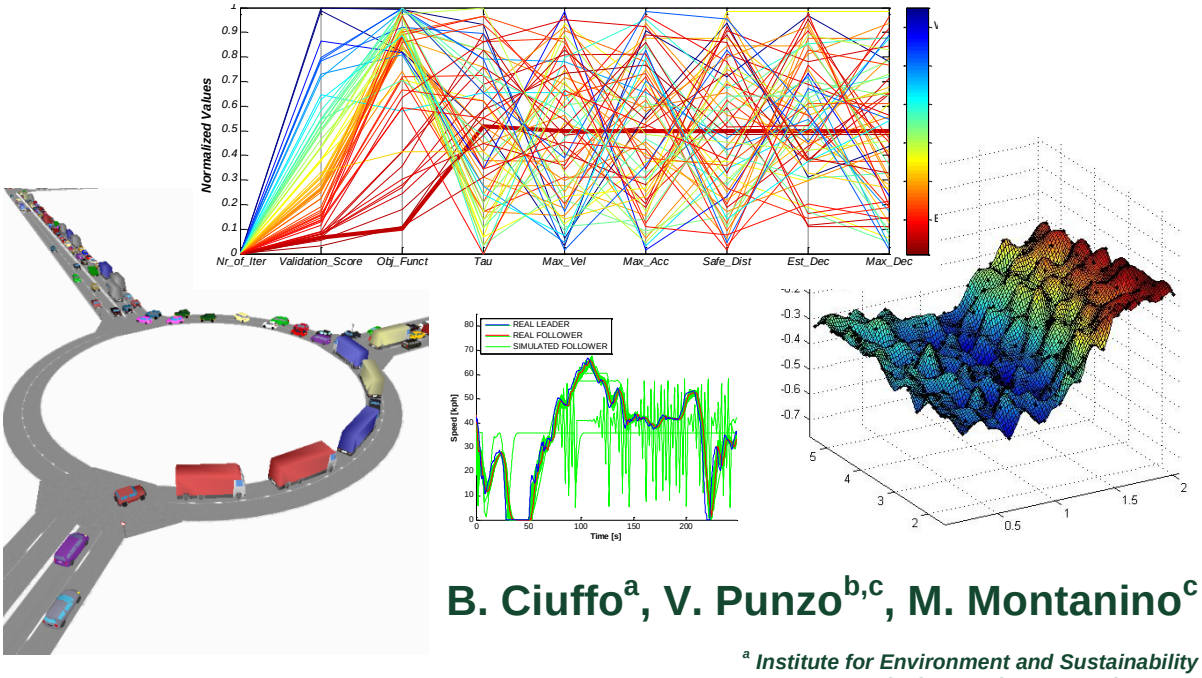
JRC Scientific and Technical Reports

The Calibration of Traffic Simulation Models

Report on the assessment of different Goodness of Fit measures and Optimization Algorithms

MULTITUDE Project – COST Action TU0903



B. Ciuffo[a], V. Punzo[b,c], M. Montanino[c]

[a] Institute for Environment and Sustainability
European Commission – Joint Research Centre

[b] Institute for Energy and Transport
European Commission – Joint Research Centre

[c] Department of Transportation Engineering
University of Napoli Federico II

EUR 25188 EN 2012

JRC EUROPEAN COMMISSION

Institute for Environment and Sustainability

Institute for Energy and Transport

The mission of the JRC-IES is to provide scientific-technical support to the European Union's policies for the protection and sustainable development of the European and global environment.

European Commission
Joint Research Centre
Institute for Environment and Sustainability
Institute for Energy and Transport

Contact information
Biagio Ciuffo
Address: Via Enrico Fermi, 2749 - TP 441 Ispra (VA) - Italy
E-mail: biagio.ciuffo@jrc.ec.europa.eu, biagio.ciuffo@gmail.com
Tel.: +39 (0332) 786782
Fax: +39 (0332) 785236

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

JRC68403

EUR 25188 EN
ISBN 978-92-79-22812-4
ISSN 1831-9424
doi:10.2788/7975

Luxembourg: Publication Office of the European Union, 2012

Printed in Italy

Executive Summary

This report summarizes the last experiences carried out on the calibration of traffic simulation models. This topic is only a specific case of the broader problem of simulation optimization.

