modellers regardless of the micro-simulation package used.

1. Introduction

Traffic micro-simulation is used to evaluate and compare traffic management policy outcomes. The criticality of this analysis task requires that such studies provide robust and defensible information.

Experienced simulation practitioners state that typically most resources in a simulation study are devoted to model building and validation. Comparatively fewer resources are devoted to analyzing the results in a valid statistical framework. This results in simulations that may not be statistically valid and "as a result, these estimates could, in a particular simulation run, differ greatly from the corresponding true characteristics for the model. The net effect is, of course, that there could be a significant probability of making erroneous inferences about the system under study" (Law 2007, p.485).

The Roads and Traffic Authority of New South Wales, Australia (RTA) engaged "MASCOS" The Australian Research Council's Centre of Excellence for Mathematics and Statistics of Complex Systems (MASCOS, 2010) now known as the Australian Centre for Commercial Mathematics (ACCM, 2011) to support a research project to develop: (1) a methodological framework for analyzing simulation outputs, and (2) a framework to inform the design stage of a simulation study. The project was based on an analysis of thirty-six PARAMICS models supplied by the RTA (Quadstone Paramics, 2006) and resulted in a statistical framework as a set of guidelines to guide modellers in traffic simulation studies.

The originating stimulus for the study was the need for defensible simulation modelling of the Sydney Coordinated Adaptive Traffic Control System (SCATS) (Chong-White et al 2010) when used with the SCATSIM (RTA 2010) platform, and other associated advanced traffic control (ATC) systems (RTA 2009). This type of modelling often requires investigation of traffic performance changes of small magnitude at a localized level which drives the particular need for a robust and defensible statistical analysis approach. However, the project aimed for a general framework that can be applied to any traffic simulation.

Existing simulation guidelines that are industry standard for example the FHWA's *Traffic Analysis Toolbox* (vols. I-IV) and Transport for London's "*Micro-Simulation Modelling Guidance Note*" contain general "rules of thumb" statements but not specific instructions with detailed statistical techniques that guide the data exploration, statistical analysis and interpretation of output results.

Furthermore, previous standard in modelling practice has used a fixed number of simulation runs - usually set at five - irrespective of the variability in the results. Defensible simulation modelling requires the *specification of precision and statistical confidence*. Confidence intervals are required to due to the uncertainty that is a feature of applied traffic analyses. This uncertainty results from both intended model variability – reflecting applied observations, and unintended model error. This uncertainty should be considered in the design of a modelling study, interpretation of modelling results, and any subsequent decision-making that results from the interpretation.

2. LAYOUT OF GUIDELINES: OVERVIEW

The guidelines are divided into three parts, each corresponding to the three stages of a simulation study:

- **Part 1:** Prior to running
- **Part 2**: Output analysis
- **Part 3:** Cost prediction

The three part process is outlined in Table 1 below.

Table 1: Layout of Guidelines

Part	Detail	
1 – Prior to Running	Select summary measure, specify test statistic, specify precision level	
2 – Output analysis	*Step 1*	Automated output analysis, Exploratory Data Analysis (EDA), graphics
	Step 2	Runs for precision and confidence levels: one and two scenario
	Step 3	Power analysis
	Step 4	Diagnostic tests using outliers and other indicators
	Step 5	Test for bias due to incomplete and unreleased vehicles
3 – Cost prediction	Planning/Design: Interactions between critical inputs and effect on variability	

2.1 Part 1 – Prior to running

Prior to running the simulation study the guidelines' first steps are to:

a) **Choose the summary response and its measures -** for example Vehicle hours travelled (VHT) which is a single summary of the total network performance and includes the effect of delays.

It is recommended to start with standard statistical measures such as the mean, standard deviation, median, 95th percentile, range of results and skewness that will summarize the total set of runs and how the results are distributed.

b) **Choice of statistical parameter for between-model comparison-** it is recommended to use the coefficient of variation (*CV*) as the standard measure of comparison of variability between models (which will also be used in the cost prediction part of the study):

$$CV = \frac{sd}{mean} \qquad (1)$$

sd = standard deviation of summary response over total number of runs

c) **Set precision measures and significance level –** typically precision is set as a % of Mean, for example, VHT to be within +/-1%, 2% or 5% of the Mean but can also be an absolute value of VHT. It is also necessary to set an acceptable significance level, as a percentage for future confidence intervals, chosen as one of 90, 95 or 99%, typically.

2.2 Part 2 – Output analysis

2.2.1 Step 1 - Automated output analysis, Exploratory Data Analysis (EDA), and graphics

This involves the automated statistical analysis of output (using EXCEL or any standard "off the shelf" statistical package) and the more complex task of applying Exploratory Data Analysis (EDA) techniques to identify the most informative features of the simulation's output and performance (Mast and Trip, 2007).

