

DAVID SHTEINMAN
MANAGING DIRECTOR
INDUSTRIAL SCIENCES GROUP

david@stamen.com.au

Two Methods to Improve the Quality and Reliability of Calibrating & Validating Simulation Models

1. Introduction

The process of calibrating and validating all types of transport simulation models (micro, meso or macro) is a labour intensive and costly process. There have been a number of theoretical advances in methods for efficient testing and running simulation experiments that could be applied to the calibration and validation process. However they have not yet penetrated to actual modelling or transport simulation practice.

Our detailed review of the transport research literature showed that many of the state-of-the-art methods for validation and calibration are focused on the following: improving the Origin-Destination (OD) matrix estimation process; optimization algorithms to automate calibration; and Sensitivity Analysis on the input-output relationships of the model's parameters. While these methods are scientifically correct they are complex to understand and apply in practice, as the modeller needs an advanced level of mathematics to use them. Our consultation with simulation practitioners and our own experience in implementing changes to modelling practice [Shteinman 2010, 2011, 2012] lead us to expect a low probability of acceptance by practitioners of these methods.

As an alternative we draw on simpler statistical tools previously used in general simulation, and apply them to traffic simulation models. They can be easily applied to two related fundamental problems in validation and calibration:

Problem 1) Calibrating a model to base data this is itself highly variable. Usually base data has a high degree of uncertainty in how representative it is of the system being modelled; and

Problem 2) During validation: Uncertainty in how well the model represents the underlying structure of the traffic dynamics in reality. For example: the formation and dissipation of congestion.

The use of these statistical tool is illustrated by a detailed discussion of selected case studies from the transport research literature. For Problem 1, drawing from the work of Kim et al. [2005], we show how non-parametric distributional tests are used to compare the *actual distributions* of travel times in order to achieve better model calibration. This approach is contrasted to the usual practice of calibration of traffic variables to their Mean values.

Although the example uses travel times, the same approach can be applied to any other traffic metrics of interest. To emphasize the importance of considering the whole distribution,

instead of only its mean, we present and discuss visual and numerical statistical for travel times and traffic flow distributions, which also includes a case study of Melbourne travel times data.

Problem 2 is tackled by employing so called “meta-models” from the work of [Toledo, 2004]. Meta-model in this context is a model with a structure outside the structure of the validated model. The meta-model for the data generated by the model (simulated data) and the meta-model for the actual field data are compared for similarity using formal statistical tests.

This approach is illustrated by employing the Fundamental Diagram as a meta-model, and estimating Fundamental Diagram equations of observed data and simulated outputs. Then they are compared in terms of the similarity of important relationships between traffic parameters. E.g. the similarity in the “Flow-Speed” relationship between the observed data and the simulation. [Toledo, 2004].

Using the tools of simple visual data inspection (Explanatory Data Analysis), non-parametric distributional test and Meta-modelling does not require knowledge of advanced statistics. Such methods can be immediately introduced into modelling practice and yield improvements in the efficiency and reliability of model calibration and validation, at very low cost.

The rest of the paper is organized as follows: In Section 2 we give a background on the two problems mentioned above and how they have a negative effect on the quality and reliability of calibration and validation. The empirical properties of traffic volumes and travel times distributions, which are one of the most important traffic variables are discussed.

A more detailed case study of travel times distributions from the Melbourne arterial network, conducted by the Industrial Sciences group for Vicroads , using GPS-based travel time data, is presented. It clearly illustrates the complex the nature of travel times distributions and gives support to the case for using actual distributions (not point estimates) when calibrating a model to base data.

Section 3 has a discussion of calibrating models to the whole distributions of traffic variables (Problem 1). Section 4 has a discussion of using Meta-models in validation (Problem 2).

2. Calibration and Validation- Existing Methods

There is general agreement amongst simulation researchers and practitioners that the reliability of any simulation model depends on how well the field conditions are captured by the parameters in the model. Calibration is the process by which the parameters of various components of the simulation model are set so that the model will replicate observed traffic conditions as accurately as possible.

Calibration and Validation can be defined respectively, as:

Calibration is defined as the process of adjusting the model parameters, network and demand to reflect observed site data and/or conditions. The calibration process aims to produce a

model that is sufficiently refined to provide reliable forecasts that will be able to satisfy the study objectives.

The validation process compares model outputs with observed data that has not been previously used in the calibration process. Validation is therefore an independent verification to confirm that the model has been accurately calibrated.

Hence calibration has the objective of finding the values of the parameters that will produce a valid model, while validation provides an answer to the question: *Does the models' predictions faithfully represent reality?*

Many case studies describe the process of calibration and validation as a sequential and iterative process. Calibration requires the following four steps:

- (1) Calibration of driving behaviour models;
- (2) Calibration of route choice model;
- (3) OD estimation;
- (4) Model fine-tuning.

Austrroads [2006] adapts methods from the Modelling Guidelines of the RMS , and FHWA recommended practices, to describe the process as:

1. Network depiction
2. Calibrating capacity
3. Calibrating demand
4. Calibrating performance

One of the most time consuming and difficult jobs in calibration is the OD Matrix estimation process. However a significant factor contributing to this difficulty occurs *prior* to matrix estimation – the collection of base data. Matrix estimation requires adjustment of the Origin-Demand matrices to match traffic counts. Then a separate and independent set of traffic counts is used for validation. *Both traffic count data sets have inherent variability and uncertainty which is then propagated through the calibration and validation process.*

Model validation is the final stage of checking if each component adequately reproduces observed travel characteristics and if the overall performance of the model is reasonable. This is done by comparing traffic flows outputted by the model against observed counts. To quantify the comparison two measures of “goodness-of-fit”, GEH and RMSE criteria are widely used by practitioners, and recommended in Guidelines in the UK, Australia, and the US.

The GEH criterion is defined as:

$$GEH = \sqrt{\frac{(M - O)^2}{(M + O)/2}}$$

where M and O , are the estimated and observed volumes, respectively. Although the formula above is similar to the well-known chi-square statistics, GEH is not a formal statistical test but a measure of goodness-of-fit which is less sensitive to large discrepancies between predicted and observed volume than other measures (e.g. absolute difference). As such, GEH has been proven to be useful in practical applications.

