European Commission

JRC SCIENCE AND POLICY REPORTS

Traffic Simulation: Case for guidelines

COST Action TU0903 MULTITUDE

C. Antoniou, J. Barcelo, M. Brackstone, H.B. Celikoglu, B. Ciuffo, V. Punzo, P. Sykes, T. Toledo, P. Vortisch, P. Wagner

Edited by: M. Brackstone, V. Punzo

2014

Report EUR 26534 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for Energy and Transport

Contact information
Biagio Ciuffo
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 441, 21027 Ispra (VA), Italy
E-mail: biagio.ciuffo@jrc.ec.europa.eu
Tel.: +39 0332 789732
Fax: +39 0332 786627

http://iet.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

This publication is a Science and Policy Report by the Joint Research Centre of the European Commission.

JRC88526

EUR 26534 EN

ISBN 978-92-79-35578-3 (pdf)
ISBN 978-92-79-35579-0 (print)

ISSN 1831-9424 (online)
ISSN 1018-5593 (print)

doi: 10.2788/11382

Luxembourg: Publications Office of the European Union, 2014

Case for Guidelines

Prepared for MULTITUDE

Preface

The MULTITUDE Project (Methods and tools for supporting the Use, caLibration and validaTIon of Traffic simUlations moDEls) is an Action (TU0903) supported by the EU COST office (European Cooperation in Science and Technology) and focuses on the issue of *uncertainty* in traffic simulation, and of *calibration* and *validation* as tools to manage it. It is driven by the concern that, although modelling is now widespread, we are unsure how much we can trust our results and conclusions. Such issues force into question the trustworthiness of the results, and indeed how well we are using them.

The project consists of 4 Working Groups (WGs) which hold short, focussed, meetings on topics of interest and propose work items on key issues. Additionally the project holds an annual meeting, as well as training schools, where the latest thinking can be passed on to young researchers and practitioners.

This report covers much of the technical work performed by Working Group 4 'Synthesis, dissemination and training', and has been contributed to by:

- Costas Antoniou, NTUA, GR
- Jaume Barcelo, UPC, ES
- Mark Brackstone, IOMI, UK
- Hilmi Berk Celikoglu, ITU, TR
- Biagio Ciuffo, JRC, IT
- Vincenzo Punzo, JRC/UNINA, IT
- Pete Sykes, PS-TTRM, UK
- Tomer Toledo, Technion, IL
- Peter Vortisch, KIT, DE
- Peter Wagner, DLR, DE

This document assesses the current situation regarding guidelines for traffic simulation model calibration and validation worldwide, discusses the problems currently faced, and suggests potential ways in which they can be addressed, both directly, and indirectly through the development of the overall field of traffic simulation as a whole.

Jan., 2014.

Summary

As part of the MULTITUDE project a survey was undertaken in 2011 regarding how practitioners used traffic simulation models. This revealed that 19% of those polled, conducted no calibration of their models, and of those that did, only 55% used guidelines during this process. To investigate this issue further a second survey was performed to identify which documents were being used most, and areas of concern, where it was felt that further/better guidance was needed.

In this report we have examined these areas, their strengths and weaknesses, and have isolated five gaps, where improvements would allow better overall guidance to be produced:

- Data, where a greater quality and quantity needs to be available not only for the performance of calibration and validation but also to allow a greater understanding of the variability of conditions likely to be encountered.
- Standardisation and definitions in basic methodology, where greater clarity is required as to what (for example) MoPs are acceptable and more importantly, essential.
- Illustration, Comparison and Evaluation, with a greater need for comparable case and meta studies.
- Variability, where guidance is needed as to which (for example) parameters effect differing macroscopic observables, the so called 'hierarchy of parameters' which can be uncovered through greater sensitivity and uncertainty analysis.
- Assisted calibration, where automated codes would aid in sensitivity analysis, batch analysis and, through reduced project 'run times', potentially an increase in the number of runs undertaken.

In addition, and through stakeholder engagement, it has been revealed that there are a number of non technical issues slowing advances in this area which are related to a lack of clarity of the purpose of guidelines, their target/required audience and simple economics. This has lead to the production of five recommended cross-cutting actions:

A. Agencies need to consider a better communication programme as to what is expected of contractors as regards adherence to their guidelines, when and how to depart from them.
B. A greater education is needed among practitioners as regards simulation and traffic fundamentals and as such there may be a need for a core qualification of some description.
C. Agencies need to ensure that contractual expectations as regards guideline observance are not in contradiction with budgetary constraints.
D. Manufacturers need to be encouraged to provide software to expedite all stages of the simulation project life cycle.
E. A heavily practitioner centric, pan-European forum, is needed for the ongoing debate of simulation issues with, as far as possible face to face meetings and the performance of short term working groups.

Table of contents

1. Introduction

While much of the technical work within the MULTITUDE project has focussed on methodology and tools for calibration, and focussed on the scientific advancement of the area, a vital element has also been to attempt to tie this in with current practice, and to try and direct research more toward issues of importance to the everyday user of traffic simulation models, whether they are macro, meso or microscopic, bridging the acknowledged gap between academia and industry. As part of this process the project undertook a 'state of the practice' survey in 2011 (1) to examine how practitioners were using models. This uncovered that 19% of practitioners polled, conducted no calibration of their models, and of those that did, only 55% used any guidelines in the process, the rest basing their decisions on personal experience. It is possible that these figures are due to the lack of a coherent set of guidelines for this process, although this is all the more surprising considering that a range of documents already exist which are reviewed in Section 2. One may hypothesis therefore that there may be a gap between what is needed in practice and what is actually available and, while tempting to address this, this has not been possible due to the pro-bono nature of the MULTITUDE project. Instead, the project has sought to 'set the scene' in this area, highlighting strengths, weaknesses and gaps, so that further work on this topic not only has justification, but also a roadmap as to what to address, and how.

As part of this process, Working Group (WG) 4 first sought to examine how widespread existing guidelines are known and how much they are used, and this is summarised in Section 3, presenting a top level review of a web survey undertaken in 2012. Most importantly however the working group has examined a range of key issues/questions relevant to this topic, and following validation and prioritisation in Section 3, these are summarised in Section 4 and expanded in detail in Appendix A, with each having been explored by one of the working group members. (It is important to bear in mind that these explorations are not meant to serve as introductory texts, nor is it intended as a review of 'State of the Art' in modelling techniques which has been provided elsewhere within the project (2)). In undertaking these examinations, authors have been asked to be as focussed as possible, with each contribution reflecting to a certain extent the authors own experiences, views and background (no one author can be expected to have an extensive knowledge of more than a few guidelines, or indeed even a cursory knowledge of most of them within a project of this type, where time is contributed for free).

Subsequently, in Section 5, a range of cross-cutting suggestions are made regarding how to go about tackling these issues and these are found to fall into five broad categories. Implementing these solutions however is far from straightforward and requires a number of supporting, non technical actions, and these are reported in Section 6.

2. A review of existing guidelines

In order to analyze the state of the art of the application of traffic microsimulation models and uncover the degree to which this topic has been already examined, data was collected on national guidelines or similar documents on how to apply simulation models. Experts were contacted in all countries with which the project had links and asked for information about guideline documents. In case there was a guideline, the contact person was also asked to send the document including a translation of the table of contents in English.

The availability of guidelines was found to vary strongly from country to country. A pattern however seems to be that, originating from the U.K., a certain philosophy of application has spread throughout the English speaking countries (USA, Australia). An indicator is that the measures for calibration and validation used in the U.K. guidelines are found also in the guidelines of those countries influenced by them. The U.K. itself has several guidelines which to some extent correlate hierarchically. In contrast, there are almost no guidelines available in 'Romanic' countries.
Additionally, some countries – while not having guidelines issued by a national agency – show some interest and are undergoing efforts to create guidelines. Another observation is that the necessity for calibration and validation in simulation projects seems to be perceived more clearly in the world of travel demand modeling than in the more traffic engineering oriented traffic flow simulation community, reflecting the far greater associated scale of investment. Eight countries were found to have some manner of guidance documents available, although these varied in nature:

- In the ***United Kingdom***, the application of transport planning models has a longstanding tradition and accordingly, procedures for application of such models are quite elaborate and influential. The U.K. has a hierarchy of guidelines, of which the leading one is WebTag issued by the Department for Transport (3), which is in parts comparable to the Highway Capacity Manual (4) as it defines the necessary infrastructure standards to meet a given travel demand. Since infrastructure requirements are primarily determined by demand, WebTAG is primarily concerned with demand models and with social, environmental and economic assessment Its recent update to include more discussion of highway assignment modelling (Section 3.19, Aug 2012) makes little reference to microsimulation being more oriented to specifying general principles of model scope, and model calibration and validation rather than the specific technology underpinning the model. For the application of microscopic traffic flow simulation models there are more specialised documents issued by other UK government agencies, such as the Highway Agency (HA, 5*)* or Transport for London (TfL, 6).

- In the ***United States***, "Guidelines for Applying Traffic Microsimulation Modelling Software" exist as Volumes 2 and 3 of the "Traffic Analysis Toolbox" published by the Federal Highway Agency FHWA (7, 8)). The Guidelines are written like a textbook on simulation and therefore is probably one of the easiest to read. It covers the whole process of using traffic simulation, starting from the question of which tool to use and whether microsimulation is appropriate, and ending with advice how to write the project report and what it should contain. The guideline contains a section on calibration, but no differentiation is made between calibration and validation.
  Additional guidelines have been compiled by several State DOTs which are sometimes more specific to a certain software or simplified or in some way extended versions of the Federal

guideline. One example is the "Microscopic Simulation Model Calibration and Validation Handbook" distributed by the Virginia DOT and originally developed by the University of Virginia (9). This handbook is more detailed and concrete on calibration than the Federal guideline, with validation explained here as the final step of the calibration process when simulation output is compared to field data not used for setting model parameters. The handbook gives model specific advice for CORSIM and VISSIM users, guiding them step by step through the process including much background information on statistical treatment of data.

- In ***Australia***, a guideline has existed since 2006: "The Use and Application of Microsimulation Traffic Models" (10), covering calibration and validation in detail, giving targets to meet explicitly referring to the U.S. Federal guideline, while network modelling is dealt with more specifically in "Guidelines for selecting techniques for the modelling of network operations" (11). Treatment of field data and model results by statistical methods is explained and in two appendices, theoretical background on traffic flow simulation is given and AIMSUN, PARAMICS and VISSIM are presented. The Roads and Maritime Services of New South Wales also produces its own guidelines with a strong emphasis on modelling dynamically controlled signalised junctions in its sections on microsimulation (12).

- In ***Germany*** (13), a guideline on the application of microscopic traffic flow simulation models was published by the German Transportation Research Board (FGSV) in 2006. The guideline is written around a suggested project workflow. The importance of calibration and validation is stressed and a larger section is devoted to the topic. As a measure for model fitness the RMSE (Root Mean Square Error) is given (no references to the 'GEH' measure are made), but the guideline does not give concrete thresholds to meet. The stochastic nature of simulation is explained as well as the need for multiple replications and the statistical treatment of simulation results, even a formula for computing the minimum number of replications is given.

- In ***Canada***, a guideline for the application of planning models exists ("*Best Practices for the Technical Delivery of Long-Term Planning Studies in Canada"* (14)). This document includes transport as part of the wider planning process and touches on microsimulation, data collection and with a brief section on calibration and validation.

- In ***New Zealand*** (15), the NZ Transport Agency's Economic Evaluation Manual contains official model transport model checks and validation guidelines. However the calibration/validation target levels in this document are only really applicable to regional strategic planning. The "New Zealand Modelling User Group" is a sub-group of NZ's professional engineering institute (IPENZ) and is currently producing an observed and modelled data comparison guideline i.e. focused on calibration/validation targets. The guideline is not specific to the specification and development of microsimulation models or any other form of transport model, i.e. it does not contain 'how to' guidance, but does contain calibration and validation targets which are applicable to a range of common microsimulation model applications. A live, ratified, document exists giving advice on calibration and validation approaches, matrix adjustment processes, and critically observed data comparisons with model values, including concrete GEH thresholds, scatter plots and correlation measures etc.

- In the ***Netherlands*** (16), a guideline for using models in transport planning exists, with a focus on project organization. It contains only a brief section on calibration and validation mentioning that model values and counted values should not differ for more than 5% in 90% of the cases. For more detail on calibration and validation, a reference is made to a handbook on applying models in water engineering. Sensitivity analysis is mentioned as is the need for multiple simulation runs.

- In ***Japan*** (17), a guideline on "verification" of traffic flow simulation models exists, however, this is not a typical guideline for applying simulation models. Instead, the focus is on testing which phenomena are reproduced by the models, i.e. the simulation is tested against theoretical considerations instead of empirical data, for example, a model is only verified if it is able to reproduce shockwaves in traffic flows realistically. The guideline contains a section on validation in which advice is given for comparing model data to empirical data.

According to information provided by project contacts, the U.K., Germany and Holland aside there are no national guidelines in any other EU countries. Work is in progress however in France, where the Transport Ministry has instituted a simulations application group to set up a guideline during the next two years led by CETE de Lyon (PCI RDRT) and IFSTTAR (LICIT), and also in Finland, where work is now underway in providing basic guidelines for the use of microsimulation models (18).

Analysis of these documents reveals that there are many commonalities between them. For example, the GEH targets for calibrated models most probably have been set originally in the UK in DMRB Vol 12 (19), now carried over to WebTAG Vol 3.19 (3). These are appropriate to large transport planning models and have been imported later into other guidelines as the US Traffic Analysis toolbox with no reflection if the same targets are appropriate for microsimulation models, which typically model much smaller networks or do not have route choice at all. Another example is the proposed structure of a simulation project. Here exactly the same text blocks can be found in several guidelines. However, it is important to bear in mind that the inclusion of a value or a procedure in more than one guideline should not be taken as a proof of its validity.

3. The need for guidelines

From the discussion in the preceding Section it is clear that many documents exist on this topic, however their coverage and scope is clearly variable. In order undertake a clearer comparison all these documents were reviewed in terms of how well they addressed thirteen key issues which were isolated by WG4 participants as core issues that are faced in performing traffic simulation projects. These issues were:

- How to structure a simulation project/calibration activity.
- How to handle model 'warm up'/run duration.
- What number of runs should be performed.
- Sensitivity analysis, and how to perform it.
- Appropriate definitions of calibration and validation.
- Calibration methodologies.
- Differences in procedure according to scale /model purpose.
- Model specific issues.
- What to do in the absence of appropriate calibration data, potential 'fall-back strategies', transferability of data between calibrations etc.
- The relative effect of parameters (and different types of parameters) on output.
- Which indicators to use/not use, for calibration and validation assessment.
- What data to use for validation and when to perform it.
- The need for more reference materials/case studies and what form should these take.

The review, while subjective, found that coverage of many of these issues was sporadic and in many cases lacking in depth and rigour, and that the production of new guidelines/ elements of guidelines was something for which there was a clear need. However, as project participants were primarily from the academic sector, a concern existed that these views and interpretations may not potentially reflect the views of other simulation user groups, in particular, model users such as consultants and government agencies, working with different needs and objectives. To that end, in 2012, a validation phase was undertaken in order to gauge practitioner opinion and examine potential differences. This consisted of two activities, a questionnaire, and a range of face to face stakeholder meetings.

The questionnaire was distributed widely within Europe and the USA, initially through targeted distribution to known 'expert' individuals, and subsequently to the wider constituency through distribution to mailing lists and on-line access. These included for example, UTSG and TMF in the UK, Traffic Flow Theory and SimSub mailing lists in the USA, ITE national groups in Greece and Italy, as well as through contact lists through model suppliers TSS and PTV. A detailed analysis of the responses may be found in the associated report on this activity (20), however in summary:

- A total of 412 responses were received, 46% of which were from North America, 34% from Europe and 9% from Australasia.
- Respondents had on average, 15 years of experience in Transportation, with wide variation found according to country and sector (Consultancy, Academia and Government).
- Responses from the primary target groups (non-Academic) were on the whole high (at worst, 70% for respondents from continental Europe).

Respondents were asked to rate the identified issues in terms of how important they felt it was for more guidance to be available, rating more guidance as being either: essential, appreciated, that enough already existed, that the issues was already over regulated, or, was unimportant. A high degree of need was expressed for guidance on all the issues (shown in one format in Figure 1), with minimal variation found due to Sector or Region.

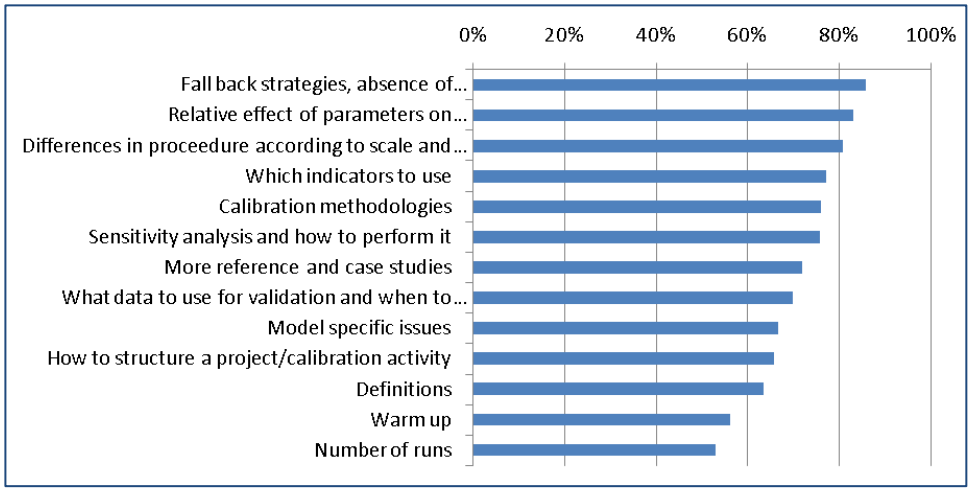

Figure 1: Issue importance by fraction of respondents rating as further guidance being either 'Essential' or 'Appreciated'.

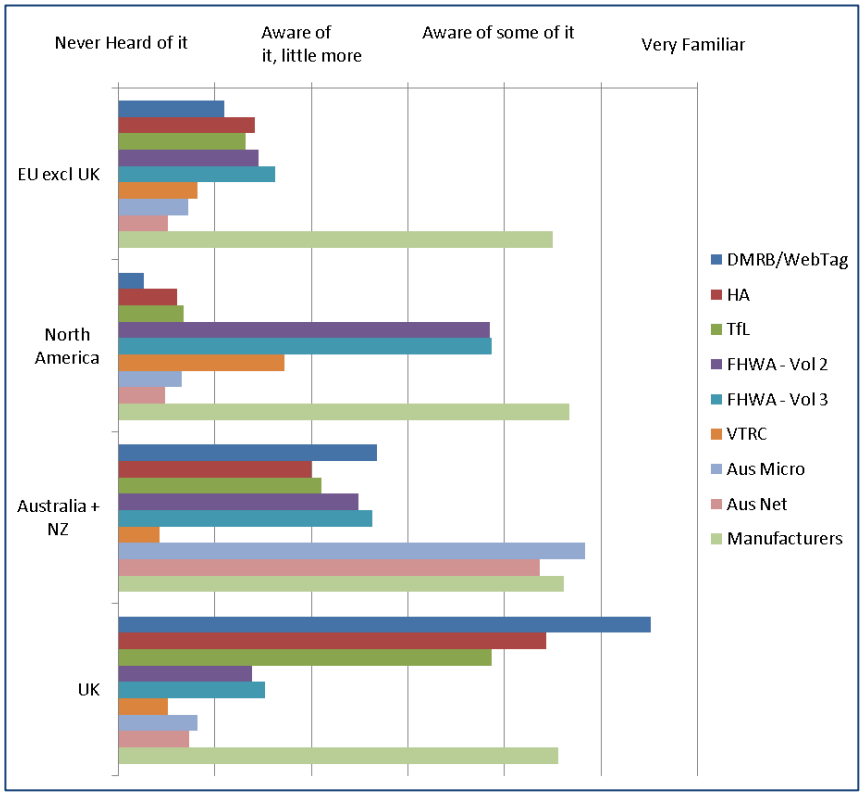

Figure 2: Awareness of existing documents by Region

Subsequently, respondents were asked about their awareness and use of key documents highlighted in the previous sections. As can be seen from Figures 2 and 3, awareness and use of guidelines from regions other than ones own was poor. While this is perhaps unsurprising, use and awareness of documents from ones own region is also lower than one would perhaps expect on first examination. (Indeed those in most widespread use are likely to be those provided by software manufacturers). (The UK has, was, for the purposes of the survey analysis, treated as a separate 'region' as it is the only EU country that has a distinct and formal document on this topic, and interest and awareness is notably far higher than other European countries).

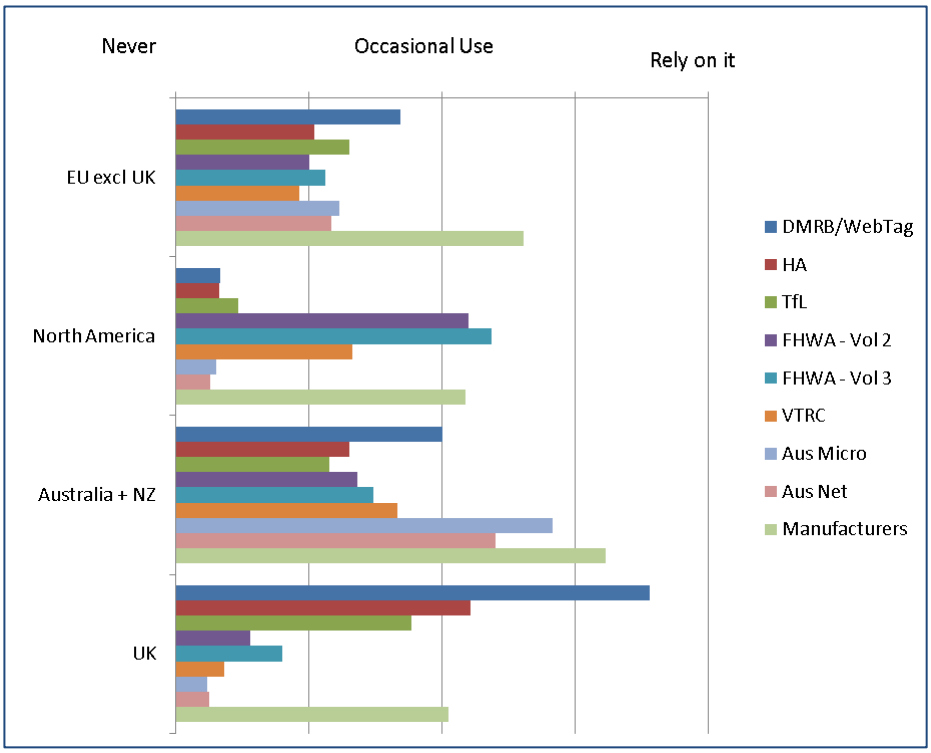

Figure 3: Use of existing documents by Region

In parallel, stakeholder meeting were held in the UK (hosted by DfT, February, 2012), Germany (as part of the meeting of the FGSV AK3.10.04 working group, February), as well as France (hosted by IFSTTAR, September) and Holland (hosted by TU Delft, October). These meetings, allowed the project to canvas views from the simulation modelling community, both on topics of immediate relevance to the work, and also on more peripheral (but equally important) aspects of both performance of simulation, and execution of commercial projects. (For example, commercial and management considerations that effect how, and when, calibration and validation are performed).
Having attained an understanding of the 'landscape' regarding the views and needs within the simulation community, the project has attempted to build on this and consider in depth, those issues (/gaps) that are not currently addressed and these are covered in the next two sections.