The elements of EDA are:

1. Display the data
2. Identify salient features
3. Interpret the salient features :

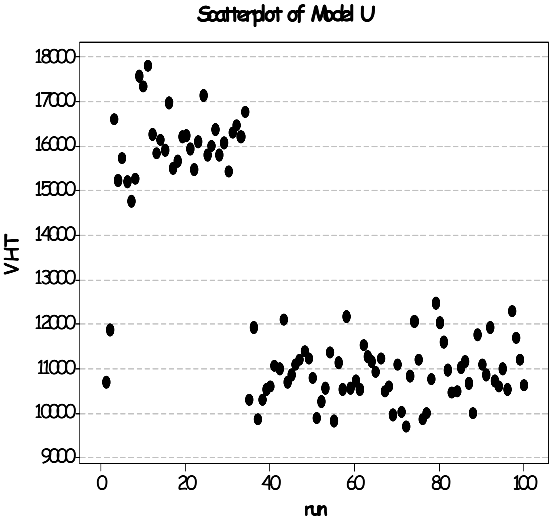
The entire process of EDA as applied to traffic microsimulation is explained in (Shteinman et al, 2010). The background to EDA and its methods are given in (Mast & Trip, 2007 and Chatfield, 1986).

The three main aspects of EDA that are critical to traffic simulation studies are:

a) **Screen data to check for independence between results** – The standard assumptions of statistical independence across simulation runs must be satisfied so that standard statistical tests can be used (Law 2007, p.486). This is possible only by using a different random number seed for each run
Evidence that the independence assumption holds is when the data shows no evidence of correlation; when the data shows random variation only between simulation runs. Random variation implies no sudden or large shifts in value due to a special or non-random event.

Graphical techniques such as scatter plots are used to screen the data. Figure 1 below is an example where screening using a scatter plot depicts a special or non-random event has occurred (at run number 34) causing serious model error. If such screening shows the assumption of statistical independence is violated, the analysis should not proceed as all subsequent analyses rest on this assumption. The models should be checked for error.

Figure 1: Scatter plot of Model U of runs 1-100 showing non-random variation in VHT



b) **Define and screen for outliers:**

The data should be screened for extreme values or outliers using boxplots which graphically display the quantity and location of outliers, as well as the median (measures of the centre of the data) and variability (using the width of the quartiles - those values that divide the data, sorted into ascending order, into four equal parts).

Boxplots are useful for outlier detection as they graphically display the quantity and location of outliers, as well as the median and variability (using the width of the quartiles). The median and quartiles are used for outlier identification as they are largely insensitive to outliers (compared to the mean). The median and quartiles are thus called ‘resistant’ statistics.

Outliers are often the most interesting feature in a simulation data set (Chatfield 1986) so it is important to note that wrongly classifying, ignoring or eliminating them can cause significant problems. For example, a very high value of VHT may reflect a highly saturated network that, due to random events, produces delays at a level far from the mean, and thus is truly representative of the network under those conditions. Such delays indicate a high variability in travel times which will not occur in only one run of the model and as such should not be classified as outliers. Hence it is necessary to introduce an objective, scientific criterion for outlier classification and elimination.

Statistical science has developed a wide range of parametric outlier tests that look at some measure of the relative distance of a particular data point to the mean and assesses what the chance is that a point of data occurred by chance. For more background on outlier classification see (Seaman & Allen, 2010).

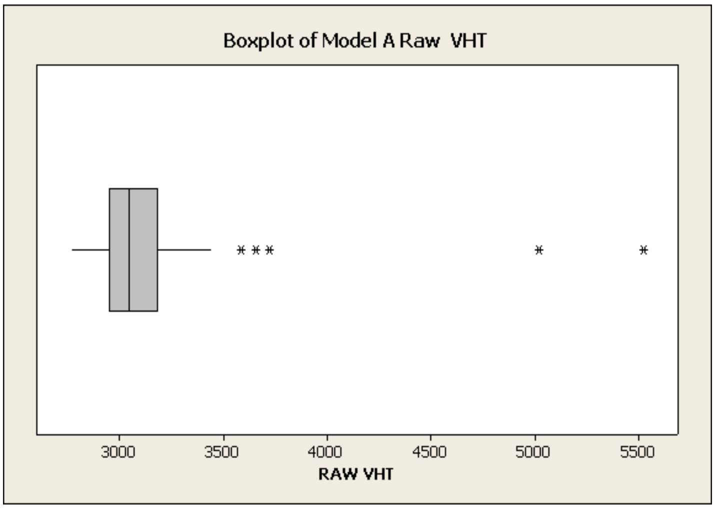
The inter-quartile method shown below in (2) is recommended for calculating box end points, and defining outliers:

Outlier definition; Q3 = 3rd quartile 75% of data, Q1 = 1st quartile 25% of data (2)

- *data beyond upper limit = Q3 + 1.5 (Q3 - Q1)*
- *data below lower limit = Q1 - 1.5 (Q3 - Q1)*

It is recommended to present outlier results using boxplots with and without outlier runs.