The Root Mean Square Error (RMSE) is a statistical measure of the correlation between the entire count data set and the predicted model volumes:

$$RMSE = \sqrt{\frac{\sum (V_o - V_m)^2}{C - 1}} \times \frac{\sum V_o}{C} \times 100$$

where V_o and V_m are the observed and modelled vehicles per hour, respectively and C is the total number of count locations [RMS, 2013, p.84]. Other calibration targets used by RMS, FHWA, and endorsed by Austroads are link flows on cordon/screenline, individual link flows as well as travel times [Austroads, 2006, p.19].

For calibration acceptance criteria or targets, most guidelines have the following consensus:

a GEH index of less than 5.0 in a measurement point is considered a good match between the modelled and observed volumes, and 85% of the volumes in a traffic model should have a GEH less than 5.0 for all measurement points [Austroads, 2006, p.19]

Unlike the GEH statistic, the RMSE applies to the entire comparison data set and is expressed as a single value. Although the GEH mathematical form is similar to a chi-squared test, it is not a true statistical test. Rather, it is an empirical formula that has proven useful for a variety of traffic analysis purposes. The main trait of the GEH formula is that, unlike the relative error, differences in high volumes are relatively more important than in lower volume. Thus the GEH measure is a form of Chi-squared statistic that is designed to be “tolerant” of larger errors in low flows. The reason for introducing such a statistic is the inability of either the absolute difference or the relative difference to cope over a wide range of flows.

2.1 Traffic Volume Variability and Calibration/Validation

The RMS guidelines note that “Traffic volumes are often used as a key statistical indicator that the model is sufficiently calibrated as they provide an easily measurable dataset both in the model and on site” [RMS, 2012, p.103].

However traffic volumes and travel times are highly variable and point estimates is not a good indicator of “similarity “between model and reality. Empirical evidence suggests [Ref Guide to benchmarking operations and performance measures, TARB, Appendix D, available at http://www.catt.umd.edu/sites/default/files/documents/final_report_compiled_v26.pdf] that the traffic volumes have complex dynamics, with a non-trivial relationship between mean traffic volumes and the variability of traffic volumes.

The report uses data for several arterial links in Northern Virginia, USA and traffic volume measured on the basis of traffic counts in 15-minutes intervals.

Volume (veh/hr/ln)	Coefficient of Variation of 15-minute Volume Counts			
	Average	Median	70 th Percentile	80 th Percentile
0-200	0.92	0.93	0.97	1.00
200-400	0.25	0.24	0.28	0.29
400-600	0.16	0.15	0.17	0.18
600-800	0.13	0.12	0.13	0.14
800-1000	0.12	0.11	0.12	0.13
1000-1200	0.13	0.08	0.09	0.13
1200-1400	0.15	0.11	0.14	0.20
1400-1600	0.06	0.06	0.07	0.07
1600-1800	0.06	0.06	0.06	0.06

Table 1 Coefficient of Variation of 15-minutes volume counts for arterials

Credit : Tarnoff et al. (2008):

http://www.catt.umd.edu/sites/default/files/documents/variance_analysis_v1.pdf

The table above gives statistics for the Coefficient of Variation (COV) of traffic volume from the report. The Coefficient of Variation is defined as the ratio of Standard Deviation of traffic volume to the mean value of the traffic volume. COV statistics are presented for different traffic conditions ranging from light (volume between 0 and 200 veh/hr/ln) to heavily congested (volume between 1600 and 1800 veh/hr/ln).

It can be seen that the COV and correspondingly, the Standard Deviation of traffic volume *varies substantially between traffic conditions*. In uncongested states, the Standard Deviation is high relative to the mean traffic volume, but the relative difference between the two sharply decreases when the traffic conditions become more congested. This suggests that the traffic volumes have complicated behaviour and accounting for some of their empirical features might be of benefit when calibrating models.

The outcome of validating a properly calibrated model can be quantified as [Barceló, 2011]

$$P\{ | \text{ "reality" } - \text{ simulation prediction} | < d \} > a$$

where

d= tolerable difference = how close is the model to reality

a=level of assurance = how certain are we of this?

We are concerned with determining the factors in simulation modelling that have most effect on the level of assurance; That is the measure of how certain we are of the observed traffic counts that are used to represent reality. This in turn is a quality measure of the base data. Quality in terms of level of variability between days of the data period, variability within the day, seasonality, outliers, and errors in the data collected

It is important to note that the observed data which is used to represent “reality” is random as it is a randomly selected sample from a population of possible days and traffic conditions. The simulated output is also randomly generated.

2.2 Features of Travel Time Distributions and Calibration/Validation

Travel time are another very important metrics used in model calibration. Again, the usual practice is to concentrate on fitting the mean of the travel time distribution (in one form or another) and completely disregarding the shape of the distribution.

However, travel time distributions have complex shape and additionally they might change with traffic conditions. Hence, summarizing the dynamics of travel times by a single value (e.g. mean) would be a great simplification and would bias the calibration process.

The example from Van Lint et al. (2008) neatly illustrates this point. The authors consider the travel times on a 6.4 km freeway stretch in Netherlands in four different conditions: free-flow, congestion onset, congestion and congestion dissolution.

The corresponding empirical distributions are shown in the Figure 1 below. It can be seen that in free-flow and congested conditions, the travel time distribution is symmetric (with short width in free-flow condition and wider width in congested condition). In this case standard deviation would be sufficient to fully describe the dispersion of the travel times. However, in the onset of congestion as well as in the moment of congestion, the travel times are clearly skewed on the right.