4. Issues of Importance

Further to identification and prioritisation of the issues raised in Section 3, a detailed examination was undertaken of each of these as regards how it is currently being approached worldwide, and what weaknesses (if any) exist in its treatment. Each of these has been undertaken by different members of the MULTITUDE team. In some cases, it has been found that the issue is 'well understood', and although the subject of debate, is in the main quantifiable and addressed in many different documents By contrast, other questions are more complex, with less existing in the literature and to some extent are still 'ill posed' from a scientific stand point, if only at times due to their large scope. Taking this further, there are others which are, while acknowledged to be items of concern, still almost exploratory in nature, where treatment can be discursive at best. A full examination of each issue may be found in Appendix A (although it should be obvious that an exhaustive treatment is beyond the scope of this project) and these are summarised below.

Issue 1 - How to structure and manage a simulation project

The primary purpose of microsimulation guidelines is to facilitate the production and use of better microsimulation models. While the issues of data, calibration, sensitivity analysis etc. are discussed later, this section reviews how the guidelines present the structure of a project that links those issues together and consequently how to manage a project for a successful outcome. Guidelines are written from different viewpoints and this section classifies them by bounds, i.e. are the boundaries set to be the model build, calibration and validation only, or do they include the preliminary specification of the project and the final process of assessment using its results. The guidelines are also classified as generic, specific to a particular project type, or specific to a particular software product. The implications in applying the guidelines are that the agency written documents are less able to include the advanced or unique features of a particular software product while the software specific guidelines will tend to focus on their own methodologies and strengths.

One interesting point on which guidelines differ is in their treatment of "innovative solutions" and how these may be justified and accepted. Not every situation can be readily simulated and observed behaviour on the road network sometimes breaks the rules. Modellers innovate to provide pragmatic modelling solutions in such circumstances. The role of guidelines is to provide advice on where the threshold lies between a justifiable model adjustment and an unacceptable fabrication. The related topic of model audit is also treated differently in the guidelines where it is alternately defined as either a final "sign off" task or as an ongoing process of checking as a model is built. Once again the issue of software specific versus generic advice arises as while advice on milestones and quality processes is software agnostic, advice on technical audit is necessarily product specific.

Issue 2 - How to handle model 'warm up'/run duration/'cooling off' period

The warm up period in a simulation study is needed, because the initially empty network needs some time to fill with vehicles: the first vehicles to enter the study area is, in effect, driving at 'midnight', with no other vehicles on the road and nobody waiting in front of them, e.g. at traffic lights, thereby causing travel times to be shorter as they do not have to wait for vehicles in front to clear the intersection. Data from the start of any run therefore have a strong bias toward smaller delays, which does not reflect the real situation.

The exact length of such a warm up period is a different question. A good hint is to wait for the longest travel time of the system under consideration, and then double this. This is often sufficient, and similar recommendations can be found in some of the handbooks. For oversaturated conditions this might not be enough, since in such a case a system might be out of any kind of equilibrium. Here, as a general rule of thumb, the warm-up period should be long enough such that traffic is getting onto all, or the majority, of the links in the network before the main simulation starts. Another practical recommendation is to simulate, in addition, the time before and after saturated conditions and use the time before the peak period as the warm-up period. However, in the case of oversaturation, the statistics to be sampled from the simulation must in addition use a cooling-off time period, since the queues created at the simulated intersections may need a long time to dissolve. However, since they have been created during the time of (over-)saturation, their effect needs to be accounted for, and this will happen in the subsequent cooling down period.

Issue 3 - Number of runs to perform

Traffic is a stochastic, highly dynamic phenomenon, resulting from the actions and interactions of large numbers of travellers, along with various exogenous events. Traffic simulation models reflect this stochasticity, as they use random variables and sample from random distributions to represent decisions made by the simulated agents (e.g. route or lane choice). The drawback of this is that multiple runs of the simulation program are needed to obtain reliable results. This allows the computation of mean, and standard deviations and from this the derivation of confidence intervals.

However, the key question is: how many replications are needed? Surprisingly, a systematic treatment of this topic is not easily found in the traffic literature. Usually, the number of replications needed is not discussed in detail, and a number between 5 and 10 is chosen, either arbitrarily, or based on some simple formula. As the number of replications increases, then the contribution of an outlier/extreme value to the average decreases. While a small number of observations may lead to biased results, as the number of observations increases the average quickly converges to the "correct" value. The general idea is that the number of replications must be increased as the standard deviation of a set of simulation runs is higher. The exact number of replications is determined using the level of significance desired of the results and allowable error of the estimate.

A number of guidelines address this topic, at least to some degree and this ranges from a simple ad-hoc recommendation on a minimum number of replications, to the provision of a methodology for the determination of a minimum number of runs and some guidelines for the application.

Issue 4 – Sensitivity analysis and how to choose parameters to calibrate

In the application of traffic simulation models a crucial question arising to the analyst is: which model parameters really need calibration? In fact, traffic simulation models are often over-parameterized and the most of the model output variance - a proxy for the output uncertainty - is explained only by a handful of these parameters. These two considerations make the calibration of all the parameters of a model neither practical nor necessary, and turn the initial question into: which are the parameters that can be fixed at whatever value without sensibly affecting the outputs? Sensitivity analysis can help answering this question. An introduction to sensitivity analysis and to its established techniques is therefore provided in this section. Besides the identification of the most influential parameters, other useful purposes of sensitivity analysis for the economy of modelling are presented herein. It is explained why sensitivity analysis should be applied since the very early

phases of model- and software-development and testing, and specific new functionalities of traffic simulation software are suggested in order to make the application of such techniques affordable also for practitioners. In the end, it is clarified that sensitivity analysis is not covered at all by existing guidelines. Indeed, these documents misleadingly call 'sensitivity analysis' what, in the broader simulation modelling community, is referred to as 'uncertainty analysis'.

Issue 5 - Appropriate definitions of calibration and validation

Simulation is a useful technique to provide an experimental test bed to compare alternate system designs, replacing the experiments on the physical system by experiments on its formal representation in a computer in terms of a simulation model. Simulation may thus be seen as a *sampling experiment* on the real system through its model. The reliability of this experiment depends on the ability to produce a simulation model representing the system behaviour closely enough or, in other words, "how close the model is to reality".

The process of determining whether the simulation model is close enough to the actual system is usually achieved through the validation of the model, an iterative process involving the calibration of the model parameters and comparing the model to the actual system behaviour and using the discrepancies between the two, and the insight gained, to improve the model until the accuracy is judged to be acceptable. Papers and guidelines provide a variety of definitions of validation and calibration, in this guidelines an overview of some of the most common definitions is provided and from them it is selected a formal definition providing technical grounds to quantify the concept of closeness between model an reality in terms of distances and statistical methods to measure it.

Issue 6 - Calibration methodologies

The final goal of calibration is minimizing the difference between reality, as measured by a set of observed data, and the model results, described by another set of data that has been produced or constructed from the simulation model. This is done mostly by adapting the parameters of the simulation until some minimum (best fit) has been reached. While simple in principle, this application of this process is surprisingly complex with (arguably) four key stages.

Firstly, ensuring key input data (digital road network, traffic signs, intersection layout, traffic control algorithms) is available and accurate is vital). The second step is the calibration of the demand needs to be calibrated. This is not a simple procedure and it often needs a lot of manual correction and massaging. One good quality measure is that 85% of the link counts (can be measured with loop detectors) should have less than 15% error in them.

Only after these steps it makes sense to correct the parameters of the simulation model, such as the car-following and lane-changing parameters, or the distribution of the maximum speeds. Validation is the final step, and by using a data-set that has not been used for calibration, it can be checked that a model not only reproduces the data used for calibration, but also data that are different from those used for calibration. It should not come as a surprise, that the validation error is very often larger than the calibration error. Nevertheless, the simulation model's output should be within acceptable thresholds that have been determined in ideal case before the actual simulation has been run.

Issue 7 - Differences according to scale and model purpose

Transport models are built at widely varying scales for many different purposes. Models which are applied inappropriately, i.e. by providing aggregate measures when detailed outputs are required, or by reporting outputs that do not permit differentiation between schemes do not serve their purpose; which is to provide decision support to transport planners. A small number of the guidelines assist in making this choice by providing methods and decision trees to determine which class of modelling tool to use from complex microsimulation to simple table look-up. Microsimulation tends to be advised by these processes in congested situations with complex junction interactions, where there are active control systems, or where the assessment of the output requires detail such as variance between individuals. It is notable that no guidelines recommend microsimulation solely for its visual outputs, or that it should only be used in small area models, but all note that the perceived higher cost of microsimulation models is a factor in their use.

This section then moves on to discuss the requirements of different scales and applications of microsimulation models with respect to data requirements calibration methods and criteria and the selection and use of output measures of effectiveness where overall there is little in the guidelines to advise the analyst. For example, the UK Webtag GEH criteria, designed to be applied to wide area assignment models, are commonly used to judge the level of calibration of a junction or corridor model where other measures might be more appropriate. The level of guidance provided is variable and may be summed up as advice to use the well-established aggregate measures for network performance; but to select a measure of effectiveness for localised issues according to the particular needs of the decision makers.

Issue 8 - Model specific issues

The questions asked in this section are how relevant are these measures described in the guidelines to the task of calibrating and validating a microsimulation model, and how relevant are they to quantifying the differences between the different infrastructure design variants or road traffic management options to be tested. The tendency to carry practice over from one modelling paradigm to another demonstrated by the ubiquitous use of the GEH measure illustrates the problem of applying a coarse criteria to a detailed model.

There are more appropriate measures depending on the application of the microsimulation model i.e. capacity analysis at junctions but the essential pre-requisite is that terms such as “delay", "queue length" and "stops" are clearly and precisely defined. Although these terms are generic and would be recognised by the transport modelling community, each software developer interprets these in ways appropriate to the algorithms they employ. Independent definitions from agency led guidelines which do not rely on any one behavioral model for their derivation are required to allow the modelling community to make consistent measures of effectiveness.

Once these measures are defined and the software developers have demonstrated that their product outputs conform to these definitions, the role of guidelines is to advise the analysts as to which measure to apply in different analyses.

Issue 9 - What to do in the absence of appropriate data, potential 'fall-back strategies', transferability of data between calibrations etc

Unfortunately, the absence of data is a matter of fact. Albeit no longer true for network data (e.g. open streetmap) but much more for infrastructure data (traffic signals), demand data, and data that may help to test the driving parameters of the model to be used (loop data, queue-lengths, trajectories). This lack of data is often due to cost or difficulties in the acquisition of data. Some recommend trying to transfer data from other, similar studies, and all too often this is simply the only possible solution. But one has to be careful, and should state all the assumptions that had been made, and still there is hardly any guarantee that it will work out.

In a certain sense, MULTITUDE's focus on sensitivity studies may help in these transferability issues. Sensitivity analysis may give hints about which parameters are important for the current situation and which ones are not, helping to decide what can be transferred and what not. It is worth mentioning that there are some situations, where no data is needed. Often, it is enough to check for so called stylized facts, there is no need for a detailed comparison to reality. Also, if a so called scenario study is performed, data can often be omitted, since only different scenarios are compared against each other. Finally, there are studies where surrogate data are sufficient. Again, this is often used in scientific papers, and there it is sometimes even a necessity, for example, to test and assess the quality of a method that estimates an OD-matrix from loop detector data needs surrogate data, because the real data are almost nowhere to find in the quality that is needed for the test of such an algorithm.

Issue 10 - Which indicators to use/not use, for calibration and validation assessment

A large number of indicators are available for the assessment of calibration and validation efforts. Each of these has different strengths and limitations and the question of which one to use very often arises. The answer to this is usually complex and may depend on a number of aspects, related to the problem, the available data (and their characteristics) and many other parameters. Square statistics are preferable in calibration as, besides theoretical advantages, they penalize the highest errors and make the response function less smooth around the minimum. Usually, when validating a model, a reasonable approach is to consider multiple indicators, thus elucidating more aspects of the model and providing a more complete assessment of its performance. The problem of how to combine different measures in the same objective function (multi-objective optimization), however, has yet to be addressed satisfactorily.

The choice of suitable goodness-of-fit (GoF) measures is an important topic, but it is not the complete picture however, and another relevant question is which measures of performance should be used for the assessment of the calibration and validation quality. One obvious answer is that only available data can be used for this purpose. But when more than one type of data are available (e.g. flows, densities, speeds and travel times), which one(s) should be used? One way to approach this question is to look at the available data and make a choice depending on their quality.

The topic of GoF measures is mentioned in general terms in many guidelines, most of which can be traced back to guidelines in the UK, which suggest to use the GEH and usually mention that the model outputs should be within 5% of the observed values.

Issue 11 - What data to use for validation and when to perform it

This section describes model validation, the process of checking to what extent the model replicates reality, and in particular the effects of changes in the system and its inputs. The importance of validation is in that it provides the modeler confidence that the responses to changes in the transportation system that are observed in the simulation model are representative of those that would have been found in the real system. Thus, a valid model will be able to accurately predict the effects of changes in the current system. Most current simulation guidelines briefly define validation, but do not give it the level of attention and detail that they do in discussing calibration procedures. Some of the guidelines (in particular those from the UK and Australia) propose specific measures of performance for the validation and thresholds and confidence intervals for their evaluation.

The section presents a conceptual framework for the calibration and validation tasks. It consists of two phases: Initially, disaggregate estimation, in which the behavioural models that make up the traffic simulation model are each estimated separately, outside the framework of the simulation model. In the second phase, which is the relevant one for simulation users, the simulation model as a whole is calibrated and then validated, using readily available aggregate data. Emphasis is placed on two main issues: (i) input modeling, and in particular estimation of the trip demand matrices, and (ii) the criteria that govern the selection of measures of performance that will be used to compare the simulation results with the real-world observations.

Issue 12 - Reference materials

The phrase “reference materials” is used to describe a wide range of documents; from software developer’s marketing materials, to example studies used to make concrete the abstract concepts involved in model calibration and validation, to ex-post studies reporting on the accuracy of the predictions made in a modelling exercise.

The marketing materials category has evolved since microsimulation was first introduced to the marketplace as software developers move from describing what could be done using what was then a new approach to transport modelling to describing how it should be done in the expectation that in describing examples of the application of best practice, better models would be built. Naturally, suppliers focus their materials on the perceived strengths of their own software, on innovation unique to themselves, and on high value successful projects.

Agency led reference materials used in guidelines to illustrate application of the practices embodied in the guidelines are less likely to be parochial in selecting which cases to describe but also, in being independent, tend to lag behind in innovation. One attempt is described in Australian guidelines to provide a constantly evolving repository of projects reporting on the purpose of the model, l conclusions from the model application, the extent of the variation from default parameter settings, and on the general robustness of model outputs. However, since creating the resource and initially populating it in 2006, no further contributions have been made from the industry. Only mandatory ex-post evaluation appears to be robust in gathering such material such as the UK Highways Agency POPE (Post Opening Project Evaluation, 21) reports made one year and five years after a road scheme is opened. These reports are growing to become a large reference material resource independent of selection bias by software supplier or consultant, albeit subject to the selection bias of a positive case for the project.

5. Suggestions for development

As has been shown in the previous Sections, many guidelines exist and valuable advice can be obtained from these already. Section 4 has introduced some of the key issues that are in need of development in these, indeed many such improvements are already underway, with new versions of Guidelines being under development by TfL in the UK, and preparatory work being undertaken by the FHWA in the U.S.A. (22), while efforts in France and Finland look to produce first documents over the course of the next year. While development is clearly needed across the board, in both breadth and depth of advice, it is also possible to isolate a number of 'cross cutting' themes, where improvement should also be made, many of which would help to address more than one of the issues covered earlier, and a number of these are provided below.

The need for data

One of the fundamental requirements in undertaking calibration and validation, is having data of sufficient quality and quantity against which to perform these processes. Although more data is now available than ever before and growing annually, they are still distributed in a suboptimal manner: while some research institutions have a large quantity, others (especially consultants) often only have poor data for the investigations they are working on. This does not allow for good calibrations to be routinely performed, and as such particular steps are needed to redress this imbalance.

One suggestion is the establishment of data resource libraries, containing data describing demand and traveller's trip based reactions to transport schemes in terms of situation based demand or mode choice, or even typical driver behaviour. While this is mentioned in some guidelines it is heavily country specific and few EU datasets are available. Such libraries could include more detailed demand profiles, a subject some guidelines mention nothing about, while other state only that these be time varying, but give no suggestion as to how this be generated or what is expected. Likewise dependable and realistic estimates of time-dependent OD matrices are crucial to the successful performance of many modeling projects. Data is also vital in order to be able to understand transferability of models, calibration and outcome, as well as enabling sensitivity analysis, which is hampered by our lack of understanding of how traffic and/or behaviour varies from (comparable) location to location, and with time.

While not an element of guidelines per-se the dependence of a modeling project on data and the significance of this is a matter which needs to be recognized by agencies explicitly, namely, that the quality of a calibration/validation exercise, and how far can it go, strongly depends on the availability of the required data, and that if adequate data cannot be sourced then project objective are likely to be compromised.

Standardisation and definitions in basic methodology

It has been made clear in the previous Section that there is a lack of real guidance from Agencies as to exactly what is required and acceptable in the performance of projects and what is said does not account for the flexibility available through microsimulation. Additionally many guidelines are vague as to their exact definition of calibration and methodological approaches that they consider best practice, validation criteria, and the reason for that selection (that should be based on an established body of theory and not intuitive beliefs). For example, there is often no clear statement of which GoF indicator should be used for calibration or validation (that is adopted as an objective

function) and this can lead to the inappropriate use of (insensitive) indicators such as GEH3 or GEH5 for example (19). In fact, GEH is inappropriate for calibration, as the usual values adopted for the statistic (i.e. GEH=3 or 5) can be obtained with a multiplicity of parameter combinations, that means having a poor calibration. On the other hand more stringent i.e. lower values of the statistic cannot be defined a priori, being case specific. There is additionally a distinct lack of advice on localised measures (that consider only small windows of assessment either spatially or temporally). Indeed the TfL document (6), which is focussed on simulation of small numbers of linked junctions, makes no reference to localised measures of effectiveness in option comparison. A rating process to select the most appropriate outputs similar to that used to select the most appropriate tool may be a useful addition to the guidelines. At the other end of the scale, it would be useful if guidelines could be developed to propose specific MoPs to be used for validation for specific classes of studies, based on their geographic and temporal scope, the type of project (construction, traffic control, ITS etc).

Lastly in terms of output, most guidelines describe results from simulation in terms of aggregate measures or statistical comparisons, the FHWA guidelines are a good example of such an approach (8), while others, for example the UK TAM, include the model results in further assessments of economic and environmental benefits. However a microsimulation model, disaggregated to individual vehicle level, can provide far more detail, e.g. the single measure of average journey time can be replaced with the distribution. In addition, as such models are stochastic, the distribution of the output measures of performance over the replications has to be retained and used. In fact, comparison of scenarios only on averages for example, produces choices based on partial or biased information. A very simple tool in this respect would be the substitution of an average value by the five numbers that are being used in a box-whisker plot. Doing so is not dramatically more effort but adds a large amount of additional and useful information. Such an approach would also enable and encourage uncertainty analysis, with results from a simulation scenario being drawn and presented as probabilistic, taking into account the uncertainty in the inputs, whether they are model parameters, demand or network. Such an approach falls in the broader field of uncertainty analysis, with techniques taking into account the uncertainty in the inputs, and these should be applied in traffic simulation in order to draw inference on the reliability of results.

Illustration, Comparison and Evaluation

One factor that is hampering progress toward obtaining reliable calibrations is the lack of clear illustrative examples of what is expected, for example reference cases/best practices illustrating how a project should develop. Such cases would allow sometimes abstract concepts to be made plain and act as templates for other similar projects, providing vital companion information to guidelines themselves (e.g. 5). These, ideally, should reflect the increasing capabilities of software systems such as the capability to model the entire, multi-modal, journey of each individual and the interactions between modes, which in turn would promote new capabilities, and encourage wider and more diverse use of modelling along with newer and more innovative transport concepts.

Comparison or meta studies are also essential in order to allow practitioners to learn more readily from each other and to evaluate the differences in applying different modelling techniques or different software to the similar problems. This would aid in understanding how sensitive studies are to the initial choices, and this “high level information” is something that should be known to the modeller before commencing any sensitivity analysis. Such work would be a significant effort and require a representative sample of models to be described and each one built by a number of

modellers who are experienced in the use of the chosen software systems, and working to guidelines from different agencies. The UK HA exercise offers a glimpse of what could be achieved using this approach (5). Finally, there is also a need for post-hoc evaluation, which evaluate the accuracy of the simulation project some time after the scheme that it tested has been implemented, thereby validating its predictive ability. Raw data for this process is available for example in the UK in the form of the HA POPE documents (21). A meta-study to combine the results of different micro-simulation projects would pull together the learning points of each study and merge them to identify common patterns amongst their results and to identify sources of difference.

Variability

One of the principal advantages and reasons that microsimulation is used is its stochastic basis and its ability to account for variations in behaviour and performance. This strength however is also its biggest weakness in that to fully benefit from this the practitioner must have not only a good understanding of its theoretical basis, but of how traffic performance depends on exo and endogenic factors and how they inter-relate. Even for the most experienced and well trained, these relationships are sometimes unclear and no modeller can be expected to be an expert on all the aspects relating to the formulation of the behavioural cores of the models they use and therefore more information on sensitivities and behavioural relationships need to be explained. Ideally this would be through documentation stating the most important parameters according to network characteristics/typical situations, and the HA guidelines (5) have made some early steps toward establishing this hierarchy (23). It is clear however that the whole process cannot stand on the shoulders of practitioners or agencies and sensitivity analysis, in particular, should be first applied by software developers to their basic models and to the ensemble of the same models. This would allow a better understanding of model capabilities and behaviours to be achieved, and the simplification of models to be undertaken (i.e. many 'unnecessary parameters' could be made fixed and invisible to the final users or flagged as 'do not modify unless in specific situations'). With the help of practitioners, their data and their networks, it would be also possible through these techniques to understand the influence of network characteristics, typical situations and meteorological conditions on the ranking and importance of parameters.

Additionally, there are several fundamental considerations that must be addressed and one of the foremost of these is the need for multiple runs of a model to demonstrate the variability of output likely in the real world and while now acknowledged the importance of considering such distributions in decision making is still not fully appreciated by many. The principle challenge here is to educate stakeholders sufficiently as regards what to expect/require, as well as modellers in terms of what is needed and the implications to the validity of their results.

Assisted calibration

With the steady increase in computing power many would argue that the calibration and validation process could (and should) be made more straightforward and faster through the development and use of new software. This has certainly been the case in sensitivity testing which uses many similar principles and methods however, there is still some work needed to ensure that such processes and software can be used on a routine basis for applications, and it is hoped that traffic simulation software will soon be developed with built-in functionality for executing experimental design that allows combining variability of parameters, demand and network elements. This should be combined with the development of appropriate procedures and templates to automatically calculate

and display the values of validation MoPs of interest, directly from simulation results which would in turn greatly facilitate the use of these statistics. Going further, and making calibration 'fully automatic' however would be unwise, with end users likely to become overly dependent , or overly trusting of output, resulting most likely in less of an even worse understanding of the inherent uncertainty in results.

6. Conclusions

In this report we have examined the degree of use of guidelines in traffic modelling, their strengths, weaknesses, and in the previous section, suggested five streams of improvements that should be made in order to facilitate the development of guidelines and the ease with which they can be implemented by practitioners, thereby raising the overall quality of modelling being undertaken. These have included:

- Data, where a greater quality and quantity needs to be available not only for the performance of calibration and validation but also to allow a greater understanding of the variability of conditions likely to be encountered.
- Standardisation and definitions in basic methodology, where greater clarity is required as to what (for example) MoPs are acceptable and more importantly, essential.
- Illustration, Comparison and Evaluation, with a greater need for comparable case and meta studies.
- Variability, where guidance is needed as to which (for example) parameters effect differing macroscopic observables, the so called 'hierarchy of parameters' which can be uncovered through greater sensitivity and uncertainty analysis.
- Assisted calibration, where automated codes would aid in sensitivity analysis, batch analysis and, through reduced project 'run times', potentially an increase in the number of runs undertaken.

Obtaining such improvements however are far from straightforward and a number of structural barriers exist within the discipline that slow this process. These 'cross-cutting issues' essentially enable the actions of Section 5, and are discussed below.