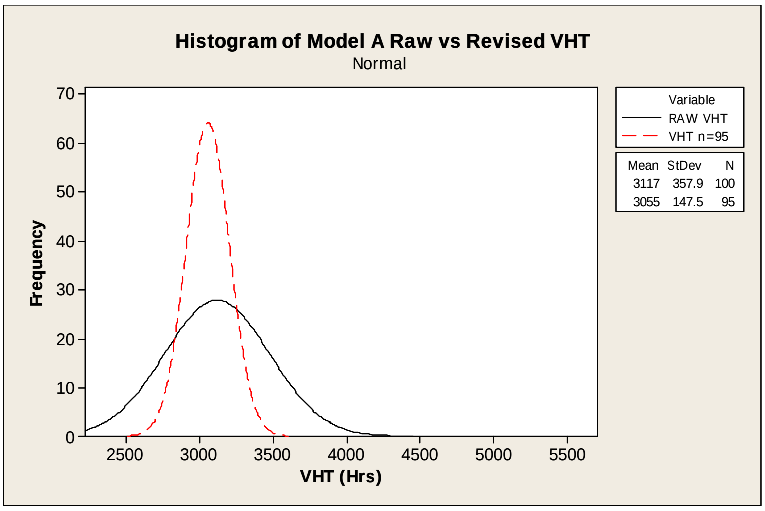
Figure 2: Boxplot of VHT for Model A showing Median, 75% of data range and quantity and location of outliers.



Note the median VHT is 3045, the inter-quartile range is 2950 to 3184, 5 outliers are present, and the range is 2767 to 5523 hrs. The longer upper whisker and large box area to the right of the median indicate that the data have a slight positive skewness the right tail of the distribution is longer than the left tail.

Figure 3 illustrates how failure to remove outliers can inflate measures of variability. The wider ‘Raw VHT’ histogram has a standard deviation (St Dev) of 357.9. When the five outliers are removed, the St Dev becomes 147.5.

Figure 3: Comparison of two curve fitted histograms of data with & without outlier



.

c) **Checking for normality** - Probability plots are used to determine whether a particular distribution (in this case the normal distribution) fits the output data (Vining, 2010). While the actual data will not be exactly normally distributed we do need to know *how far from normal it is.* The probability plot shows this graphically. This information is important if we wish to test for significant differences between two scenarios which require a t-test. The t-test assumes that i) the data are independent between results and ii) the data *approximately* follow a normal distribution.

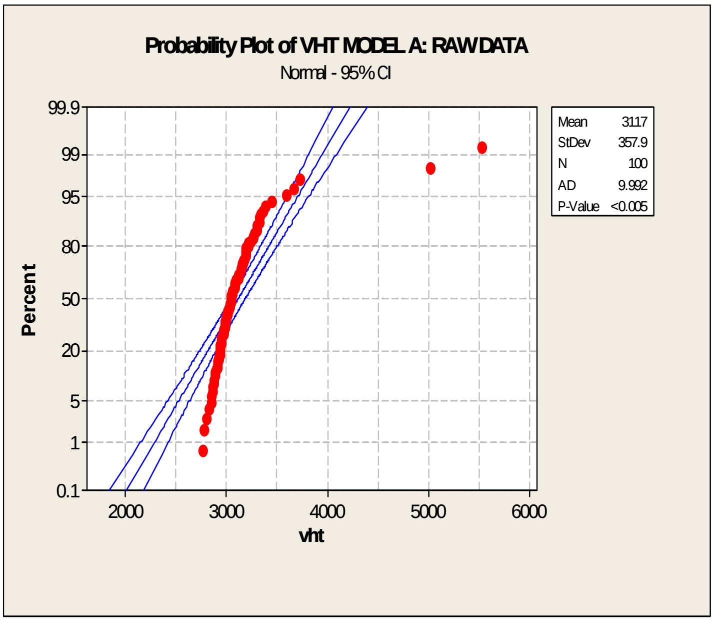
A probability plot is formed after outliers are deleted. If the data are approximately normally distributed then the probability plot will show these features:

- the plotted points will roughly form a straight line
- the plotted points will fall close to the fitted line
- the *P*-value will be larger than the chosen α-level. (commonly chosen level for α is 0.05)

Note that the smaller the sample size (number of runs in the simulation) the less robust the t-test will hold to the normality assumption. Hence the minimum sample recommended in the guidelines is *n*=30. *At n=5*, deviation from normality will affect the reliability of the test. If the data are far from normal the data points will fall well outside the confidence bands (outer blue lines) and the *P*-value will be lower than 0.05.

A highly non-Normal distribution for VHT is shown in figure 4 for 'Model A' VHT hours (with outliers included). Note that the plotted points do not form a straight line, are well outside the 95% confidence bands and the P-value of .006 is far less than the common significance level of 0.05

Figure 4: Normal Probability Plot of Model A with outliers included.



2.2.2 Step 2 - Precision, sample size: one and two scenario

Single scenario

Standard statistics are used to calculate a confidence interval on the estimate of the Mean of a summary output (Shteinman et al, 2010, pp.9-10, Kelton 1997). In brief, the precision of the estimate, expressed in the form of a 95% confidence interval is:

$$\hat{x} \pm t_{n-1}(0.95)\frac{s}{\sqrt{n}} \qquad (3)$$

Where $t_{n-1}(0.95)$ is the upper 95% critical point for the *t* distribution with *n*-1 degrees of freedom. Note that the critical factors determining precision of a simulation are the variability of the sample runs, *s*, and the sample size, *n*.