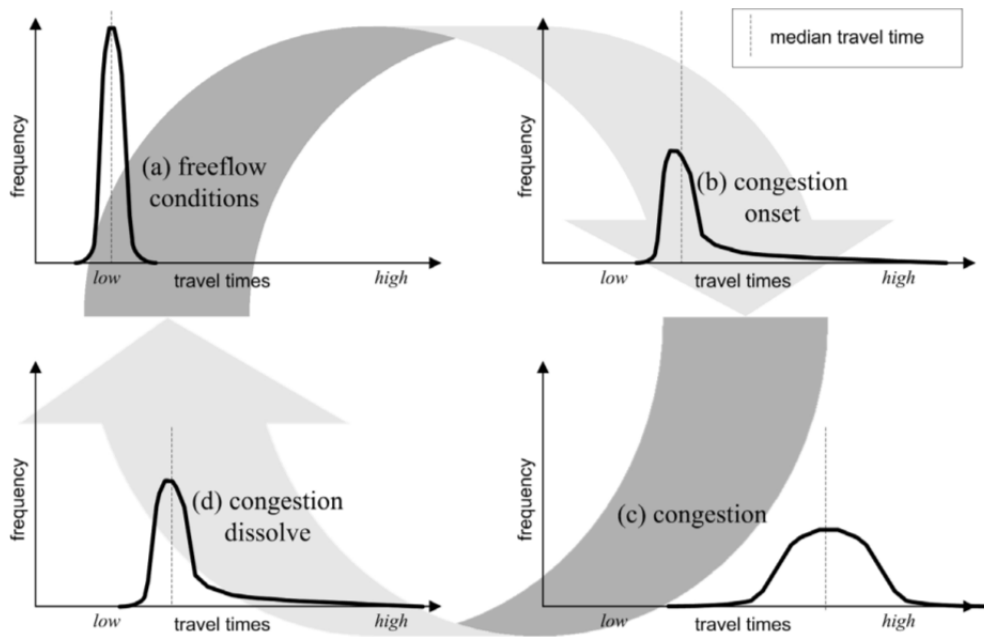


Figure 1 Travel Time Distributions in Different Traffic Conditions
Credit: van Lint et al. (2008)

2.3 Case Study of Travel Times

In another example of the complexity of travel times, more relevant to Australian context, we present a brief analysis of a travel time data sample which closely follows Shteinman and Lazarov [2014]. The original data is collected by Intelomatics and is derived from vehicles with GPS's. Our intention is to show how to use simple visual statistics (Explanatory Data Analysis), when considering the whole travel time distributions in the calibration process.

The data consists of travel times for 6 links along the South Road in Melbourne defined as intersections of the South Road with 6 roads (in order): Nepean Highway, Station Street, Jasper Road, Tucker Road, Chapel Road, East Boundary Road, Chesterville Road. The data spans the interval 1 Jan 2013 to 18 Apr 2014 for each day of the week (total of 447 days) with six days of missing data. Average travel times as well average speeds are available for each of the 96 15-minutes intervals in the 24 hour day.

First, sample statistics for the intra-day mean and Standard Deviation are computed. Visual inspection of the data reveals that it is noisy and has numerous spikes, which might bias the Standard Deviation (SD) estimation. For the sake of brevity, results for East Boundary to Chapel Road link are shown in Figure 2.

In order to account for noise, we also compute a robust measure of the dispersion of the distribution, called Mean Absolute Deviation (MAD) which is analogous to SD, but it is more resistant to outliers (For a set of observations x_1, \dots, x_N Mean Absolute Deviation (MAD) is defined as $\frac{1}{N} \sum_{j=1}^N |x_j - M|$).

There is a clearly observable difference between the standard deviation (SD) and the mean absolute deviation (MAD). The SD plots are noisier, while MAD plots are smoother and also exhibit very similar intra-day pattern to the mean travel times plot. This suggests the presence of outliers which bias the SD estimates.

Also, the mean travel times as well as the standard deviation clearly exhibit intra-day seasonality, both having markedly larger values during the time period [6:00 AM; 8:00 PM] and even higher values during peak hours. Interestingly, for other links we didn't find evidence for higher mean/ standard deviation of travel time during peak hours relative to the time period [6:00 AM; 8:00 PM].

We also found that the shape of travel time distributions can vary significantly from link to link and time of the day. Overall, we found out that the log-logistic distribution fits the data reasonably well.

For example, in Figure 3, the histogram of the travel time distribution in the interval [9:00 AM; 9:15 AM] as well as best fit to the normal, log-normal and gamma distributions are shown for the link Nepean Highway to Station Street. The graph clearly shows that the normal distribution doesn't fit the data very well. In particular, there is a pronounced skew on the right-hand side. This is partially captured by the log-normal distribution, however the gamma distribution provides the best fit.

The other feature of the graph is the observed spikes in the histogram. In this case, a very pronounced frequency of observations of mean travel times with value 35 seconds is observed (there are 34 such observations) while the number of observations with values between 32 and 35 seconds, and between 35 and 37 seconds drops to 2 and 16, respectively. This clearly suggests an artificial filling of the data by the number 35.