In the last decades, simulation optimization has received considerable attention from both researchers and practitioners. Simulation optimization is the process of finding the best values of some decision variables for a system whose performance is evaluated using the output of a simulation model. In the calibration process, the decision variables are the model parameters and the performance to be evaluated is the capability of the model reproducing the reality.

In traffic modelling this topic is particularly relevant. Indeed the capability of a model to correctly estimate the effect of whatever policy or measure on the performance of a transportation system is dramatically affected by the value of the model parameters. Therefore, without a careful model calibration it would not be possible rely on its results. In addition, it is easy to be proven that, the solutions to the methodological issues arising when setting up a calibration study cannot be posed independently. This calls for methodologies able to check the robustness of a calibration framework as well as further investigations of the issue, in order to identify possible "classes" of problems to be treated in a similar way. Therefore in the present work, first a general methodology for verifying a traffic micro-simulation calibration procedure (suitable in general for simulation optimization) is described, based on a test with synthetic data. Then a preliminary analysis, based on the response surface technique, of the different effect that sixteen measures of Goodness of Fit have on the mathematical properties of the objective function included in the optimization problem is presented. Then the methodology of calibration verification is applied to two different case studies to draw inferences on the effect that different combinations of parameters to calibrate, optimization algorithms, measures of Goodness of Fit (seven out of the sixteen preliminarily included in the analysis) and noise in the data may have on the optimization problem.

In one of the case studies, the time required to perform all the simulations needed to the study would have made the present applications unfeasible. For this reason, a Kriging surrogate of the simulation model was used instead.

Results clearly showed how the sources of uncertainty related to the optimization problem strongly influenced the results and thus the importance of verifying the calibration procedure with synthetic data. In addition they ascertain the need for global optimization solutions, giving new insights into the topic.

Research contained within this paper benefited from the participation in EU COST Action TU0903 MULTITUDE – Methods and tools for supporting the use calibration and validation of traffic simulation models.

Contents

List of Tables

List of Figures

1 Background and introduction

In the last decades simulation optimization has received considerable attention from both researchers and practitioners. Simulation optimization consists in the use of simulation models to maximize the efficiency of the real system which is simulated. In other words, it is the process of finding the best values of some decision variables for a system whose performance is evaluated using the output of a simulation model (Olafsson et al., 2002). After all, "we have the reason to believe that the system, whether natural or artificial, is governed by rules which we have the ambitious to uncover, or to use to our advantage" (Saltelli et al., 2008).

In simulation optimization a simulation model is then jointly used with a mathematical programming model. However it is common opinion that a gap exists between academic theory and commercial practice, especially in software implementation of optimization approaches (Fu, 2002, Fu et al., 2008) in several application and scientific fields.

In traffic applications, simulation optimization has been used in several contexts. These include wide area signal optimization (Gartner et al., 1991, Sanchez-Medina et al., 2010), pricing (Gardner et al., 2010), work zone configuration (Chen et al., 2005), traffic optimization (Chang and Li, 2002, Stevanovic et al., 2008) and models calibration, only to provide few examples. However, these different applications have been usually treated separately without considering that they all pertain to the same field of the simulation optimization. Furthermore, they have only occasionally focused on the optimization side of the problem (e.g. to answer to various methodological questions as which is the best approach to be followed and why, how good is the solution found in the space of the possible solutions, global vs. local optimization and so on). In this way the several studies carried out have only marginally contributed to the enrichment of the state-of-the-art on this topic and to increase the consciousness about the complexity of the problem.

