Use, Calibration, and Validation of Traffic Simulation Models in Practice: Results of Web-Based Survey

Mark Brackstone, Marcello Montanino, Winnie Daamen, Christine Buisson, Vicenzo Punzo

Use, Calibration and Validation of Traffic Simulation Models in Practice: Results of a Web based Survey

Mark Brackstone, PhD (corresponding author)
IOMI Consulting
Southampton, SO40 7FF, U.K.
Phone +44 7501 470849
e-mail mark.brackstone@iomi.eu

Marcello Montanino
Department of Transportation Engineering
Università di Napoli "Federico II"
Via Claudio, 21 – 80125 Napoli – Italy
Phone +39 81 7683770
Fax +39 81 7683946
e-mail marcello.montanino@unina.it

Winnie Daamen, PhD
Department of Transport & Planning
Faculty of Civil Engineering and Geosciences
Delft University of Technology
Stevinweg 1, PO Box 5048, 2600 GA Delft – The Netherlands
Phone +31 15 278 59 27
Fax +31 15 278 31 79
e-mail w.daamen@tudelft.nl

Prof.dr. Christine Buisson
Université de Lyon, F-69622, Lyon, France
IFSTTAR, LICIT, F-69675 Bron, France
ENTPE, LICIT, F-69518 Vaulx-en-Velin, France
Phone +33 4 72 04 77 13
Fax +33 4 72 04 77 12
e-mail christine.buisson@ifsttar.fr

Vincenzo Punzo, PhD
IET Institute for Energy and Transport
European Commission – Joint Research Centre
(On temporary leave from
Università di Napoli "Federico II")
Via E. Fermi, 2749 – 21027 Ispra (VA) – ITALY
e-mail vincenzo.punzo@jrc.ec.europa.eu

Word count	
Abstract	234
Main text	6476
Figures (3 x 250)	750
Tables (0 x 250)	0
Total	**7460**

Key Words: Microscopic simulation models, Traffic simulation, Calibration, Validation, Web-based survey.

ABSTRACT

This paper reports on the results of a web based survey conducted as part of the MULTITUDE project (Methods and tools for supporting the Use caLibration and validaTIon of Traffic simUlation moDEls). The project, being performed in Europe, is attempting to focus and drive forward the issue of 'model validity' and thereby calibration and validation methods and their usage in simulation. As part of this, an assessment of the state-of-the-practice has been performed, examining among other things how models are applied, what procedures are used for calibration and what guidelines are followed. The basis for this state-of-the-practice, is a web based survey which ran in the latter quarter of 2010 and obtained 215 responses. The paper details the findings of this survey which, while lengthy, substantiates two clear messages.

Firstly, that Calibration, Validation and Sensitivity analysis although now more widespread than was previously thought, are still far from being required procedure, and secondly that while guidelines are increasingly in use, governing for example number of simulation runs, there are still many regions where this does not occur and indeed there is still much reliance on personal experience and habit, especially in situations where network characteristics become large. The results of the survey underpin the need for further focus on the use of models and procedures for calibration and validation, and indicate, more specifically, the areas that MULTITUDE will focus on in the near future.

INTRODUCTION

Traffic and transportation applications are rapidly expanding in scope given their potential impacts on community and environmental decision making. These applications range from planning and assessment of road infrastructures to evaluation of advanced traffic management and information systems (e.g. dynamic hard-shoulder running) and to testing technologies and systems that aim to increase safety, capacity and environmental efficiency of vehicles and roads (e.g. Cooperative Systems and Intelligent Speed Adaptation). The complexity and scale of these problems dictate that accurate and dynamic traffic simulation models, rather than analytical methods, are increasingly being used for these purposes.

Many commercial traffic simulation models are currently available, and even more models have been developed by research institutes and research groups all over the world. However, simulation results need to be interpreted with great care. First of all, the quality of the simulation models must be considered. In addition, the reproducibility of the simulation results is important i.e., the ability for the results to be accurately replicated by someone else working independently, using the same (or different) traffic simulation model.

One of the most important steps in this field is therefore, to develop methods and procedures to help developers and users to apply simulation models correctly, effectively and reproducibly. Motivations and solutions to these problems should ideally be found in the traffic models themselves and in the way they are applied, following an approach which is often half-way between deductive and inductive, “whereby one first develops (via physical reasoning and/or adequate idealisations and/or physical analogies) a basic mathematical modelling structure and then fits this specific structure (its parameters) to real data” (*1*). The latter phase is generally referred to as model calibration (with validation described as a test whether the model gives a sufficiently accurate representation of reality (*2*) using independent data from that used for calibration).

Unfortunately, adequate calibration and validation against suitable observed data, are far from common practice in the field of traffic simulation (*3*) and at present, no standardised methods exist, with most efforts and resources having been focused on model (and software) development. The aim of COST action TU0903 “Methods and tools for supporting the use, calibration and validation of traffic simulation models” (also called MULTITUDE) is therefore to develop, implement and promote the use of methods and procedures for supporting the use of traffic simulation models, especially on the topics of model calibration and validation, to ensure their proper use and the validity of the results and the decisions made on them (*4*). (COST is an intergovernmental framework for European Cooperation in Science and Technology, allowing the coordination of nationally-funded research on a European level (*5*). A COST Action is a network of (mostly) European scientists (*6*), to cooperating and exchanging expertise with financial support for joint activities such as conferences, short-term scientific exchanges, training schools and publications).

One of the first activities of MULTITUDE has been to undertake a state-of-the-art review of traffic simulation research and practice. This paper presents some first results from a web survey undertaken to assess the current ‘state of practice’ and to identify, arguably for the first time, common practice in the use of traffic simulation tools, covering in particular, issues of model calibration and validation against actual traffic data. In the survey, respondents were asked which models they are using and how they have set up and used the model in their last

model application. Using the information on the respondent (experience, type of company, etc), type of application, the simulation model and how the application is performed, we can draw conclusions on how models are used in practice. It is important to note however, that the purpose of the paper is not to catalogue existing commercial tools, e.g. (*7*), nor to compare the modelling methods used in traffic simulation, e.g. (*8*), or to compare the tools features, but to examine which tools are used and most importantly, how.

