



HAL
open science

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

► **To cite this version:**

Mark Brackstone, Marcello Montanino, Winnie Daamen, Christine Buisson, Vicenzo Punzo. Use, Calibration, and Validation of Traffic Simulation Models in Practice: Results of Web-Based Survey. TRB 2012 - Transportation Research Board 91rd Annual Meeting, Jan 2012, WASHINGTON D.C, United States. 12 p. hal-01411970

HAL Id: hal-01411970

<https://hal.science/hal-01411970v1>

Submitted on 7 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3 Mark Brackstone, PhD (corresponding author)
4 IOMI Consulting
5 Southampton, SO40 7FF, U.K.
6 Phone +44 7501 470849
7 e-mail mark.brackstone@iomi.eu
8

9 Marcello Montanino
10 Department of Transportation Engineering
11 Università di Napoli “Federico II”
12 Via Claudio, 21 – 80125 Napoli – Italy
13 Phone +39 81 7683770
14 Fax +39 81 7683946
15 e-mail marcello.montanino@unina.it
16

17 Winnie Daamen, PhD
18 Department of Transport & Planning
19 Faculty of Civil Engineering and Geosciences
20 Delft University of Technology
21 Stevinweg 1, PO Box 5048, 2600 GA Delft – The Netherlands
22 Phone +31 15 278 59 27
23 Fax +31 15 278 31 79
24 e-mail w.daamen@tudelft.nl
25

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

34 Vincenzo Punzo, PhD
35 IET Institute for Energy and Transport
36 European Commission – Joint Research Centre
37 (On temporary leave from
38 Università di Napoli “Federico II”)
39 Via E. Fermi, 2749 – 21027 Ispra (VA) – ITALY
40 e-mail vincenzo.punzo@jrc.ec.europa.eu
41
42
43

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

44
45 **Key Words:** Microscopic simulation models, Traffic simulation, Calibration, Validation, Web-based survey.
46
47

1 ABSTRACT

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

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

17 INTRODUCTION

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

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

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

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

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

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

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

14 **TARGET AUDIENCE AND DISSEMINATION**

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

25 **SURVEY OVERVIEW**

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

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

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

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

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

49 **STATISTICAL METHODS**

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

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

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

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

$$31 \quad \text{Standardized Residual} = \frac{O - E}{\sqrt{E}} \quad (1)$$

32 where O and E are, respectively, the observed and expected counts from the two-way table. Standardized residuals
 33 allow us to see the direction and strength of the association between categorical variables. A large standardised
 34 residual provides evidence of association in that cell. Thus, if the standardized deviate value is between -1.5 and 1.5 ,
 35 observed counts agree with the independence. If it is between -2.0 and -1.5 (or 1.5 and 2.0), observed counts give
 36 mild evidence against independence (showing negative or positive correlation). If it is less than -2.0 (or greater than
 37 2.0), there is a strong evidence against independence (negative or positive correlation).
 38
 39

40 CHARACTERISTICS OF RESPONDENTS

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

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

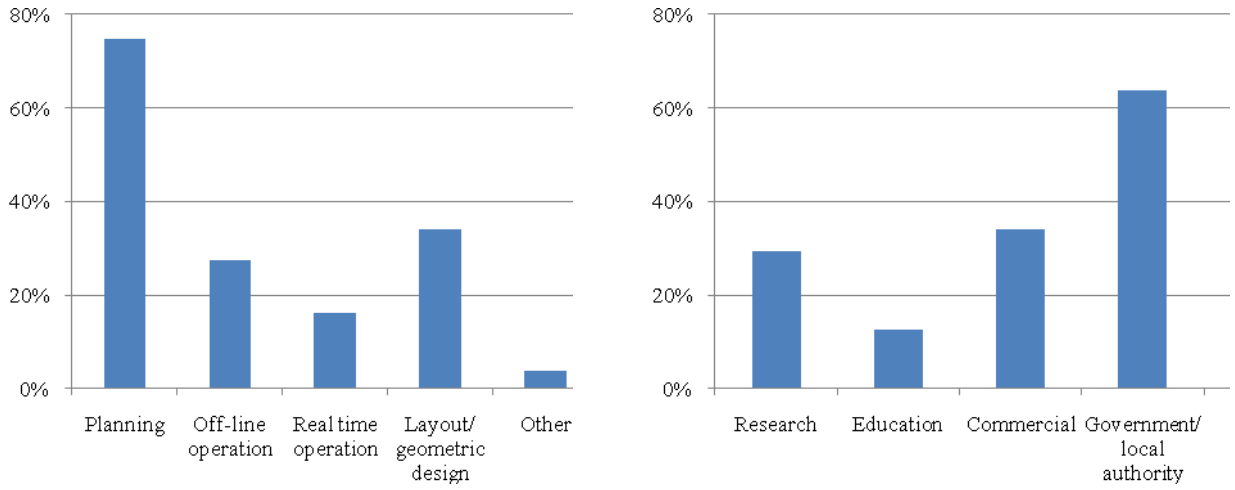
1 Average number of traffic projects undertaken per year and Number of years spent in the field (experience). From
2 these responses it was observed that overall, 64% of respondents came from Consultancies, 14% from Educational
3 establishments, 10% from Research establishments and 9% from Central or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