Complexity of such problems is instead well known in optimisation theory and is enlightened by so-called "no free lunch" theorems (Spall et al., 2006, Wolpert and MacReady, 1997): they basically state that "the expected performance of any pair of optimization algorithms across all possible problems is identical", that is to say that there is no algorithm that outperforms the others over the entire domain of problems. In other words, the choice of the most appropriate algorithm depends upon the specific problem under investigation. As a consequence, since setting up an optimization problem means tackling a number of methodological issues, solutions to such questions can be posed neither independently of each other nor of the optimization algorithm. We are referring, in

particular for the application to transportation studies, to the choice of i) the measure of performance (MoP) that depicts traffic behaviour, iii) the location and the traffic conditions for data collection, iv) the sub-set of parameters whose optimal value is searched, given v) the specific traffic simulation model applied, as well as vi) the traffic scenario to be simulated (demand pattern, network, etc.). In the subject of micro-simulation model calibration this dependence has been recently pointed out by Punzo and Ciuffo (2009, 2010) who showed, for instance, how the sub-set of sensitive parameters to calibrate depends upon the chosen measure of performance

As a consequence of this complexity, most studies did not investigate these issues in depth, mainly focusing on the performance of the optimization algorithm. More importantly, in such studies algorithm performances are evaluated in terms of convergence time or model fitting by directly calibrating models against real traffic data (see e.g. Hollander and Liu, 2008b, Balakrishna et al., 2007, Hourdakis et al., 2003, Ben Akiva et al., 2004, Dowling et al., 2004, Vaze et al., 2009). However, calibrating models against real traffic data does not allow the effectiveness of the algorithm or of the whole calibration procedure to be evaluated. Indeed, given unavoidable modelling errors in the data as well as in the inputs (e.g. the O/D demand), such distance may never vanish. As a result, on looking at the solution found by an algorithm, there is hardly anything to be said about the correctness of the optimization problem posed, or about the effectiveness of the procedure applied. Whether it is a global or local minimum, whether calibrated values are really those which best capture drivers' behaviour, is not easy to recognize, and comparative evaluations are the only ones that can be confidently carried out. The only attempt to understand how far the minimum found was from the true solution is (Ma et al., 2007), in which the authors performed a preliminary exhaustive search on the domain of the two parameters considered in order to find the global solution of the calibration problem and then used this information to compare the performances of two optimization algorithms. However, the preliminary exhaustive search to find the global minimum of the function was quite expensive to be carried out, leaving in any case some probabilities of errors in the definition of the global solution (using real data the global solution is unknown and also an exhaustive search can leave some uncertainty).

By contrast, the use of synthetic measurements, that is of data obtained from the model itself, allows all the above errors to be removed and the effectiveness of the calibration procedure to be verified. Such an approach is adopted e.g. by (Ossen and Hoogendoorn, 2008), who test the effects on car-following model calibration results of applying errors to synthetic vehicle trajectories, or by (Vaze et al., 2009), who compare the performances of different optimization algorithms on a synthetic network. (Ciuffo et al., 2007) proposed the use of tests with synthetic data as a way to drive the set-

up of traffic micro-simulation calibration procedures, in a trial-and-error fashion, while in (Ciuffo et al., 2008) synthetic data were applied to compare model-based and simulation-based calibration of traffic micro-simulation models.

In this paper, the use of tests with synthetic data is proposed to verify the effectiveness of the calibration procedure. The choice of the goodness-of-fit function and of the optimization algorithm to be used are fundamental, because the former defines the mathematical properties of the objective function, while from the latter depends the possibility to solve the optimization problem. Yet, to the best of our knowledge, it is hard to find studies dealing with these issues jointly. In (Hollander and Liu, 2008a) while the authors present an exhaustive list of the GoF used in traffic micro-simulation calibration tests, also offering some tips for their use, thorough investigation seems lacking. For this reason, considering the MoP and the sensitive parameters of the AIMSUN 6.0 micro-simulation model (TSS, 2008) obtained in Punzo and Ciuffo (2009), in this paper, using the same case study, we first use synthetic measurements with and without the addition of synthetic noise to understand the impact of choosing different GoF measures (we analyzed sixteen different GoF among the most commonly used) on the objective function of the calibration problem and to study properties of the calibration problem itself.