This paper starts with a description of the target audience, the dissemination of the survey and the responses received, followed by an overview of the survey itself. The next section describes the statistical methods that have been used to analyse the survey responses. The subsequent three sections show the results, starting with the characteristics and statistical analyses of the respondents and followed by the analyses of the reasons why applicants used a specific model and analyses on the actual use of the model in their last application (including calibration and validation). Due to the on-going nature of the MULTITUDE project, we follow this with a range of results from analysis of interdependence between some of the key variables, along with conclusions and an overview of further research.

TARGET AUDIENCE AND DISSEMINATION

With many of the participants in the MULTITUDE project coming from the academic and research sectors, it is one of the goals of the project to ensure that there is sufficient outreach to the practitioner community, i.e. those that use the models for commercial purposes, with a view to bringing (the often estranged) communities closer together. In order to do so, the survey was 'advertised' in both the academic press and 'trade press', from mid-October 2010, to the end of that year. Dissemination routes included a range of LinkedIn groups, national email lists (e.g. UTSG in the UK, HITE in Greece and SIDT in Italy), ITS-America, TRB Committees (a total of 6 were approached in order to attempt to obtain a multi-disciplinary response) as well as newsletters of model manufacturers (PTV, TSS and SIAS) and in publications both international, such as Traffic Engineering and Control (TEC), and national, e.g. Local Transport Today (LTT) in the UK. It should be noted that while a global response was sought, the project is, first and foremost, and due to its funding, European in nature.

SURVEY OVERVIEW

The survey is structured in five parts, discussed in more detail below. In the first, respondents were asked to list up to five traffic simulation models/software tools they use in practice. This enables an understanding of whether respondents have a choice between various models and may be choosing the most appropriate model for a specific assignment, or whether only a single model is available and being used for all projects.

The second part aims at investigating the level of awareness and comprehension of respondents about the model/software they used in their last application/assignment. This part includes questions on the model approach (such as static or dynamic), on the level of detail of the traffic models (such as macroscopic, mesoscopic or microscopic), on the type of model performance functions and on the model type (such as stochastic or deterministic). Findings from this part of the survey are not reported in this paper due to space restrictions, and are instead, the focus of a separate publication submitted elsewhere for print in the latter half of 2011.

The third and fourth parts of the survey both aim to examine how respondents use the simulation model, including how (and if) calibration and validation was performed. Part three looks at the characteristics of the last model application and motivations which lead to the model choice. This includes general information on the application type (such as planning or real-time operations), on the impact assessed, (such as throughput and safety), on the type of project and on the type of client.

The fourth part is more technical and addresses specific questions about the use of models in the last application, including technical information and descriptions of the simulation scenarios (e.g. number of links, nodes, and OD pairs), the number of replications/runs performed and the outputs typically used. In addition, it deals with MULTITUDE's core issues of model calibration and validation. Questions are designed to provide an understanding of to what extent, and how, these processes are performed by researchers and practitioners, the sub-models and parameters involved, as well as the measurements available and made.

Personal information is collected in part five of the survey regarding the type, size and focus of the respondents' organisations etc.

STATISTICAL METHODS

For each question, respondents were asked to select the most suitable answer (or answers) from a list of choices thereby providing categorical data. Some questions had an option 'other', in which the respondent could fill out a dedicated response. The text of this response is typically analysed separately. While space restrictions prevent a

detailed presentation of all derived statistics behind our findings, we include below a synopsis of the approach used in order to substantiate our conclusions, presented in brief in subsequent sections.

The first stage of the analysis was to undertake an overview of the distribution of the answers to each question. This was performed through the use of frequency cross tabulation tables, providing information on the percentage of respondents who selected the different choices and allowing the identification of the most frequently used models, and general trends. However, this basic analysis neither provides the significance of the differences between the answers' frequencies, nor correlations between the answers that individuals gave. To explore the correlation structure among the categorical variables, a multiple correspondence analysis was therefore performed and results of this analysis significant at the 5% (p=0.05) level or greater are reported in this paper.

Subsequent analysis involved correspondence analysis, a descriptive/exploratory technique designed to analyse two-way tables containing some measures of correspondence between the rows and columns (i.e. the observed frequencies) (*9*). According to the well-known formula for computing the Chi-square statistic for a two-way table (*10*), the expected frequencies in a table, where the column and rows are independent of each other, are equal to the respective column total times the row total, divided by the grand total. Any deviation from the expected values (expected under the hypothesis of complete independence of the row and column variables) will contribute to the overall Chi-square. Thus, another way of looking at correspondence analysis is to consider it a method for decomposing the overall Chi-square statistic by identifying a small number of dimensions in which the deviations from the expected values can be represented. In this framework, the analysis of the associations between variables was based on the Chi-square test for independence, resulting in the Pearson Chi-square coefficient and the p-value of the test. The Pearson Chi-square measure is based on the fact that one can compute the expected frequencies from a two-way table. If there is no relationship between two variables (i.e. questions), one would expect an equal number of choices of the different responses to the questions. The Pearson Chi-square test becomes increasingly significant as the observed counts deviate further from this expected pattern. However, the Chi-square coefficient does not tell anything at all about the nature of the association, only if there is evidence of it (or not). When such an association appears to be present, the strength of this association can be identified. For this purpose, several statistical methods can be used, for example, residual analysis. In the context of a two-way table, a residual (or deviate) is defined as the difference between the observed frequency and the expected frequency. To make comparisons more straightforward, statisticians generally prefer to use standardized deviates, which are calculated by dividing the residual value by the square-root of the expected count.

$$Standardized\ Residual = \frac{O - E}{\sqrt{E}} \quad (1)$$

where O and E are, respectively, the observed and expected counts from the two-way table. Standardized residuals allow us to see the direction and strength of the association between categorical variables. A large standardised residual provides evidence of association in that cell. Thus, if the standardized deviate value is between –1.5 and 1.5, observed counts agree with the independence. If it is between –2.0 and –1.5 (or 1.5 and 2.0), observed counts give mild evidence against independence (showing negative or positive correlation). If it is less than –2.0 (or greater than 2.0), there is a strong evidence against independence (negative or positive correlation).