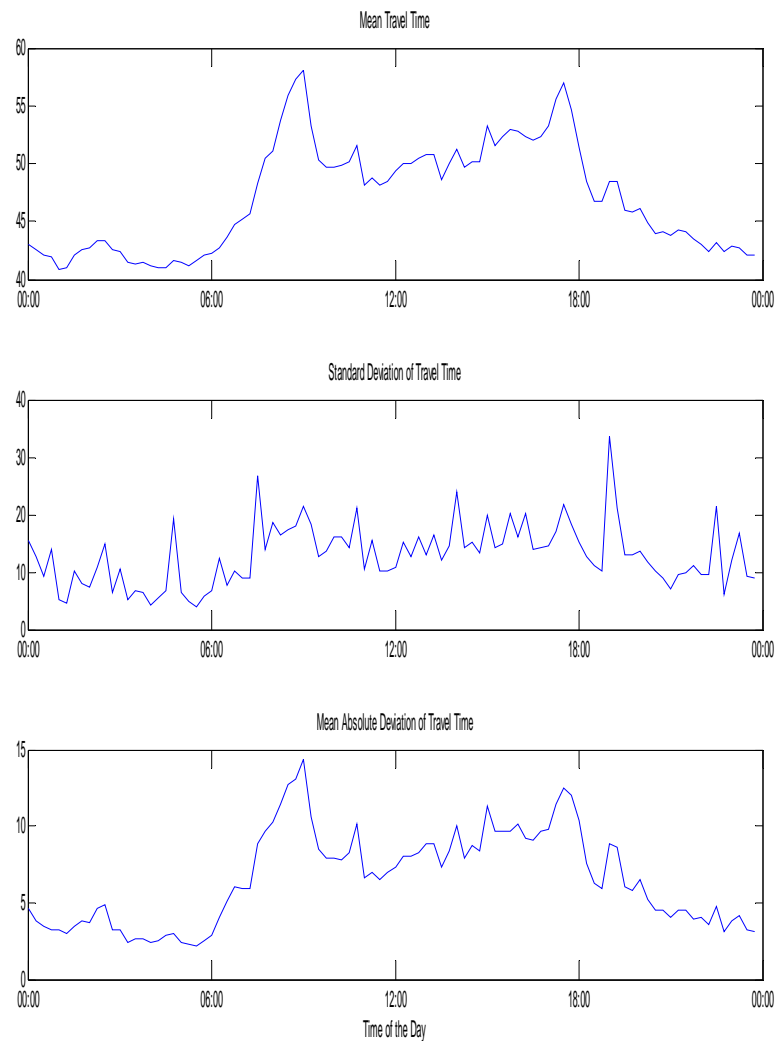


Figure 2: Mean Travel Time, Standard Deviation and Mean Absolute Deviation for East Boundary Rd to Chapel Rd to Link (688 mt.).

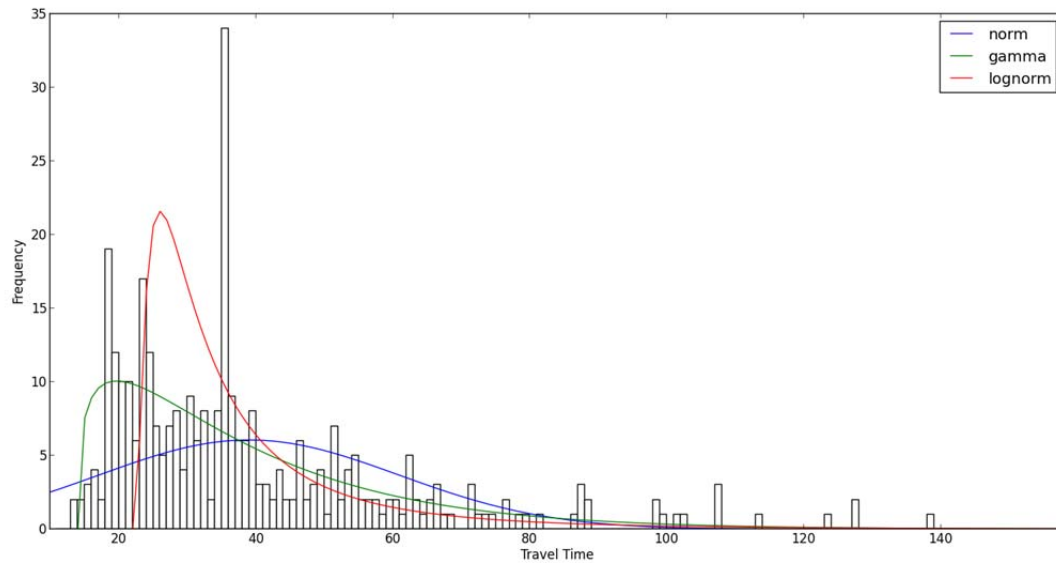


Figure 3: Histogram of the Mean Travel Time in the interval [9:00 AM; 9:15 AM] for Nepean Highway to Station Street Link

From our brief discussion of the travel times data above, it is clear that there are (at least) two main issues which have to be dealt with: outliers and repeated observations. This emphasizes the importance of EDA (Explanatory Data Analysis) and careful travel times data pre-processing before using it for practical applications, such as model calibration and validation.

3. Problem of Comparing Point Estimates – and a Potential Solution

Historically, traffic simulation models have been calibrated to small data samples due to the expense of obtaining large data sets. Various metrics for the closeness between the simulated and observed data have been proposed and used (as discussed above). Most of them are based, in one form or another, on comparing the central tendencies. Central tendency of a random variability is loosely defined as its typical value (e.g. mean, median and so forth).

The degree of which the model is deemed to represent “reality” is determined by how well it can match some measure for the Mean of the field data. This approach is only valid if the simulated output and the observed data come from the same probabilistic distribution, with possibly different mean values.

However, there is no reason that this should be the case, and the recent increase in availability of traffic data, more specifically travel times data, allows the modeller to take into account the *full distributional properties* of the data of interest during calibration.

It is shown below in Figure 4 that two sets of parameters can have a comparable degree of fit to the empirical data on the basis of some criterion for central tendency (mean), and yet they

can produce widely different distributions. Such an example is given in the paper by Kim et al. (2005) where two parameter sets of a VISSIM micro-simulation model are compared in terms of how well they can reproduce the actual travel time distribution. Both parameter sets produce comparable fit to the empirical data given by about 1% MAER error. The travel time distributions are very different however. (Mean Absolute Error (MAER) is defined as

$$MAER = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i^s - x_i^o}{x_i^o} \right|$$

where x_i^s is the simulated data and x_i^o is the observed data.

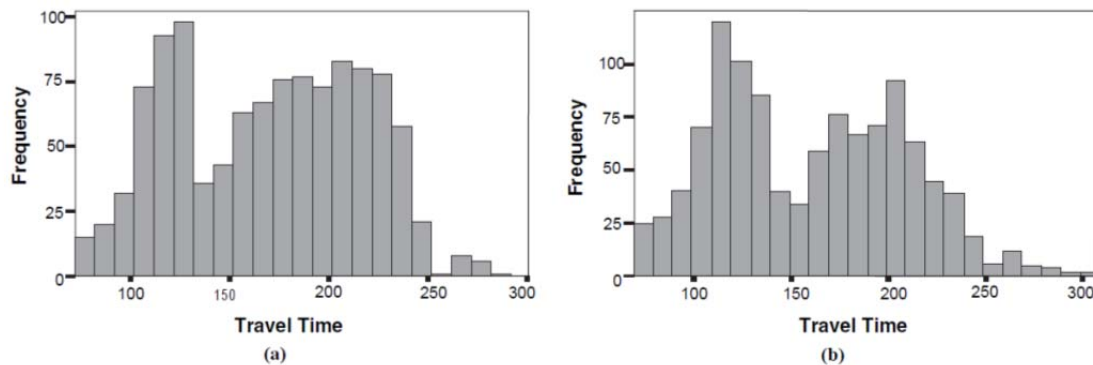


Figure 4: Comparison of travel time distributions: a) simulated travel time distribution (1% MAER) – rejected and b) simulated travel time distribution (1% MAER)– accepted
Credit : Kim et al. (2005)

The first travel time distribution is uni-modal and it is rejected by statistical tests for the closeness to the actual travel time distribution, while the second travel time distribution is bi-modal and can't be rejected by standard distributional tests.