Then, we investigate the joint effect on the calibration problem of using 7 different GoFs (those found suitable at the previous step) with 5 different optimization algorithms. The study has been carried out using the synthetic measurements without and with two different levels of artificial noise. The time required to perform all the simulations needed to the study would have made the present applications unfeasible. For this reason, a Kriging surrogate of the simulation model was used instead. In the paper we also provide details about the GoF measures and the optimization algorithms selected and about the kriging meta-model.

Then we adopted the same approach to draw inference about the uncertainties connected to the calibration of a well known car following model (i.e. the Gipps model, Gipps, 1981) against vehicle trajectory data, and to question the reliability of any further analysis, based on these results, directed to depict the properties or the inadequacy of the model itself

It is worth underlying that the approach followed throughout this paper for the calibration of a microscopic traffic simulation model, can be applied in the same way to the analysis of whatever simulation optimization study.

1.1 Verification approach for optimization simulation in traffic micro-simulation applications

As already anticipated, in simulation optimization one looks for best values of some decision variables for a system whose performance is evaluated using the output of a simulation model. In this light it can be formulated in the following optimization:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} f\left(\boldsymbol{M}^{sim}\right) \tag{1}$$

possibly subject to the following constraints:

$$\boldsymbol{l}_{\boldsymbol{\beta},i} \leq \boldsymbol{\beta}_i \leq \boldsymbol{u}_{\boldsymbol{\beta},i} \quad i = 1 \dots m$$

$$\boldsymbol{l}_{\boldsymbol{\gamma},i} \leq \boldsymbol{\gamma}_i \leq \boldsymbol{u}_{\boldsymbol{\gamma},i} \quad i = 1 \dots m$$

and potentially also to other constraints:

$$\boldsymbol{\beta}_i \leq \boldsymbol{\beta}_j \quad \boldsymbol{i} \neq \boldsymbol{j}$$

$$\boldsymbol{\gamma}_i \leq \boldsymbol{\gamma}_j \quad \boldsymbol{i} \neq \boldsymbol{j}$$

where $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$ are, respectively, the vectors of continuous and discrete decision variables, potentially belonging to *m* different classes of simulation subjects (e.g. different vehicles classes in the traffic micro-simulation); $f(\cdot)$ is the objective function (or fitness or loss function) to be maximized, which is function of the vector of the simulated traffic measurements $\boldsymbol{M}^{sim}$; $\boldsymbol{l}_{\boldsymbol{\beta},i}$, $\boldsymbol{l}_{\boldsymbol{\gamma},i}$, $\boldsymbol{u}_{\boldsymbol{\beta},i}$, $\boldsymbol{u}_{\boldsymbol{\gamma},i}$, are model parameter lower and upper bounds. $\boldsymbol{M}^{sim} = S(\boldsymbol{D_1}, \dots, \boldsymbol{D_H}, \boldsymbol{TS}, \boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_m}, \boldsymbol{\gamma_1}, \dots, \boldsymbol{\gamma_m})$, in which $S(\cdot)$ is the micro-simulation model, $\boldsymbol{D_h}$(*h=1,…,H*) the vectors of OD flows departing in interval *h* and **TS** is the vector of the transportation supply system characteristics.

In the case of model calibration, equation 1 becomes:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} f\left(\boldsymbol{M}^{obs}, \boldsymbol{M}^{sim}\right) \tag{2}$$

in which the objective function to be minimized, $f(\cdot)$, measures the distance between the simulated and the observed and traffic measurements, $\boldsymbol{M}^{obs}$ and the decision variables of the model are the parameters to be calibrated.

The practical specification of the optimization problem thus relies upon several factors: i) the measure of goodness of fit (GoF) used to create the objective function, ii) the traffic measurements

(i.e. the measure of performance, MoP) and thus where and how they are collected, iii) the quality (in terms of error presence) of the observed data, iv) the traffic simulation model, v) the transportation system (i.e. the traffic scenario), vi) the demand pattern and the accuracy of its estimation, vii) the parameters to be calibrated and, viii) the constraints adopted. Changing one of these elements will result in a modification of the optimization problem. Therefore, also the algorithm to use may be different. Indeed, according to the NFL theorems, "for any algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class" (Spall et al., 2006). This is the reason why a universal recipe cannot yet exist at this stage of our knowledge. On the other hand, further analyses are required at least to identify "classes" of problems which could be tackled approximately in the same way.