8 In terms of company size, 25% of the responses were received from persons in organisations of more than
9 2000 employees, 29% between 100 and 2000 and 28% from small organisations of less than 20 employees.
10 Considering the large number of respondents from large organisations, the number of persons using traffic models
11 was perhaps, surprisingly low with 47% belonging to groups of 5 or less, 81% to groups of 20 or less and all except
12 4% to groups of less than 100. The number of projects actually performed in a year was also low, with 32% stating
13 that 5 or fewer projects were being performed, 53% less than 10 projects, while only 21% were in organisations
14 where 30 or more projects were being undertaken annually. (These low percentages might be caused by a
15 misinterpretation of the question: instead of mentioning the total number of projects performed by the total company,
16 the respondents might have returned the number of project they were involved in themselves). Lastly, it was
17 observed that 39% of the responses came from individuals with 5 or less years of experience in the field and 32%
18 with 10 or more.

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

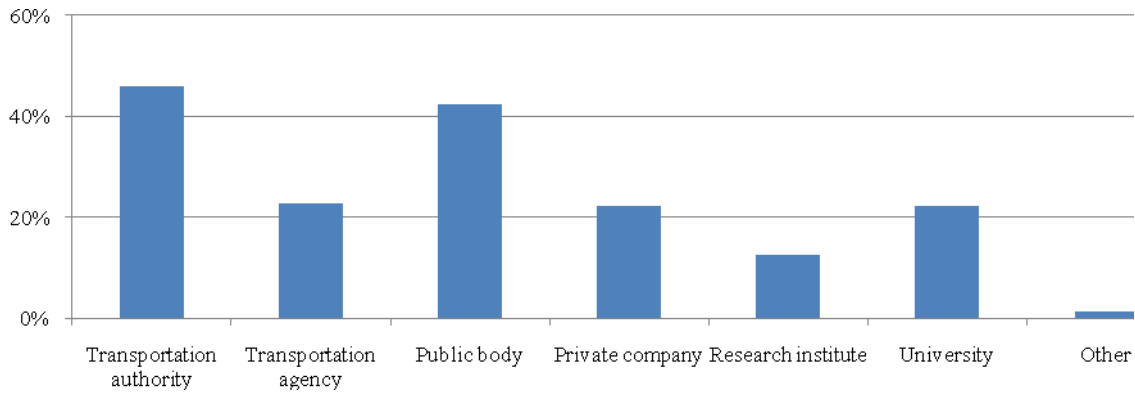
37 ANALYSES ON LAST MODEL APPLICATION OF RESPONDENTS

38 Questions in the third part of the survey were designed to take a ‘snapshot’ of where the focus currently lies in the
39 simulation world, regarding what models are being used for, and for example where most backing is coming from
40 for work ‘in the field’. Here, respondents were allowed to choose as many applicable answers as they wished, and
41 findings indicated (FIGURE 1a) that while 75% were involved with planning, there was perhaps a surprisingly large
42 number of persons working in/on real-time operations (16%), and, while 93% had a primary interest in assessing
43 throughput as their main impact of interest, a significant minority were also interested in safety (18%) and/or
44 environment (22%).
45