A simple way of avoiding choosing models' parameters which result in distributions for the output variables that are substantially different from the actual one, is to perform a test or a series of tests for the closeness of the two distributions (simulated and observed). The results of the distributional test(s) could serve as a filter of the pool of possible candidates for parameters (see figure below):

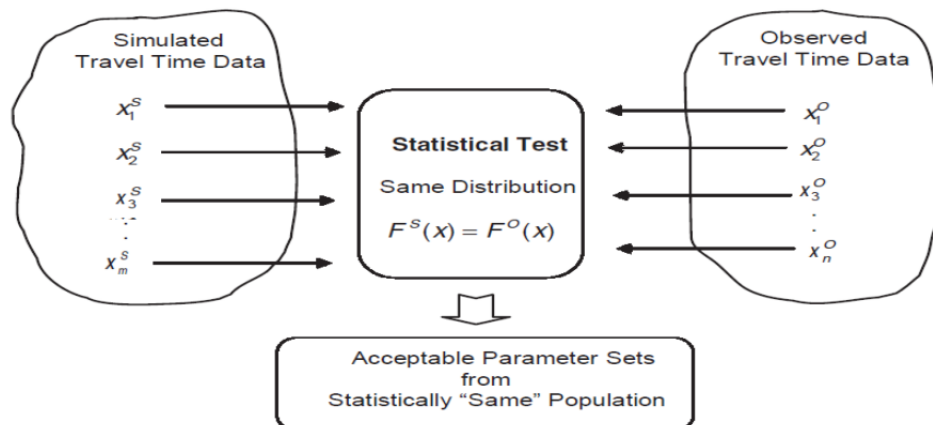


Figure 5: Conceptualization of disaggregated performance measures in calibration
Credit : Kim et al. (2005)

The most widely used statistical test for similarities of distributions is the Kolmogorov-Smirnov (K-S) test, which is also used in the paper by Kim et al. (2005). The idea of the K-S test is to build the empirical cumulative distribution functions $\bar{F}^Y(x)$ and $\bar{F}^Z(x)$ from the observed realizations y_1, \dots, y_n and z_1, \dots, z_m of two different random variables Y and Z , respectively.

We can test whether y_1, \dots, y_n and z_1, \dots, z_m come from the same distribution by computing the magnitude of the maximal absolute difference D between the empirical distribution functions. More formally, the K-S statistics D is given by:

$$D = \max |\bar{F}^Y(\cdot) - \bar{F}^Z(\cdot)|$$

Note that, the empirical distribution functions $\bar{F}^Y(x)$ and $\bar{F}^Z(x)$ are approximations for the cumulative distribution functions $F^Y(x)$ and $F^Z(x)$, which are defined as:

$$F^Y(x) = \text{Prob}[Y \leq x]$$

$$F^Z(x) = \text{Prob}[Z \leq x].$$

That is, the value of $F^Y(x)$ ($F^Z(x)$) is the probability that the random variable Y (Z) will be smaller or equal to x .

The empirical distribution functions are built from the observed observations in a straightforward way. For example, if $y_1^* \leq \dots \leq y_n^*$ is the ordered sequence of observations y_1, \dots, y_n , the empirical distribution function $\bar{F}^Y(x)$ is a step function which starts at 0 and increases by a fraction of $1/n$ at each point y_i^* :

$$\bar{F}^Y(y_i^*) = \frac{i - n}{n}$$

The function $\bar{F}^Z(x)$ is defined in a similar way. In the figure below (Figure 6), a graphical representation of the K-S statistics is shown. On the basis of the K-S statistics D we can test whether the two distributions are identical at certain significance level α . Intuitively, the higher the K-S statistics is, the most likely is that the two distributions are different.

More formally, we can reject the hypothesis that the two distributions are the same (at significance level α) if:

$$D > \sqrt{\frac{n + m}{nm}} K(\alpha)$$

Here $K(\cdot)$ is a function of α and pre-calculated tables for it exist. The most common choice for significance level is 5 % ($\alpha = 0.05$), in which case $K(\alpha) = 1.36$.

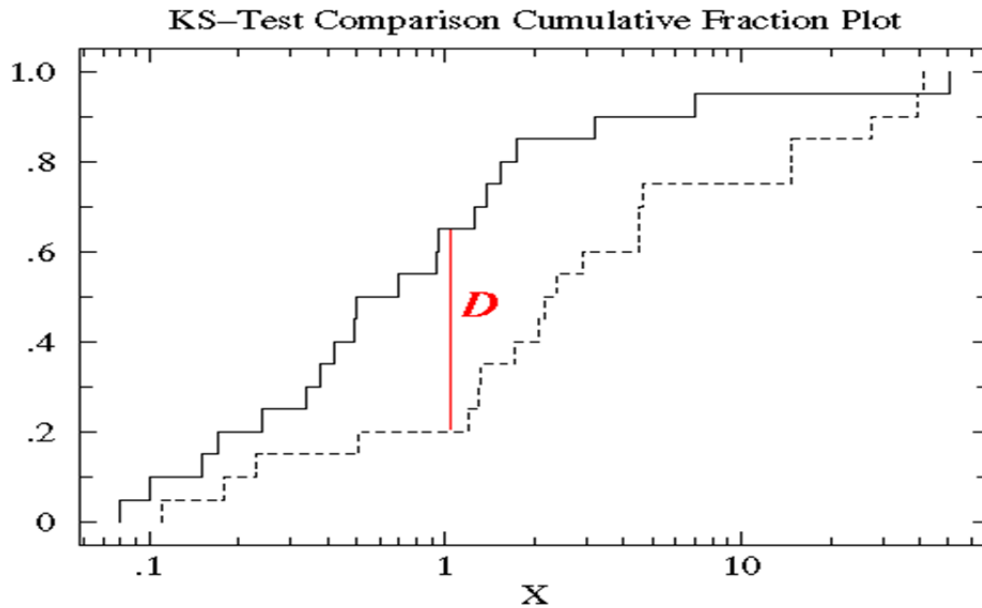


Figure 6. Graphical representation of the K-S test. Source: Kirkman, T.W. (1996) Statistics to Use, <http://www.physics.csbsju.edu/stats/>

It can be seen that the K-S test has a clear intuitive interpretation and also easy to perform. It can be used to filter out a simulation model's parameter values which lead to erroneous outputs for the whole distributions of a variable.