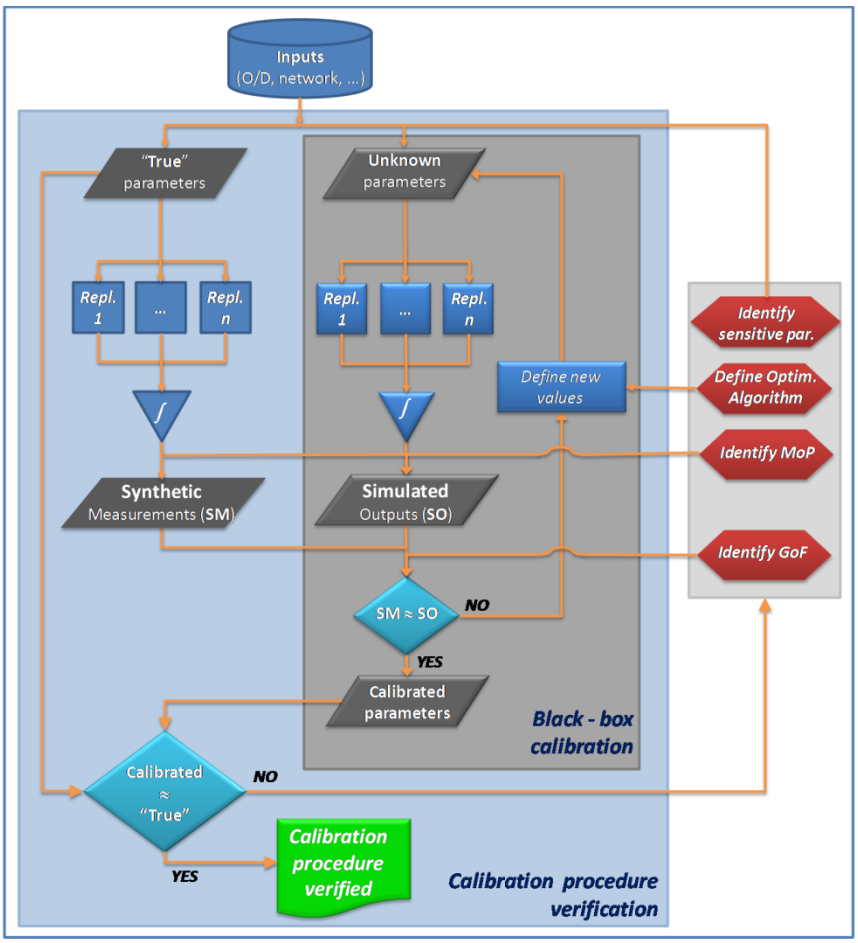


Figure 1. Flow chart of the verification approach to the calibration procedure: parameters calibrated by means of synthetic measurements (see the inner box, black-box calibration) are compared to "true" parameters (i.e. to those which generated the measurements). When the comparison is not satisfactory, the calibration procedure needs amending.

The methodology here presented is developed in the context of such uncertainties and, in the light of the results achieved with past research, is presented in Figure 1. For a given traffic scenario (i.e.

having defined the transportation network and the corresponding demand pattern), once the sensitive parameters of the model have been identified (by means of a sensitivity analysis of the model as shown in Punzo and Ciuffo, 2009) the calibration verification procedure can start. This verification involves at least the following variables: the MoP, the GoF and the optimization algorithm. The verification is based, as shown in Figure 1, on a calibration with synthetic data. As already underlined, the effectiveness of a calibration process can be evaluated at best only by means of "controlled" experiments, in which all the model inputs and the "true" values of parameters are known.