The purpose of guidelines

In the UK, and several other nations where modelling is (and long has been) a mainstream tool in traffic engineering and planning, guidelines exist in order to ensure that projects undertaken for an agency conform to their basic standards thereby (in theory) ensuring a degree of quality. There are two schools of thought as regards this approach, the first is that such guidelines are not complete enough and not as strongly enforced as they should be and that quality would be further enhanced if guidelines were replaced/supplemented with some manner of ISO standard or benchmark. However, raising any document to such a level may prove difficult as in order to be enforceable it may have to be seen as showing a clear 'state of the art', a definition which is constantly evolving, exposing any agency not only to commitment to maintain this status but also potentially to liability. Conversely, it is possible to argue that even current guidelines, if rigorously enforced, may actually stifle innovation and discourage practitioners from thinking for themselves, and while raising standards for some, may actually lower the standard of modelling that is performed by others. In short, if a core basic standard can be met, what is the incentive to exceed this? In short, it may be better to have a 'weaker' document that is 'widely observed', than a stronger document that is not/cannot be enforced, and may, potentially enforce incorrect guidance. A last issue related to this, concerns how strict guidelines are (and indeed how stringently they are enforced). For example in the HA guidelines parameters are allowed to be changed outside or proscripted bounds if there is a good defensible reason (5) and this makes logical sense, guidelines are indeed just that, a guide, not

a stricture. However this can be misunderstood, and it essential that enforcement policy is clearly communicated between client and contractor.

Who is the audience?

While we have assumed there is a need for guidelines, we have also assumed that all practitioners need the same level of guidelines, but this is unlikely to be the case. For example many core concepts are not properly detailed as it assumed that practitioners already have this knowledge, however many junior staff will not, and potentially a surprising number of middle and even senior staff too. (In fact it may be argued that senior staff may actually have less knowledge as they are more remote from current developments). This is most likely due to the wide variety of backgrounds and qualifications that practitioners (and even government staff) have in this field, with some coming even from social sciences or management and having learned 'ad hoc' simulation or traffic modelling skills, or as part of training courses, many of which will perforce be related to how to use particular simulation tools, as opposed to how they are formulated and their theoretical constructs.

If a distinction is to be made between experienced users and starters, an additional question is whether junior staff actually need more guidelines, or potentially less? While perhaps a surprising assertion, it should be born in mind that basic users only want the basic information required to perform a simulation study without detail, while it is the experienced user that may want more detail. (It is of course a matter of debate whether middle grade staff actually have knowledge much beyond that of their junior counterparts).

While a case could be made for the provision of such 'primer texts', it is beyond the remit of any agency to provide this, however a potential solution could be the adoption of some manner of recognised qualification throughout the industry from certified training courses, as opposed to software based courses, thereby minimising the chances of such knowledge gaps occurring in the first place. Such a qualification would not only be of use to consultants but also (and maybe more importantly) to agency staff to ensure that clients do not become technically 'outgunned' by contractors, and find themselves in an exposed position. Such a concern has already been expressed in (non UK) stakeholder groups, where guidelines are already viewed as being documents that are needed in order to 'protect' agencies, as opposed to the conventional view as regards imposing a basic quality control filter.

Economics – the final arbiter of calibration?

While the above discussions are important, there is perhaps one underlying consideration that presents a far greater problem, and that is the economics of undertaking commercial projects. In essence, while we acknowledge we need more guidelines, there is a danger that (for example) requiring more detailed testing, more simulation runs and sensitivity analysis will drive up the resources needed for any project, arguably disadvantaging the more diligent when it comes to winning contracts, or indeed making compliance with guidelines impossible while maintaining profit margin. Care is needed therefore from the client side, not to 'price out' calibration (one may argue that this element of any quality submission should potentially be used as a counter to cost based calculations in funding decisions).

One way to minimise the cost implications of extra guidelines however, in addition to greater overall awareness of what is currently done in practice, would be to take advantage of, and encourage, new

software that would enable many of the technical processes, data analysis, calibration and sensitivity testing to be under taken faster, at the 'click of a button'. Such advances toward 'auto calibration' are very appealing and would be to the benefit of all concerned. However there are both contractual and diligence related concerns associated with this. Firstly, there would be a concern over who would be liable for a poor calibration, the consultant or the software supplier? Secondly, and perhaps most importantly, there is the danger that if not coupled with a greater emphasis on education and understanding of modelling basics this could effectively encourage practitioners to 'model without thinking', driving quality down, instead of up and requiring agency staff to scrutinise projects still further to maintain quality.

A way forward

In Section 5 we have suggested technical improvements that can and should be made, however we acknowledge that it is unlikely that any agency will be able to implement these in one fell swoop, both through simple economics, and also due to differing policies and interests that will take precedence in differing countries. (It is noteworthy that concerted efforts even in producing coherent modelling guidance for regional models, have met with limited success, for example the MOTOS project funded by the EU under the FP6 programme (24), despite having a budget of over half a million Euro, and producing almost 600 pages.

The items discussed earlier in this section however may be easier to address as they do not necessarily require dedicated funding streams and can be considered as 'easy to implement' non technical actions from which all would benefit. These may be summarised as follows:

A. Agencies need to consider a better communication programme as to what is expected of contractors as regards adherence to their guidelines, and when and how departures from these can be made. The enforcement of such documents needs to be consistent within economic bounds.
B. A greater education is needed among practitioners in all sectors as regards simulation and traffic fundamentals and as such there may be a need for a core qualification of some manner in this discipline.
C. Agencies need to ensure that contractual expectations as regards guideline observance are not in contradiction with budgetary constraints.
D. Manufacturers need to be encouraged to provide software to expedite all stages of the simulation project life cycle, thereby enabling action C. It is vital that if pursued, this is undertaken in combination with Action B in order to offset any potential loss of skills that this may create.

Lastly, findings as regards guidelines and their observation and status in the continental EU, leads the project to also suggest a fifth action:

E. Establish a practitioner centric, pan-European forum, for the discussion and debate of simulation issues with, as far as possible occasional face to face meetings for dialogue and the performance of short term (pro-bono) working groups.

This last action may be the most important to implement for many European countries for while 'advanced' guideline countries already have such fora in place, there is no equivalent within the EU

despite the existence of an extensive transport planning market, and it is believed that was a primary reason for the low response (and comparative lack of experience) of respondents from the EU to the WG4 survey. (For example in the U.S.A primarily academic discourse with practitioner involvement is taken forward by TRB, while more practical topics are advanced through the SimCap action of the ITE (26) and elsewhere (eg AASHTO, 27), while in the U.K, the topic is taken forward by the TMF (28)). In short, the lack of an ability to discuss these and related issues may result in a profession that does not consider the issue of calibration and validation to be as important as their counterparts in other regions.

In Conclusion, the MULTITUDE project has provided a first avenue for the exploration of calibration, validation and other modelling issues, for participants from some countries for the first time, and has been pivotal is raising the need for discussion on this topic and has re-invigorated progress in Holland, France and Finland. While Transport Planning, Engineering and even ITS are seen as key enablers for enhancing mobility, safety and economy worldwide, and are the subject of many governmental funded initiatives worldwide, simulation, one of the core methodologies on which these disciplines rests (and in particular the science underlying it) is being increasingly overlooked. Maintaining the momentum set by MULTITUDE in this area is therefore crucial to ensuring the gap between simulation science/academia and practice, is not allowed to widen any further.

Acknowledgements

The authors would like to thank the following WG4 participants for their contributions and comments over the course of the last 3 years and in providing direction to this work:

- Christine Buisson, IFSTTAR, FR
- Jordi Casas, TSS, ES
- Winnie Daamen, TU Delft, NL
- Axel Leonhardt, PTV, DE
- Monica Menendez, ETHZ, CH
- Ronghui Liu, University of Leeds, UK.

This document has also benefitted from comment received from a range of external reviewers, including agencies, consultants and academics from the UK, Germany, France, Holland, the USA, Australia and New Zealand.

References

1. Brackstone, M., Montanino, M., Daamen, W., Buisson, C. and Punzo, V. (2012). Use, Calibration and Validation of Traffic Simulation Models in Practice: Results of a Web based Survey. Proc. of the 90th Transportation Research Board Annual Meeting. Paper 12-2606. TRB, Washington, D.C. U.S.A.
2. Daamen, W., Buisson, C. and Hoogendoorn, S. Eds. (Forthcoming). Traffic Simulation and Data: Validation Methods and Applications. CRC Press, Oxford, U.K. 2014.
3. UK Department of Transport. Transport Analysis Guidance - WebTAG. Unit 3.19. Last accessed 29/10/13.
4. Highway Capacity Manual: HCM2010. Transportation Research Board. (2010). Accessed July 30, 2013.
5. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.
6. Transport for London (2010). Traffic Modelling Guidelines, TfL Traffic Manager and Network Performance Best Practice. V3.0. Transport for London, London, UK.
7. FHWA (2004). Traffic analysis toolbox volume II: decision support methodology for selecting traffic analysis tools. FHWA-HRT-04-039. Federal Highway Administration (FHWA), Washington, DC.
8. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
9. VTRC (2006). Microscopic Simulation Model Calibration and Validation Handbook. Virginia Transportation Research Council Technical Report VTRC 07-CR6, Traffic Operations Laboratory, Center for Transportation Studies, University of Virginia, Park, B. and Won, J.
10. Austroads (2006). The use and application of microsimulation traffic models. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
11. Austroads (2010). Guidelines for selecting techniques for the modelling of network operations. Austroads Research Report AP-R350/10. Austroads, Australia. ISBN 978-1-921709-07-4.

12. Road and Maritime Services, NSW, (2013). Traffic Modelling Guidelines. V1.0. Roads and Maritime Services, NSW, Australia.
13. FGSV Verlag Köln (2006). Hinweise zur mikroskopischen Verkehrsflusssimulation - Grundlagen und Anwendung [FGSV Nr. 388], FGSV Verlag Köln, ISBN 3-939715-11-5. Germany.
14. Transport Association of Canada, (2008). Best Practices for the Technical Delivery of Long-Term Planning Studies in Canada. Final Report. Transport Association of Canada, Ottawa, Canada. ISBN 978-1-55187-261-7.
15. New Zealand Modelling User group. (2013). NZ Transport Modelling: Observed and Modelling Comparison Criteria. Last accessed 10/1/14.
16. Adviesdienst Verkeer en Vervoer. (2002). Leidraad model – en evaluatiestudies benuttingsmaatregelen. Holland.
17. Japan Society of Traffic Engineers, (2002). Standard Verification Process for Traffic Flow Simulation Model. Traffic Simulation Committee, Japan Society of Traffic Engineers.
18. Aalto University (2012). Calibration of Simulation Tools. Last accessed 20/8/13.
19. Department for Transport, U.K. (2013). Design Manual for Roads and Bridges (DMRB): Volume 12, Section 2. Last accessed 30/7/13.
20. Brackstone, M. (2013). Findings from MULTITUDE Guidelines Questionnaire. IOMI, UK. Available at: www.multitude-project.eu.
21. UK Highways Agency. Post Opening Project Evaluation (POPE) of Local Network Management Schemes (LNMS). Last accessed 29/10/13.
22. Wunderlich K., Vasudevan, M., and Deurbrouk, T. (2013). Traffic Analysis Tools Volume III Update: Key Topic Prioritization Exercise. TRB SimSub Annual Report, December 2013.
23. Ge, Q. and M. Menendez, M. (in press). An Efficient Sensitivity Analysis Approach for Computationally Expensive Microscopic Traffic Simulation Models. *International Journal of Transportation.*
24. Transport Modelling: Towards Operational Standards in Europe. Last accessed 26/11/13.
25. MOTOS Handbook. (2007). MOTOS Project. European Commission, FP6, DGTREN Last accessed 26/11/13.
26. SimCap The Simulation and Capacity Analysis User Group (SimCap). ITE. Last accessed, 8/11/13.
27. American Association of State Highway and Transportation Officials (AASHTO) (2010). Best Practices in the Use of Micro Simulation Models. Final Report of NCHRP 08-36/Task 90. AASHTO. Last accessed 6/11/13.
28. Transport Modelling Forum. (2013). PTRC Education and Research Services ltd. Last accessed 6/11/13.

Appendix A – Issues of Importance. Discussions.

Issue 1 - How to structure and manage a simulation project

By Pete Sykes (PS-TTRM, UK).

The topic of project management and its role in the successful outcome of an endeavour of any kind is covered in numerous textbooks and generic advice available to cover the involvement of stakeholders, the selection of the project team and the skills they require, the budget controls to be implemented, and the role of quality control procedures. This advice is freely accessible, and as applicable to a transport microsimulation project as to any other, and it is assumed that these basic principles are in place as a matter of course.

Existing guidelines

The existing technical guidelines for simulation projects are aimed at different readerships and hence vary significantly in how they set their bounds of what constitutes a 'project' and how much emphasis they place explicitly on microsimulation. For example, The Canadian "Best Practice" guidelines (1) and the UK WebTAG (2) are both concerned with long term transport planning studies and discuss the role of the transport model in the decision making process with emphasis tending towards the demand side of the transport model rather than the assignment model. The Canadian document places the transport project in the context of the wider strategic planning decisions, while the UK TAM document focuses on environmental, sociological and economic evaluation measures in a narrower more transport based context. At the other end of the scale, software developer's guidelines focus solely on the production of the microsimulation model and make the assumption that the wider context of the project is described elsewhere. As the various software products each have different methods of representing the road network and the traffic demand in the network etc, these guidelines are necessarily product specific.

The spectrum of guidelines is completed by those lying between these two extremes. Some are wide ranging in scope but describe only one software product, e.g. the Oregon DoT VISSIM protocol (3), while others focus on one application and are more agnostic about the preferred product, e.g. the UK HA guidelines (4) is solely concerned with application of microsimulation to trunk highway applications and TfL (5) discuss microsimulation primarily in the context of signalised junction optimisation. Guidelines may also limit their remit to a particular issue in the modelling process such as the VTRC guide to calibration which is highly focussed on that one specialised task.

It is also important to be aware of the intention of the authors of each guideline. In general there are two classes of author: the software suppliers and the agencies that use the simulation models. The former will naturally focus on the capabilities of their own products and can be very specific with respect to how to undertake various tasks with their software. However bias can creep in with respect to what tasks can be undertaken and what calibration inputs or criteria are most significant. Agency led guidelines will be less prescribed with respect to software capabilities and more oriented to general methodologies but with the limitation that the general view may omit some of the advanced or unique features of a more product specific method. The analyst may well be advised to consult both sets of guidelines and negotiate with the stakeholders over a project specific management methodology derived from the best practices of both.

Project structure

Almost all major existing guidelines describe a basic structure of the project and discuss a broadly similar sequence of events: Inception, data gathering, model build, model calibration and validation, use of the model in assessment, and final reporting. The emphasis, the detail, and the identification of points unique to a microsimulation project differentiates each of the documents.

The inception stage of a project covers the identification of stakeholders, the setting of the project bounds and expectations of what can, and, according to the FHWA guidelines, what cannot be tested (6). Some guidelines simply state bounds must be set but offer no advice on how to set them, while others go into the detail of specific technical considerations such as geographic bounds to ensure all realistic diversionary routes are included and time period settings to ensure current and future congestion peaks are modelled.

Similarly, all guidelines to some degree discuss data acquisition. The differentiating features are in the processes described to clean and condition data, and the collection of data specific to microsimulation. Pointers to data sources are described but these are inevitably local to the country of origin of the guidelines. Perhaps the largest issue in data acquisition is in the interface to other transport models to re-use existing data. Inevitably these models are built with different intentions, at different scales, or to different standards of accuracy. For example, zone schemes for a wide area strategic model will be coarse compared with those required for a small area simulation model. No guidelines refer to the problems inherent in taking data from one class of assignment model (i.e. an aggregate model) to another (i.e. a microsimulation model).

The model build process is covered in a generic manner in some guidelines or with explicit reference to particular software products in others, similarly the discussions of methods for calibration and validation. The FHWA recommends calibration of the network capacity, the route choice and finally overall system performance, while the Australian guidelines (7), derived from the FHWA, add a demand calibration step before route choice while the UK HA guidelines (4) discuss probable parameters to adjust, but as the suppliers of the software referred to in that document show little agreement in the order or methodologies to adjust these parameters, the guidelines can go no further than itemising them.

Reporting on the model outputs and the relative merits of the alternative options tested, once again is covered in different levels from a tutorial on statistical hypothesis testing to differentiate between options in the FHWA guidelines, to brief treatments in both the Australian guidelines and the SIAS Good Practice Guide (8). Similarly the use of the outputs in further social or environmental assessment is discussed by some, notably the UK WebTAG, while others simply state what raw outputs are available.

Use of Innovative Solutions

A common issue in simulation models is when a situation is encountered that the software cannot reproduce and an "innovative" method of introducing that behaviour is required. This covers practices such as artificially extending off ramp lanes into the hard shoulder area when queues regularly overflow, or using signals to modify junction priorities in high congestion to reflect observed rather than intended behaviour. These modelling 'tricks of the trade' are intended to introduce artificial physical changes to the network to accommodate observed behaviour that would otherwise be absent from the model. The question asked in some guidelines is: "*When are these*

acceptable pragmatic solutions, and when are they abuse of model parameters to achieve an artificial level of calibration?" The Australian RMS/NSW guidelines (7) explicitly mention speed adjustments on uncongested remote links that enable the modelled journey times to calibrate with observations with no further justification as unacceptable practice which will result in rejection of the model, whereas the FHWA sanctions capacity adjustments on individual links where required with the caveat that as these are "non-behaviour based" they should be used sparingly. The UK HA document offers advice on resolution of this issue by requiring that changes to parameters are based on observation, can be justified in terms of human behaviour, and are documented with an adequate evidence base to enable the model auditors to judge whether a model is fit for the purpose of future scheme appraisal. The documentation of parameter choice with the reason for that choice is regarded as an essential part of the model validation process, the interpretation of that choice and the judgement as to its suitability is however quite subjective.

Audit and error checking

Model checking and audit is described by all guidelines, with the role of quality control being an essential part of project management. The document from the RMS/NSW gives a clear and succinct differentiation between the two activities by defining "audit" as a final external report to "sign off" a completed model and "model checking" as an on-going process of internal peer review as work progresses (7). This guideline also comments that while an audit may be the cheaper option if the model is approved, it is more expensive if it reveals errors which may have been found much earlier by a peer review process.

Both audit and peer review, require guidance as to what to look for in a simulation model and this is supplied in the guidelines. Those which are software product specific are able to give detailed lists of model artefacts and parameter ranges to check and these can run to several pages long. Guidelines that are not software specific are restricted to more general lists of items to check and to expounding a philosophy of justification for all changes from the software supplier defaults or from normal modelling practice.

Project milestone charts and the criteria to state if a milestone has been reached may be found in many guidelines either as guidance for project managers or as a formal mandatory requirement for project acceptance (5). The milestones refer to distinct events such as completed data collection and verification, achieving the calibration goal. They may also refer to more nebulous goals such as "50% complete model" or "skeleton model complete" which lack clear definition and are therefore less useful in describing project progress. In terms of stakeholder control of the audit and review process, most guidelines are advisory only. However, the UK TfL document (5) and the ODoT VISSIM protocol (3) merge audit and peer review in their audit process by describing a formal procedure with pro-forma checklists to describe mandatory sign-off points at set milestones in the project.

References

1. Transport Association of Canada, (2008). Best Practices for the Technical Delivery of Long-Term Planning Studies in Canada. Final Report. Transport Association of Canada, Ottawa, Canada. ISBN 978-1-55187-261-7.
2. Transport Analysis Guidance - WebTAG. Unit 3.19. UK Department of Transport. Last accessed 29/10/13.

3. Oregon Department of Transportation (ODoT), (2011). Protocol for VISSIM Simulation. Oregon Department of Transportation.
4. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.
5. Transport for London (2010). Traffic Modelling Guidelines, TfL Traffic Manager and Network Performance Best Practice. V3.0. Transport for London, London, UK.
6. FHWA (2004). Traffic analysis toolbox volume II: decision support methodology for selecting traffic analysis tools. FHWA-HRT-04-039. Federal Highway Administration (FHWA), Washington, DC.
7. Road and Maritime Services, NSW, (2013). Traffic Modelling Guidelines. V1.0. Roads and Maritime Services, NSW, Australia.
8. SIAS Ltd (2006). The Microsimulation Consultancy Good Practice Guide. SIAS Ltd.

Issue 2- How to handle model 'warm up'/run duration

Peter Wagner (DLR, DE)

In microscopic traffic simulation, as in any dynamic model, the 'state' of the system at the initial instant of the simulation has to be known and fed into the model. In fact, the evolution of the system depends on its initial state i.e. the traffic on the network at t_0, as well as on the inputs for all the following instants such as the OD demand. Since the state of the traffic on a network at time t_0 is rarely known – for a microscopic model it would entail knowing the state and features of all the vehicles on the network (position, speed, route and destination) the ideal way of solving such an 'initial conditions' problem is simply to start the simulation from an instant in which the network is empty, that is simply to simulate from midnight, or at least, from off-peak. While necessary, this is rarely done however and in practice, a warm up period of the microscopic traffic simulation is made. The aim of the warm up period, therefore, is just to obtain a reasonably realistic state of the traffic on the network at the beginning of the simulation.

For example, consider a study area of 10 km in diameter, with a maximum speed of 20 m/s and a static demand. Under the assumption that all the trips are 5 km long, for a trip to complete its journey, it will take 5000/20=250 seconds. This is the time the simulation needs to reach a first equilibrium at which at least in principle as many vehicle enter the network as are leaving it. For a small demand level, this is the shortest warm-up period that is required: i.e. it is given by the longest travel time of all the routes of the study area under consideration. Typically, it is a good idea to add a safety margin to this value, e.g. to make the warm-up time period twice as long as the time computed from the geometry of and the routes chosen in the study area. (The first vehicles to enter the study area is, in effect, driving at 'midnight', with no other vehicles on the road and nobody waiting in front of them, e.g. at traffic lights, thereby causing travel times to be shorter as they do not have to wait for vehicles in front to clear the intersection. Data from the start of any run therefore has a strong bias toward smaller delays, which does not reflect the real situation).

This text will very often use the word demand. In general, demand is described by a deterministic function $q(t)$ on top of which some noise is added. This is due to the fact that the process that describes the entering of a vehicle into the study area can in most cases be modelled as a Poisson process. Often, a peak hour is simulated with a fixed demand, i.e. the deterministic function for this peak hour is just a constant, while there may be still the noise on top of this constant. Almost all simulation tools do also have the option to fix the demand for this peak hour to be exactly the pre-specified amount (say 3000 vehicles). In applications, this may make a difference: e.g. a traffic signal is sensitive to the stochastic process that is used to put traffic into the network.

For larger demand, which is usually the case under consideration, the answer is not so simple and can sometimes be found only by trial-and-error: given that the demand during warm-up is constant and small enough so that the system does not run into oversaturation, then by analysing the difference between vehicles entering the study area and the number of vehicles that have completed their route, a time-series can be easily constructed which must finally settle around zero. In more complicated cases, it might not be possible at all to state that warm-up has been long enough. In particular, traffic signals at intersections may switch between free and congested, causing strong fluctuations and non-equilibrium behaviour. By analysing those data, one typically sees strong fluctuations, e.g. the travel time can easily vary by a factor of two or three between the

shortest and the longest. As a general rule of thumb, for large networks with high demand levels, the warm-up period should be long enough such that traffic is getting onto all, or the majority, of the links in the network before the period of sampling results from the simulation starts.

If possible, an additional recommendation is to simulate, in addition, the time before and after saturated conditions and use the time before the peak period as the warm-up period. However, in the case of oversaturation, the statistics to be sampled from the simulation must in addition use a cooling-off time period, since the queues created at the simulated intersections may need a long time to dissolve. However, since they have been created during the time of (over-)saturation, their effect needs to be accounted for, and this will happen in the subsequent cooling down period.