For a given desired confidence interval half width, *w*, and assuming *s* is known, we can then calculate a required sample size using Eqn 4: (Abdy & Hellinga ,2008)

$$n = (t_{n-1}(1 - \frac{\alpha}{2})\frac{s}{w})^2 \qquad (4)$$

The sample standard deviation *s* can be estimated from an initial set of runs and then we recalculate the number of rune requited to obtain the required precision eqn 4). Hence we recommend to set the initial sample run of *n* = 30. This will give a reasonably good estimate of sample standard deviation, which can then be used to determine the final required sample size.

Determining the sample size required to meet a specified precision can therefore be viewed as an iterative procedure. The iterative procedure is:

1. **Set *n* =30 and run *n* simulations**
2. **Calculate $\bar{x}$ and *s***
3. **Set precision as a percentage of the mean (say, ±1% or ±2%)**
4. **Calculate n^***
5. **Compare n^* to *n* and run n^*-*n* additional simulations**

Note that it is possible that there will be large variation in the n^* between different simulations.

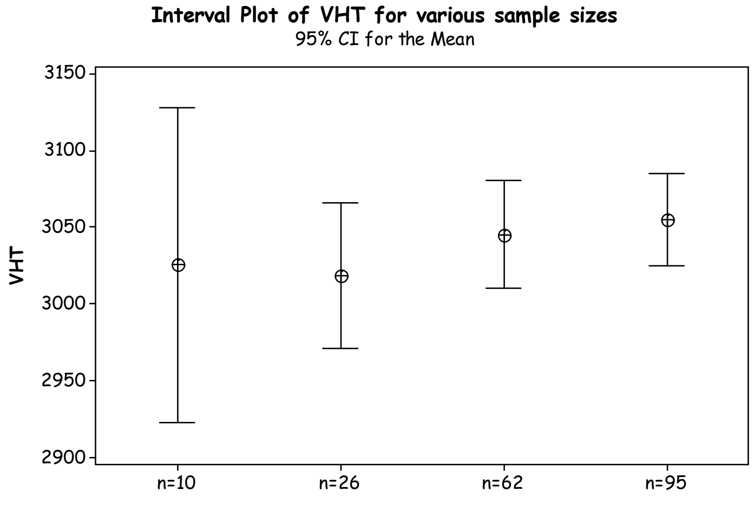
Robustness of this approach

An important feature of the framework to guide simulations is that it ought to be able to be applied to a very wide range of simulation outputs. That is, the analysis method used must be robust or insensitive to data with distributions that are not normal. According to the Central Limit Theorem (Niles 2009, Nakayama 2008, Law, 2007, pp.232-233) using “sufficiently large” samples sizes - we recommend starting with n>30 - these methods should be robust to deviations from normality.

Effect of sample size

It is important for simulation practitioners to consider the effect of sample size. The intervals for four sample sizes are plotted below in Figure 5, showing improvement in precision as run size increases

Figure 5: Confidence intervals for “Model A” Mean VHT at four sample sizes, n=10, 26, 62 and 95



Two scenarios

As in the single scenario case the Mean is often the parameter of interest for comparison. A paired analysis is the appropriate method. The confidence interval for the difference can be constructed in the same way as the mean for a single scenario:

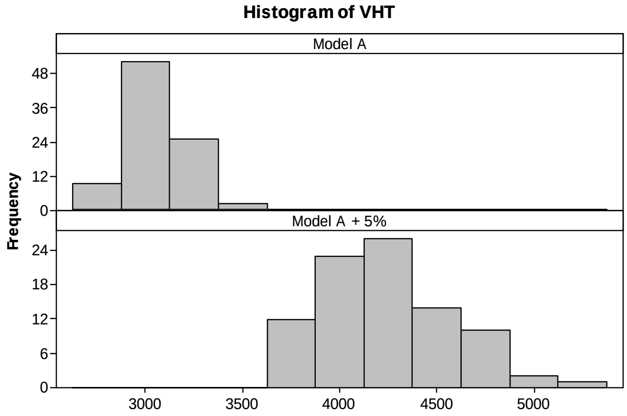
$$\bar{x}_d \pm t_{n-1}(0.975)\frac{s_d}{\sqrt{n}}, \qquad (5)$$

In this case, $\bar{x}_d$ is the mean difference and s_d is the standard deviation of the differences.

The paired t-test allows for comparison of the means between two scenarios. However to compare the *difference in the spread or variability* between two scenarios, the 'F-test' is employed.

The histograms in figure 6 below demonstrate the difference in the spread for the VHT of Model A as a result of increasing the demand by 5%. This illustrates how the increase in demand affects not only the mean but also the spread.

Figure 6. Histograms of Model A - comparing variability when demand increased by 5%



There was a highly significant difference between the standard deviations of these models (*P*<0.001). The observed ratio of the standard deviations was 2.31 -indicating that the 5% increase in demand caused a two-fold (200%+) increase in variability.