4. Use of Meta Models in Model Validation

As described above, validation is the process whereby model outputs are compared to observed data that has not been previously used in the calibration process. Hence validation provides an answer to the question: *Does the models' predictions faithfully represent reality?*

To "faithfully represent reality" suggests the model can reproduce the underlying "behavioural" features of the modelled network not just an average behaviour. A formal approach how to use this insight in practice is proposed in a theoretical paper by Toledo and Koutsopoulos hereafter the "TK method" [Toledo, 2004]. Their approach has not yet penetrated modelling practice (as far as is known by this author). We believe it has the potential to greatly enhance the validation process. Our aim here is to explain their approach so it can become a useful tool for modellers in validation, without needing advanced mathematical skills.

The TK method is based on the use of Meta-models. In this context, a Meta-model is a model which captures an important feature of the behaviour of the network being modelled and at the same time, it is not part of the validated model.

In other words, the meta-model is chosen to capture important relationships between traffic variables, which are not explicitly represented in the base model. Another useful feature to

have in mind when choosing appropriate meta-model for validation is that is relevant to the task at hand.

To test whether the validated model is “correct”, the meta-model for the data generated by the model (simulated data) and the meta-model for the actual field data are compared for similarity using formal statistical tests.

Examples of "behaviours or underlying structure" of a traffic network can include the relationships of flow-density or speed-density of a link or set of links in the network. These relationships are described by empirically derived equations and graphically displayed in the Fundamental Diagram of traffic flow. The TK paper has a detailed example of employing the speed-density Fundamental Diagram of traffic flow as a Meta-model for validation.

The Fundamental Diagram is well known as a model of motorway traffic flows and recently as a way of describing macroscopic flow behaviour in some types of arterial networks [Geroliminis, 2008]. However it has not yet been used as a tool for validating simulations.

First, the meta-model (e.g. the equation for the Fundamental Diagram) is estimated from the simulated data. Then, the same meta-model is estimated from the observed data. Then a metric for “closeness” between the simulated and observed data is given by the difference between estimated parameters of the Meta models for the observed and simulated data.

The benefit of the Meta-model approach is that it can overcome the problem caused by limited data. In simulation studies observational data is usually only available for 1 or 2 so-called “representative days” and aggregated statistics are used (E.g. GEH and RMSE).

Disaggregate data (e.g., observations of individual vehicles, 1-min sensor readings) can be used when fitting the Meta-model without having to aggregate observations (e.g., to 5-min intervals) as would be required if real and simulated outputs were directly compared [Toledo, p.147]. Thus the Meta-model approach uses more of the limited data available than the traditional approach which aggregates the data to compare Means (for example).

For validation purposes, the test of “closeness” between Meta –models of the observed data and the simulation is given by the F-test (a simple statistical test for the equality of the estimated parameters of the two meta-models.). In outline the method works as follows:

$$\begin{aligned}
 \textit{Simulated Data} &\Rightarrow \textit{Meta Model Estimation (e.g. FD)} \\
 &\Rightarrow \textit{parameters of meta model } p^S \\
 \textit{Historical Data} &\Rightarrow \textit{Meta Model Estimation (e.g. FD)} \\
 &\Rightarrow \textit{parameters of meta model } p^0 \\
 \textit{Degree of Closeness of Model to Reality} \\
 &: F - \textit{test statistics for equality of } p^S \textit{ and } p^0
 \end{aligned}$$

4.1 Choosing a Form of Fundamental Diagram to Use

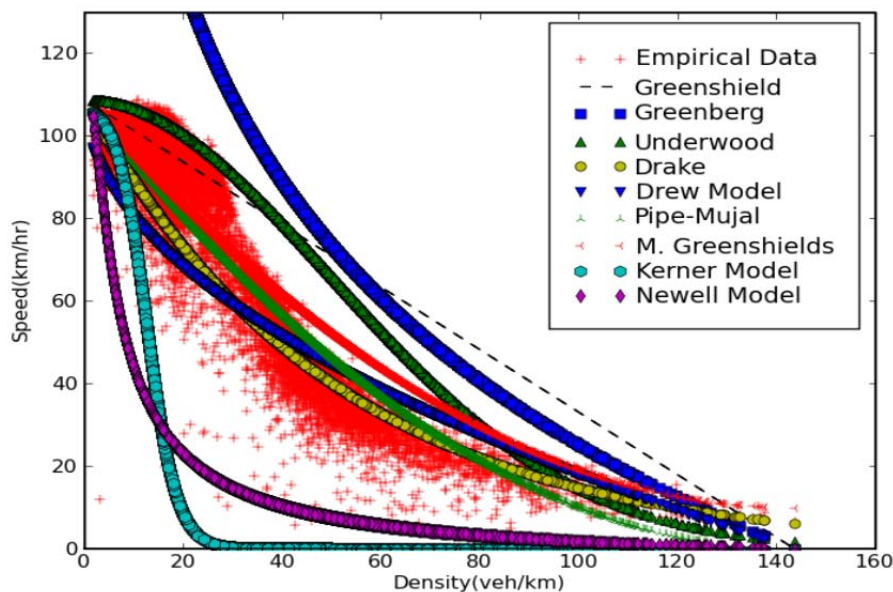
The simplest type of FD in steady state for uninterrupted traffic flow on a motorway is the flow-density FD. From the relationship

$$\text{Flow} = \text{density} * \text{space mean speed}$$

The flow-density FD can be represented as a speed-density or flow-speed FD in a straightforward way. Historically, the concept of FD was proposed from experimental observations by US traffic engineer Bruce Greenshields in 1933, who postulated a *linear relationship* between speed and traffic density. However it has long been observed that a linear specification often describes the data poorly. For example, quite often the FD has a crucial point at which the traffic changes from free-flow to congested condition and correspondingly the functional relationship changes. Various other functional relationships between the speed V and density k have been proposed to account for the observable data features (see Table 2 below for summary). Figure 7 shows the empirical performance of some of the models in Table 1 using a data for a route in Atlanta, Georgia. The original model of Greenshields clearly doesn't fit the data very well. More complex models with higher number of parameters are a better fit.