Over-saturation

The issue of setting warm up period for over-saturated conditions is far from completely resolved. In case of congestion, it may spill back into the entry links and block incoming traffic which can trick the above mentioned method to determine the end of the warm-up period. In addition, a large network may hold a lot of simulated vehicles at the end of the simulation run, that have accumulated a large delay already and would need a considerable amount of time until the whole network is cleared, but they have picked-up this delay during the simulation time. By ignoring them, delay times will be seriously underestimated.

Note in addition, that micro-simulation programs have a tendency toward grid-lock which is eventually stronger than the corresponding real-world behaviour. One can easily construct a situation where vehicles block each other so, that a whole block in the network is locked, where real drivers still find a place to go ahead. If this happens, then the results from the simulation have to be taken with extreme care. This will be signalled by a strong increase in travel time. One solution to this is to extend the simulation time around these times of oversaturation by running a whole day. Alternatively, it is possible to run a cooling-off period. Traffic is still being generated onto the network, but the purpose of the cooling off period is to allow the traffic from the main simulation period to complete their journey. Depending on the congestion level of the simulated network, the length of the cooling off period is therefore varied. A practical recommendation is to set a large cooling off period and to allow the simulation to run until all traffic in the main period has reached their destination. This would make it much easier to compute the correct delay time, and other performance indicators (PI) as well, for a given scenario.

Existing guidelines

Where this issue was discussed in existing guidelines, their recommendation has been similar to the ones formulated above. TSS and PTV user manuals recommend the longest travel time as warm-up time, and so does the German guidelines (1, although it is formulated differently) while UK-HA and AU (2, 3) recommend twice this time, while the FHWA guidelines (4) recommend the extension of the simulation time-span by a warm-up and cool-down period, and, if this cannot be adopted, then it suggests for the longest travel time version. Other guidelines contain such recommendations too, but they are embedded into more complex additional settings such as 15 min (uncongested) or 15 min to 30 min for congested conditions in the UK TfL guidelines for example (5).

References

1. FGSV Verlag Köln (2006). Hinweise zur mikroskopischen Verkehrsflusssimulation - Grundlagen und Anwendung [FGSV Nr. 388], FGSV Verlag Köln, ISBN 3-939715-11-5. Germany.

2. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.
3. Austroads (2006). The use and application of microsimulation traffic models.. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
4. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
5. Transport for London (2010). Traffic Modelling Guidelines, TfL Traffic Manager and Network Performance Best Practice. V3.0. Transport for London, London, UK.

Issue 3 - Number of runs to perform

By Costas Antoniou (NTUA, GR) and Peter Wagner (DLR, DE)

Traffic is a stochastic, highly dynamic phenomenon, resulting from the actions and interactions of large numbers of travellers, along with various exogenous events. For instance, upon arriving at a traffic light the split of a second decides whether to stop or not, with a profound consequence on the overall travel time. An unloading truck for example may block a road partly or entirely, with a considerable impact on flow and travel times. Traffic simulation models reflect this stochasticity, as they use random variables and sample from random distributions to represent decisions made by the agents simulated in the models (e.g. route or lane choice decisions). The drawback of this is that multiple runs of the simulation program are needed to obtain reliable results. This allows the computation of mean, and standard deviations and from this the derivation of confidence intervals. However, the key question is: how many replications are needed?

Importance

By using a single run of a simulation, there is a danger that, while the value found is possible, it does not represent the desired result properly. Using such an instance for decision making could jeopardize the validity of the results and lead to bad planning. Performing multiple runs and averaging their outputs has the advantage of steering the results more towards the expected values of the 'true' distributions. For example, the effect of a replication near the tail of the distribution would be muted by other replications closer to the expected value. When the outputs from two or more replications are averaged, the likelihood of approaching the expectation value increases. This is reflected also in a statistical treatment: a small number of replications implies that there is uncertainty in the obtained outputs beyond the inherent stochasticity of the simulator. As the number of replications increases, then the contribution of an outlier/extreme value to the average decreases. This is demonstrated in Figure 4, in which 50 random draws from a uniform distribution between 7 and 13 (therefore an expected value of 10) are shown as bullets. The grey line shows the expected value, while the solid like shows the average value obtained as from when considering the observations up to the current one. While a small number of observations may lead to biased results, as the number of observations increases, then the average quickly converges to the "correct" value. Naturally, the magnitude of the measurements that is being considered may also help determine the number of replications required for a specific metric. For example, large travel time values might be able to absorb some of the uncertainty, thus allowing a smaller number of replications. On the other hand, density values might require a larger number of replications. In any case, a simple statistical test could be constructed to determine whether the number of replications is satisfactory for a given situation. Appendix E of FHWA (1) provides an overview of hypothesis testing that is useful in this respect.

Surprisingly, a systematic treatment of this topic is not easily found in the traffic literature. One exception to this may be the work of Shteinman et al. (2), however the procedure described in this paper starts with at least 30 replications. Usually, the number of replications needed is not discussed in detail, and a number between 5 and 10 is chosen, either arbitrarily (based on experience, perhaps an empirical rule like that shown in the previous figure), or based on some simple formula (eg, 3, 4, 5). The general idea is that the number of replications must be increased as the standard deviation of a set of simulation runs is higher. The exact number of replications is determined using the level of significance desired of the results and the allowable percentage error of the estimate.

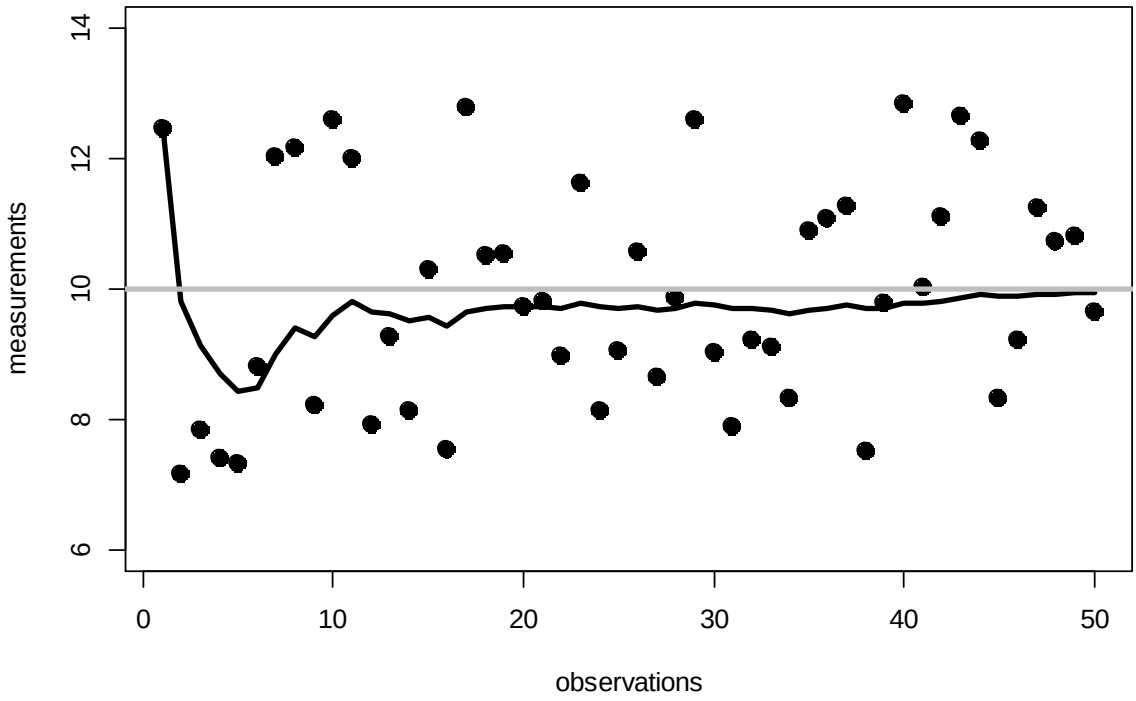

Figure 4: Demonstration of average value convergence as the number of observations increases.

There are two additional points however that the practitioner should consider during any project. In a simulation, it is easy to use the same demand curve $q(t)$ with the additional noise added by the stochastic nature of the demand process. However, in reality it is not clear, that one has the same demand curve for each day, in addition to the fact that demand is varying stochastically around this demand curve. So, there is in fact a difference between the day-to-day variability and the additional short-term fluctuations. This is definitely still an open research question, and no clear advice can currently be provided. Secondly, while we have spoken of drawing averages from multiple runs, this will result in the loss of key information regarding the variability of output, and could perhaps be better dealt with by reporting distributions and/or a set of quantiles instead.

The number of runs is a key element also in the performance of Sensitivity analysis which also involves running multiple replications and processing the results. While in this case the results are generally not averaged, the procedure for selecting the number of runs and – perhaps more importantly - automating the replications could possibly be shared or at least coordinated.

Current treatment

A number of guidelines address this topic, at least to some degree and this ranges from a simple ad-hoc recommendation on a minimum number of replications, to the provision of a methodology for the determination of a minimum number of runs and some guidelines for the application. For example VTRC (6) addresses the issue explicitly and provides both theoretical background and guidance to the modeller to determine the number of replications, as well as some indicative numbers, while the related report (7) indicates that during the process of calibration/optimization (performed using a Genetic Algorithm) only a small number of replications are considered (“usually

less than five"). Therefore, it is recommended the final calibrated parameter sets are evaluated using a larger number of run ("say, 100") and the obtained distributions are compared with field measurements. Alternately, the FHWA (1) presents an example using 10 replications for capacity calibration, and background and guidance on hypothesis testing in several Appendices, while the Oregon DoT (8) mentions using a minimum of 10 runs with different random seed and proposes a process for the determination of the number of runs (similar to that outlined above), and a 95% confidence interval is suggested. Another aspect relates to which outputs should be considered while determining the minimum number of runs, with ODOT stating that it is not practical to test the statistical significance of the average of every data output. Therefore, a recommendation is made to select one or two key measures of effectiveness, based on the project objectives, and report these. It is stated that the selection of the points/corridors for which to test this can be selected by the modeller, but they should be approved by the client.

References

1. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
2. Shteinman, D., Chong-White, C. and Millar, G. (2011). Development of a statistical framework to guide traffic simulation studies. Australasian Transport Research Forum 2011 Proceedings, 28 - 30 September 2011, Adelaide, Australia.
3. Toledo, T., Koutsopoulos, H., Davol, A., Ben-Akiva, M., and Burghout, W. et. al. (2003). Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study. *Transportation Research Record*, **1831**, 65-75.
4. Bourrel, E. (2003). Modélisation Dynamique De L'ecoulement Du Traffic Routier: Du Macroscopique Au Microscopique, PhD Thesis thesis, L' Institut National des Sciences Appliquées de Lyon.
5. Ahmed, K., (1999). Modelling Drivers' Acceleration and Lane Changing Behaviour, PhD Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
6. VTRC (2006). Microscopic Simulation Model Calibration and Validation Handbook. Virginia Transportation Research Council Technical Report VTRC 07-CR6, Traffic Operations Laboratory, Center for Transportation Studies, University of Virginia. Park, B. and Won, J.
7. VTRC (2006). Simulation Model Calibration and Validation: Phase II: Development of Implementation Handbook and Short Course. Virginia Transportation Research Council Technical Report VTRC 07-CR5, Traffic Operations Laboratory, Center for Transportation Studies, University of Virginia. Park, B. and Won, J.
8. Oregon Department of Transportation (ODoT) (2011). Protocol for VISSIM Simulation. Oregon Department of Transportation.

Issue 4- Sensitivity analysis and how to choose parameters to calibrate

By Vincenzo Punzo (JRC/UNINA, IT) and Biagio Ciuffo (JRC, IT).

To understand the influence of parameters or other inputs (e.g. OD demand) on model outputs is one of the most challenging activities for simulation practitioners and developers and allows us to understand the effects of differing parameters. It is not unusual to encounter parameters that have more influence on some specific outputs than on others or, that are meaningful only in specific operating conditions. Sometimes a parameter seems to have no influence on simulation results, but only because it affects the outputs exclusively in conjunction with other parameters (i.e. it has only interaction effects). Other times, a parameter may have marginal influence on the outputs so that it can be fixed at any value of its range of variation without affecting results. Similarly, analysts are often concerned with the uncertainty in the estimation of the transport demand and seek to understand the relative importance on model outputs of different OD relationships or evaluate the sensitivity of uncertain future demand patterns on model predictions.

The simplest, commonest and most intuitive way to test the influence of a 'factor' - being a model parameter or another input - is to let it vary while keeping all others at the fixed default values, an approach referred as One factor at A Time (OAT) Sensitivity Analysis (SA). Similar to this is the calculation of partial derivatives in a point of the input space. Unfortunately, both these approaches can provide biased information on the influence of inputs and parameters as *i)* they have local validity: they explore only few points in the neighbourhood of the default values (for non-linear models, results at a point cannot be extrapolated elsewhere) and *ii)* they do not take into account the impact on the outputs of the inputs interaction: it can happen that a parameter shows its capability of affecting the outputs only when it is varied simultaneously with other parameters.

In the last decades, however, new cross-disciplinary techniques have been developed, generally referred as Global Sensitivity Analysis, that aim to perform quantitative sensitivity analysis without the limitations of the 'local' methods. In general, applying these techniques involves constructing an input-output relationship and this can be achieved in a Monte Carlo framework, that is, by running the model a large number of times. At each run the model is fed with a different combination of inputs or parameters which are randomly drawn from the corresponding ranges or probability density functions. Therefore, a proper sensitivity analysis involves analyzing the experimentally drawn input-output relationship. An output of the analysis might be, for instance, the so-called sensitivity indexes that allow identifying the share of the output variance that is explained by each of the uncertain parameters or inputs involved in the analysis. Quoting Saltelli (1): "Sensitivity Analysis is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be apportioned to different sources of uncertainty in its inputs".

Global sensitivity analysis techniques can therefore be used for a number of useful purposes:

- to prioritize parameters according to their importance, that is, their influence on a specific output (also referred as 'factor prioritization' setting);
- to identify the subset of parameters to calibrate, or conversely, to identify the parameters that can be fixed at whatever value, within their range of variation, without affecting the outputs ('factor fixing' setting);

- to identify the inputs that make the outputs exceed a specific threshold ("factor mapping" setting).

It is worth mentioning that, in a sensitivity analysis the selection of the inputs is not defined a-priori but depends on the analysis's objective. All the parameters of one sub-model e.g. the route choice model, could be chosen as inputs to the analysis if the aim is to investigate their relative influence on the outputs. The analyst could choose instead to group parameters in order to investigate the relative influence on the outputs of the entire groups, that is, of the corresponding models (e.g. the group of car-following parameters vs. the group of lane changing parameters, etc.). As the number of inputs to investigate drives the choice of the sensitivity technique to apply (together with the computational cost of model run) the analyst can choose different strategies. For example, one can rely on established knowledge in traffic flow theory in order to identify a smaller subset of inputs already recognized as influential to which directly apply more demanding and powerful techniques. Or, one can first apply less demanding screening techniques to the larger set and, then, more demanding ones to the smaller subset identified through the first screening (e.g.2, 3). The application of meta-modelling is also another option in order to reduce the computational burden of the model (3).

It is also important to note that the results of a sensitivity analysis are conditional on the inputs that remain fixed. In other words, if a sensitivity analysis of a traffic flow model is run on a specific network with a specific demand, results are likely to change for a different network or if a different demand is adopted.

The importance of such techniques for the economy of modelling suggests their adoption from the early phases of model and software development and testing. Commercial software developers should use them to simplify, enhance and defend their models against misuse. Practitioners would benefit from this work through an improved understanding of the models response and from the simplification of models; in fact, like most of law-driven models, traffic flow models are often over-parameterized (3, 4).

Finally, it is also worth mentioning that, besides the software code needed to perform the sensitivity analysis (for instance, to calculate the sensitivity indexes from the experimentally drawn input-output relationship), a code to run the model iteratively i.e. in a Monte Carlo framework, is also needed. At the moment, this is not a built-in functionality in currently available commercial traffic simulation software, but requires additional coding by the final users. At the end, this requirement is what really makes the application of such techniques cumbersome for most of the practitioners and, that hinders their massive utilization, although progress is now being made in this area (5). Given the relative novelty of the topic, in Appendix B a brief overview and some insight on sensitivity analysis techniques suitable for traffic simulation models are reported.

The importance of understanding Sensitivity

According to Hornberger and Spear (6), "*most simulation models will be complex, with many parameters, state-variables and non-linear relations. Under the best circumstances, such models have many degrees of freedom and, with judicious fiddling, can be made to produce virtually any desired behaviour, often with both plausible structure and parameter values*."

This poses serious questions on the reliability of results of simulation studies as well as on the transparency of the whole modelling process. In fact, it is reasonable to claim that results of a study are mostly driven by the way in which the model parameters and the transportation demand are estimated. These two elements are crucial for a correct representation of the transportation system. However, their estimation - either carried out simultaneously or sequentially - is a complex high-dimensional problem. Therefore, either relying on a simple trial and error approach, or applying an automated procedure based on optimization algorithms, it is hard to find a solution which is reliable and robust, given the large number of unknowns.

However, as mentioned before, often it happens that just a subset of the inputs drives the overall variability of the outputs: in most of the cases, complex high-dimensional models present a strong asymmetry in the way the inputs influence their outputs. The identification of these inputs is therefore crucial in order to simplify the problem and to make it tractable and affordable in practice as well. Global sensitivity analysis represents the family of techniques to be used also for this purpose. In particular, the analysis (analytical or numerical) of the input-output relationship carried out using one of the available techniques, allows the identification and the ranking of the most important inputs or parameters.

The reader, at this point, may think that this has a scarce applicability to the real world, given its (apparent) complexity, and until some decades ago this was the general opinion. Nowadays, however, after several mistakes caused by erroneous modelling practice (the most famous being the modelling arguments that neglected the existence of climate-change at the end of the last century, or the failure in predicting the economic crunch starting in 2008), sensitivity analysis has rapidly spread in the scientific community as a means of achieving a more robust modelling practice. For the same reason, in the last years, important international institutions have started to ask for evidence from sensitivity analyses to support the modelling methodologies applied (this is the case, for example, of the European Commission, the Intergovernmental Panel for Climate Change, etc.). Transportation modellers and practitioners are aware that the same request is starting to arrive also in their field. Another important aspect that should encourage practitioners to carry out sensitivity analysis is its economic implication. When the estimation of the inputs of a model requires the implementation of experimental investigation or surveys (like in the case of traffic demand), restricting the analysis to a more limited number of inputs may provide significant economic benefits, or at least, may help using the available resources in a more efficient way. If, for example, in a transportation study, results of a sensitivity analysis show that just a dozen of the hundreds of origin/destination flows affect the variability of the outputs of interest, the traffic demand estimation can be reasonably restricted to those ones. The reader is referred to Saltelli et al. (1, 7), Daamen et al. (8) and (9).

Lastly, it should be kept in mind that the output of this topic is an input for the calibration and validation phases and therefore should be considered in conjunction with the corresponding topics. SA can also be useful in case of lack of data, or in providing information to drive the future collection efforts, or provide the sensitivity about alternative scenarios or parameter configurations. For example, if the calibration objective was to reproduce travel times with high fidelity and the sensitivity analysis showed that route choice parameters are more influential on travel time than the car-following ones, one could decide to devote more effort in the calibration of route choice than of car-following.

Steps in the sensitivity analysis of model outputs

Saltelli et al, (1), describes this process as consisting of eight distinct steps. Firstly, 1) defining the goal of the SA and therefore the output of the model that is better suited to achieve it. Next, 2) decide which are the input factors to include in the analysis (considering our introductory discussion, depending on the case study and on the specific application, inputs to be included can be either the OD pairs, or the parameters of some or all the sub-models of a traffic model, or all together). 3) Each input needs to have a distribution function chosen for it (this can be taken from the literature, derived from empirical data, based on expert opinions, or, in case no a priori information, hypothesized by the analyst) and define, if relevant, the correlation structure for the inputs (in case no information is available, the analyst can think to perform a SA in the hypothesis that no correlation is in place, checking later the assumption correctness). 4) Choosing a sensitivity analysis method (depending on the specific application) is the next, followed by 5) designing the input sample (this is connected to the method chosen). Subsequently, 6) the evaluation of the model over all the inputs' combinations defined in the experimental design is performed, and 7) the SA method is applied to the input-output relationship. Lastly, 8) if the results do not appear satisfactory, change the settings and perform a new SA until results are satisfactory. For example one could be interested in understanding the influence of the lane changing parameters on the lane counts, and being surprised by the fact that none of them has a sensible influence on the output chosen. It could be the case that such parameters have influence in their interactions with the car-following parameters that means one needs to include the latter in the analysis, or verify whether they have influence on other output, such as lane speeds.

Many practical difficulties are hidden behind these steps and the practitioner will easily realize that only experience will make the entire process smoother. In an attempt to guide the reader throughout the selection of the method which fits the objectives of the analysis, the following table summarizes some characteristics/requirements to execute the methods proposed in the columns.

From the table it is clear that different methods have very different rationales and for this reason they are tailored for different applications. For example, variance-based methods which are very time consuming need N(k+2) model evaluations, and can be reasonable only when one single evaluation lasts less than 1 minute. In fact, for all the methods the requirement on the model evaluation time is strictly related to the number of runs needed by the method. A brief introduction to each method is provided in Appendix B.

	Regression coefficients	Scatter plots	Factorial (fractional)	Elementary effects	Variance-based	Meta-model based	Monte Carlo filtering
Non-linear models?	No	Yes	Yes	Yes	Yes	Yes	Yes
Inputs' interactions?	No	Yes	Yes	Yes	Yes	Yes	Yes
Sample from?	Distrib	Distrib.	Levels	Levels	Distrib.	Distrib.	Distrib.
Number of inputs	<100	<10	>100	20-100	<20	20-100	<20
Evaluation time of the model	1min-1h	<1h	<10h	<1h	<1min	<1h	<1h
Number of runs of the method	500-1000	1000	k-2k	r(k+1)	N(k+2)	100-1000	500-2000
SA setting	FP	FM	FF	FF	FP,FF	FP,FM	FM

Symbols:
k: number of inputs (factors)
N: Monte Carlo Size of the experiment (typically 500-1000)
r: number of trajectories in the sampling strategy (typically 4-10).
FP: factor prioritization (inputs are ranked on the basis of their influence on the outputs)
FF: factor fixing (those inputs with no impact on the outputs are individuated; these inputs can be fixed to any value without affecting the results of the analysis)
FM: factor mapping (a setting aimed at individuating which factor makes the output of the model falling in certain regions)

Table 1. Table of when to use what (Saltelli et al., 7, page 273). Indications provided in the table should not be considered as prescriptive and are based on the experiences of the author.

(Misconception of) Sensitivity analysis in existing guidelines

At present, in the existing guidelines, there is a total lack of coverage of Sensitivity Analysis, as defined in the previous sections. In fact, what is described as this' is generally called 'Uncertainty Quantification' (UQ) or Uncertainty Analysis in the wider simulation modeling community. "*UQ is the science of quantitative characterization and reduction of uncertainties in applications. It tries to determine how likely certain outcomes are if some aspects of the system are not exactly known*". This misconception of sensitivity analysis can be firstly found in the "Guidelines for Applying Traffic Microsimulation Modeling Software" (10) of the California Department of Transportation and, successively, resumed in other documents (11, 12 and 13). For instance, the FHWA guidelines (11) state "… *A sensitivity analysis is a targeted assessment of the reliability of the microsimulation results, given the uncertainty in the input or assumptions. The analyst identifies certain input or assumptions about which there is some uncertainty and varies them to see what their impact might be on the microsimulation results".* It should be clear then that what is called 'sensitivity analysis' here is

something different from the techniques introduced earlier and nearer instead to an uncertainty analysis.