2.2.3 Step 3 - Power analysis

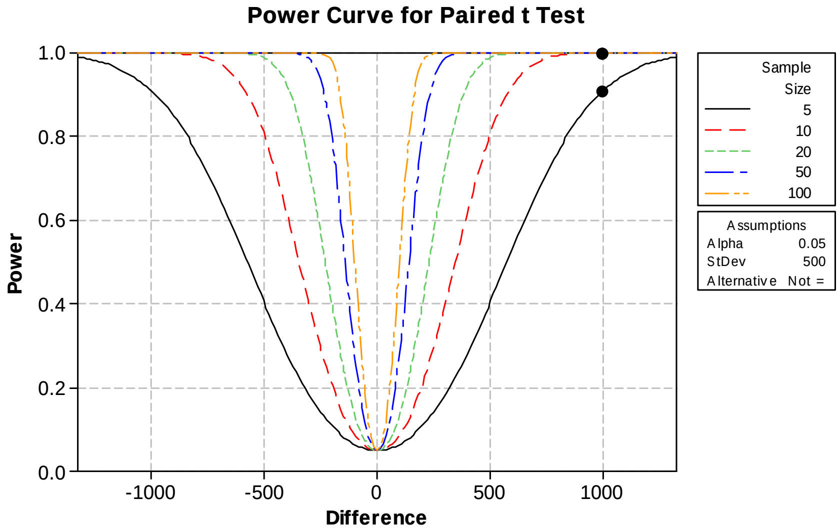
When the objective is to determine whether the observed differences between two scenarios are likely to be genuine (significantly different) and not just due to chance, then the sample size calculation is not a question of precision but *statistical power.*

The ability to detect a statistical significant difference between two scenarios depends on the underlying mean difference between the groups, the standard deviations in the groups and the size of the sample *n* taken. The chance of detecting a difference for a test is also known as the *power* of the test and is usually set at 0.8 or 0.9 (Montgomery 2001, p. 34).

Power in this case then may be defined as *the probability of correctly detecting a difference between two scenarios (accepting the alternative hypothesis to the null hypothesis that states there is no difference).*

For example Model A and Model Q differed only in demand, where the latter had the observed demand level increased by 5%. In order to compare the VHT it is appropriate to use a t-test. The power curves in figure 7 illustrate how the larger sample size, the smaller the difference detectable between the two scenarios

Figure 7: Power curves showing the power by a variety of different underlying mean differences and corresponding sample sizes at a significance level of 0.05



The guidelines for power and sample size calculations are:

a) Determine the smallest effect size (the difference to be detected) that is of practical interest, noting that the smaller the effect size the greater number of observations will be required to detect that difference.

b) Choose the desired level of power to detect this difference (usually 80 or 90%).

c) Choose the desired significance level (usually 5%).

d) Determine the sample size required for this comparison.

e) Display power curves to assess the consequences of the choices in a),b),c).

2.2.4 Step 4 - Diagnostic tests using outliers and other indicators

The information inherent in the outliers can be used as a diagnostic tool to assess model performance and detect model errors.

In a simulated environment outliers can only occur due to one of three causes:

a) **Simulation running error -** for example, simulation stopped too early so not all runs completed.

b) **Under congested conditions -** the simulated network may be very sensitive to small changes in input values. This sensitivity is shown by a large and sudden increase in total vehicle delay in the network, resulting in an extreme or outlier value of VHT (Dowling et al, 2002, p.63).

c) **Model error -** due to the model not having been properly calibrated.

Furthermore, it is necessary, for diagnostic purposes, to distinguish between l*ow or high* value outliers or travel time. *Low* value outliers indicate simulation or model error since it is highly unlikely that it was caused by an actual network event since extremely high vehicle speeds (well above the speed limit) are not part of the simulation input. *High* value outliers indicate an event that caused extreme delays. Thus, we can suggest a framework for diagnostic tests using the outliers that are *the extremely high values* of travel time, delay or distance travelled. These are:

a) Identify the run number of the outlier run
b) Using the same random number seed for the run, repeat the simulation
c) Focus on the time and location of the outlier run event to diagnose the cause
d) Drill down into the time of the outlier event and its location in the network

For model diagnosis under congested conditions the quantity and timing of unreleased and incomplete trips should be used as a symptom of the root problem to be diagnosed. In the case of high value VHT outliers it is recommended to check for correlations with the following inputs:

- **Boundary congestion effects -** unreleased vehicles that could not enter network due to congestion at the entry zones. A result of high unreleased vehicle count is the full demand planned for the network was not simulated, and thus not included in the travel time total.
- **Incomplete trips -** vehicles that did not exit the network at the end of a run.
- **The total number of vehicles** that did get through the network.

Generally VHT outlier runs *(high values*) will be correlated with the following results which can be used as diagnostic indicators:

- Very high incomplete trips
- Very high unreleased vehicles
- Low total throughput of vehicles

2.2.5 Step 5 Quantifying and correcting bias due to congestion

A statistical framework must take into account the effect of vehicles not counted in summary measures due to congestion. This refers to vehicles that were not loaded onto the network due to a queue build up to the entry zones, ("unreleased vehicles") as well as vehicles that could not exit the network as the simulation terminated before they could exit ("incomplete trips").