CHARACTERISTICS OF RESPONDENTS

In total, 215 responses were obtained from 37 countries (70% from Europe, 14% from the USA and Canada, 7% from Asia and Australasia, with the remainder from Latin and South America, and Middle East). Of the 150 responses from Europe over half were returned from three countries, the UK (47), Italy (16) and Spain (15). From the total pool of respondents, 71% identified that the main micro-simulation product that they were using was one of either VISSIM (27%), AIMSUM (25%) or PARAMICS (15%). Respondents were subsequently classified as falling into one of four Regions: North America (NoA), UK, Europe excluding the UK (EXUK) and Rest of the World (RoW). Not only does this classification ease analysis (proportions are respectively 14%, 22%, 48% and 16%), but this split also allows us to investigate potential hypotheses regarding prevalent cultures on model use, where for example, it is known that modellers in the UK have, and use, extensive guidelines, as do (to a lesser extent) those in North America, while those in the rest of Europe (predominantly) do not (*11*).

In the Demographic section of the survey a range of background information was sought to enable an analysis of how issues explored in other sections may vary according to respondent background. These included Type of respondent, Size of respondents organisation, Number of persons in their organisation using traffic models,

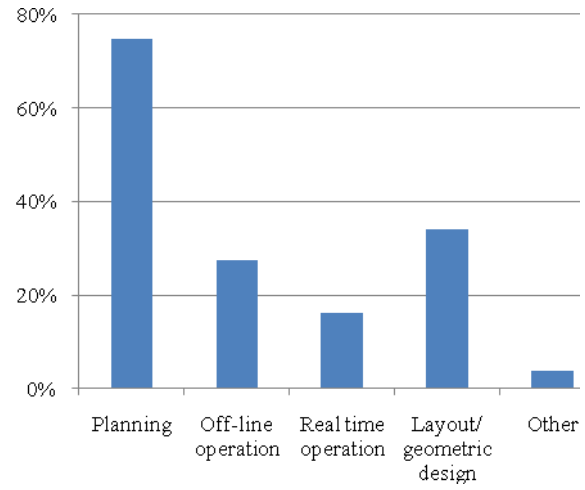
Average number of traffic projects undertaken per year and Number of years spent in the field (experience). From these responses it was observed that overall, 64% of respondents came from Consultancies, 14% from Educational establishments, 10% from Research establishments and 9% from Central or Local Government organisations. As we aim to get an overview of how (primarily) practitioners (i.e. consultants) use models, this distribution meets our desires and expectations. In addition, the respondents from research institutes and governments enable us to compare the behaviour of respondents working for different types of employers, including the possibility of different approaches of 'practitioners' and academics.

In terms of company size, 25% of the responses were received from persons in organisations of more than 2000 employees, 29% between 100 and 2000 and 28% from small organisations of less than 20 employees. Considering the large number of respondents from large organisations, the number of persons using traffic models was perhaps, surprisingly low with 47% belonging to groups of 5 or less, 81% to groups of 20 or less and all except 4% to groups of less than 100. The number of projects actually performed in a year was also low, with 32% stating that 5 or fewer projects were being performed, 53% less than 10 projects, while only 21% were in organisations where 30 or more projects were being undertaken annually. (These low percentages might be caused by a misinterpretation of the question: instead of mentioning the total number of projects performed by the total company, the respondents might have returned the number of project they were involved in themselves). Lastly, it was observed that 39% of the responses came from individuals with 5 or less years of experience in the field and 32% with 10 or more.

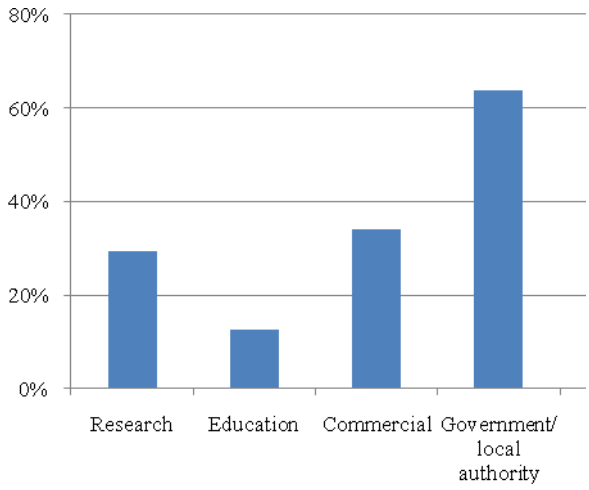
A number of interesting relationships were however found between respondent demographics and their Region. For example it was clear that from the UK a significantly greater number of Consultancies and fewer Research organisations responded, while throughout RoW the converse was true ($p=0.0013$). While European respondents tended to come from smaller organisations (mostly from 6-20 persons, for 1 in 3 of the sub-sample), UK respondents primarily were drawn from larger companies with over 2000 persons (50% of the sub-sample). This was reflected in the fact that a significant proportion (37%) were also part of groups where 21-100 persons were employed in the use of models with a significant proportion (29%) in organisations that undertook more than 30 modelling projects a year. Conversely, 51% of the respondents from RoW were involved in a very small number of projects (2-5 per year, $p=0.00086$). Experience in the UK however was loaded away from those with 20 or more years (with all respondents having less), while in NoA the converse was true with 42% of respondents claiming 20 or more years ($p=0.0255$). It is clear then that those responding in the USA are likely to have been more senior individuals than from elsewhere, with UK responses drawn from larger consultancies perhaps involving those at an operational level more, and with RoW respondents having been disproportionally drawn from smaller Research Institutes. Surprisingly, no significant correlations were found between any of the demographic variables and model chosen. Lastly, no significant variation was found in the distribution of Type of Respondent according to simulation model ($p=0.16$) however the low number of responses for some of the classifications, means that this finding should be viewed with caution. The distributions however do meet the broad expectations of the model manufacturers who were consulted regarding whether the sample adequately represented their expected customer base.