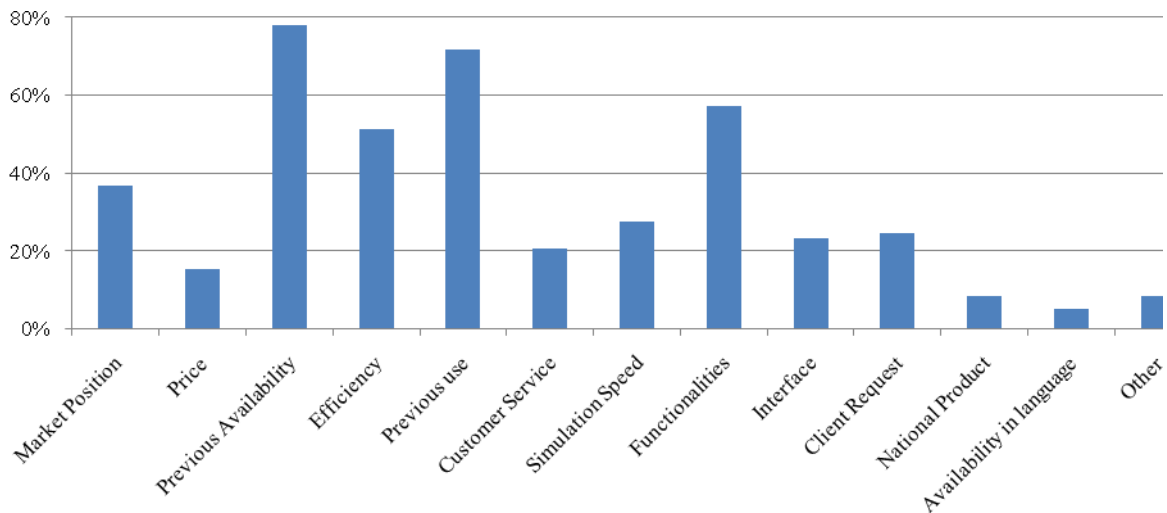


a. Application type.

b. Type of project.



c. Client



d. Model choice motivation.

1
 2 **FIGURE 1 Information on last model application (% of respondents classing their activity in each category,**
 3 **more than 1 choice allowed) by a) Application type, b) Type of project, c) Client, d) Model choice motivation.**
 4

1 While over 64% of the respondents were engaged on government projects and over 34% on commercial projects, the
 2 29% response for research projects (FIGURE 1b) would seem unusual (given that only 10% of those responding
 3 came from such institutes), and, while the majority of clients were private and/or government, a significant minority
 4 (10-20%) were either Universities or Research Institutes themselves (FIGURE 1c).

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

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

22 ANALYSES OF SIMULATION MODEL USE IN LAST APPLICATION

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

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

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

- 40 • A 'Warm up period' was found to be used by 76% of the respondents. Since a warm up period is necessary
 41 to fill the network (in reality the network is almost never empty) and to let the system reach a steady state,
 42 in 24% of the replies it could be questioned whether the results of this warm up period are included in the
 43 final results, thus affecting the outcomes of the simulation.
- 44 • For the Duration of the warm up period [<15 , 15-30, 30-60, 60-120, >120 mins], 53% of the respondents
 45 were found to be using periods of 15 minutes or less, and a total of 83% less than 30 minutes. While
 46 seeming quite short, these should be viewed in the context of the overall simulation time, which is
 47 examined later.
- 48 • Simulation duration [<30 , 30-120, >120 mins] was generally found to be 30-120 minutes of simulated time
 49 (64% of the sample), with only 7% undertaking shorter simulation runs, and 29% longer.
- 50 • Regarding Number of runs performed for each simulation scenario [1, 2-5, 6-10, 11-20, >20], 30% of the
 51 respondents were found to perform 2-5 runs for each scenario in order to establish convergence and 35% 6-
 52 10 runs. However, this varied heavily with 17% only performing 1 run, and 18% performing more than 10
 53 runs. Obviously, when performing only a single run when using a stochastic simulation model the results
 54

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

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

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

14 Significant variations were also found in simulation duration, firstly by Region ($p=0.00001$), where the UK
15 was found to focus on longer runs (44%) and the RoW substantially on shorter runs of less than 30 mins (25%).
16 Model was also found to play a role ($p=0.00017$) with PARAMICS users focusing on runs longer than 120 mins
17 (52% of the sub-group) and VISSIM users on periods longer than 30 mins (100%). This might be explained by the
18 fact that PARAMICS users on average had applications with larger networks which would require longer durations
19 and this point will be explored more fully later.