Hence, once "realistic" inputs to the simulation model have been assumed (i.e. time varying O/D matrices, realistic model parameter values, etc.), the model is run and outputs from several replications (each using a different seed in the model random numbers generation process) are averaged to reduce the effects of stochasticity, thus becoming the synthetic "true" measurements. The parameters used in the simulation are therefore our "true" parameters. The black-box calibration then tries to find the parameter values that minimize the distance between true and simulated outputs (see Figure 1). The calibration then "verifies" whether at the end of the process the calibrated parameters equal the "true" ones. Should they not, then the calibration is repeated by changing one of the three variables and possibly also the sensitive parameters considered.

This methodology, described here for the calibration of a traffic simulation model, can be applied in the same way in any simulation optimization application.

The main problem of the proposed methodology is the large amount of time required by each attempt to calibrate the traffic model. This is why it is not realistic to evaluate all the possible combinations of GoF/algorithm/MoP etc. As regards the MoP and the parameters, a model sensitivity analysis is a fundamental preliminary step to drive the selection of the parameters to calibrate and choose the measure of performance to use. Regarding the GoF and the optimization algorithm, the choice is much harder, since there are many possible alternatives and very few elements to use to guide one's choice. Hence in the following section a list of the possible GoFs is reported (also mentioning the calibration experiments in which they were used) together with a possible method for evaluating their suitability as fitness functions of the calibration process. This will lead to a first reduction in the choice of possible GoF measures. However, also in this case, the analysis of all the remaining combinations GoF/algorithm would lead to an unfeasible number of simulations. For this reason, the verification of the calibration procedures has been carried out on a surrogate of the simulation model, namely on its Kriging approximation. As will be detailed in the

remainder, this choice is widely acceptable in this field, as Kriging meta-models are considered as reliable tools used for the calibration of expensive simulation model. For the moment, we do not aim here at calibrating the model using Kriging, but only at verifying the effectiveness of a calibration procedure, thus requiring a lower degree of precision.

The paper is structured as follows: in the next two sections the goodness of fit measures and the analysis to study their reliability in the calibration process are described. Then the basics of the optimization algorithms chosen and of Kriging models are provided. The description of the case study and of the analyses carried out, together with the results achieved will then precede the conclusive section of the paper.

2 Goodness-of-fit measures in the calibration problem

In the following, the goodness-of-fit measures used to compare real and simulated traffic measurement in this study is presented. The selection is made on the basis of the results obtained in Ciuffo and Punzo (2010), which also propose a state-of-the-art of the possible options used so far. In brackets are indicated previous works which have used it in the model calibration.

In all the following expressions, *x* and *y* represent, respectively, the simulated and true traffic measures (M^{sim} and M^{obs}) used in the calibration problem, while *N* represents the total amount of available data (considering all the possible dimensions, spatial, temporal, etc.) and *i* represents the generic single observation. In addition, $\bar{x}$, σ_x, $\bar{y}$ and σ_y represent, respectively, the mean and the standard deviation of simulated and true measurements, F_x and F_y the cumulative probability density function for *x* and *y*, while *X* and *Y* represent the area of the speed-flow diagram respectively covered by simulated and true data (for further explanations see Menneni et al., 2008).

1. Squared Error SE (Jayakrishna et al., 2001, Toledo et al., 2003, Hourdakis et al., 2003, Ben Akiva et al., 2004, Dowling et al., 2004, Ciuffo et al., 2008)

$$SE(x,y) = \sum_{i=1}^{N} (x_i - y_i)^2 \tag{3}$$

1. Mean Error ME

$$ME(x,y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i) \tag{4}$$

2. Mean Normalized Error MNE

$$MNE(x,y) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - y_i)}{y_i} \quad (5)$$

3. Mean Absolute Error MAE (Ma and Abdulhai, 2001)

$$MAE(x,y) = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \quad (6)$$

4. Mean Absolute Normalized Error MANE (Ma and Abdulhai, 2001, Schultz and Rilett, 2005, Kim et al., 2005, Lee and Ozbay, 2009)

$$MANE(x,y) = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - y_i|}{y_i} \quad (7)$$