ANALYSES ON LAST MODEL APPLICATION OF RESPONDENTS

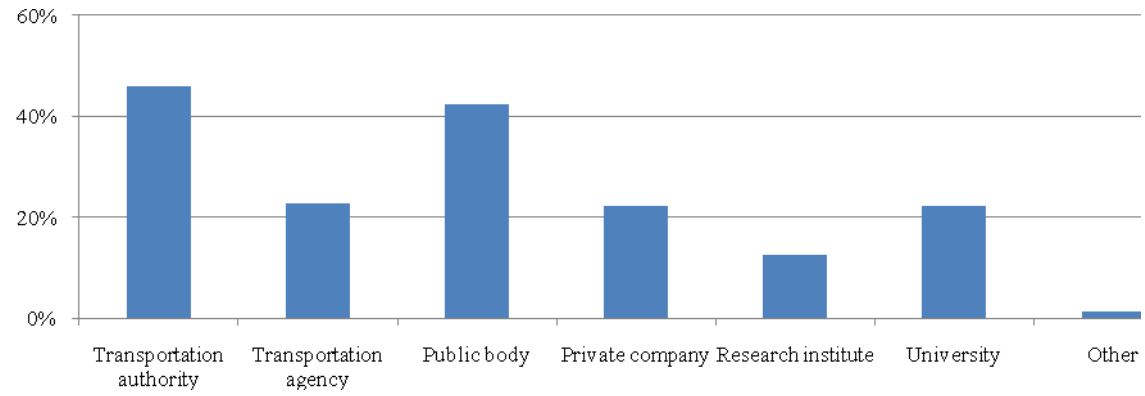
Questions in the third part of the survey were designed to take a 'snapshot' of where the focus currently lies in the simulation world, regarding what models are being used for, and for example where most backing is coming from for work 'in the field'. Here, respondents were allowed to choose as many applicable answers as they wished, and findings indicated (FIGURE 1a) that while 75% were involved with planning, there was perhaps a surprisingly large number of persons working in/on real-time operations (16%), and, while 93% had a primary interest in assessing throughput as their main impact of interest, a significant minority were also interested in safety (18%) and/or environment (22%).



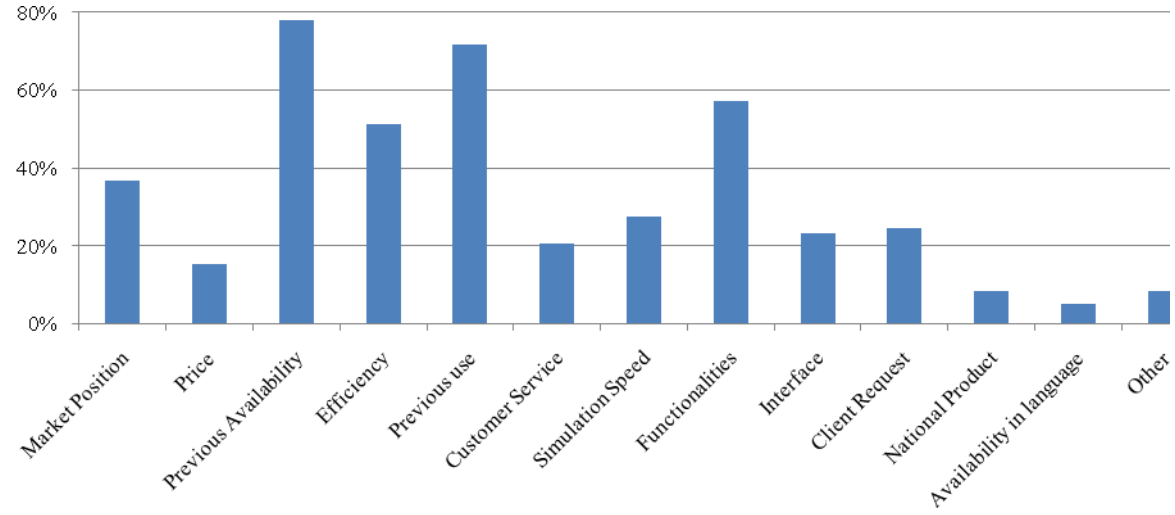
a. Application type.



b. Type of project.



c. Client



d. Model choice motivation.

FIGURE 1 Information on last model application (% of respondents classing their activity in each category, more than 1 choice allowed) by a) Application type, b) Type of project, c) Client, d) Model choice motivation.

While over 64% of the respondents were engaged on government projects and over 34% on commercial projects, the 29% response for research projects (FIGURE 1b) would seem unusual (given that only 10% of those responding came from such institutes), and, while the majority of clients were private and/or government, a significant minority (10-20%) were either Universities or Research Institutes themselves (FIGURE 1c).

Model choice motivation was established through a multi choice question with on average, each respondent citing 4 factors (the format of the question however, precluded establishing any priority between these responses for each respondent (no ordering of the answers)). A wide spectrum of responses were found, with most citing the importance of functionalities and previous availability and use, few were interested in price, speed or interface and, in only 25% of the cases was the product specified by the client (FIGURE 1d). Examination of common pairings reveals that Functionality and Previous use are indeed almost always cited together with Functionality and Efficiency with Previous Use also forming a key pairing with Previous availability and Efficiency. Analysis by Region revealed a range of significant trends with:

- Impact assessed (p=0.017), where NoA respondents placed a significantly higher emphasis on assessing safety and 'other' impacts including for example fuel economy or performance and economic factors and issues**.**
- Type of Project (p=0.001), with the UK placing a far greater emphasis on Government/local authority projects at the expense of Research and Education projects and the RoW conversely, having a higher emphasis on Research projects.
- Client (p=0.0001), with work in NoA coming far more from Transportation Agencies (the biggest single client for that Region) and in RoW far more from Universities than in other regions.