Before commenting further, it helps to clarify that SA and UQ have in common the phases of the experimental design and the Monte Carlo simulations. These steps are functional to build the aforementioned input-output relationship and are generally referred as 'uncertainty propagation'. In the existing guidelines, however, what is called sensitivity analysis (to make "additional model runs with changes in the demand level and key parameters…") and that should be properly named uncertainty analysis, is made through a one factor at a time approach. As clarified in the introduction, such an approach is likely to produce biased results. On the contrary, an experimental design that allows varying all the inputs simultaneously has to be made in order to explore the whole input space and account for the interactions of the various inputs (e.g. demand, parameters, the network).

The ensemble of parameter estimation/calibration, sensitivity analysis and uncertainty analysis, is what is generally referred with the name of 'uncertainty management' of the modeling process. It would be desirable therefore that future guidelines properly address all these phases and help developers and users to migrate existing techniques from other fields to the traffic simulation field. A crucial requirements, however, is that traffic software would be provided with built in functionalities to make a design of the experiments and run multiple simulations accordingly.

References

1. Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. (2004). Sensitivity Analysis in Practice. A guide to Assessing Scientific Models. ed. John Wiley & Sons.
2. Ge, Q. and Menendez, M. (2013). An Efficient Sensitivity Analysis Approach for Computationally Expensive Microscopic Traffic Simulation Models. Submitted to *Int. J. Transp.*
3. Ciuffo, B., Punzo, V., Montanino, M. (2013) "Global sensitivity analysis techniques to simplify the calibration of traffic simulation models. Methodology and application to the IDM car-following model". IET Intelligent Transport Systems. Forthcoming.
4. Punzo, V., Montanino, M., Ciuffo, B.. Which parameters of the Intelligent Driver Model (IDM) do really need calibration? Variance-based sensitivity analysis of traffic flow models. Submitted to IEEE Transactions on Intelligent Transportation Systems.
5. Hale, D. (2013). CORSIM Self-Calibration (patent pending): Users Guide. McTrans Center, University of Florida.
6. Hornberger G. and Spear, R. (1981). An approach to the preliminary analysis of environmental systems. *Journal of Environmental management*, **12**, 7-18.
7. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., and Cariboni, J. et. al. (2008). Global Sensitivity Analysis. The Primer. ed. John Wiley & Sons.
8. Daamen, W., Buisson, C. and Hoogendoorn, S. Eds. (Forthcoming, 2014). Traffic Simulation and Data: Validation Methods and Applications. CRC Press, Oxford, U.K.
9. Sensitivity Analysis. EU-JRC. Last accessed 24/10/13.
10. Dowling Associates (2002). California Department of Transportation Guidelines for Applying Traffic Microsimulation Modelling Software. Dowling Associates.
11. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.

12. Austroads (2006). The use and application of microsimulation traffic models. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
13. Road and Maritime Services, NSW, (2013). Traffic Modelling Guidelines. V1.0. Roads and Maritime Services, NSW, Australia.

Issue 5 - Appropriate definitions of calibration and validation

By Jaume Barcelo (UPC, ES)

From a methodological point of view it is widely accepted that simulation is a useful technique to provide an experimental test bed to compare alternate system designs, replacing the experiments on the physical system by experiments on its formal representation in a computer in terms of a simulation model. Simulation may thus be seen as a *sampling experiment* on the real system through its model (1). In other words, assuming that the evolution over time of the system model imitates properly the evolution over time of the modelled system, samples of the observational variables of interest are collected from which, using statistical analysis techniques, conclusions on the system behaviour can be drawn. The reliability of this experiment depends on the ability to produce a simulation model representing the system behaviour closely enough. A common statement shared by almost all papers dealing with traffic simulation starts by the assertion "*Microscopic traffic simulators are useful tools for designing, evaluating and optimizing transportation systems*", and is followed by a statement on the acceptability of the simulation results to support such decisions, in other words "how close the model is to reality". This can be summarised (2) as: "*To successfully apply a simulation model, the "correctness" or "credibility" of the model is crucial and some testing processes have to be resorted to in order to ensure the quality of the model through model validation, a critical testing process that compares the model output with real-world system behaviour*".

The process of determining whether the simulation model is close enough to the actual system is usually achieved through the validation of the model, an iterative process involving the calibration of the model parameters and comparing the model to the actual system behaviour and using the discrepancies between the two, and the insight gained, to improve the model until the accuracy is judged to be acceptable. Papers and guidelines provide a variety of definitions, leaving aside that in some cases the term verification is also used, sometimes as a third additional concept and sometimes as a synonym for validation, also called evaluation in some cases. Examples include:

- **Calibration,** where the analyst selects the model parameters that cause the model to best reproduce field measured local traffic operation conditions. Model parameters can be clustered in two general classes: network and traffic parameter. In most professional software network calibration is assisted by automatic tools. Since large amount of network data are available from GIS, including turning at intersections when navigation GIS are available, and control plans can also be imported from digital files, as well as many other network attributes, it has been possible to implement utilities to check the network consistency. Therefore the main calibration effort is usually concerned by traffic calibration.
- **Model calibration,** the process of assuring that a model reproduces real-world traffic conditions reasonably well.
- **Model calibration,** defined as the process of adjusting the values of the simulation model parameters and other user-controlled variables such that the observed data is consistent with the simulated data.
- **Calibration** (3) the process in which the model parameters of the simulator are optimized to the extent possible for obtaining a close match between the simulated and the actual traffic measurements, which primarily include volume, speed and occupancy. Generally, calibration is an iterative process in which the engineer adjusts the simulation model parameters until the

results produced by the simulator match field measurements; the comparison part is often referred to as validation.

- **Calibration** (4) implies that the input parameters (e.g. driver behaviour, desired speed) allow the model to recreate the specific network under certain circumstances (i.e. replicate observations, field measurements and other empirical data)
- **Validation** (5) is concerned with determining whether the conceptual simulation model (as opposed to the computer program) is an *accurate representation of the system under study*. If a model is valid, then *the decision made with the model should be similar to those that would be made by physically experimenting with the system (if this were possible).*
- **Validation** (6) means the process of testing that the model does actually represent a viable and useful alternative means to real experimentation. This requires the exercise of *calibrating the model,* that is, adjusting model parameters until the resulting output data agree closely to the system observed data. Validation of the simulation model will then be established on basis to the comparison analysis between the observed output data from the actual system and the output data provided by the simulation experiments conducted with the computer model.

As one can easily see, these definitions share the same concepts formulated in similar terms. Bayarri et al. (7) did a first step to formalize these concepts and in (8) proposes a formal definition, *"A clear statement of what "validation" means is rarely set forward. Usually, the question is put as "does the model faithfully represent reality?" But, the answer to this question is simple: no, models are not perfect. But models can make useful, reliable predictions in particular settings; they may be useful for some purposes, useless for others. We can state this as:*

$P\{ |"reality" - simulated\ output | \leq d \} > \alpha$

where we must specify d = tolerable difference (how close) and α = level of assurance (how certain), say what is meant by "reality" and what is needed to make sense of P and how to compute the probabilities involved. What we mean by "reality" is, operationally, a feasible measure of actual performance of a particular network. For example, it may be a system queue time measure in an urban traffic network under a current signal plan or, perhaps, under a proposed timing plan. To compare actual performance with simulation prediction will require access to field data and simulation output that relate to a performance measure. A review of the literature indicates that little attention has been paid to the characterization of the uncertainty in simulation model inputs. Rather the focus appears to be on the analysis of the stochastic outputs of the performance measures derived from various models".

Unfortunately the definitions stop here and there is no a consensus with respect to *how to do it*. Each paper or guidelines make their own proposal of a variety of statistical methods and measures of how close are the reality and the model, see (9) for an overview. The way generally acknowledged to cope with the approximations of a model, is that of fitting parameters to the real data. This is expected to partly cover both the modelling errors, due to modelling assumptions, and approximate resolution methods, etc. and the uncertainty in the inputs. The process of fitting model parameters to the real data is generally referred as (model parameter) calibration and means identifying the values of model parameters which make the description provided by the model as close as possible to the reality.

This raises two formal questions. Traffic models, as formal representations of traffic systems, have two main components: the supply and the demand. Traditionally calibration and validation have addressed the supply side, but, taking into account the discussion in Bayarri et al. (8), the analyst's perception of reality relies on the information gathered through data collection and subsequent data processing in order to account for uncertainties. *The available data and its uncertainties will determine what can be said about d and* α*.* Surprisingly, the last assertion has received little attention. In almost all the discussions and methodological approaches for calibration and validation of traffic simulation models, attention has been focused on the processes' ability to accurately estimate model parameters and on the statistical methods for assessing model validity, with the implicit assumption has been that the available data for comparison is reliable enough. Data inputs to traffic models can be classified in two categories:

- Directly observable data, i.e. measurements of traffic variables affected by errors (flows, speeds etc.), based on available technologies which must be filtered and processed before using them in the applications.
- Data not directly observable, such as demand modeled in terms of time sliced into Origin-Destination matrices. This input process asks for sound, indirect estimation procedures in order to generate the suitable inputs. The question as to whether demand and supply should be calibrated independently or in a common algorithmic framework is still open to debate.

Another question that has recently drawn the attention of the traffic simulation community concerns the increasing number of parameters available with which to conduct calibration. As Ge and Menendez point out (4), *"commercial traffic simulators usually contain a huge number of parameters to cover various kinds of simulators (e.g. vehicles, public transport). As an example, VISSIM (10) has 192 parameters …, and this figure will most likely continue to grow with new updates".* This means that in practice that calibration is limited to a selected subset of parameters, the problem then is how to make sure that the selected parameters are the most relevant ones with respect to the objectives of the simulation study. However, usually there is no formal procedure to select such relevant parameters and in most cases the ones selected are those subjectively appearing as the most influential, supported usually by the analyst experience. The formalization of a procedure has been the object of recent research based on the application of Sensitivity Analysis to traffic simulations models and this is dealt with elsewhere.

From a practical point of view we can conclude that calibration and validation are essential to determine how reliable a simulation model is in reproducing the observed reality, and they are formally two complementary steps of a process, which depends on the data availability and the selection of the model parameters, in terms of their relevance for the simulation objective, and the adequacy of data to estimate their values. The first step, usually called calibration, assumes that the perception of the reality is provided the available data and tries to determine which model parameters can be adjusted to reproduce as closely as possible the observed data, in terms of a GoF measure. The measure and its levels of significance should be previously determined by the analyst in terms of his objectives. The second step, validation, does essentially the same exercise with an independent sample of data and quite frequently requires some further fine tuning of parameter values. A complementary aspect is the way in which the values of the parameters are estimated, the current trend being to use a simulation-optimization procedure.

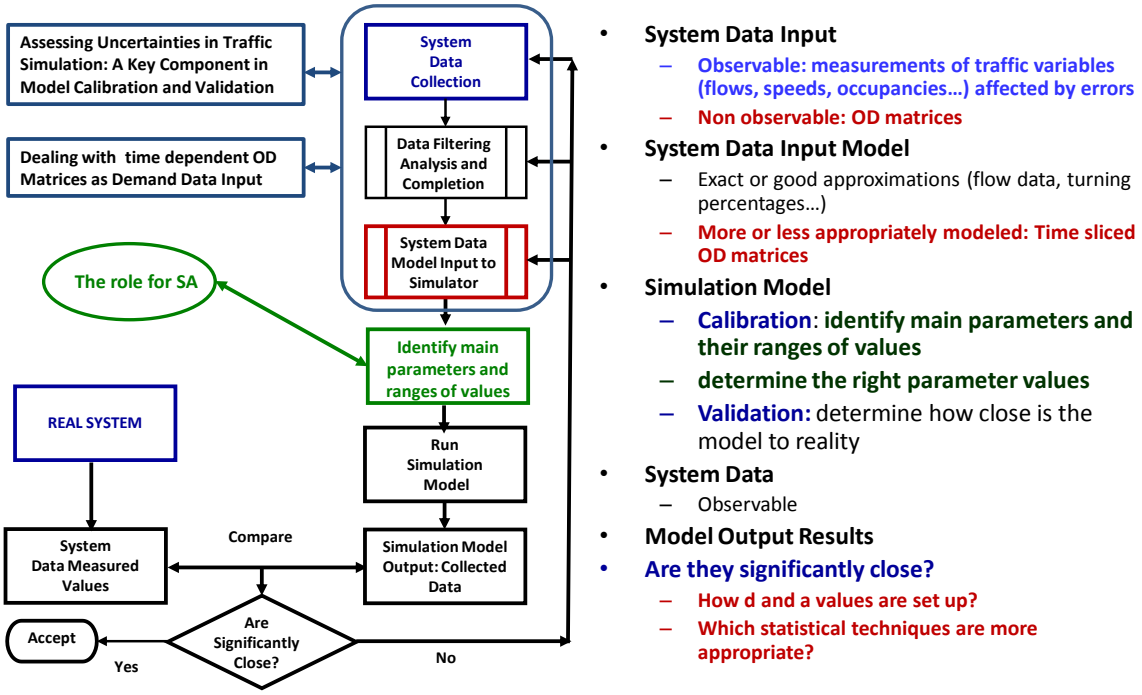

Figure 5: Conceptual methodological diagram for the calibration and validation of traffic simulation models

References

1. Pidd, M. (1992). Computer Simulation in Management Science. John Wiley.
2. Ni et al., (2004). A Systematic Approach for Validating Traffic Simulation Models. Presented at TRB 2004 Annual Meeting.
3. Hourdakis, J. et al., (2003). A Practical Procedure for Calibrating Microscopic Traffic Simulation Models. Presented at TRB 2003 Annual Meeting.
4. Ge, Q. and Menendez, M. (2013). An improved approach for the sensitivity analysis of computationally expensive microscopic traffic models: a case study of the Zurich network in VISSIM. Presented at TRB 2013 Annual Meeting.
5. Law, A. and Kelton, W. (1991). Simulation Modelling and Analysis. McGraw-Hill.
6. Standard Verification Process for Traffic Flow Simulation Model, Version 2, 2002, Traffic Simulation Committee, Japan Society of Traffic Engineers
7. Bayarri, M., Berger, J., Higdon, D., Kennedy, M. and Kottas, A. et.al. (2002). A Framework for Validation of Computer Models, NISS Technical Report Number 128.
8. Bayarri, M., Berger, J., Molina, G., Rouphail, N. and Sacks J. (2004). Assessing Uncertainties in Traffic Simulation: A Key Component in Model Calibration and Validation, Paper # 04-2504, 83rd TRB Meeting, Washington
9. Hollander, Y. and Liu, R. (2008). The principles of calibrating traffic microsimulation models. *Transportation*, **35**, 347-362.
10. PTV, VISSIM 5.40-03 User Manual. PTV Plannung Transport Verkher AG., Karlsruhe, 2012.

Issue 6 - Calibration methodologies

By Peter Wagner (DLR, DE) and Costas Antoniou (NTUA, GR)

Issue 5 has made clear that from a methodological point of view it is widely accepted that simulation is a useful technique to provide an experimental test bed to compare alternate system designs. To reach this goal, calibration and validation, as described in issue 5, is usually used to increase the likelihood that a model behaves as in reality, and that requires a good knowledge of the reality, that can only be reached by an exhaustive traffic data collection, to make sure that the analyst has the necessary input parameters and insight on driver behaviour to guide him/her in defining potential ranges of parameter values. Given a certain measure that depends on the application under consideration, test the model against a given data-set to see to which degree it fits the equation:

$$\text{Prob}(|\text{model} - \text{reality}| \leq d) > \alpha$$

Ideally, the values d and α have been determined beforehand by the analyst. So far, nothing has been said about the parameters of the model. If the models were models with physical processes behind them, they may have been measured independently by a completely different process unrelated to the current data-set under consideration. If this is not appropriate, and this is the normal case for many socio-demographic models including transport models, then one must determine these parameters by a process called calibration. The final goal of calibration therefore may be viewed as minimizing the difference between reality, as measured by a set of observed data, and the model results, described by another set of data that has been produced or constructed from the simulation model. This is done mostly by adapting the parameters of the simulation until some minimum (best fit) has been reached. While simple in principle, this application of this process is surprisingly complex with (arguably) four key stages.

Firstly, ensuring key input data is available and accurate is vital. This ranges from the digital road network data regarding the geometry and layout of the roads to the detailed working of any traffic control algorithms is needed, from simple speed advisory signs to adaptive traffic signal controllers. Note, that when it comes to data, distrust and care is needed. Even with digital road networks, the actual lane and sign layout of an intersection may be very different from the data, the same holds for other input data as well. The actual algorithms applied to in an intersection controller are not necessarily the ones in the documentation, and so on. In particular, getting the capacities right can be a tricky issue, because for a microsimulation model capacity is a function of external conditions (intersection layout, speed limits, interaction with buses, pedestrians and the like) as well as of the internal parameters of the simulation model (if applicable: reaction-time, preferred headway etc.).

Very often, traditionally calibration and validation have only addressed the supply component of the traffic system assuming that the demand was an already calibrated input. Next, demand needs to be calibrated, and this usually means in the case of dynamic traffic models (micro or meso) a time-dependent origin-destination trip matrix. This is still a critical question for which unfortunately none of the available professional software provides yet a reliable solution. Nevertheless, it has recently been the subject of an intense research work, see for instance (1 - 4). However, taking into account that the proposed approaches to estimate a time dependent OD matrix assume that a calibrated traffic model is available, the question as to whether demand and supply should be calibrated independently or in a common algorithmic framework is still open to debate.

After the network configuration and demand have been determined correctly, one may correct the parameters of the simulation model, such as the car-following and lane-changing parameters, or the distribution of the maximum speeds. Validation is the final step, and by using a data-set that has not been used for calibration, it can be checked that a model not only reproduces the data used for calibration, but also data that are different from those used for calibration. It should not come as a surprise, that the validation error is very often larger than the calibration error. Nevertheless, the simulation model's output should be within acceptable thresholds that have been determined in ideal case before the actual simulation has been run. It is important also to pay attention to over-fitting, i.e. calibrating too aggressively against the data at hand, in a way that does not describe the average traffic conditions. One indication of this is a good calibration fit, but a far worse validation fit. Validation is examined in detail in another sub-section.

In terms of the process of calibration itself there are a number of key elements identified by (for example) AP-R286/06 AUSTROADS (5) and FHWA Vol III (6), including:

- Identification of necessary model calibration targets, such as travel time, traffic flows, local speeds or even capacities. Different targets may need different sets of parameters. The choice of the targets to use depends on the application itself.
- Allocation of sufficient time and resources to achieve calibration targets. Sometimes, it may take longer than expected, however; this depends strongly on the experience of the user. Proper calibration of reasonable sized networks can require effort in the order of several person-months.
- Selection of the appropriate calibration parameter values to best match locally measured street, highway, freeway, and intersection capacities.
- Selection of the calibration parameter values that best reproduce current route choice.
- Try to limit calibration to a workable set of parameters. Unless there is a clear understanding of the contribution of each parameter, or very detailed data, then simply adding degrees of freedom many lead to a less stable calibration process. This points to a kind of sensitivity analysis, in order to select a reasonable set of calibration parameters, but of course it is sometimes difficult to know in advance which parameters are sensible and which ones are not.

If a study is following these recommendations seriously, then it will be most likely do the correct and sensitive things with respect to calibration. However, there are a number of other issues that need to be considered.

- Calibration methodologies may depend on the type of data that are considered: aggregate vs. disaggregate. Disaggregate data include detailed data on the driver behaviour, such as trajectory data, while aggregate data include flow counts or speeds, which are the manifestation of the behaviour of multiple drivers. Disaggregate data can be used to calibrate individual models, while aggregate data can be used to calibrate the entire traffic simulator at the same time.
- Another distinction is sequential vs. simultaneous calibration. In sequential calibration each model or time period may be calculated independently, while in simultaneous calibration all models and time intervals are calculated concurrently. Sequential calibration is more practical, as it corresponds to a smaller problem, with fewer parameters and fewer data to consider. On

the other hand, simultaneous estimation offers more efficient use of data, and it better captures the interactions between models and time periods.

Finally, one should briefly mention calibration algorithms in the context of systematic calibration. In principle, calibration is formulated as an optimization problem and therefore any suitable algorithm could be used. However, obtaining the gradient is usually problematic, as the simulator functions are not usually analytically differentiable, requiring the use of a numerical derivative or a heuristic. Considering that each function evaluation requires a run of the simulation software (which can take from a few seconds to hours, depending on the problem size and characteristics), directly obtaining these derivatives may be impractical. Therefore, more efficient algorithms such as the Simultaneous Perturbation Stochastic Approximation (SPSA) have been used (e.g. 7, 8). This optimization problem that minimizes an objective function, expressing the "distance" between an observable traffic variable and its simulated value, constrained by the set of feasible values of the model parameters on which the simulated variable depends, has also been solved by other optimization methods (9), while other researchers have used Genetic Algorithms (10, 11).

To conclude, from a methodological point of view, calibration procedures can be summarized as follows:

- The process starts by collecting the available data, it would be desirable that data collection was the consequence of a detailed design of which data to collect and which sampling procedures to use. In the case of network simulation, this should also include the demand data.
- Collected data are not usually directly input into the model, the modeller instead defines the input in terms of assumed probability distributions, the quality of the input should be assessed in terms of the quality of the assumed distributions to suitable fit the observed data, otherwise we risk to provide a flawed input to our model and this will strongly affect the quality of the whole process.
- The provision of the input data becomes an even more critical question when concerns time dependent origin destination matrices as discussed above.
- Next it would be desirable to identify the relevant model parameters and determine whether the available data allow to estimate their values or not, this could be the subject of a Sensitivity Analysis.
- Methodologically, as highlighted above the analyst should split the process in two independent phases conducted with independent data samples, one for the so called calibration, fixing the initial values of the selected parameters and the second one, the validation, to check for practical purposes the degree of validity of the model in reproducing the observed system.
- In both cases the process requires the determination of the degree of significance meaning that the model is acceptable and an optimization process finding the values best fitting a GoF defined in terms of the observed and simulated data.

References

1. Ashok, K. and Ben-Akiva, M. (2000). Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows. *Transp. Science,* **34**, 21-36.
2. Antoniou, C., Ben-Akiva, M. and Koutsopoulos, H. (2007). Non–linear Kalman Filtering Algorithms for On–line Calibration of Dynamic Traffic Assignment Models. *IEEE Transactions on Intelligent Transportation Systems*, **8**(4), 661 - 670.

3. Barcelò, J., Montero L., Bullejos, M., Serch O. and Carmona, C. (2013). A Kalman Filter Approach for Exploiting Bluetooth Traffic Data When Estimating Time-Dependent OD Matrices. *JITS Journal of Intelligent Transport Systems*, **17**(2), 1-19.
4. Barceló, J., Montero, L., Bullejos, M., Linares, P. and Serch, O. (2013). Robustness and Computational Efficiency of a Kalman Filter Estimator of Time-Dependent OD Matrices Exploiting ICT Traffic Measurements. Paper 13-3919, accepted for publication in TRR Transportation Research Records: Journal of the Transportation Research Board.
5. Austroads (2006). The use and application of microsimulation traffic models. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
6. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
7. Balakrishna, R., Ben-Akiva, M. and Koutsopoulos, H. (2007). Off-line Calibration of Dynamic Traffic Assignment: Simultaneous Demand and Supply Estimation. *Transportation Research Record*, **2003**, 50-58.
8. Ma, J., Dong, H. and Zhang, M., (2007). Calibration of Microsimulation with Heuristic Optimization Methods, *Transportation Research Record*, **1999**, 208-217.
9. Hourdakis, J., Michalopoulos, P. and Kottommannil, J. (2003). A Practical Procedure for Calibrating Microscopic Traffic Simulation Models. Presented at the 82nd TRB Annual meeting, Washington D.C.
10. Ma, T. and Abdulhai, B. (2002). A genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters. *Transportation Research Record*, **1806**, 6-15.
11. Kim, K. And Rilett, L. (2004). A genetic algorithm based approach to traffic micro-simulation calibration using ITS data. 83th TRB Annual Meeting.