Model	Functional Relationship	Parameters
Greenshields (1935)	$V = v_f(1 - k/k_j)$	v_f, k_j
Greenberg (1959)	$V = v_m \log(k/k_j)$	v_m, k_j
Underwood (1961)	$V = v_f \exp(-k/k_j)$	v_f, k_j
Northwestern (1967)	$V = v_f \exp\left(-\frac{1}{2}(k/k_0)^2\right)$	v_f, k_0
Drew (1968)	$V = v_f \left(1 - (k/k_j)^{n+\frac{1}{2}}\right)$	v_f, k_j
Pipes-Munjal (1967)	$V = v_f(1 - (k/k_j)^n)$	v_f, k_j
Newell (1961)	$V = v_f \left(1 - \exp\left(-\frac{\lambda}{v_f} \left(\frac{1}{k} - \frac{1}{k_j}\right)\right)\right)$	v_f, k_j, λ
Modified Greenshields	$V = v_0 + (v_f - v_0)(1 - (k/k_j)^\alpha)$	v_0, v_f, k_j
Van Aerde (1995)	$k = 1/\left(c_1 + \frac{c_2}{(v_f - v)} + c_3 v\right)$	c_1, c_2, c_3, v_f
MacNicholas (2008)	$V = v_f \left(\frac{k_j^n - k^n}{k_j^n - m k^n}\right)$	v_f, k_j, m, n

Table 2 Single-regime Speed-Density FD Models Credit: Wang et al. [2010]



**Figure 7: Empirical Fit of Single Equation Models for the Fundamental Diagram
(Credit: Wang et al. (2010))**

Toledo suggests using the Pipes-Munjal FD model (see Munjal and Pipes (1971) for detailed discussion). However our review of the literature showed that the FD model of Wang et al. [2010] proposes a flexible specification for the speed-density fundamental diagram. It has major advantages over other existing models:

- 1) It can fit the data with reasonable accuracy although it is a single regime model.
- 2) The curve that produces has desirable mathematical properties (e.g. it is smooth and bounded – doesn't go to infinity or zero).
- 3) The parameters of the model have a clear interpretation.

The equation for the speed-density relationship is:

$$(2.1) V = v_b + \frac{v_f - v_b}{\left(1 + \exp\left(\frac{k - k_t}{\theta_1}\right)\right)^{\theta_2}}$$

The parameter v_f is free flow speed and v_b is the average travel speed at stop and go condition. θ_1 is a scale parameter which describes how the curve is stretched out over the whole density range, θ_2 is a parameter which controls the lopsidedness of the curve. The parameter k_t is the turning point that the speed-density curve makes the transition from free-flow to congested flow.

4.1 Using the Fundamental Diagram in Model Validation

Toledo gives an illustration of the approach by using it to validate a micro-simulation mode of a section of the M27 motorway in Southampton, UK. The observational data includes traffic

counts and speeds for 5 days at 1 minute intervals using loop detectors on a 4.3 km length of the motorway. The micro-simulation outputs are the same measures (counts and speeds) from 10 runs.

For our purposes we will summarize the main results and expand on what is relevant for further implementation of the approach: The standard goodness of fit tests were used on the speed outputs of observed vs. simulated showed good overall fits. However, as stated above, a positive Goodness of Fit result *does not indicate* how well the simulation is replicating the dynamics of traffic flows in the network being modelled.

For implementing the following 3 –step process is recommended:

Step 1: Decide which meta-model to use for validation.

TK used the Fundamental Diagram with Pipes-Munjal specification for the speed-density:

$$V = v_f \left(1 - \left(\frac{k}{k_t} \right)^n \right)$$

where v_f is the free-flow speed and k_t is the critical density, n is an exponent parameter.

For improved representation of traffic dynamics we recommend a three-regime FD model such as the Wang-Li (2010) FD model described above.

Step 2: Estimate the meta-model for simulated and observed data, and plot the graphical representation of both meta-models outputs, if possible.

Simple visual inspection of the meta-models is always advisable before proceeding with more complicated statistical techniques.

In the current case, the parameters of the meta-models are given by the parameter vector $\beta = (v_f, k_t, n)$. The value of the vectors is chosen on the basis of minimizing the sum of squared differences between the meta-model outputs for the speed $\hat{V}_1, \dots, \hat{V}_n$ and the values of the speed which are used for the meta-model calibration V_1, \dots, V_n :

$$(\hat{V}_1 - V_1)^2 + \dots + (\hat{V}_n - V_n)^2.$$

Efficient algorithms for solving this type of problems (non-linear least squares) such as Levenberg-Marquart method exist (Press et al. [2007])

The Fundamental Diagrams for the observed and simulated data are shown below.

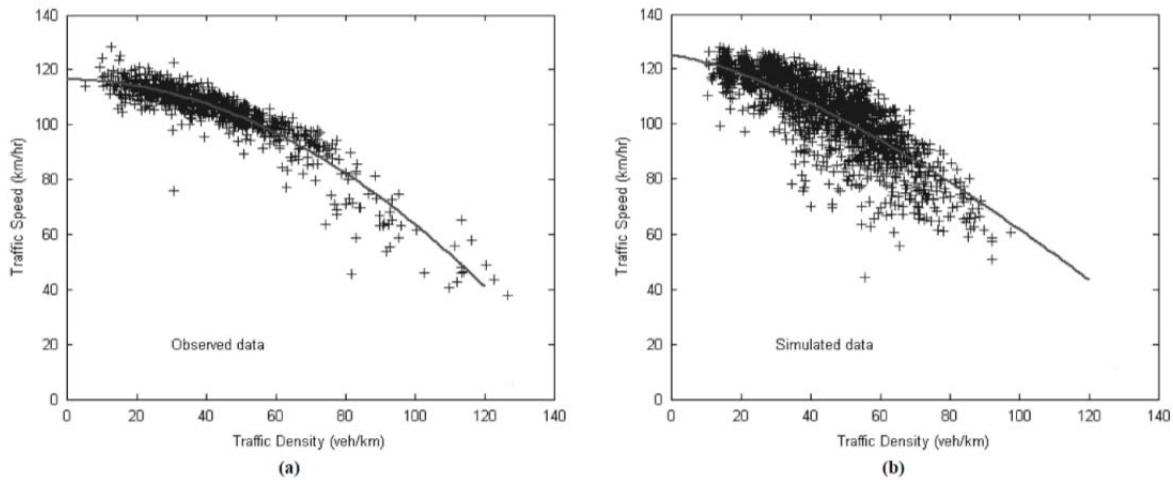


Figure 8: Observed and simulated speed–density FD Meta –models
Credit: Toledo [2004, p.148]