ANALYSES OF SIMULATION MODEL USE IN LAST APPLICATION

This part of the survey consisted of a total of 25 questions regarding technical aspects of the last simulation performed by the respondent, including set up details and calibration and validation procedures used. These are addressed below. Firstly, questions were asked regarding characteristics of the simulation itself:

- Network type: Urban (59%), Rural (4%), Motorway (8%), Mixed (29%).
- Diameter of simulated area: <1Km (24%), 1-5Km (30%), 5-25Km (26%), >25Km (20%).
- Number of links: <100 (37%), 100-500 (22%), 500-2500 (20%), 2500-10,000 (14%), >10,000 (7%).
- Number of Intersections: <10 (21%), 10-100 (43%), 100-500 (12%), >500 (24%).
- OD pairs: <10 (15%), 10-50 (33%), 50-100 (16%), 100-500 (16%), >500 (20%).
- OD trips: <1000 (11%), 1000-10,000 (41%), 10,000-100,000 (29%), >100,000 (19%).

The first three of these variables proved to be independent of Model used or Region. A clear correlation was found however between Number of Intersections and Model chosen (p=0.013) with VISSIM usage being focused on smaller Numbers of Intersections (36% of the sub-group responded with less than 10), reflecting VISSIMs known background as a tool traditionally offering sophisticated junction design options and controller interfaces. As regards OD characteristics, while no variation was found regarding numbers of OD trips, a Model (but not a Region) dependency was found for the number of OD pairs implemented, with PARAMICS exhibiting significant variations from the average (p=0.008), being used more for the 100-500 band (37%) and 2500-10,000 (6%). The next five questions focused on run time issues and outputs and how they were addressed (if at all):

- A 'Warm up period' was found to be used by 76% of the respondents. Since a warm up period is necessary to fill the network (in reality the network is almost never empty) and to let the system reach a steady state, in 24% of the replies it could be questioned whether the results of this warm up period are included in the final results, thus affecting the outcomes of the simulation.
- For the Duration of the warm up period [<15, 15-30, 30-60, 60-120, >120 mins], 53% of the respondents were found to be using periods of 15 minutes or less, and a total of 83% less than 30 minutes. While seeming quite short, these should be viewed in the context of the overall simulation time, which is examined later.
- Simulation duration [<30, 30-120, >120 mins] was generally found to be 30-120 minutes of simulated time (64% of the sample), with only 7% undertaking shorter simulation runs, and 29% longer.
- Regarding Number of runs performed for each simulation scenario [1, 2-5, 6-10, 11-20, >20], 30% of the respondents were found to perform 2-5 runs for each scenario in order to establish convergence and 35% 6-10 runs. However, this varied heavily with 17% only performing 1 run, and 18% performing more than 10 runs. Obviously, when performing only a single run when using a stochastic simulation model the results

should be questioned, since one does not know whether the results correspond to an average situation, a worst case situation or a positive situation. Due to the many stochastic factors in a simulation model, even performing 6-10 runs may be limited, although this may depend on the uncertainty in the model and inputs as well as on the aim of the study (i.e. the outputs of interest).

- The Method used to calculate the number of runs performed [Personal experience, Statistical test, Habit, Other] was mostly found to be personal experience (52%) with only 26% calculating this number through statistical test, and 13 arriving at this number purely by habit.

A number of interesting variations were found in these responses. The use of a warm up period was found to be more likely among respondents from NoA (93%, p=0.03), or VISSIM users (95%) at the p=0.03 level. As regards warm up duration, RoW respondents were found less likely to favour times over 30 minutes (only 4%), while NoA respondents favoured 30-60 mins (39% as opposed to the average of 14%), and those from the UK preferred more than 60 mins (8% as opposed to the average of 3%) at the p=0.09 level.

Significant variations were also found in simulation duration, firstly by Region (p=0.00001), where the UK was found to focus on longer runs (44%) and the RoW substantially on shorter runs of less than 30 mins (25%). Model was also found to play a role (p=0.00017) with PARAMICS users focusing on runs longer than 120 mins (52% of the sub-group) and VISSIM users on periods longer than 30 mins (100%). This might be explained by the fact that PARAMICS users on average had applications with larger networks which would require longer durations and this point will be explored more fully later.

While the Number of runs performed was found to be model independent it was found (p=0.007) that the 11-20 band was favoured in the UK (21% as opposed to the average of 10%). This may in part be due to guidelines provided by the Highways Agency (*12*) which suggest more than 10 simulation runs should be performed. While the number of persons working on projects to which these guidelines directly apply is unlikely to be large, it is possible that they are now being more widely adopted for use on projects funded by other agencies or authorities. No variation was found in the Method used to calculate this number by either Region or model.

The subsequent three questions centre around the key issues of the survey (and indeed the MULTITUDE project) and concern whether calibration was usually undertaken and if so, how? In summary: 45% were found to use guidelines of some form; 81% were found to perform calibration and of those performing calibration, 74% were found to have calibrated the underlying traffic flow model and 32% the underlying route choice model. While the performance of calibration, and the typical models calibrated, were found to be independent of Model or Region there was substantial variation in the use of guidelines. These were found to be substantially (p=0.0000) more used in the UK (81%) and less in EXUK (29%). Of additional note (p=0.001) is the fact that PARAMICS users were more likely to use guidelines (76%) and AIMSUN users less (35%).

Features of the calibration process (both for traffic flow (TF) and route choice (RC)) were examined as regards the number of variables that the user attempts to calibrate (this is not necessarily the same as the number that the user knows to need calibrating but is instead a number governed by lack of resource or data). For both of these models users most commonly attempted to calibrate 2-5 parameters (70% of users in the case of the TF model, 68% in the case of the RC model) with 51% of users for the TF model and 47% of users for the RC model assuming that all other parameters are calibrated and suitable for use. Twenty five percent of the respondents were found to be using automatic optimization procedures for TF calibration and 17% for the RC model, with counts and travel times being the most common types of data being used (85% and 65% of the responses for the TF model and 61% and 65% for the RC model, FIGURE 3). Of note is the fact that 10% of TF calibrators and 22% of RC calibrators stated that they were using trajectory data and that 38% for TF and 47% for RC used ‘personal experience’. From a scientific point of view, depending on the level of detail sought, for an extensive calibration of the traffic flow model (in more specifically a car following or lane changing model), trajectory data would be very useful, however whether such data is fully and globally applicable considering national variations in driving behaviour is an open question (*13*). Calibrating a model on personal experience is obviously very qualitative and does not give any guarantee for accurate simulation results.