Not counting unreleased vehicles and the time they are queued will result in a misleading reduction in VHT. In order to adjust for this in the assessment of network congestion it is necessary to add back the hours that vehicles spent queuing to enter the network.

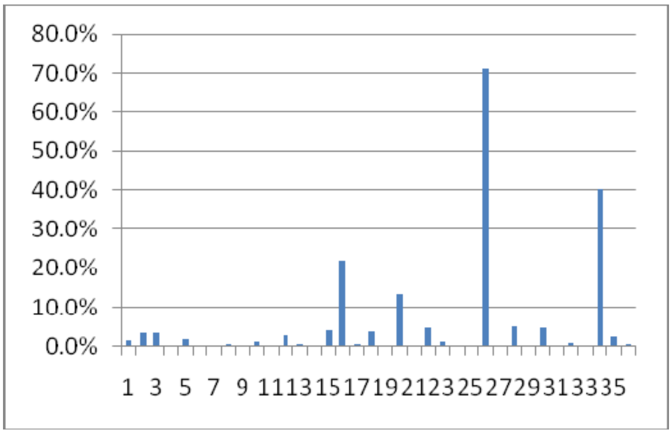
To solve the problem an appropriate 'add back 'of hours of VHT is added to the initial VHT. (here known as dVHT). As the time that vehicles spent queuing is only available in blocks for each time interval, so the total time to be added to the initial VHT is:

dVHT = Mean of vehicles blocked per time interval (hrs)* x *qty of time intervals (6)

The critical measure for assessing the effect of this is the ratio of dVHT to original VHT.

The following bar graph gives the average add-back for 36 PARAMICS models as a ratio of dVHT to original VHT. Typically the effect was small (less than 10%), though there were four models where the add back was above 10% of the originally counted VHT. This is a strong indication of a significant effect due to congestion – when a large number of vehicles could not be loaded onto the network due to congestion for these four models.

Figure 8: Mean Add Back VHT Due Unreleased Vehicles As % Of Original VHT: 36 Models



Models for which the number of unreleased has a large impact

A major characteristic of the models with a very high ratio of "addback" was the high demand. Two of the models correspond to already busy networks to which an extra 5% demand was added. In these cases, the modelled demand has clearly exceeded the effective modelled road capacity, leaving a large number of vehicles unable to enter the network.

Another two models differ from their corresponding models in behaviour only with increased driver aggression and awareness levels.

It is recommended to use the total adjusted VHT (that includes the add back figure dVHT) in all cases where there are any unreleased vehicles. Using the adjusted VHT will also change the sample standard deviation "*s*" of the VHT over all runs and hence change the confidence interval of the Mean estimate of VHT, as per equations (3) and (4) above.

.

2.3 Cost prediction

Cost prediction for simulation requires a valid predictive model to give the number of simulation runs to meet a specified precision as a function of the model's critical inputs, *before the simulation is run.*

Such a model can be used by decision makers since it allows determination of whether the simulation (and specified precision) is economically viable before incurring the expense of model building, validation, calibration and analysis.

To explore the effect that various input factors have on variability (and thus precision), it was necessary to examine *simulation model complexity* as well as, individual input factors' effects on output variability. A "Complexity index" was introduced to rate a network. The index is a first attempt to rate complexity of a model using measures that are common to all the major simulation models (PARAMICS, VISSIM, AIMSUN, COMMUTER).

Statistical analyses using regression and analysis of variance (ANOVA) was then used to identify the simulation inputs that had greatest effect on variability. A Regression model is formed to *predict the number of runs required for a given precision based on model complexity.*

As seen above, the precision of an estimate depends on the sample size and the sample standard deviation, *s*. In this context, where the size of the networks varies greatly, it is most appropriate to consider *s* in relation to the mean, not its absolute value. For this reason, the quantity considered here was the Coefficient of Variation for VHT, as explained in (1) above. The CV was computed only for those simulation runs that were not considered outliers.

Model complexity

A complexity index was developed based on the full range of simulation inputs covering behaviour, network types, control, and demand profile and vehicle assignment. The index was divided into individual indices, as well as a composite index. The indices are described in table 2 below.

Table 2: Complexity index parameters

Index	Level
Behaviour	1. using single vehicle behaviour
	2. using normal distributed behaviour from Paramics
	3. Paramics default behaviour plus plugins to change the vehicle behaviour
Network	1. Linear
	2. Parallel routes
	3. Network (grid or similar)
Control signal types	1. Paramics fixed time coding
	2. Azalient Playback Fixed Time (may contain conditional phasing)
	3. SCATSIM
Demand	2.2-4 matrices per demand file
	3. more than 4 matrices per demand file
Assignment	1. All or nothing technique to load OD matrix
	2. Perturbation only or Dynamic Feedback only
	3. Combined Dynamic Feedback and Perturbation

The complexity index was termed BNCDA for Behaviour, Network, Control, Demand, Assignment (Millar 2010) as it uses those five inputs which are common to all simulation modelling techniques, as the basis of measuring model complexity. The indexes, as measures of complexity, are defined as follows:

Behaviour: *Vehicle behaviour ranging from single behaviour for all vehicles, through normal randomised vehicle behaviour, to normal plus plugins to apply further behaviour changes.* Hence complexity levels range from the simplest (Level 1 single vehicle) through randomly distributed behaviour through to randomness with plug ins (Level 3).