5. Root Mean Square Error RMSE (Punzo and Tripodi, 2007, Punzo and Simonelli, 2005; Ciuffo et al., 2008; Ciuffo et al., 2010; Punzo et al., 2011)

$$RMSE(x,y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \quad (8)$$

6. Root Mean Squared Normalized Error RMSNE (Ma and Abdulhai, 2001, Brockfeld et al., 2005)

$$RMSNE(x,y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - y_i}{y_i}\right)^2} \quad (9)$$

7. Customized version of the Geoffrey E. Havers statistic GEH (Ma et al., 2007, Ciuffo et al., 2010; Punzo et al., 2011)

$$-GEH_k(x,y) = -\frac{1}{N}\sum_{i=1}^{N}\delta_i \quad (10)$$

with $\delta_i = \begin{cases} 1 \; if \; GEH_i(x,y) \le k \\ 0 \qquad elsewhere \end{cases}$ where $GEH_i(x,y) = \sqrt{\frac{2\cdot(x_i - y_i)^2}{x_i + y_i}}$

8. Correlation Coefficient *r*

$$r(x,y) = \frac{1}{N-1}\cdot\sum_{i=1}^{N}\frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x\sigma_y} \quad (11)$$

9. Theil's Bias Proportion U_m

$$U_m(x,y) = \frac{N\cdot(\bar{x} - \bar{y})^2}{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (12)$$

10. Theil's Variance Proportion U_s

$$U_s(x,y) = \frac{N \cdot \left(\sigma_x - \sigma_y\right)^2}{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{13}$$

11. Theil's Covariance Proportion U_c

$$U_c(x,y) = \frac{2 \cdot (1-r) \cdot N \cdot \sigma_x \sigma_y}{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{14}$$

12. Theil's Inequality Coefficient *U* (Punzo and Simonelli, 2005; Brockfeld et al., 2004; Brockfeld et al., 2005; Ossen et al., 2008; Ossen et al., 2009; Kim et al., 2011, Ma et al., 2002; Ciuffo et al., 2010; Punzo et al., 2011)

$$U(x,y) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} {x_i}^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N} {y_i}^2}} \tag{15}$$

13. Kolmogorov-Smirnov Test *KS* (Kim et al., 2005, Hollander and Liu, 2008b)

$$KS(x,y) = \max_{x,y}\left(\left|F_x - F_y\right|\right) \tag{16}$$

14. Speed-Flow Graph *SFG* (Menneni et al., 2008)

$$SFG(x,y) = Y - (Y \cap X) \tag{17}$$

The following observations can be made on these GoF measures:

a. the most widely used GoF is the Squared Error (SE). Like other estimators (e.g. RMSE and U) SE uses the squared difference between simulated and real measurements to measure goodness of fit. Low values show a good fit. It strongly penalizes large errors. It is the estimator at the basis of the famous Least Squares Method, which, according to the Gauss-Markov theorem (Plackett, 1950), provides the best parameter estimation for linear models with zero-mean, unbiased and uncorrelated errors. There is no other apparent reason to prefer SE to the other GoF measures like RMSE and U as a fitness function of a traffic model calibration problem (which is not a linear model);

b. Mean Error (ME) and Mean Normalized Error (MNE) are useful to indicate the presence of systematic bias, but cannot be used in the calibration because low values do not ensure a good fit (the same high errors with opposite sign will result in zero ME and MNE);

c. using absolute values (as in MAE and MANE) would result in using the same weight for all errors, while it would be preferable to assign more importance to high errors than to small

ones. In addition, the gradient of the absolute value analytical function has a discontinuity point in zero. Despite this, the MANE is the second most widely used GoF;

d. normalized measures like RMSNE and MANE are very attractive GoFs, since they allow a model to be calibrated using different measures of performance (only relative error is considered). However instabilities due to low values among the measurements in the fraction's denominator advise against their use (e.g. if the traffic counts during a certain time interval are very low, even small absolute errors can result in high relative errors, unlike what happens when the traffic counts are very high, which is undesirable);