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

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

34 Features of the calibration process (both for traffic flow (TF) and route choice (RC)) were examined as
35 regards the number of variables that the user attempts to calibrate (this is not necessarily the same as the number that
36 the user knows to need calibrating but is instead a number governed by lack of resource or data). For both of these
37 models users most commonly attempted to calibrate 2-5 parameters (70% of users in the case of the TF model, 68%
38 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
39 all other parameters are calibrated and suitable for use. Twenty five percent of the respondents were found to be
40 using automatic optimization procedures for TF calibration and 17% for the RC model, with counts and travel times
41 being the most common types of data being used (85% and 65% of the responses for the TF model and 61% and
42 65% for the RC model, FIGURE 3). Of note is the fact that 10% of TF calibrators and 22% of RC calibrators stated
43 that they were using trajectory data and that 38% for TF and 47% for RC used 'personal experience'. From a
44 scientific point of view, depending on the level of detail sought, for an extensive calibration of the traffic flow
45 model (in more specifically a car following or lane changing model), trajectory data would be very useful, however
46 whether such data is fully and globally applicable considering national variations in driving behaviour is an open
47 question (13). Calibrating a model on personal experience is obviously very qualitative and does not give any
48 guarantee for accurate simulation results.

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

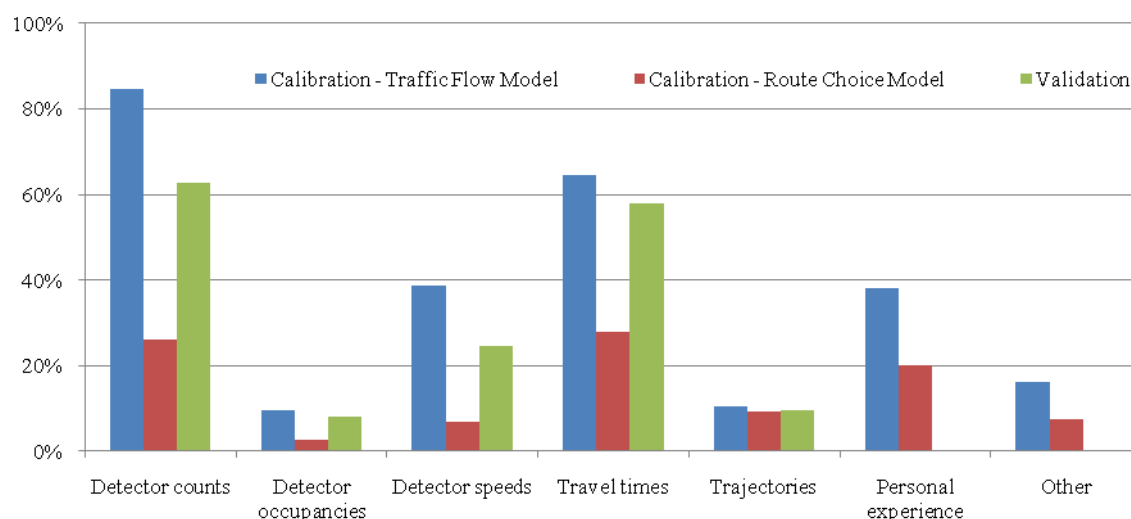


FIGURE 2 Percentage of respondents using differing types of data in the Calibration of Traffic Model, Calibration of Route Choice model and, Validation.

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).