Issue 7 - Differences according to scale and model purpose

By Pete Sykes (PS-TTRM, UK)

The use of modelling in transport planning covers a range of tasks varying widely in complexity and size, for example, microsimulation models may be focussed on optimisation of a single junction, they may be used to model highway corridors, or they model wide urban networks with many junctions and complex route choice. Model outputs may also be used in several different types of assessment. Outputs may be required to describe average speeds and flows, they may be required to quantify the congestion in the system measured by queue length or by journey time, or they may be used to give estimates of the environmental effect of transport proposals by quantifying the changes in vehicle emissions with an emphasis on aggregating the instantaneous emissions from each individual rather than aggregating flows and estimating emissions from those flows.

If a model is applied in an inappropriate manner, the results may not truly reflect the scenario to be examined. Poorly selected calibration criteria, such as average hourly flows in a model with short peaks of congestion may make calibration meaningless in that model context. Similarly, outputs may be reported at a level of aggregation that does not allow differentiation between schemes for example intra-zonal trips being omitted in a model with a coarse zoning scheme.

Model tool Selection

Microsimulation is just one of many transport modelling methodologies which range from look up table estimations, to agent based land use planning and travel demand models. The UK HA guidelines (1) comment that selection of a more sophisticated tool, such as microsimulation, is justified only when it significantly reduces the risk of the wrong decision being made and therefore that careful consideration is given to the choice of model type and how it is used. The main source document for advice in selection of a modelling method is the FHWA “Traffic Analysis Toolbox Volume II: Decision Support Methodology for Selecting Traffic Analysis Tools” (2). This is a comprehensive guide to selection of an appropriate method from “sketch planning” to application of microsimulation. The process it recommends analyses the planning task using seven criteria: 1) The geographic scope, 2) The transport facility types to be modelled i.e. urban roads, highways, ramps, , 3) The transport mode to be modelled, i.e. bus, freight, HOV … 4) The management strategy i.e. active signals, ITS … 5) The traveller responses such as route change, departure time change … 6) The performance measures required in the analysis, and finally, 7) The cost of applying a particular methodology.

Other guidelines use a very similar procedure with similar categories, albeit with different weightings. In particular, the Australian guidelines adopt a “strength and weakness” rather than an aggregate scoring system and get the same results in four of five examples (3). The anomaly is due to different scoring of the effect of flow interactions between closely spaced junctions. The Oregon DoT document (4) proposes a simpler procedure asking limited questions about the required outputs, the complexity of the situation to be modelled and the level of flow saturation.

In all guidelines, the most persuasive reasons for selecting microsimulation are where flow breakdown is imminent, where the situation is complex with junctions in close proximity, where there are active control measures to manage the network, or where the outputs required from the analysis contain detail such as variance between individuals, which cannot be otherwise obtained.

Many guidelines contain caveats about the cost of microsimulation which must be weighed against the requirements for it. No guidelines recommend microsimulation solely for its visual outputs, or that it should only be used in small area models.

Differences in application - Data

The differences in application of microsimulation depending on the scale and purpose of the model can be distinguished by three factors and the first of these is data. The FHWA makes the obvious statement that the data to build and calibrate the network must support the project objectives but offers no further guidance on what type of data is required to support any particular objective. The only area where any distinction in type of data required is made is in the demand data for a model where OD matrices may be used as an alternative to a method of balancing junction counts. Note though that the FHWA guidelines first present this in the context of the requirements of a particular software suite and only later mention that OD data is essential if route choice is to be modelled. The RMS/NSW guidelines (5) also discuss the same issue but simply ask for documentation of the chosen method.

Differences in application - Calibration

The FHWA guidelines describe the calibration process in three steps. (1) calibration for junction capacity, (2) calibration of route choice and (3) calibration for overall performance. The latter stage effectively implies iteration around steps 1 and 2, hence the FHWA advise it is used with caution. The FHWA guidance assumes that the OD demand is a known quantity derived from an independent source and is fit for the purpose of the project. The Australian Guidelines (3), and the software specific SIAS good practice guide (6) counter that assumption and widen the calibration process to include the OD matrix estimation. (It is interesting to note that the FHWA guidelines, the document most oriented to generic projects while being microsimulation specific, offers extensive guidance on capacity calibration of junctions including search algorithms to optimise calibration, yet it offers very little advice on route choice calibration stating that this is too software specific).

Calibration requirements identified in many documents refer to the UK WebTAG criteria of 85% of flows being within 5 GEH as one of the criteria which must be met to describe a model as "calibrated". This measure is commonly used and well suited to wide area assignment models where route choice is an important facet of the model, but of less interest in other contexts. For example, models that examine a single junction where the turn count is the sole determinant of demand on each link will inherently meet the GEH criteria. Similarly, in highway corridor models one of the key factors in determining the performance of the network is in the amount of merging and weaving between lanes, this will depend on the trip length as shorter trips between junctions have different lane use patterns to longer trips through the corridor. There may be many OD matrices that will satisfy the GEH turn count criteria but will not model the observed behaviour in the network.

Looking to the guidelines which are more focussed on specific project types, the UK HA document (1) does not refer to flow or turn data in calibration, instead it refers to matching observed attributes such as headway distribution, speed distribution and lane change behaviour as being more applicable in the case of highway corridors where there is little route choice and few junctions. Validation methods described in this document vary with model intent but include comparison with travel time data, queue length data and also with the count data (if route choice is used) which is more commonly used in model calibration. TfL's guidance on microsimulation (7) is more tightly

focussed on single junctions or small numbers of linked junctions and once again turn count criteria are less important in calibration. Here, the analyst is directed to use observed travel times, queue lengths and driver behaviour to calibrate the model. Validation is carried out using essentially the same class of data but with the addition of flows and, unique to the task of modelling signalised junctions, verifying that the pattern of calls to demand dependent signal stages is replicated.

Differences in application - Outputs

The discussion of outputs of simulation models mirrors that of the calibration criteria; the guidelines look first to the established measures. The UK HA guidelines and the SIAS Good Practice Guide refer to the WebTAG procedures for economic assessment using the same journey time data as is used to assess scheme options using traditional assignment models. In this case, while the modelling method may produce more accurate output, the nature of the output is the same. The US FHWA guidelines meanwhile adopt a two part approach to outputs with measures of overall system performance, such as vehicle miles travelled or average speed, allied to localised outputs based on congestion hotspot detection and on allowing the analyst to determine the reason for the congestion. The Australian guidelines (3) and the Oregon DoT guidelines (4), which are both in part derived from the FHWA document, mirror this approach and add more detail to the localised outputs referring to environmental outputs such as tail pipe emissions and quantifying the detailed merge and weave behaviour. However, other Australian guidelines from the RMS/NSW (5) refer only to Vehicle Hours Travelled as the overall system performance measure, to be applied to microsimulation models of all scales and complexity.

Overall there is little to guide the analyst in the selection of appropriate measures and the guidance may be summed up as advice to use the well-established aggregate measures for network performance and to select a measure of effectiveness for localised issues according to the particular needs of the decision makers.

References

1. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.
2. FHWA (2004). Traffic analysis toolbox volume II: decision support methodology for selecting traffic analysis tools. FHWA-HRT-04-039. Federal Highway Administration (FHWA), Washington, DC.
3. Austroads (2006). The use and application of microsimulation traffic models. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
4. Oregon Department of Transportation (ODoT), (2011). Protocol for VISSIM Simulation. Oregon Department of Transportation.
5. Road and Maritime Services, NSW, (2013). Traffic Modelling Guidelines. V1.0. Roads and Maritime Services, NSW, Australia.
6. SIAS Ltd (2006). The Microsimulation Consultancy Good Practice Guide. SIAS Ltd.
7. Transport for London (2010). Traffic Modelling Guidelines, TfL Traffic Manager and Network Performance Best Practice. V3.0. Transport for London, London, UK.

Issue 8 - Model specific issues

By Jaume Barcelo (UPC, ES) and Pete Sykes (PS-TTRM, UK)

In general, existing guidelines can be classified in three categories: guidelines from software developers, agency provided guidelines explicitly addressing microsimulation, and generic modelling agency guidelines independent of the modelling approach. Those in the first category are oriented towards the best use of their own software. They make specific reference to calibration, but tend to focus discussion on those aspects that they consider more relevant from the point of view of their algorithms and models. In general, the aim of these guidelines is to enable the user to optimally calibrate the model, given the respective software tool - it is more or less natural that guidelines provided by software developers is meant to be guidance to the user, rather than an instrument to evaluate (the boundaries of calibration guidelines and traditional user manuals become indistinct in this case). As a result, this type of guidance is, and has, to be the most application specific among the three types of guidelines. Agency derived microsimulation guidelines (FHWA, etc.) are more specific with respect to the general methodological aspects and some of them (e.g. FWHA) set up a basic generic methodology and data requirement for calibration. Finally generic agency guidelines which are independent of the modelling approach tend to propose indicators whose purpose is to define quite coarse measures to describe how close the models are to the observed reality. The GEH indicator is a relevant example.

However, frequently, software developer guidelines, as well as agency guidelines, with the objective of simplifying the process, either to adapt it to the specificities and/or utilities provided by the software, or reducing it to practical rules to ensure the acceptance of the results by public agencies, do not take explicitly into account that microsimulation, frequently requires more sophisticated procedures. In order to use the model as an experimental substitute for the actual system, the reliability of this decision-making process depends on the ability to produce a simulation model that represents the system's behaviour closely enough, as discussed in Issue 5.

For the proposed measures of quantification of the difference between models and reality guidelines should be model independent or, in other words, product agnostic. They should be defined by the modelling agencies and not necessarily by the software producers. There are two questions to be asked therefore, the first is - how relevant are these measures to the task of calibrating and validating a microsimulation model and how relevant are they to quantifying the differences between the different infrastructure design variants or road traffic management options to be tested. The second is - how can each software product provide these measures and how can the software producers assure the modelling agencies that their measures conform to their definitions.

This is an issue that affects both simulation and analytical models but, as the ubiquitous use of the GEH measure shows, the tendency is to carry practice over from one modelling paradigm to another and neglect the task of defining more relevant measures. The present microsimulation guidelines, with very few exceptions, continue to use the same calibration, validation and measures of effectiveness as have been habitually used with traditional aggregate assignment models and network wide measures and do not identify more detailed measures that would be used to quantify the performance of microsimulation models. One example (1), considers a useful method of model verification which could be used to form these microsimulation specific measures. It merges

compatibility between microsimulation methods and established analytical techniques used in traffic engineering. It considers some applications of traffic models; for example:

- the use of simulation for capacity analysis, including the dependence of capacity on demand flow rates;
- modelling of queue discharge (saturation) flow rate, queue discharge speed and other queue discharge parameters at signalized intersections, and relating them to the general queuing, acceleration and car following models used in microsimulation;
- modelling of gap-acceptance situations at all types of traffic facilities, and
- estimation of lane flows at intersection approaches, and relating this to lane changing models used in microsimulation.

An essential requirement for the method is to define traffic performance measures such as "delay", "queue length" and "stops" clearly and precisely. For example, delay could mean control delay, stop-line delay, and queuing delay or stopped delay. Furthermore, additional types of delay such as geometric delay, queue move-up delay and major stop-start delay could be identified for various purposes. Distinguishing between delay based on the queue sampling method vs. delay based on the path trace (instrumented car/probe vehicle) method is also important in oversaturated conditions especially when these are experienced within time slices employed in variable-demand modelling.

Similarly, queue length could mean back of queue, cycle-average queue, queue at the start of green period, or overflow queue. Measures of queuing could be the average or the percentile values of each of these differently defined queues. The mean of flagging when a vehicle is in a queued state is equally critical, with some definitions relying on a simple measure of headway and speed and others including hysteresis in these measurements to include vehicles moving within a queue.

All of the above indicator definitions are generic and would be recognised by traffic engineers in current practice albeit with the proviso that the definitions are loose and inconsistent between agencies and between software products. Bringing consistency to these definitions and measurement methods for traffic performance variables is essential to allow analysts to use different microsimulation models to make comparable decisions. It is the role of the agency based microsimulation guidelines to formalize these definitions, and the role of the controlling agencies to mandate their use in a manner similar to that by which the present network wide measures are required.

The role of the guidelines provided by software developers would then be to provide guidance on how their microsimulation packages are able to derive these performance measures. As derivation of some types of delay would be likely to involve some level of calculation beyond aggregating time in different modes of the driver behaviour model underlying the software, some software development could be required or a fall back option, as advocated in the US NCHRP-385 project (2), to derive these measures by post processing vehicle trajectory files may be used. The former option, in which the agency guidelines set the definitions of the performance variables and the software suppliers demonstrate their ability to produce these outputs would be the most rational approach. The key point is that the performance measures are derived from the definitions which originate from the controlling agencies and not derived from logging a time in a behavioural mode of a particular software package. The measurements should be independent of behavioural model and as

already mentioned, software agnostic and as such relieve the model user from the need to examine the detailed algorithms used by each software supplier to understand their specific performance measures.

An example of an agency setting the standards for capacity evaluation and then bringing the software tools in accordance can be observed at the moment in Germany where the Federal Highway Research Agency (BASt, 3) has initiated a project in which guidance has been developed how the common simulation tools (including AIMSUN, Paramics and VISSIM) are to be applied to generate exactly these traffic engineering values necessary to assess Level of Service compatible with the German HCM. Within this project, e.g. the method is defined how to determine capacity from the microsimulation so that it is in line with the methods used in the German HCM.

Finally as an example to illustrate the issues surrounding consistency of measurement, we can compare two global model indicators, the GEH statistic which uses an empirical formula based on a chi squared test to compare link and turn counts, and the Relative Gap Function, Rgap(t), which measures the progress towards the equilibrium in dynamic assignment, and therefore qualifies the goodness of the solution. Rgap (t) estimates, at time t, the relative difference between the total travel time actually experienced and the total travel time that would have been experienced if all vehicles had the travel time equal to the current shortest path (4, 5). However, the analyst should be aware that models in which the forecasted traffic flows are within the threshold of acceptance with respect to the measured flows under the GEH criteria – which is defined by the UK TAM to be 85% of counts within 5 GEH - may be simultaneously unacceptable from the Rgap(t) point of view. This would however very quickly lead to a discussion over which is more appropriate a measure long established in transport planning or a newer, less well understood, though potentially more relevant to a microsimulation project.

The inference for the writers of guidelines is that they must not only make precise and detailed definitions of measures of effectiveness but must also describe how they are to be used, how they relate to similar measures that may be derived from the model and which measures are most appropriate for each purpose.

References

1. Akçelik R. and Besley M. (2001). Microsimulation and Analytical Methods for Modelling Urban Traffic, Conference on Advance Modelling Techniques and Quality of Service in Highway Capacity Analysis, Truckee, California, USA.
2. NCHRP (2010). NCHRP Project 3-85: Guidance for the Use of Alternative Traffic Analysis Tools for Highway Capacity Analysis. Transportation Research Center University of Florida.
3. HBS-compatible simulation of the traffic flow on motorways (03.460). BASt, Germany. Last accessed 24/10/13.
4. Florian, M., Mahut, M. and Tremblay, N. (2001). A Hybrid Optimization-Mesoscopic Simulation Dynamic Traffic Assignment Model, Proceedings of the 2001 IEEE Intelligent Transport Systems Conference, Oakland, pp. 120-123.
5. Janson, B. (1991). Dynamic Assignment for Urban Road Networks, *Transpn. Res. B*, **25**(2/3), 143-161.

Issue 9 - What to do in the absence of appropriate data, potential 'fall-back strategies', transferability of data between calibrations etc

By Peter Wagner (DLR, DE).

We have learned within the MULTITUDE project that roughly two thirds of all studies are done without ever using any data (1), and there is even reason to suspect that this figure may be positively biased. Despite much advice and many recommendations, it seems then that microsimulation simulation studies are often done without data, and any calibration, or validation apart from a visual inspection. Unfortunately, the absence of data is a known and common problem. This is true not so much for network data (in these times of open streetmap and Google Earth, there is almost no excuse in not having such information) but much more on infrastructure data (traffic signals), demand data, and data that may help to test the driving parameters of the model to be used (loop data, queue-lengths, trajectories).

Lack of vehicle behaviour data must also be considered for some projects where the changes to be tested include the "rules of the road", examples include a possible change to the rules on passing on one side or both on a German autobahn, abolishing the (unique) New Zealand left turn priority rule[1] in 2012, or the slow introduction of roundabouts to drivers in the USA. In these cases there is no directly relevant driver behaviour data available related to that scenario in that driving regime and the analyst must judge whether parameters derived from other similar scenarios or from other countries can be appropriately applied.

Unfortunately no studies have addressed the issue of transferability of data directly, and strictly speaking, a calibration performed for one study cannot be transferred to another. While for demand data, this is obvious – the data of one region cannot be transferred to another one, for old data from the same region it is questionable but in principle possible. For behavioural data, the issue is still more involved because drivers in one region may behave differently than in another. For example, in the case of the UK HA guidelines (2), it is stated that behavioural parameters should not be changed without empirically justifiable reason. While some guidelines contain information on how to collect and prepare data (3, 4, 5) very little more is said. In a certain sense, MULTITUDE's focus on sensitivity studies is a first step into this direction, since it may give some hints, for example, if a situation is stable and robust with respect to behavioural parameters, then it can be expected that in a different situation the parameters will serve well, although there can be no guarantee.

No data needed

It is worth mentioning that there are some situations, where no data is actually needed. In a scientific endeavour that investigates the character of a new model, it is often enough to check for so called stylized facts (6), and there is no need for a detailed comparison to reality. Note however, that some discussions in the past could have been avoided if the researchers would have had good data at hand to perform a more serious comparison to reality, e.g. the discussion around the three-phase-theory of Kerner might serve as an example (7). This issue is still not resolved. (Three-phase theory expects that traffic flow can display three phases of traffic: the familiar free flow, jammed flow, and a third called synchronized, which is in between the two extremes).

[1] Where a vehicle turning right takes precedence over one turning left- rule abolished in 2012

Additionally, no data may be needed when one compares via simulation scenarios the outcome of different policies or different technologies. For instance, a new traffic light control algorithm is often evaluated at a real intersection, but it might even be better to compare it for different intersections and different demands than those observed at a single special intersection. In essence, one compares the outcome of different scenarios (for the same input data), but the data itself can be almost completely artificial (within reason).

Surrogates

Another issue to discuss is that of surrogate data. Again, this is often used in scientific papers, and there it is sometimes even a necessity, for example, testing and assessment of the quality of a method that estimates an OD-matrix from loop detector data needs surrogate data, because the real data are almost nowhere to be found at the quality that is needed. In this case, a simulation might be used to generate the data that can be used to test the method, in this case a simulated ground-truth is known against which to test the OD-estimator.

Another work-around for real data that may sometimes work might be to go for a thorough test of a large number of potential input values. E.g., for a given new intersection control system where the demand is not known, it might be possible to generate a large number of vectors of realistic demand inputs, and for all these it can be investigated whether the new control scheme is superior to the old one. If this is the case, then there is good reason to hope that it will be so even in reality. While such testing is presented here as a work around, it is also important for stability of the results, since it considers variation in the data, which, if dealt with correctly, adds a greater stability and reliability of the results. Such an approach may be used also in accounting for temporal variability such as time of day.

Averaging and transferability

While averaging of data and output as hinted at above can ensure some transferability (or understanding of transferability) it raises the question of why micro-simulation is actually being used. While it is tempting to say that such an approach obviates the need for micro modelling, this is not actually true. To highlight this by an almost classical example: if a microsimulation result has 1 hour free traffic at 110 km/h and 1 hour congested traffic at 20 km/h, then is the average is 65 km/h? That cannot be the case, but it demonstrates first that mean values are good only when accompanied by a small standard deviation. So, a microsimulation study even using aggregated results, gives a much more detailed account of what happens. Of course, it may depend on the question behind the study whether this advantage is important or not, however in most cases, it (again) adds reliability if the results can state an additional confidence level to a mean value.

Conversely, in trying to ensure models are transferable there is the temptation with many analysts to (defensibly) ‘tweak’ driver behaviour but at the expense of the ‘hard part’ of getting O-D matrices right or road geometry properly described. In general it is vital to get the situational description right before starting to change how drivers react. In addition, it might make sense to look harder at our models. For example, reaction times, gap acceptance, acceleration all vary at junctions but not arbitrarily. Consider a person familiar with the network who is about to go through a junction where they know there is a very short green time. They will expect the regular commuters to pre-empt the usual green time delay and the usual reaction time to the green light reduced to zero through anticipation of the change. Similarly on a roundabout, few gaps may be present so they may have to

accept a smaller gap and accelerate more at that location in the peak period than they would otherwise. This is not junction specific, it is congestion specific. The question is - are we over calibrating junctions (for example) to compensate for inadequate behaviour representation and then giving ourselves problems in transferring those calibration parameters? If that was the case then we would have a better case for transferability of parameters with better models. As it is we probably should look at the congestion levels to pick parameters but all too often we parameterise the location, not the situation.

References

1. Brackstone, M., Montanino, M., Daamen, W., Buisson, C. and Punzo, V. (2012). Use, Calibration and Validation of Traffic Simulation Models in Practice: Results of a Web based Survey. Proc. of the 90th Transportation Research Board Annual Meeting. Paper 12-2606. TRB, Washington, D.C. U.S.A.
2. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.
3. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
4. SIAS Ltd (2006). The Microsimulation Consultancy Good Practice Guide. SIAS Ltd.
5. Road and Maritime Services, NSW, (2013). Traffic Modelling Guidelines. V1.0. Roads and Maritime Services, NSW, Australia.
6. Treiber, M., Kesting, A. and Helbing, D. (2010). Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts.
7. Treiber, M. and Kesting, A. (2012). Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns. *Transportation Research Part C: Emerging Technologies*, **21,** pp 31-41. DOI: 10.1016/j.trc.2011.09.002

Issue 10 - Which indicators to use/not use, for calibration and validation assessment

By Costas Antoniou (NTUA, GR) and Vincenzo Punzo (JRC/UNINA, IT).

A large number of indicators are available for the assessment of calibration and validation efforts (e.g. 1). Each of these has different strengths and limitations and the question of which one to use very often arises. The answer to this is usually complex and may depend on a number of aspects, related to the problem, the available data (and their characteristics) and many other parameters. Usually, when validating a model, a reasonable approach is to consider multiple indicators, thus elucidating more aspects of the model and providing a more complete assessment of its performance. The remainder of this sections aims to provide some insight into the problem of selecting the appropriate indicators this issue.

Overview, guidance and caveats

A number of goodness-of-fit (GoF) measures can be used to evaluate the overall performance of simulation models and the reader is referred to Hollander and Liu (2), Ciuffo and Punzo (3) and Punzo et al. (4), which are summarized in the Appendix C. In principle, all these GoFs can be used both for calibration and validation, however, some of them can present disadvantages when used in calibration, and, although, in general it is difficult to distinguish between "good" and "bad" indicators some observations can be made concerning effectiveness, for example:

Non square indicators like Mean Error (ME) and Mean Normalised Error (MNE) cannot be used in the objective function of a calibration problem as errors with opposite sign compensate with each other (i.e. high errors of opposite sign yield low values of the indicators). In Ciuffo and Punzo (3), the number of iterations required by the Mean Average Error (MAE) turned out to be much higher than that for the other GoFs, which discourages the use in calibration of MAE as well. Normalized, or percent errors (Mean normalized error, MNE, Mean absolute normalized error MANE, Root Mean Square Normalised Error, RMSNE), are very tempting as they allow building multi-objective functions, e.g. on counts and speeds. However, their use in calibration is generally discouraged. In fact, the same absolute error between observed and simulated measurements (e.g. a difference in speed of 10 km/h) yields higher relative errors for small values of the measurement (e.g. 20 km/h) than for high values (e.g. 100km/h). As a result, the calibration is mainly driven by the lowest measurement values. Also Theil's inequality coefficient can be used to combine different measures of performance in a multi-objective function, without suffering from the above inconvenience. However, in the context of calibration of a car-following model, Punzo et al. (4) showed that a multi-objective Theil's statistic combining vehicle speed and spacing performed worse than the same statistic applied separately, to speed or spacing.