Network*: Physical attributes of the model including size, complexity and type (e.g. linear, parallel, network etc).* Complexity ranges from low (Level 1 for a linear model) through to high (level 3 for a network model)

Control Signal type*: Traffic signal control type – low complexity (Level 1 fixed time), fixed time with Playback, through to high complexity (level 3 SCATSIM)*

Demand: *The number of demand periods used in the model and the number of Origin-Destination matrices per demand file. Refers to all traffic assignment–type models.* Complexity Levels range from Low (Level 2) corresponding to 2-4 matrices per demand file, to high (Level 3) corresponding to more than 4 matrices per demand file.

Assignment*: The assignment techniques used to load the OD demand matrices onto the model.* Complexity ranges from low (Level 1) for the “All or Nothing” method to load demand, through to high *(*level 3) for a combination of dynamic feedback and perturbation.

The 'BNCDA Index' quantifies the complexity of a model by summing the index values for Behaviour, Network, Control, Demand and Assignment. The smallest possible total value is 5 (low complexity), the highest is 15 (high complexity).

The composite index is formed simply by adding the Level values above. For example: A model with Paramics default behaviour plus plugins has level = 3, Network of Parallel routes has level = 2, Control of Azalient Playback Fixed Time has level = 2, Demand of 2 to 4 matrices per demand file has level =2, Assignment of combined Dynamic Feedback plus Perturbation has level = 3:

BNCDA Index = 3 + 3+2+2+3=13

Analysis of the data from 36 PARAMICS models showed a close to linear trend in the effect of the indices on CV. In order to generalize the results, a linear relationship was constructed (equation 7) using a regression model which enabled estimation of CV for a given level of the index:

$$CV = -0.044 + 0.0069 * BNCDA \tag{7}$$

While there was a significant relationship between the BNCDA index and CV, it must be stated there was still a lot of variability in the models (R2-adj=23.9).

The estimation of CV was used to estimate the required sample size. Expressing the required precision as a confidence interval of half-width equal to 1% of the mean, we have the following relationship between CV and runs:

$$\text{runs} > \left(CV\frac{1.96}{0.01}\right)^2 \tag{8}$$

For sample values of the Complexity index “BNCDA”, we estimate a minimum number of runs required, using equations (4) and (5). As (4) gives an estimate of CV we use a prediction intervals (PI) for CV with a width approximately ±0.033. The minimum number of simulation runs required is shown in Table 3.

Table 3: BNCDA complexity index and minimum number of runs to obtain a precision of 1% of mean VHT

BNCDA	Mean *CV*	Approx. 95% PI for *CV*	min runs
9	0.018	(0, 0.051)*	13
11	0.032	(0.000, 0.065)	40
13	0.046	(0.013, 0.079)	81

Even though the variation found in CV translated into great uncertainty about the required number of runs, it was suggested that the average be used as a guide for the initial number of runs to try, with the option of increasing this in a second phase once a measure of variability for that network is available.

Individual factors

While the composite complexity index "BNCDA" provided a possible summary of each model, we also considered the different input factors separately, in terms of their effect on variability and hence run size requirement for a specified precision. Analysis results found positive relationships were present between CV and: signalized nodes (P<0.0001); and number of zones (P=0.003). Less significant relationships were found between CV and: demand level (P=0.14); and model type (P=0.056).

A combined model

While considering each factor separately is informative, the aim is to produce one model for predicting variability (CV) from the model characteristics. The factors that demonstrated a relationship with CV were: *demand level, complexity, model type, number of signal nodes and number of zones.* Given the limited number of models (36) and the fact that not all combinations of variables were available, only main effects and not interactions between these factors were considered.

Of these combinations, the best model in terms of statistical significance and the proportion of variability explained was one *with demand level and the number of signal nodes*. This model accounts for 40% of the variability in CV (R^2-adj=39.7%) and is given by:

$$CV = 0.01297 + 0.00117 * \text{signal nodes} + \begin{cases} -0.01740 & \text{if demand } 95\% \\ -0.00681 & \text{if demand } 100\% \\ 0 & \text{if demand } 105\% \end{cases} \tag{9}$$

The estimation of CV was used to estimate the required sample size. For a confidence interval of half-width 1% of the mean, we have the relationship described in Equation (9).

The predictive power of the model depended on the demand level and how far the number of signal nodes was from the mean of the input data (25 nodes) but the prediction interval widths were generally around ±0.03. Table 4 lists the estimates formed with a sample of levels of these factors.