Lastly 68% of users claimed to undertake sensitivity analysis and 77% of users claimed to have performed validation on their last model application with detector counts and travel times being the most commonly used sources of data (63% and 58%, FIGURE 2). While the use of differing types of data was not found to vary, a slight dependence was found in the performance of sensitivity analysis and validation itself with Region (p=0.018), with both being found to be more prevalent in the UK (81% and 93% respectively).

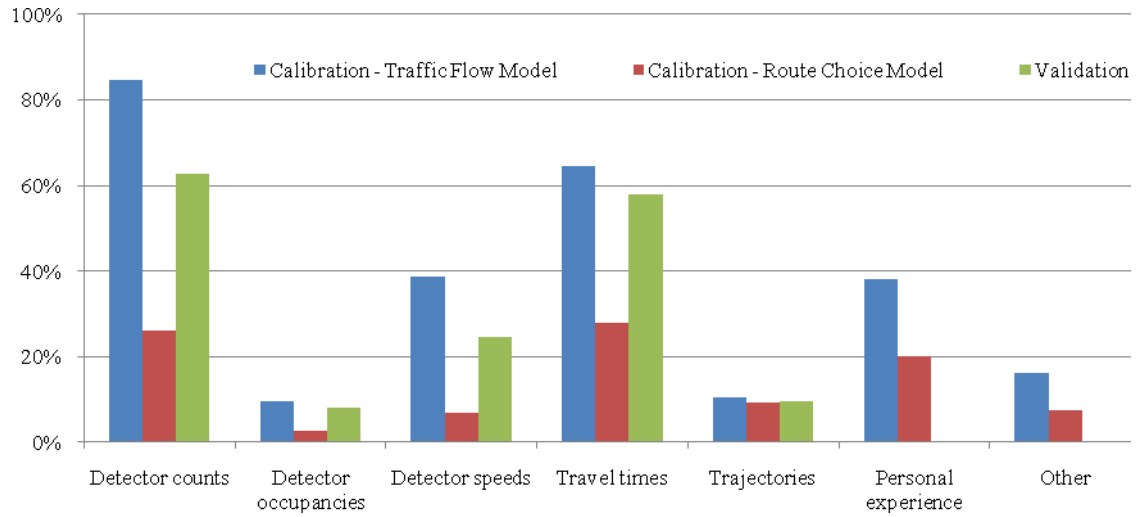


FIGURE 2 Percentage of respondents using differing types of data in the Calibration of Traffic Model, Calibration of Route Choice model and, Validation.

RELATIONSHIPS BETWEEN VARIABLES AND FUTURE WORK

While it has been instructive to examine how many of the survey variables vary according to Region or Model used, it is also highly likely that there are correlations between the answers to many of the individual questions, and to that end we have also examined eight key relationships. (At this stage these have not been subject to the more rigorous analyses presented earlier, but rather these are presented as potential relationships, that may be confirmed in subsequent investigations).

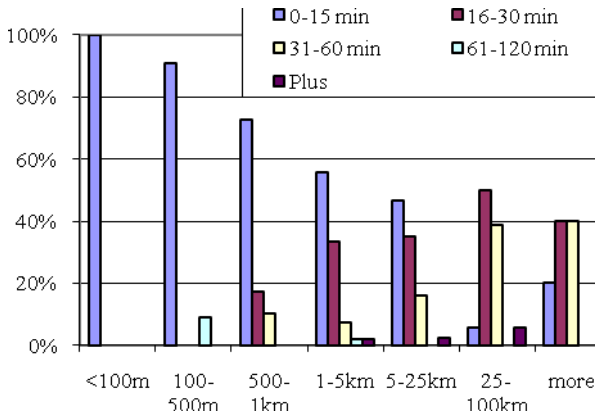
Firstly, it is highly likely that simulation duration and other practical issues that may affect the logistics of performing a simulation project on time and within budget, are related. Here, one finds that the fraction of respondents using longer warm up times increases with Network size (FIGURE 3a), as does the fraction using longer Simulation runs (FIGURE 3b). Clearly the larger the model, the longer the run needs to be in order to ensure that that the network is filled with interacting vehicles. Additionally, it may be seen that as the length of each simulation run increases then the percentage of respondents using longer warm up periods does also (FIGURE 3c).

Next, one may hypothesise that as the size of the network increases, the data and resource required in order to undertake Calibration or Validation will in-turn increase, as does indeed the importance of undertaking these processes. FIGURE 3d shows the percentage of respondents that stated that they undertook either of these processes, which exhibit similar levels of prevalence. While these levels seem to rise with Network size, a slight dip seems to occur for the largest networks, indicating potentially that the resources available, are outstripped by those ideally needed. (Correlation between undertaking Validation, given Calibration is high with 67% of respondents having undertaking both, 9% neither, 14% Calibration but not Validation, and 10% Validation but not Calibration).

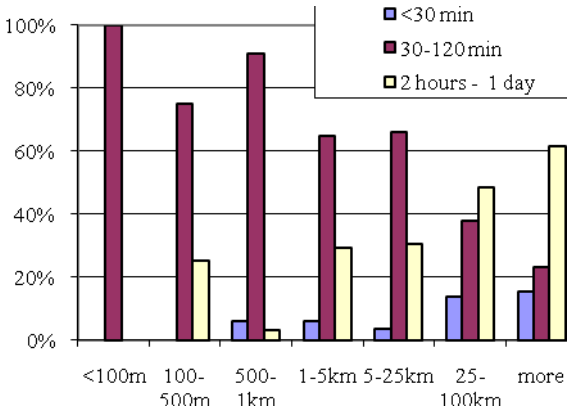
Examining the Number of runs and Network size (FIGURE 3e) one can observe a very wide spread including a large number that only perform 1 run for very large networks. Putting such respondents to one side and assuming that their responses are driven by particular needs/resource requirements, one can see a gradual rise in the Numbers of runs used, including most notably the 20+ group for the largest of networks. The very occurrence of such networks and simulation projects is a clear demonstration not only of the power of modern computing, but also of the ambition, scope, and indeed scientific rigour of an increasing fraction of the work performed in the traffic and transportation domain.