According to its original concept, the GEH statistic could be used in calibration by maximizing the number of occurrences for which the statistic is below a predefined threshold e.g. GEH>5. However, depending on the threshold, such an objective function can lead to multiple solutions and ineffective calibration (4). As an alternative, the sum of the GEH statistic for all measurements has been proposed in Zhang et al., (5).

Square statistics are preferable in calibration as, beside theoretical advantages, they penalize the highest errors and make the response function less smooth around the minimum. Therefore, in the

calibration of traffic flow models, results from previous studies and the above considerations suggest to use mainly the Root Mean Square Error (RMSE) and Mean Square Error (MSE) statistics. To lesser extent, the Theil's inequality coefficient U - applied to a single performance measure - and the GEH - only in the form of the sum of the GEH values for all the observation - are suggested. The problem of how to combine different measures in the same objective function (multi-objective optimization), however, has yet to be addressed satisfactorily. In any case, a good practice is to verify the efficacy of any of the chosen indicators for calibration by running a laboratory experiment with synthetic data on the problem at hand (4).

Different considerations arise when the same GoFs are used in model validation. In such a case, the GoFs advised against for calibration can actually become valuable. In addition, it is useful to use more statistics at the same time as they can capture different aspects of the obtained results. For example, Vaze et al. (6) and Balakrishna et al. (7) (in the context of a large-scale microscopic application) present a discussion on the benefits of using multiple statistics and their relative benefits. Considering the minimal cost of computing multiple goodness-of-fit measures, it is recommended to compute multiple such measures and report all of them. In most cases, the results will be similar, i.e. concur that a "better" calibration outcome can be easily identified. There are cases, however, in which one calibration measure might indicate lower overall error for case A (over case B), while another measure might indicate a larger bias for case A. For example, it is possible to have two different calibration results with the same RMSN or RMSPE results, leading the analyst to assume that both have a similar performance. If, however, the one approach has a higher MPE value, then that would result in a higher overall bias. Conversely, two calibration results might have the same MPE, giving the impression that the overall calibration results are equivalent, while one of the two might have a much higher RMSN or RMSPE value. To demonstrate this point, consider the synthetic example in the following figure. The black filled points correspond to synthetic observations of a specific measure (e.g. density). The two other datasets are synthetically computed, so that the one presented by blue triangles has higher variability around the mean values, but its errors are well balanced between over- and under-estimations of the "true" data, while the second, represented by the red squares has a smaller variability and smaller deviations, but its values are consistently lower than the "true" observations. Considering a measure such as the RMSN, which penalizes large deviations would suggest that the second set (red squares) of calibration results is preferable (with an RMSN of 5.7% vs 11.2% of the calibration result of the first dataset). If, on the other hand, a measure such as MPE is considered (which penalizes bias in the estimation) then one would conclude that the first data set (blue triangles) is superior (with MPE of -1.5% versus -4.8% for the first data set).

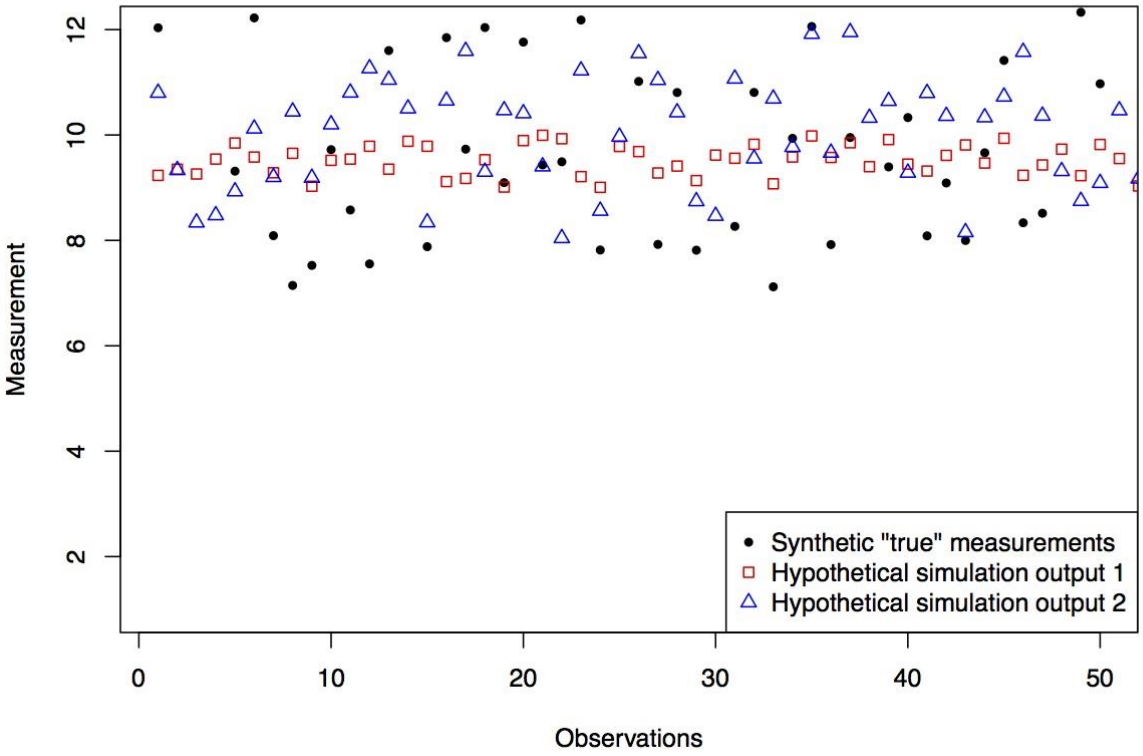

Figure 6. Synthetic data to demonstrate use of GoFs in validation

Selecting measures of performance (MoP)

The choice of suitable GoF measures is an important topic, but it is not the complete picture however, and another relevant question is which measures of performance should be used for the assessment of the calibration and validation quality. One obvious answer is that only available data can be used for this purpose. But when more than one type of data are available (e.g. flows, densities, speeds and travel times), which one(s) should be used?

One way to approach this question is to look at the available data and make a choice depending on their quality. For example, if flow counts are more reliable than the speed measurements, then perhaps the calibration should be guided by flow counts. If on the other hand (e.g.) flow counts are obtained from unreliable loop detectors, but point to point travel times are available from a tag-based automated vehicle identification system (which are very reliable, as they are used for tolling), then travel times would be more suitable.

Another approach would be driven by the time of application for which the model will be developed. For example, if the objective is to forecast flows, then perhaps link flows might be the measure of performance of choice. If, on the other hand, the objective of the model would be to generate guidance based on predicted travel times, then it might be more effective to base the calibration and validation process on speeds or travel times. So saying, in the same way that it is better to use more than one GoF indicator, it is arguably better to use more than one MoP. More information can provide better calibration and validation and help capture aspects of uncertainty that may not be evident by a single data source. This raises the problem of how to combine the multiple measures of

performance and one way is to consider the resulting GoF measures independently and make a choice. A more practical way could be to combine into multi-objective functions more measures. This, however, did not yet achieve satisfactory results in the field literature as mentioned before.

Existing guidelines

The topic of GoF measures is mentioned in general terms in many guidelines (e.g. 8, 9), most of which can be traced back to guidelines in the UK, which suggest to use the GEH and usually mention that the model outputs should be within 5% of the observed values.

On page 18 of the Austroads Guidelines (10), it is stated that some target values are necessary for the proper calibration of a microsimulation model and it is further stated that the target values for calibration (Table 3.3) are for guidance only and represent current practices of RTA NSW and those recommended in the FHWA Toolkit (11). The values emphasize that there is no need to achieve 100% agreement between all model and observed outputs and the recommended GoF measures include GEH and RMSE. TfL (12) adopts a similar philosophy with five model outputs needing to be within 5% of observed values, also stating that modellers should use the GEH parameter to demonstrate that traffic flows within the model (i.e. internal mid-links, stop lines etc.) match traffic counts to an acceptable level of accuracy (and again references GEH <5). However, it is stated that TfL Traffic Directorate advocates GEH values of less than three for all important/critical links within the model area.

While, the FHWA toolkit (11) adopts the Wisconsin DOT freeway model calibration criteria (13), which in turn recommend the use of GEH (based on guidelines from the UK), while Dowling et al. (14, Section 6.6) recommend that the analyst seek to minimize the MSE between the model estimates of maximum achievable flow rates and the field measurements of capacity. The MSE is the sum of the squared errors averaged over several model run repetitions with each set of repetitions containing a single set of model parameter values (p) with different random number seeds for each repetition within the set. However, in the next section (6.7) a reference is made on the Wisconsin DOT freeway model calibration criteria, which use GEH. ODoT VISSIM Protocol mentions that “the best universal measure to compare simulation inputs and outputs is the GEH formula. Calibration of the model should use the GEH formula, calculated to a value of 5 or lower”. Lastly, NSW/RTA (15) reference GEH and RMSE: “Useful measures of “goodness-of-fit” generally used to compare model flows against observed counts are GEH and RMSE.”

References

1. Toledo, T., Koutsopoulos, H. and Davol, A., et. al. (2003). Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study. *Trans. Res. Rec.*, **1831**, 65-75.
2. Hollander, Y. and Liu, R., (2008). The principles of calibrating traffic microsimulation models. *Transportation*, **35**, 347-362.
3. Ciuffo, B. and Punzo, V. (2010). Verification of Traffic Micro-simulation Model Calibration Procedures: Analysis of Goodness-of-Fit Measures. Proceeding of the 89th Annual Meeting of the Transportation Research Record, Washington, D.C.
4. Punzo, V., Ciuffo, B. and Montanino, M. (2012). Can Results of Car-Following Model Calibration Based on Trajectory Data Be Trusted? *Transp. Res. Record*, **2315**, 11-24.
5. Zhang, M., Ma, J. and Dong, H. (2008). Developing Calibration Tools for Microscopic Traffic Simulation Final Report Part II: Calibration Framework and Calibration of Local/Global Driving

Behaviour and Departure/Route Choice Model Parameters. UCB-ITS-PRR-2008-7. University of California, Davis.

6. Vaze, V. and Antoniou, C., et. al. (2009). Calibration of Dynamic Traffic Assignment Models with Point-to-Point Traffic Surveillance. *Transp. Res. Rec.* **2090**, 1-9.
7. Balakrishna, R. and Antoniou, C., et. al. (2007). Calibration of Microscopic Traffic Simulation Models: Methods and Application. *Transp. Res. Record*, **1999**, 198-207.
8. Zhang, M. and Ma, J. (2008). Developing Calibration Tools for Microscopic Traffic Simulation Final Report Part I: Overview Methods and Guidelines on Project Scoping and Data Collection. UCB-ITS-PWP-2008-3. California PATH Working Paper, University of California, Davis.
9. Zhang, M., Ma, J., Singh, S. P. and Chu, L. (2008). Developing Calibration Tools for Microscopic Traffic Simulation Final Report Part III: Global Calibration - O-D Estimation, Traffic Signal Enhancements and a Case Study. UCB-ITS-PRR-2008-8. University of California, Davis.
10. Austroads (2006). The use and application of microsimulation traffic models. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
11. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
12. Transport for London (2010). Traffic Modelling Guidelines, TfL Traffic Manager and Network Performance Best Practice. V3.0. Transport for London, London, UK.
13. Wisconsin Department of Transport. (WDoT). Wisconsin DoT Microsimulation Guidelines.
14. Dowling Associates (2002). California Department of Transportation Guidelines for Applying Traffic Microsimulation Modelling Software. Dowling Associates.
15. Road and Maritime Services, NSW, (2013). Traffic Modelling Guidelines. V1.0. Roads and Maritime Services, NSW, Australia.

Issue 11 - What data to use for validation and when to perform it

By Tomer Toledo (Technion, IL)

Model validation is the process of checking to what extent the model replicates reality, and in particular the effects of changes in the system and its inputs. It is closely related to the model calibration task and ideally, the two tasks should take place before each new application. The validation importance (and difference from calibration) is that it provides the modeler confidence that the responses to changes in the transportation system that are observed in the simulation model are representative of those that would have been found in the real system. Thus, a model may be calibrated so that it replicates current traffic measurements, but, only a valid model will be able to accurately predict the effects of changes in the current system.

Most current simulation guidelines briefly define validation, but do not give it the level of attention and detail that they do in discussing calibration procedures. They commonly state that it should be done with data that was collected independently of the data used for calibration. The Australian Austroads guidance (1) states that "It is common to collect sufficient input data such that a portion of the input data is for calibration and the rest is for validation". The FHWA toolkit (2) suggests using measures of system performance such as travel times, speeds, delays and queues for validation while the UK HA guidelines (3) argue that the data used for validation should be related to the type of application, the scale of the model and the available data. Some of the guidelines (in particular those from the UK and Australia) propose specific measures of performance for the validation and thresholds and confidence intervals for their evaluation.

Figure 7 shows a conceptual framework for the calibration and validation tasks. It consists of two phases: Initially, the individual behavioral models that make up the traffic simulation model (e.g. driving behavior and route choice models) are estimated using disaggregate data, detailed driver behavior information such as vehicle trajectories for example. These individual models may be validated independently, for example, using a holdout sample. The disaggregate analysis is performed within statistical software and does not involve the use of a simulation model. The level of effort required to collect and analyze trajectory data and the limited access to modify the models implemented within traffic simulators dictate that this step is most often only performed by the model developers. They then provide users with default values for the model parameters. In the second phase, which is the one that the various guidelines discuss, the simulation model as a whole is calibrated and then validated. This phase involves fine-tuning for the specific application and site being studied of the previously estimated model parameters and setting additional parameters that may have not been previously estimated. This phase is commonly done using aggregate data (e.g. flows, speeds, occupancies). Compared to the disaggregate data used in the first phase, these data are often readily available or can be easily and cheaply collected.

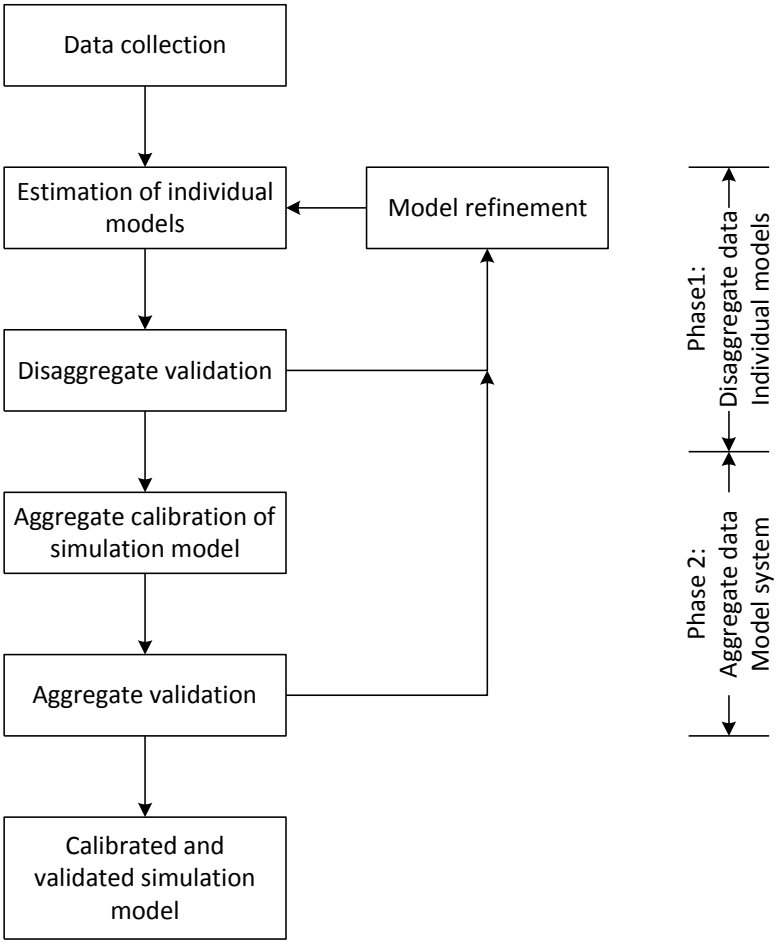

Figure 7 Overall calibration and validation framework

Ideally, the aggregate validation in the second phase should be done for every application. Figure 8 shows a conceptual framework for the aggregate validation task. The validation should be based on comparing outputs that were generated by feeding the real and simulated systems with identical inputs. Therefore, the real system should be observed not only for its outputs but also for its input variables, which are then used in the simulation study. This practice reduces the variance of the differences between observed and simulated outputs and therefore increases the efficiency of the comparison.

With respect to the input modeling, the most relevant input in traffic simulation models is often the travel demand. It is commonly given in the form of dynamic origin to destination (OD) matrices. However, in most applications OD matrices are not observed directly, and so OD flows must be estimated. There are well established methods for OD estimation. These traditionally utilize count data from loop detectors and other sensor locations. Recently, new methods that use other types of data, such as measurements of speeds and densities at sensor locations, point-to-point travel time measurements and locations matching.

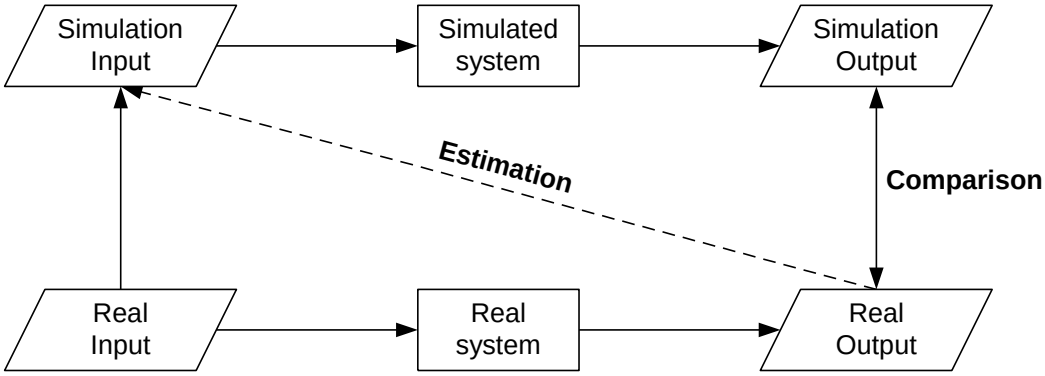

Figure 8 The aggregate validation process

With respect to the outputs from the real system and the simulation model to be used in the aggregate validation, the modeler has a wider choice of aggregate statistics and data, such as flows, speeds, densities, occupancies, travel times and delays. As noted above, the model validation is based on the similarity between the simulated and observed measures of performance (MoP). The following criteria are useful to assist in selecting MoPs:

- Context of the application. MoPs should be statistics that are important in the intended study. For example, point-to-point travel times are useful MOPs for validation when a traveller information system is to be evaluated on the basis of travel time savings. However, if a sensor-based incident detection system is studied MoPs extracted from the sensors (e.g. occupancies, flows, speeds) may be more useful. In this context a site-visit or discussion with the clients may be useful to help decide which the most relevant MoPs are.
- Independence. MoPs used for validation should be independent of any measurements used for calibration or to estimate inputs to the simulated system and differ from them in some substantial way, which would be meaningful for the intended application. For example, if the simulation study intends to evaluate various traffic control plans at intersections, the validation should address the ability of the model to replicate traffic patterns in the intersection under different control conditions (e.g. different timing plans applied sat different times of the day). Note that OD flows are commonly estimated by minimizing a measure of the discrepancy between observed and simulated traffic counts. Therefore, validation of the simulation model (only) against traffic counts may lead to overestimating the realism of the model.
- Error sources. In traffic analysis the discrepancy between observed and simulated outputs can be explained by four sources of error: Travel demand; Route choice; Driving behaviour; and Measurement errors in the observed outputs. The first three sources contribute errors to the simulated output while the last represents errors in the observed output relative to the "true" output. In most cases, the contributions of the three simulation error sources are confounded and cannot be isolated in the validation. The locations and types of MoPs to be collected should be chosen to reflect errors from all these sources and reduce the effect of measurement errors as much as possible. Measurement locations should be chosen to provide spatial coverage of all parts of the network. Moreover, measurements close to the network entry points will mostly reveal errors in the OD flows with little effect of the route

choice and driving behaviour models. As many measurement points as possible should be used.

- Traffic dynamics. MoPs and the level of temporal aggregation at which they are calculated (e.g. 15 minutes, 30 minutes) should be chosen such that they facilitate testing whether or not the model correctly captures the traffic dynamics. This is especially true in network applications where both the temporal and the spatial aspects of traffic are important.
- Level of effort required for data collection. In many cases this is the most constraining factor in practice. Point measurements (e.g. flows, speeds and occupancies) are often readily and cheaply available from the surveillance system. Other types of measurements (e.g. travel times, queue lengths and delays) are more expensive to collect. It is also important to note that data definitions and processing are not standardized. For example, statistics such as queue lengths may be defined in different ways, and surveillance systems may apply various time-smoothing techniques. It is therefore necessary to ensure that the simulated data is defined and processed the same way as the observed data.

References

1. Austroads (2006). The use and application of microsimulation traffic models. Austroads Research Report AP-R286/06. Austroads, Australia. ISBN 1 921139 34 X.
2. FHWA (2004). Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modelling software. FHWA-FRT-04-040. Federal Highway Administration (FHWA), Washington, DC.
3. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.

Issue 12 - Reference materials

By Pete Sykes (PS-TTRM, UK).

As microsimulation entered the marketplace in the late 1990s, the style of publication about road traffic simulation moved from academic research into fundamental algorithms and data structures and more application oriented materials started to appear describing the use of microsimulation on actual projects. The target audience for such project case notes was the body of transport planners who were familiar with the traditional assignment models and, in a largely conservative profession, were cautious in their acceptance of a new modelling paradigm. Reference materials were, in general, descriptive of what could be done. Later, as microsimulation gained wider acceptance, reference materials emerged in modelling guidelines to describe not what could be done, but to describe how it should be done with the intention that, by illustrating best practice through describing cases that analysts could readily relate to their own projects, better models would be built. Note that here, the term "reference materials" is here is used to describe a narrative description of a project, not a set of parameter values, prescribed option lists or approved techniques.

Marketing

Reference materials from both consultancy and software companies are usually closely related to their marketing material. Each software supplier produces a steady stream of project case notes, conference presentations, magazine articles and, to a lesser extent, academic papers which illustrate the capabilities of their software, and where their software is used. Their intention is to give their customers a feeling of safety in numbers of users and to promote the view that if a particular problem can be solved by one customer in one location with their software, similar problems can be solved by other users in other locations using the same software. Suppliers will naturally focus their materials on the perceived strengths of their own software, and on innovation unique to themselves. Consultants will focus on their skills in use of the software and their specialist knowledge However, both will inevitably focus on projects which are successful and have high net worth, projects which foundered on the limitations of microsimulation modelling will not appear in marketing materials. A strong bias towards success in the selection of materials is only to be expected.

Guidelines

The simulation guidelines make use of reference cases to illustrate the application of their prescriptions. The strength in these case studies is in their independence from any one supplier or user but their weakness is in the nature of the guidelines which will always lag behind the innovative edge, in a domain that is still developing new software technology, indeed strict adherence to guidelines may even stifle novel solutions and innovation.

Two examples of note are the VDoT (Virginia Department of Transportation) Calibration Guidelines (1, 2) which use a set of four real applications (an isolated interchange, an urban network, a simple highway segment and a highway merge weave section) to illustrate the calibration procedures using two software products. Similarly, the UK HA Microsimulation Guidelines Technical Review (3) modelled three highway situations (a congested merge weave section, a motorway link with gradient, and a signalised motorway junction) each with four software products to experiment with the different products and to produce advice on application of each product to each situation.

Consistency

The UK HA report emphasised the validation of the models in each system and described the parameter adjustments required to calibrate them, with considerable effort made to validate the recommended ranges of the calibration parameters in the software algorithms against observed data, and to advise on the features of different packages that should or should not be used. In effect in producing these guidelines, which are heavily based in these example projects, three exemplar models were produced with four different packages. Some comparison tests were run with these models to examine sensitivity to common parameters across all software packages, but as parameters, which are outwardly similar, can have different definitions within software algorithms; few complete cross product comparisons could be made.