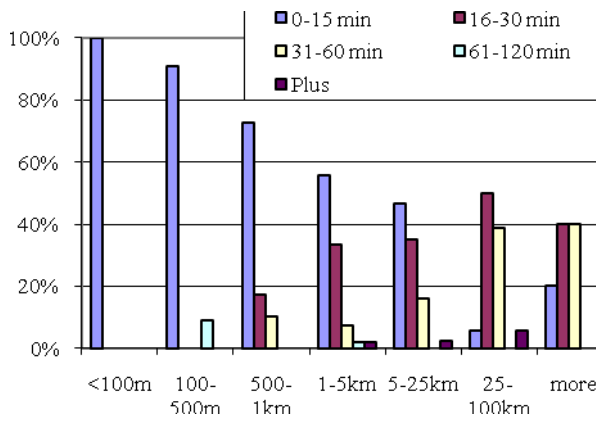
Firstly, it is highly likely that simulation duration and other practical issues that may affect the logistics of performing a simulation project on time and within budget, are related. Here, one finds that the fraction of respondents using longer warm up times increases with Network size (FIGURE 3a), 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 undertaken both, 9% neither, 14% Calibration but not Validation, and 10% Validation but not Calibration).

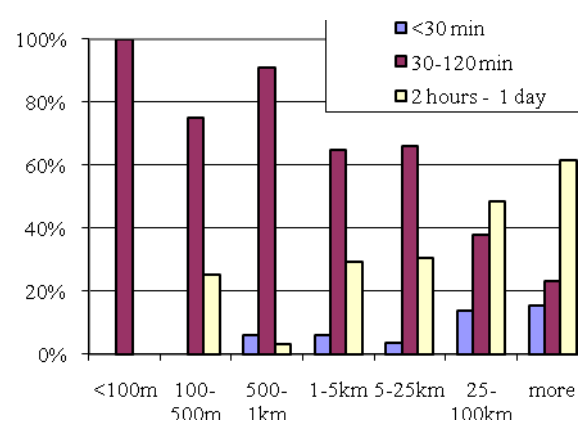
Examining the Number of runs and Network size (FIGURE 3e) one can observe a very wide spread including a large number that only perform 1 run for very large networks. Putting such respondents to one side and assuming that their responses are driven by particular needs/resource requirements, one can see 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.

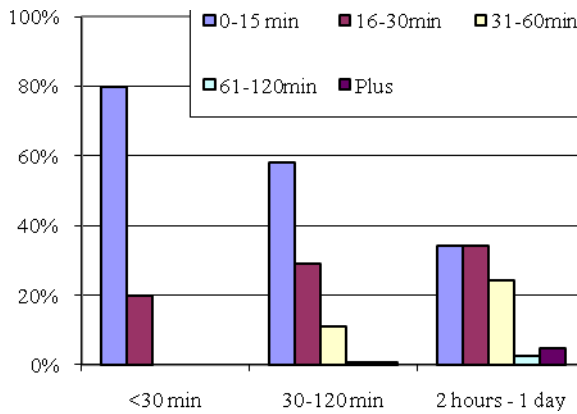
1



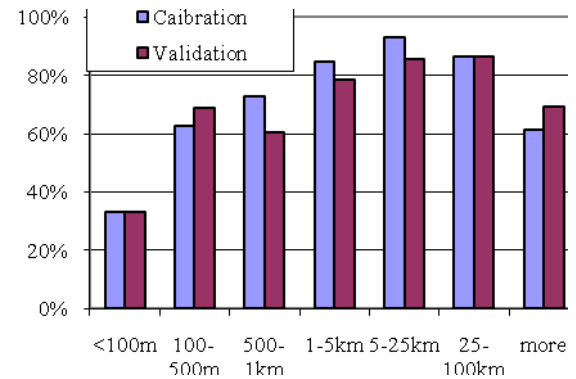
a. Duration of warm up runs by network size



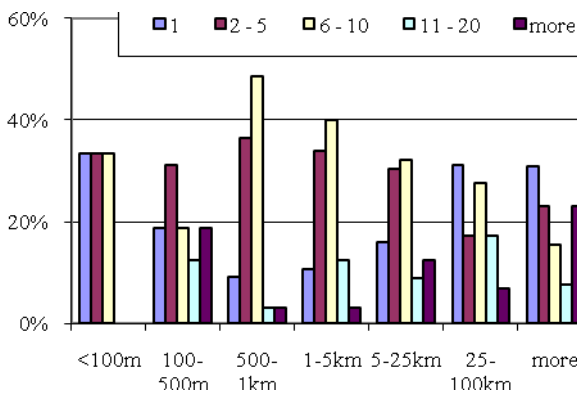
b. Duration of simulation runs by network size