Lastly, the method used in order to arrive at the number of simulation runs required, reveals that while personal experience is always much in evidence, the fraction of respondents citing this method tends to reduce as the Number of runs increases, with a corresponding increase in the prevalence of those using statistical tests (FIGURE 3f). This trend is however reversed for those citing more than 20 runs, indicating that potentially, due to scale and resource, experience may be ‘winning out’ over otherwise increasingly lengthy simulation programmes that would be dictated by statistical tests.

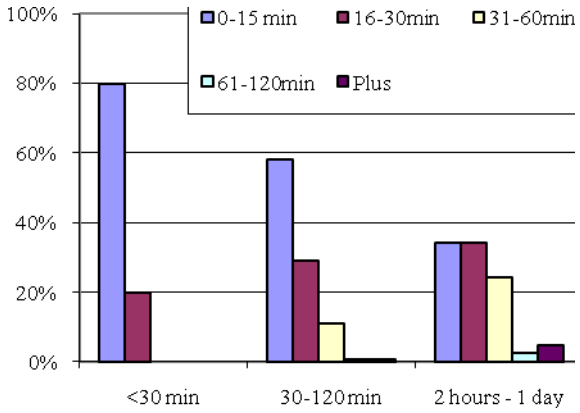
1



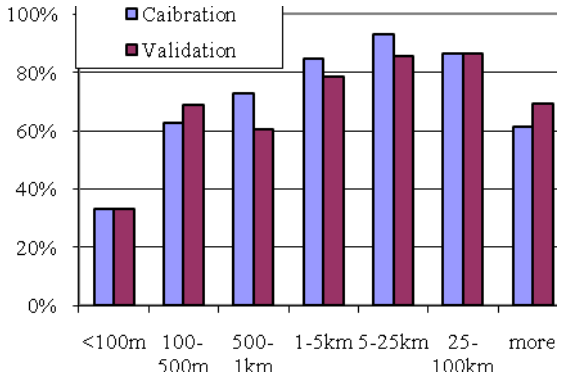
a. Duration of warm up runs by network size



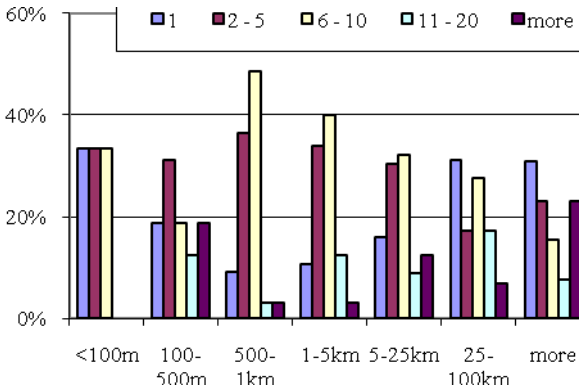
b. Duration of simulation runs by network size



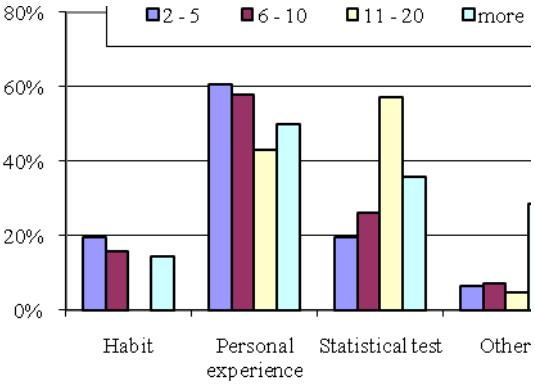
c. Duration of simulation runs by duration of associated warm up period



d. Prevalence of calibration and validation by network size



e. Numbers of simulation runs performed by network size



f. Stated preference in technique used to determine number of simulation runs

2

3 **FIGURE 3 Fraction of respondents undertaking, a) Differing duration warm up runs by network size, b)**
4 **Differing duration simulation runs by network size, c) Differing duration simulation runs by duration of**

associated warm up period, d) Calibration and validation by network size, e) Differing numbers of simulation runs performed by network size, f) Use of different techniques in determine number of simulation runs.

CONCLUSIONS AND RECOMMENDATIONS

This paper has presented initial results from a web based survey conducted as part of the MULTITUDE project which has sought to assess how traffic simulation models are being used, and what procedures are used for calibration and validation. The survey, performed in the last quarter of 2010 has obtained 215 responses from 37 countries, predominantly (64%) from Consultancies. The survey has revealed a wide range of findings with some dependencies on Region (of respondent) and Model (used). In brief, these are:

- Most respondents were found to be working on planning based projects (75%) with 93% showing a primary interest in assessing Throughput and Model choice being found to be influenced mostly by functionalities and previous availability and use.
- The majority of respondents were found to be working on Urban network simulations (59%) with a very wide variety of sizes with, most commonly, 100 to 1000 links, 10 to 100 intersections, 10 to 50 OD pairs and 1000 to 10,000 OD trips.
- A 'Warm up period' was found to be used by 76% of the respondents with 83% of these found to be less than 30 minutes in duration. The duration of the Simulation run itself was typically found to be of the order of 30-120 minutes with 65% performing 2-10 runs per scenario in order to establish convergence (this was however found to vary widely with a significant minority found to be performing 1 run, or 20 or more), and using mostly personal experience to decide on this number (52%).
- Existing Modelling guidelines were found to be used 45% of the time with 81% performing calibration of the underlying Traffic Flow model (74%) or the underlying Route Choice model (32%), typically for 2-5 variables each. Sensitivity analysis was undertaken by 68% of users with 77% claiming to have performed validation, most commonly using detector counts and travel times (63% and 58% respectively).
- A number of interrelationships were found between some of the variables (though as yet not statistically proven) that are most likely related to the physical and resource constraints governing the performance of simulation projects. For example as Network size increases there appears to be a tendency to both use longer Warm up periods, and longer Simulation runs, and indeed the longer the simulation run, the longer the warm up period. Network size also seems to effect the likelihood of the performance of both Calibration and Validation, with generally, the larger the network, the more likely the performance of these processes.
- The relationship between Network size and Number of runs performed is observed to be complex, with an increase in one, generally positively correlated to the other, and, although many cite personal experience as a key factor in deriving this number, statistical tests appear to be increasingly important as network size increases.