Visually, the two curves are very similar to each other. Formal statistical test (the step below) reveals however that we can reject the hypothesis that the parameters which determine the shape of the two curves are identical. Thus, a visual inspection of the regression curves and data points (as in Fig. 8 above) is not sufficient

Step 3: Test of the closeness between the meta-model for the observed and simulated data.

A formal statistical test can be devised to test if the parameters of the meta-models on the basis of the simulated and observed data are identical. It is based on the F-test which is a standard statistical tool. We will briefly describe its application in the current context. For more in depth explanations and discussions the reader is referred to any standard statistical text e.g. Lomax (2007).

The F-test is performed by estimating two general meta-models: both of them use the combined simulated and actual data. The first meta-model have separate coefficients for the simulated and actual data (unrestricted meta-model) and the second meta-model forces equality of these two coefficients sets (restricted model).

The F-test is based on comparing how well the unrestricted model fits the data in comparison to the restricted model. If the benefit of not restricting the coefficients is only marginal, we can conclude that the meta-models for the simulated and observed data are not statistically different.

The goodness of fit of the two models is measured by the sum of the squared residuals: SSR^U and SSR^R – sum of the squared residuals for the unrestricted and restricted models, respectively.

Since the unrestricted model has twice the number of coefficients than the restricted one it will always provide a better fit e.g. $SSR^U \geq SSR^R$. The idea is to adjust the difference in goodness of fit for the fact that the unrestricted model has more degree of freedom.

It can be shown that under certain general conditions, the ratio

$$F = \frac{(SSR^R - SSR^U)/K}{SSR^U/(N^o + N^s - 2K)}$$

has F-distribution with parameters $F_{K, N^o + N^s - 2K}$. Here N^o and N^s are the number of data points in the observed and simulated data, respectively and K is the number of parameters (three in the current case).

Comparing the value of the test statistics F to a distribution table for $F_{K, N^o + N^s - 2K}$ allows us to test the (null) hypothesis of equality of coefficients of the meta-models for the observed and simulated data. In the example, above the (null) hypothesis that both coefficient sets are identical can be rejected.

Thus, the calibrated model in the TK example doesn't pass the calibration step. This step could be performed in similar fashion to other (traffic) meta-models.

5. Conclusion

In this paper we have attempted to offer practical solutions to two fundamental problems arising in the calibration and validation of traffic simulations: namely the problems of Calibrating a model to base data this is itself highly variable which leads to a high degree of uncertainty in how representative it is of the system being modelled; and uncertainty in how well the model represents the underlying structure of the traffic dynamics in reality (the validation process).

Our proposed solution to the calibration problem is to first offer the easy to use tools of EDA to help modellers be aware of the features of the variability of base data traffic counts and travel times. This is achieved by graphing the observed data, and taking into account the shape of the distribution and effect of outliers.

6. Acknowledgments

The author would like to thank Dr Zdravetz Lazarov, Senior Economic Modeller at Industrial Sciences Group for his comments and suggestions on this paper. I would also like to thank Gary Millar of Adanner Consultants for his discussions on current calibration and validation practice.

7. References

Austrroads (2006) Research Report AP-R286/06 *Use and Application of Micro Simulation Traffic Models*, Austrroads Incorporated, Sydney

Barcelo, J. (2011) *Discussion On Calibration & Validation Of Microscopic Traffic Simulation Models*, presentation to EU Multitude Workshop

Geroliminis, N ; Daganzo , N (2008) *Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings*, Transportation Research Part B , 42(2008), pp.759-770

Kim, S; Kim W, Rilett, L (2005) *Calibration of Micro simulation Models Using Non-Parametric Statistical Techniques*; Transportation Research Record, No.1935, TRB 2005, pp.111-119

Lomax, Richard G. (2007) *Statistical Concepts: A Second Course*

Munjal P.K., Pipes L.A. (1971) *Propagation of On-Ramp Density Perturbations on Uni-Directional and Two- and Three-Lane Freeways*. Transportation Research Part B, 5(4):241-255

Press WH, Teukolsky SA, Vetterling WT, Flammig BP. (2007) *Numerical Recipes: The Art of Scientific Computing*. Third Edition. Cambridge University Press

Roads and Maritime Services (2013) *Traffic Modelling Guidelines*, Sydney, NSW

Shteinman, D, Lazarov, Z (2014); *Review Of Research In Travel Time Variability With Data Analysis*; Technical Report to Vicroads; Industrial Sciences Group, Sydney

Shteinman , D (2012) "Mathematics for Traffic Engineering " *Engineers Australia Magazine (Civil)* November 2012

Shteinman, D, Chong-White, C, Millar, G (2011) Advanced development of a statistical framework to guide traffic simulation studies, *Proceedings of the 34th Australasian Transport Research Forum Adelaide* <http://www.atrf11.unisa.edu.au/Default.aspx>

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>

Toledo, T; Koutsopoulos, *Statistical Validation of Traffic Simulation Models*, Transportation Research Record, No.1876, TRB 2004, pp.142-150

Van Lint , J.W.C., Van Zuylen, H.J., Tu, H. (2008). *Travel time unreliability on freeways: Why measures based on variance tell only half the story*. *Transportation Research 42A*, 258-277

Wang, Li et al (2010) *Representing the Fundamental Diagram: the Pursuit of Mathematical Elegance and Empirical Accuracy*, presentation at the TRB 89th Annual Meeting at Washington D. C. Jan. 2010