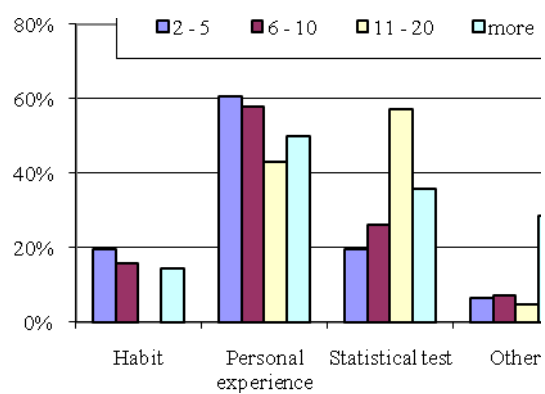
c. Duration of simulation runs by duration of associated warm up period



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

4

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

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

3 **CONCLUSIONS AND RECOMMENDATIONS**

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

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

35
36 Although providing a large number of results, the survey is not without its weaknesses and a number of factors
37 restrict the (detailed) validity of our results. For example, during the analysis questions arose regarding how the
38 respondent interpreted the questions and this is a confounding factor that is, to a certain extent, impossible to
39 overcome. Additionally it is clear that there is a bias within the survey toward respondents within the UK (and using
40 the PARAMICS model), and this (and the associated guidelines likely to be used by them) may have skewed many
41 of our findings. Additionally it is possible to question whether respondent type is statistically representative of the
42 actual distribution of model users, for each of the different models. While these criticisms may affect the magnitude
43 of our results, the contrasts observed however are sufficiently strong to indicate that there are issues that can be
44 clearly identified and these include:

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

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

19 ACKNOWLEDGMENTS

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

24 REFERENCES

- 25 [1] Papageorgiou, M. Some remarks on macroscopic traffic flow modelling, *Transportation Research Part A*,
26 Vol. 32, No. 5, 1998, pp. 323-329.
- 27 [2] Kleijnen, J.P.C. Verification and validation of simulation models, *European Journal of Operational*
28 *Research*, Vol. 82, 1995, pp. 145-162.
- 29 [3] National Cooperative Highway Research Program (NCHRP 3-85): "Guidance for the Use of Alternative
30 Traffic Analysis Tools in Highway Capacity Analyses", Presentation at Transportation Research Board
31 Annual Meeting, Washington D.C., Jan., 2007.
- 32 [4] MULTITUDE. *Introduction to MULTITUDE*. <http://www.multitude-project.eu/>. Accessed July 6, 2011.
- 33 [5] COST. *About COST*. http://www.cost.esf.org/about_cost. European Cooperation in Science and Technology.
34 Accessed July 6, 2011.
- 35 [6] COST. *COST Countries*. http://www.cost.eu/about_cost/cost_countries. Last accessed July 9, 2011.
- 36 [7] Federal Highway Administration. http://ops.fhwa.dot.gov/trafficanalysistools/tat_vol2/sectapp_e.htm#top.
37 Accessed July 17, 2011.
- 38 [8] Brackstone, M. and M. McDonald. Car-following: a historical review, *Transportation Research Part F:*
39 *Traffic Psychology and Behaviour*, Vol. 2, No. 4, 1999, pp 181-196.
- 40 [9] Dodge, Y. *The Oxford Dictionary of Statistical Terms*, OUP, 2003.
- 41 [10] Siegel, S. and N.J. Castellan. *Nonparametric statistics for the behavioral sciences*, 2nd edition. McGraw-Hill,
42 London, 1988.
- 43 [11] Vortisch, P. *National Guidelines for Microsimulation*. Presented at 3rd MULTITUDE Meeting, Haifa, Israel,
44 6-7 December 2010. <http://multitude-project.eu/meetings.html>. Accessed 26 July 2011.
- 45 [12] *Guidelines for the Use of Microsimulation Software*. Highways Agency. July 2007.
46 http://www.highways.gov/knowledge_compendium/assets/documents/Portfolio/Guidelines - 901.pdf.
47 Accessed 26 July, 2011.
- 48 [13] Marsden, G., M. McDonald and M. Brackstone. A Comparative Assessment of Driving Behaviours at Three
49 Sites, *European Journal of Transportation and Infrastructure Research*, Vol. 3, No. 1, 2003, pp 5-20.