Although providing a large number of results, the survey is not without its weaknesses and a number of factors restrict the (detailed) validity of our results. For example, during the analysis questions arose regarding how the respondent interpreted the questions and this is a confounding factor that is, to a certain extent, impossible to overcome. Additionally it is clear that there is a bias within the survey toward respondents within the UK (and using the PARAMICS model), and this (and the associated guidelines likely to be used by them) may have skewed many of our findings. Additionally it is possible to question whether respondent type is statistically representative of the actual distribution of model users, for each of the different models. While these criticisms may affect the magnitude of our results, the contrasts observed however are sufficiently strong to indicate that there are issues that can be clearly identified and these include:

- Calibration, Validation and Sensitivity analysis are perhaps now more widespread than were previously thought, however they are still far from being required procedure.
- While, notably in the UK, guidelines are increasingly in use, governing for example number of simulation runs, there are still regions where this does not occur and indeed there is still much reliance on personal experience and habit, especially in situations where network characteristics become large.
- While counts and travel times are still the most commonly used data sources for calibration, increasing number are using trajectory data. While in principle such an increasingly microscopic approach is to be encouraged, caution is needed that the data used is sufficiently transferable so that the chances of a mis-calibration are minimised.

Whilst it is true that these issues are to a certain extent a matter of perspective (for example, what is viewed as adequate by industrial standards is unlikely to be viewed similarly by the more stringent standards set and sought by academics), it is probable that most would agree that these issues are both important, and need to be examined, especially in light of the prevalence of simulation projects that are now pushing the boundaries of computation in the use of exceptionally large networks and/or testing real-time applications. Through the performance of this survey MULTITUDE has highlighted and substantiated a range of issues that, although known, have not previously been quantified, and it is these issues which the project will attempt to progress during 2012. In particular, we intend to pursue three stands of work, firstly continued analysis of the data (for example how do methods and results vary with sector of respondent, and what do the findings of Section 2 of the survey tell us about user comprehension and understanding of the very models they are using). Secondly there is scope for re-launching an improved version of the survey, building on the findings and weaknesses found here, ensuring that the population sample is truly representative and indeed potentially extending its scope to cover the more prevalent field of macro/meso models. Lastly, it is hoped that this body of evidence may be usable to the community as a whole in efforts to lever increased support for development of guidelines needed for addressing many of the problems faced in this sector in the coming years, indeed this is to be pursued through a range of Stakeholder meetings to be held with Government and Consultants in Europe in 2012, not only in nations where guidelines exist, but also in those where they are currently under formulation. In doing so it is hoped that the project can act as a catalyst to identify research gaps from which industry may benefit through closer interaction with the work underway in academia.

ACKNOWLEDGMENTS

This paper has been made possible through co-operation brought about by participation in COST Action TU0903 - MULTITUDE, funded in the EU by the COST programme. The Authors are indebted to the survey respondents who so willingly gave their time and views and to other MULTITUDE participants for the provision of the survey in differing languages.

REFERENCES

[1] Papageorgiou, M. Some remarks on macroscopic traffic flow modelling, *Transportation Research Part A*, Vol. 32, No. 5, 1998, pp. 323-329.

[2] Kleijnen, J.P.C. Verification and validation of simulation models, *European Journal of Operational Research*, Vol. 82, 1995, pp. 145-162.

[3] National Cooperative Highway Research Program (NCHRP 3-85): "Guidance for the Use of Alternative Traffic Analysis Tools in Highway Capacity Analyses", Presentation at Transportation Research Board Annual Meeting, Washington D.C., Jan., 2007.

[4] MULTITUDE. *Introduction to MULTITUDE*. http://www.multitude-project.eu/. Accessed July 6, 2011.

[5] COST. *About COST*. http://www.cost.esf.org/about_cost. European Cooperation in Science and Technology. Accessed July 6, 2011.

[6] COST. *COST Countries*. http://www.cost.eu/about_cost/cost_countries. Last accessed July 9, 2011.

[7] Federal Highway Administration. http://ops.fhwa.dot.gov/trafficanalysistools/tat_vol2/sectapp_e.htm#top. Accessed July 17, 2011.

[8] Brackstone, M. and M. McDonald. Car-following: a historical review, *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 2*,* No. 4, 1999, pp 181-196.

[9] Dodge, Y. *The Oxford Dictionary of Statistical Terms*, OUP, 2003.

[10] Siegel, S. and N.J. Castellan. *Nonparametric statistics for the behavioral sciences*, *2nd edition*. McGraw-Hill, London, 1988.

[11] Vortisch, P. *National Guidelines for Microsimulation*. Presented at 3rd MULTITUDE Meeting, Haifa, Israel, 6-7 December 2010. http://multitude-project.eu/meetings.html. Accessed 26 July 2011.

[12] *Guidelines for the Use of Microsimulation Software*. Highways Agency. July 2007. http://www.highways.gov/knowledge_compendium/assets/documents/Portfolio/Guidelines - 901.pdf. Accessed 26 July, 2011.

[13] Marsden, G., M. McDonald and M. Brackstone. A Comparative Assessment of Driving Behaviours at Three Sites, *European Journal of Transportation and Infrastructure Research,* Vol. 3, No. 1, 2003, pp 5-20.