The only example where a valid comparison could be made was in the differences between the predictions of different products to the effect of an exogenous and consistently described change, i.e. an increase in demand. Here significant differences in the increase in journey time as vehicle demand was increased were evident. However, investigating these differences, which were surmised to be either related to behaviour algorithms or to calibration choices, was beyond the remit of these guidelines.

Post evaluation reports

To address the issue of maintaining case studies which are topical, unbiased, and also present in sufficient numbers to cover a wide range of applications, in 2006 the Australia Roads Authority ARRB added a project repository facility as an adjunct to its microsimulation guidelines (4). Each entry in the repository presents a brief explanation on the purpose of the model, a project description and some general conclusions from the model application, i.e., results of investigating different scenarios, sensitivity tests undertaken, extent of the variation from default parameters, difficulties encountered and ways to overcome modelling issues, and comments on the general robustness of model outputs. To progress towards this goal in 2008 and recognising that information dissemination for the microsimulation community could benefit from an electronic facility for knowledge sharing, ARRB funded the creation of a Microsimulation Hub which has now developed into a technical note and case study store (4). However, as no new microsimulation projects have been documented since the original 2006 guidelines were written, it may be surmised that there is little incentive within the industry or the agency to populate the repository.

The ARRB exercise does offer a complete template for a case study report providing for both “soft” information on modelling issues and “hard” detail on calibration. The inclusion of detailed technical information on the parameters used in calibration to build a shared knowledge base of suitable values is a sentiment that has occurred during stakeholder meeting, namely, that that one purpose of guidelines is to advise users in detail on what modelling choices have worked elsewhere. The use of standardised reporting of reference cases would, assuming a reasonable sized number of cases could be collated, offer a form of crowd-sourced reference values.

Reference cases that only report on the modelling exercise do however omit one of the main questions to be asked of any modelling project; did it predict the behaviour that was subsequently observed on the road network? To answer this, the UK Highways Agency produces two POPE (Post Opening Project Evaluation) reports one year and five years after a project that they have funded is opened. In 2012, 46 POPE reports were written to evaluate local schemes (5) and 77 major scheme

reports were available (6). While these cover all forms of transport models from demand forecasting to microsimulation modelling, this constitutes a large resource which is available to examine the accuracy of the predictions made by microsimulation models and one that does not suffer from selection bias by either software suppler or consultant, although it inevitably can only contain those projects where the model reported a positive assessment in favour of the scheme.

References

1. VTRC (2006). Microscopic Simulation Model Calibration and Validation Handbook. Virginia Transportation Research Council Technical Report VTRC 07-CR6, Traffic Operations Laboratory, Center for Transportation Studies, University of Virginia, Park, B. and Won, J.
2. VTRC (2006). Simulation Model Calibration and Validation: Phase II: Development of Implementation Handbook and Short Course. Virginia Transportation Research Council Technical Report VTRC 07-CR5, Traffic Operations Laboratory, Center for Transportation Studies, University of Virginia. Park, B. and Won, J.
3. Highways Agency (2007). Guidelines for Microscopic Simulation Modelling. Vaughn, B. and McGregor, A.
4. ARRB (2013). Micro simulation resources. Last accessed 21/8/13.
5. UK Highways Agency. Post Opening Project Evaluation (POPE) of Local Network Management Schemes (LNMS). Last accessed 29/10/13.
6. UK Highways Agency. Major Projects POPE Reports. Last accessed 29/10/13.

Appendix B: Overview of main sensitivity analysis (SA) techniques

By Vincenzo Punzo (JRC/UNINA, IT) and Biagio Ciuffo (JRC, IT).

In this Annex we review some of the most common techniques for SA: i) One-At-a-Time sensitivity analysis, ii) input/output scatter-plots, iii) Elementary Effect test, iv) Sigma-normalized derivatives, v) Partial Correlation Coefficient analysis, vi) Standardized Regression Coefficient analysis, vii) Monte Carlo filtering, viii) Meta-modelling, ix) Factorial Analysis Of VAriance, and xi) Variance-based method based on the Sobol decomposition of variance. In the following some elements of each technique are presented.

One-At-a-Time (OAT) sensitivity analysis

In the OAT sensitivity analysis, one studies the variation in the model outputs due to the variation of one input parameter at a time, while the others are kept fixed to certain values. The difference in the model output due to the change in the input variable is referred to as the sensitivity or swing weight of the model to that particular input variable (Morgan and Henrion, 1). However, this approach is "illicit and unjustified unless the model under analysis is proved to be linear" (Saltelli et al., 2). It cannot detect interactions between input parameters.

Input/output scatter-plots

Let the model considered be in the form

$$Y = f(Z_1, Z_2, \dots, Z_r) \tag{1}$$

being Z_i ($i:1,\dots,r$) the model's input and Y its output. Let's perform a Monte Carlo experiment with our model. This means that, given the statistical distribution of the model inputs, we sample N possible combination of them in order to achieve the following matrix:

$$\mathbf{M} = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} & \cdots & z_r^{(1)} \\ z_1^{(2)} & z_2^{(2)} & \cdots & z_r^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ z_1^{(N)} & z_2^{(N)} & \cdots & z_r^{(N)} \end{bmatrix} \tag{2}$$

Computing Y per each row of the matrix in equation 5.2 we obtain the vector of model outputs **Y**.

$$\mathbf{Y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(N)} \end{bmatrix} \tag{3}$$

If we now plot the elements of **Y** against the correspondent elements of each column of **M**, we obtain r scatter-plots. From the visual analysis of the different scatter-plots it is possible to identify those parameters which have an influence on the model outputs and those parameters which do not. For the parameters able to influence the model outputs, the cloud of points of the scatter plot will have a more or less defined shape. For the others it will approximately resemble a rectangle. In this way it represents the simplest way to perform sensitivity analysis. The problem is that increasing the variables number, this method becomes unpractical. In addition it does not allow for the sensitivity of group of variables to be investigated.

Elementary Effects Test

The Elementary Effects method can be considered an extension of the OAT technique as it is able to overcome its main shortcoming (the weak coverage of the input space). It basically consists of an average of derivatives over the space of inputs. If the *r* input variables vary across *p* levels, the elementary effect of the i-th input variable at the level *j* is given by:

$$EE_{i,j} = \frac{Y(Z_1, \dots, Z_i + \Delta_j, \dots, Z_r) - Y(Z_1, \dots, Z_i, \dots, Z_r)}{\Delta_j} \tag{4}$$

In which Δ_j is the width of the level *j*. The sensitivity index for the i-th variable is then evaluated by the following:

$$\mu_i = \frac{1}{p}\sum_{j=1}^{p} |EE_{i,j}| \tag{5}$$

which allows for the variables to be ranked. In this way, it can be considered as a screening method, to be preferably used before the application of a more sophisticated method in order to reduce the number of input variables to consider.

Sigma-normalized derivatives

Function's derivatives seem to be the most natural way to perform sensitivity analysis, especially for analytical models. In reality, derivatives are not always suitable (3) for this aim. In their place, sigma normalized derivatives are used instead. Considering the previous example, the formulation for sigma-normalized derivatives is the following:

$$S_{Z_i}^{\sigma} = \frac{\sigma_{Z_i} \partial Y}{\sigma_Y \partial Z_i} \tag{6}$$

in which $S_{Z_i}^{\sigma}$ represents the sensitivity index for the variable Z_i and σ the standard deviation. It is worth noting that, sensitivity index as in equation (6) is recommended for sensitivity analysis by the Intergovernmental Panel for Climate Change (IPPC).

The main shortcoming of this approach is for the application with black-box models (i.e. simulation-based). In this case the derivatives' computation can be very expensive in terms of time. For this reason, they are usually evaluated only in the middle of the distribution of the single variables and some hypotheses on the function are made to extrapolate results obtained to the entire function. When the hypotheses result false, the results achieved may be misleading.

Standardized Regression Coefficient (SRC) analysis

Another possibility for black-box models is to create a regression model on the basis of the evaluations of the function. If we consider again elements of equation (2) and (3), a linear regression model can be written in the form:

$$y^{(j)} = b_0 + \sum_{i=1}^{r} b_{Z_i} Z_i^{(j)} \tag{7}$$

in which b_{Z_i} are the coefficients of the regression model. Normalizing these coefficients with the standard deviations of input and output, we obtain the sensitivity index

$$\hat{\beta}_{Z_i} = \hat{b}_{Z_i} \frac{\sigma_{Z_i}}{\sigma_Y} \tag{8}$$

For linear models the sensitivity index in equation (8) coincides with that of equation (6). This holds only in this case. In general, standardized regression coefficients are more robust and reliable than sigma-normalized derivatives, as they result from the exploration of the entire space of the input variables. However, their precision depends on the size *N* of the Monte Carlo experiment.

Partial Correlation Coefficient (PCC) analysis

A simple method for assessment of the relationship between independent and dependent variables is the calculation of the correlation coefficient for the values of input parameters and the output. The most frequently used method for linear correlation is the Pearson product moment correlation coefficient (R), which is expressed as:

$$R = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2}} \tag{9}$$

Where, x_i is an observation of the model input, y_i is an observation of the model output, and *n* is the number of observations of the model inputs/outputs.

The drawback of this simple correlation function is that a strong correlation between input parameters may influence input–output correlations. The partial correlation coefficient is calculated to represent the linear relationship between two variables with a correction to remove the linear effects of all other variables. For example, given a dependent variable Y with two independent variables, X_1 and X_2, the partial correlation coefficient between X_1 and Y has eliminated the possible indirect correlations between X_2 and Y and (X_1, X_2) and Y. PCC can be defined as follows:

$$R_{X_1 Y|X_2} = \frac{R_{X_1 Y} - R_{X_1 X_2} R_{X_2 Y}}{\sqrt{\left(1 - {R_{X_1 X_2}}^2\right)\left(1 - {R_{X_2 Y}}^2\right)}} \tag{10}$$

where $R_{X_1 Y|X_2}$ represents the PCC for X_1 and Y while removing the effects of X2.

Monte Carlo filtering

When one is not interested in studying the specific value of Y, but if Y is above or below a certain threshold (that is to say if Y creates or not a certain effect), a Monte Carlo filtering can be used. Indeed, using a Monte Carlo setting to produce matrix and vector of equations (2) and (3), and then applying the filter of interest to the values of Y, it is possible to divide the matrix M in two groups, one for the variables' values producing one effect and the other for those which do not produce it. At this point a statistical test can be carried out to check whether each of the inputs is statistically responsible for the effect to be produced.

Meta-modelling

A possible way to perform sensitivity analysis of complex black-box models is to use a meta-model able to approximate the output of the model itself. In this way the time required is used to create the meta-model, while the analysis can be then easily performed using its analytical formulation. This topic is attracting the interest of researchers. More information can be found in Chapter 5 of Saltelli et al. (3).

ANalysis Of VAriance (ANOVA)

The analysis of variance is a model independent probabilistic sensitivity analysis method used for determining whether there is a statistical association between an output and one or more inputs. ANOVA differs from regression analysis in that no assumption is needed regarding the functional form of relationships between inputs and outputs.

In ANOVA, model inputs are referred to as factors and their values are referred to as factor levels. An output, instead, is referred to as a response variable. Multifactor ANOVA studies the effect of two or more factors on the response variable and it is used to determine both the first-order and the interaction effect between factors and the response variable.

To apply this technique a number of evaluations of the responses against different values of the input parameters are required. From a statistical point of view, an appropriate way of performing these evaluations is defined by the experimental design techniques. In particular, a full factorial design can be properly applied in this case. In such a way, the full factorial experimental plan consists of *n*k* model evaluations, where *k* is the number of factors and *n* is the number of levels. The use of this experimental plan can also determine whether the factors interact with each other, that is, whether the effect of one factor depends on the levels of the others. The results that can be obtained from the ANOVA are twofold. First, they give an estimation of the model output variance explained by each parameter or by their combination. On the basis of this result, it is then possible to use a Fisher probability distribution to test the null hypothesis that the variance explained by a single parameter is negligible with respect to the whole model, that is, that the model is not sensitive (with a well-defined level of significance) to parameter changes.

Variance-based methods based on the Sobol decomposition of variance

The variance-based method based on the Sobol decomposition of variance is considered the most powerful method for SA. As reported in Table 1, this method unfortunately requires a significantly high number of model evaluations and therefore it is not tailored for expensive models with too many input factors.

A brief description of the method follows. Let us consider again the model of equation (1). We want to see what happens to the uncertainty of Y if we fix one of the input variables Z_i to a specific value z_i^*. The resulting variance of Y, that we call conditional variance, will be $V_{\mathbf{Z}_{\sim i}}(Y|Z_i = z_i^*)$. In which the symbolism in $\mathbf{Z}_{\sim i}$ means that we are considering the variance across all the variables but the *i*-th. It is expected that the higher the influence of the variable Z_i, the lower the conditional variance. For this reason the conditional variance can be considered as an index of the sensitivity for Z_i. The problem with this formulation is that the sensitivity index would depend on the specific value z_i^*

considered. For this reason, we consider the average of this measure over all possible points z_i^*, $E_{Z_i}\left(V_{\mathbf{Z}_{\sim i}}(Y|Z_i)\right)$.

Furthermore it is known that

$$V(Y) = E_{Z_i}\left(V_{\mathbf{Z}_{\sim i}}(Y|Z_i)\right) + V_{Z_i}\left(E_{\mathbf{Z}_{\sim i}}(Y|Z_i)\right) \tag{11}$$

Equation (11) shows that for Z_i to be an important factor we need that $E_{Z_i}\left(V_{\mathbf{Z}_{\sim i}}(Y|Z_i)\right)$ is small, that it is to say that the closer $V_{Z_i}\left(E_{\mathbf{Z}_{\sim i}}(Y|Z_i)\right)$ to the unconditional variance $V(Y)$, the higher the influence of Z_i. Thus the first order sensitivity index of Z_i with respect to *Y* is defined as:

$$S_i = \frac{V_{Z_i}\left(E_{\mathbf{Z}_{\sim i}}(Y|Z_i)\right)}{V(Y)} \tag{12}$$

First order sensitivity index is a very important measure to understand how much the correct definition of an input to the model may reduce the overall variance of results. From equations (11) and (12) we have $S_i \leq 1$. It is possible to define a model as additive if $\sum_{i=1}^{r} S_i = 1$. In this case, indeed, the unconditional variance of the model can be decomposed into the sum of the first order effect of each single variable. Usually this is not the case, meaning that the joint combination of some variables can be responsible for a certain share of the unconditional variance (i.e. the model is non-additive). In this case, a low first order sensitivity index does not necessarily imply that the corresponding variable has scarce effect on the output variance, since it might considerably contribute to the total output variance, by means of its combination with the other variables. A full analysis of a model with *r* variables would therefore require all the elements of the following equation to be discovered (in number of $(2^r - 1)$):

$$\sum_{i=1}^{r} S_i + \sum_{i=1}^{r}\sum_{j>i} S_{i,j} + \sum_{i=1}^{r}\sum_{j>i}\sum_{l>j} S_{i,j,l} + \cdots + S_{1,2,3,\ldots,r} = 1 \tag{13}$$

However, the characterization and practical evaluation of all the sensitivity indices in equation (13) would require a very expensive work. In order to reduce the efforts required, a synthetic indicator to be coupled with the first order sensitivity index is the total effects index, defined as follows:

$$S_{T_i} = 1 - \frac{V_{Z_{\sim i}}\left(E_{\mathbf{Z}_i}(Y|Z_{\sim i})\right)}{V(Y)} = \frac{E_{Z_{\sim i}}\left(V_{\mathbf{Z}_i}(Y|Z_{\sim i})\right)}{V(Y)} \tag{14}$$

Total effects index of the input factor *i* provides the sum of all the elements in equation (14) in which the i-th is included. When the total index is $S_{T_i} = 0$ the i-th factor can be fixed without affecting the outputs' variance. If $S_{T_i} \cong 0$ the approximation made depends on the value of S_{T_i} (4). It is worth noting that $\sum_{i=1}^{r} S_i \leq 1$ and $\sum_{i=1}^{r} S_{T_i} \geq 1$, both being equal to one only for additive models.

Computation of first and total order sensitivity indices is not straightforward. More information can be found in Saltelli et al. (3), and the software code is available through the JRC website (5).

References

1. Morgan, M.G., Henrion, M. (1990), Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press: Cambridge, NY.
2. Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F. (2006). Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering and System Safety, vol. 91, issue 10-11, pp. 1109-1125.
3. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., and Cariboni, J. et. al. (2008). Global Sensitivity Analysis. The Primer. ed. John Wiley & Sons.
4. Sobol', I.M., Tarantola, S., Gatelli, D., Kucherenko, S., Mauntz, W. (2007). Estimating the approximation error when fixing unessential factors in global sensitivity analysis. Reliability Engineering and System Safety, vol. 92, 957-960.
5. Sensitivity Analysis. EU JRC. Last accessed 24/10/13.

Appendix C: Measures of Goodness of Fit

By Costas Antoniou (NTUA, GR) and Vincenzo Punzo (JRC/UNINA, IT).

In this Annex we expand and detail many of the Goodness to fit measures from Appendix A. These are listed in the following table which is adapted from Hollander and Liu (1) and Ciuffo and Punzo (2).

The following notation is used:

- x_i: simulated measurements.
- y_i: observed measurements.
- N: number of measurements.
- $\underline{x}, \underline{y}$: sample average.
- σ_x, σ_y: sample standard deviation.
- X, Y area of the speed-flow diagram covered by simulated and observed measurements.

Name	**Measure**	**Comments**
Percent error (*PE*)	$\frac{x_i - y_i}{y_i}$	Applied either to a single pair of observed-simulated measurements or to aggregate network-wide measures.
Sum of Squared Errors (*S*) or (*SSE*)	$\sum_{i=1}^{N} (x_i - y_i)^2$	It is at the basis of the famous Linear Least Squares method which, according to the Gauss-Markov theorem, provides the best parameter estimation for linear models with zero-mean, unbiased and uncorrelated errors.
Mean Square Error (*MSE*)	$\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2$	It measures the average of the squared errors. Most widely used GoF in calibration. Low values show a good fit; strongly penalizes large errors.
Mean error (*ME*)	$\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)$	Indicates the existence of systematic bias. Useful when applied separately to measurements at each location. Useful to indicate the presence of systematic bias e.g. in validation, but cannot be used in calibration because low values do not ensure a good fit (the same high errors with opposite sign will result in zero ME)
Mean normalized error (*MNE*) or Mean percent error *(MPE)*	$\frac{1}{N}\sum_{i=1}^{N} \frac{x_i - y_i}{y_i}$	Indicates the existence of systematic bias. Useful when applied separately to measurements at each location. Useful to indicate the presence of systematic bias e.g. in validation, but cannot be used in calibration because low values do not ensure a good fit (the same high errors with opposite sign will result in zero MNE)

Mean absolute error (*MAE*)	$\frac{1}{N}\sum_{i=1}^{N} \lvert x_i - y_i \rvert$	Not particularly sensitive to large errors.
Mean absolute normalized error (*MANE*) or Mean absolute error ratio *(MAER)*	$\frac{1}{N}\sum_{i=1}^{N} \frac{\lvert x_i - y_i \rvert}{y_i}$	Not particularly sensitive to large errors. Using absolute values would result in using the same weight for all errors, while it would be preferable to assign more importance to high errors than to small. The gradient of the absolute value analytical function has a discontinuity point in zero. Second most widely used GoF in calibration.
Root mean square error (*RMSE*)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2}$	Large errors are heavily penalised. Sometimes appears as mean squared error, without the root sign.
Root mean squared normalized error (*RMSNE*) or Root mean squared percent error (RMSPE)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(\frac{x_i - y_i}{y_i}\right)^2}$	Large errors are heavily penalized. Normalized measures (also MANE) are very attractive GoFs, since they allow a model to be calibrated using different measures of performance (only relative error is considered). However, instabilities due to low values among the measurements in the fraction's denominator might affect their use.
GEH statistic	$\sqrt{2\frac{(x_i - y_i)^2}{x_i + y_i}}$	Applied to a single pair of observed-simulated measurements. *GEH*<5 indicates a good fit. According to UK Dept. for Transport (3), GEH<5 for 75% of the observed-simulated measurements indicates a good fit.
Correlation coefficient (*r*)	$\frac{1}{N-1}\sum_{i=1}^{N} \frac{(x_i - \underline{x})(y_i - \underline{y})}{\sigma_x \sigma_y}$	
Theil's bias proportion (*Um*)	$\frac{N(\underline{y} - \underline{x})^2}{\sum_{i=1}^{N} (y_i - x_i)^2}$	A high value implies the existence of systematic bias. *Um* =0 indicates a perfect fit, *Um* =1 indicates the worst fit.
Theil's variance proportion (*Us*)	$\frac{N(\sigma_y - \sigma_x)^2}{\sum_{i=1}^{N} (y_i - x_i)^2}$	A high value implies that the distribution of simulated measurements is significantly different from that of the observed data. *Us* =0 indicates a perfect fit, *Us* =1 indicates the worst fit.
Theil's covariance proportion (*Uc*)	$\frac{2N(1 - r)\sigma_x \sigma_y}{\sum_{i=1}^{N} (y_i - x_i)^2}$	A low value implies the existence of unsystematic error. *Uc* =1 indicates a perfect fit, *Uc* =0 indicates the worst fit. ***r*** is the correlation coefficient.
Theil's inequality coefficient (*U*)	$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - x_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i)^2}}$	Combines effects of all 3 Theil's error proportions (*Um, Us, Uc*). *U*=0 indicates a perfect fit, *U*=1 indicates the worst fit.

References:

1. Hollander, Y. and Liu, R., (2008). The principles of calibrating traffic microsimulation models. *Transportation*, **35**, pp.347-362.
2. Ciuffo, B. and Punzo, V. (2010). Verification of Traffic Micro-simulation Model Calibration Procedures: Analysis of Goodness-of-Fit Measures. Proceeding of the 89th Annual Meeting of the Transportation Research Record, Washington, D.C.
3. Department for Transport, U.K. (2013). Design Manual for Roads and Bridges (DMRB): Volume 12, Section 2. Accessed July 30, 2013.

European Commission
EUR 26534 EN – Joint Research Centre – Institute for Energy and Transport

Title: Traffic Simulation: Case for Guidelines

Author(s): Constantinos Antoniou, Jaume Barcelo, Mark Brackstone, Hilmi B. Celikoglu, Biagio Ciuffo, Vincenzo Punzo, Pete Sykes, Tomer Toledo, Peter Vortisch, Peter Wagner

Edited by: Mark Brackstone, Vincenzo Punzo

Luxembourg: Publications Office of the European Union

2014 – 100 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424 (online), ISSN 1018-5593 (print)

ISBN 978-92-79-35578-3 (PDF)
ISBN 978-92-79-35579-0 (print)

doi: 10.2788/11382

Abstract

The MULTITUDE Project (Methods and tools for supporting the Use, caLibration and validaTIon of Traffic simUlations moDEls) is an Action (TU0903) supported by the EU COST office (European Cooperation in Science and Technology) and focuses on the issue of uncertainty in traffic simulation, and of calibration and validation as tools to manage it. It is driven by the concern that, although modelling is now widespread, we are unsure how much we can trust our results and conclusions. Such issues force into question the trustworthiness of the results, and indeed how well we are using them.
The project consists of 4 Working Groups (WGs) which hold short, focussed, meetings on topics of interest and propose work items on key issues. Additionally the project holds an annual meeting, as well as training schools, where the latest thinking can be passed on to young researchers and practitioners.
This report covers much of the technical work performed by Working Group 4 ‘Synthesis, dissemination and training’, and assesses the current situation regarding guidelines for traffic simulation model calibration and validation worldwide, discusses the problems currently faced, and suggests potential ways in which they can be addressed, both directly, and indirectly through the development of the overall field of traffic simulation as a whole.

z

LB-NA-26534-EN-N

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

Publications Office

doi: 10.2788/11382
ISBN: 978-92-79-35578-3