Table 4: Key simulation inputs, and CV versus number of runs required to meet specified precision of 1% of Mean VHT

Demand level	Number of signal nodes	Mean *CV*	Approx. 95% PI for *CV*	min runs
95%	10	0.0073	(0, 0.037)	4
100%	10	0.0240	(0, 0.054)	23
105%	10	0.0247	(0, 0.055)	24
95%	20	0.0190	(0, 0.049)	14
100%	20	0.0357	(0.006, 0.066)	49
105%	20	0.0364	(0.006, 0.066)	51
95%	30	0.0307	(0.001, 0.061)	37
100%	30	0.0474	(0.017, 0.077)	87
105%	30	0.0481	(0.018, 0.078)	89

3. Conclusion

The joint MASCOS-ACCM-RTA project summarised in this paper has produced a methodological framework that can be used for analyzing simulation outputs, and informing the design stage of a simulation study by predicting simulation costs. The project aimed at improving the overall statistical rigor and defensibility of simulation study results. It produced a three-part set of guidelines that focus on what is needed prior to the analysis, during the analysis and a general model for cost predictions applicable to the planning and design stages of the analysis. Prior to the analysis summary measures need to be chosen (such as VHT),as well as the statistical parameter for between-model comparison such as the coefficient of variation.

During the analysis modellers are directed: to check data for independence, normality and outliers using graphical techniques; set precision and therefore sample sizes for single and two scenario comparisons. They are also recommended to perform power analysis for scenario comparisons, run diagnostic tests to check model performance and detect model errors; and lastly, quantify and correct errors due to congestion by examining the average add-back of each model.

Cost prediction in the guidelines is based on using the function that critical network features play in identifying inputs that have the greatest effect on variability. A regression model was built, based on an index of model complexity or on the combination of the most significantly interacting input factors. The regression model was limited by the small sample size of models used (a total of 36 models were available as inputs to the regression). Further research is continuing on cost prediction, using a larger set of simulation outputs and model types.

Overall, the guidelines developed have provided value to RTA traffic control modelling practice and can be used by simulation modellers regardless of the micro-simulation package used.

Acknowledgement

The authors wish to thank Dr Sandy Clarke of the University of Melbourne and Katrina Loukas, Centre Manager of the Australian Centre of Commercial Mathematics, for their many helpful suggestions on an earlier draft of this paper.

References

Abdy, Z and Hellinga, B (2008) Use of micro-simulation to model day–to-day- variability of intersection performance, *Transportation Research Record* 2088, 18-25

ACCM(2011) Australian Centre for Commercial Mathematics http://www.maths.unsw.edu.au/accm

Chatfield,C (1986) Exploratory data analysis, *European Journal of Operational Research* 23, 5-13.

Chatfield,C (1991) Avoiding statistical pitfalls, *Statistical Science 6* (3), 240-268. 1991A

Chong-White, C, Millar, G and Johnson, F, (2010) Introduction to modelling experimental design when operating SCATS within simulation, *Proceedings of the 17th ITS World Congress,* Busan , http://www.itsworldcongress.kr

Dowling , R (2002) *Guidelines for applying Traffic Micro-simulation Modelling Software,* California Department of Transportation *2002*

Federal Highway Administration (2004) *Traffic Analysis Toolbox Volume III*, U.S. Department of Transportation, Publication No. FHWA-HRT-04-040

Kelton, W D "Statistical analysis of Simulation Output", *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradottir et al

Law, A (2007) *Simulation Modelling and Analysis* (4th ed) New York: McGraw Hill

MASCOS (2010) Centre of Excellence for Mathematics and Statistics of Complex Systems (2010) http://www.complex.org.au

Mast and Trip (2007) Exploratory Data Analysis in Quality-Improvement Projects, *Journal of Quality Technology* 39 (4), 301-311, 2007

Millar, G (2010) *Private communication* Sydney: Azalient Pty Ltd

Montgomery D C (2001) *Design and Analysis of Experiments*, (5th ed) J Wiley & Sons Inc

Nakayama,M (2008) "Statistical analysis of Simulation Output", *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason et al

Niles, K (2010) Sample Wise, *Quality Progress,* May 2009*,* p.80

Quadstone Paramics Ltd., “Modeller”, *Quadstone Paramics* (traffic microsimulation software), Version 5.2.2, June 2006, http://www.paramics-online.com/product_modeller.php.

Roads and Traffic Authority of New South Wales (2009), *Paramics microsimulation modelling - RTA manual, RTA manual, Version 1.0,* , http://www.rta.nsw.gov.au/doingbusinesswithus/downloads/technicalmanuals/paramicsmanual_i.pdf [accessed 10 May 2009]

Roads and Traffic Authority of New South Wales (2010), Simulation software – SCATSIM, *SCATS software options* http://www.scats.com.au/product_family_options.html. [accessed 28 July 2010]

Seaman, J and Allen, I (2010) Outlier options, *Quality Progress, February 2010,* pp56-57

Shteinman, D, Clarke S, Chong-White, C, Millar, G and Johnson, F (2010) Development of a statistical framework to guide traffic simulation studies, *Proceedings of the 17th ITS World Congress,* Busan http://www.itsworldcongress.kr

Transport for London (2010) *Micro-Simulation Modelling Guidance Note for TFL,* Mayor of London

Transportation Research Board (2001) *Traffic flow theory a state of the art report*, The Committee on Traffic Flow Theory and Characteristics

Vining, G (2010) Quantile Plots to Check Assumptions, *Journal of Quality Engineering* Vol.22, pp.364-367, 2010