e. unlike all the others, the GEH statistic is usually not evaluated over a series of data, but over a single pair of observed and simulated measurements. According to the guidelines provided by English engineers (Highway Agency, 1996), if the value of the GEH is less than 5 for 75% of the pairs, then the two series of data show a good fit. In Ma et al. (2007), GEH is used as a fitness function for a traffic model calibration by summing all the values obtained from all the pairs of measured and observed data. Here we propose also an alternative way in which GEH is used counting the number of times its value is below a certain threshold *k*, be divided by the total number of observations. In this way the indicator would be bounded between 1 (perfect fit) and 0 (worst fit). In the analyses we considered three different values for k, namely 1, 3 and 5 to check also its sensitivity on the results;

f. to be used as a fitness function within the framework of an optimization problem, some of the GoF measures shown have to be considered with the opposite sign (namely r, Uc, and GEH if considered with the meaning suggested in the previous point);

g. the Kolmogorov-Smirnov test can also be a very suitable GoF [26]. However, it would probably require more detailed traffic measurements. In particular, in both [21, 26] it was used on vehicle point-to-point travel times. Although such data are becoming increasingly available, at present the test does not represent a real possibility for most traffic practitioners. For this reason the KS, albeit very interesting, is not included in our analysis;

h. the SFG has also been removed from the list of GoFs to analyze. Though potentially interesting, the authors are still discussing the best way of implementation (in particular, the effect of the resolution chosen to create the raster image from the speed-flow plot remains unclear).

In conclusion, the measures of goodness of fit included in our analysis consist in the first 13 listed above plus three additional ones concerning the GEH in the revised form mentioned above in point *f.*

3 Response surfaces: a visual approach

The methodology applied to draw inferences about the effect of using different GoFs in the objective function of a traffic model calibration problem follows the logic adopted in the methodology used for the calibration itself. In particular, what is desirable here is to study exactly the response surface that will be used in the calibration test with synthetic data. Of course, studying a function without knowing anything about its analytical formulation is not that simple, but some useful information can be retrieved by plotting in a 3D space an approximation of its shape obtained by evaluating it on as many as possible discrete points and then interpolating them. The main problem is that the objective function should depend at most on two parameters. In other words, the calibration should be performed on two parameters only. Of course, this is usually not the case, there being usually more than two sensitive parameters in traffic micro-simulation models. However, as it will mentioned in the remainder, in Punzo and Ciuffo (2009) the results of the model sensitivity analysis suggested the possibility of splitting the calibration of the four most important parameters of the AIMSUN microscopic traffic simulation model (TSS, 2008) into two sub-problems in which each pair of parameters is calibrated independently on different sets of data. A more detailed description of the results obtained in Punzo and Ciuffo (2009) is provided below prior to describing the application performed.

Once the plots have been created, information can be retrieved concerning the apparent *smoothness* of the function, its *linearity*, the existence of different *local minima* as well as the existence of a wide area with *constant value* of the objective function. All these elements can be useful to understand how complex and expensive (in terms of computation efforts) the algorithm to implement should be. In addition, as will be detailed in the implementation, the plots are also used to test the above-mentioned hypothesis of subdividing the calibration problem into two sub-problems and also to test the effect of random noise on the observed measurements.

4 Optimization algorithms for the calibration of microscopic traffic simulation models

As anticipated in the previous section, five different optimization algorithms have been tested in this study. They were chosen because, for different reasons, they have already been used in the

calibration of microscopic traffic simulation models, all presenting features that theoretically suggest their use for this kind of application. Pertaining to different classes of algorithms, they allowed highlighting the existing correlations between the performance of whatever algorithm and the optimization problem. All the algorithms considered here use only objective function measurements, being unavailable any direct measurements of the gradient of the objective function (in our black-box optimization problem). In this section, we will outline the main features of the employed optimization algorithms, leaving to the interested reader the possibility to use the references cited throughout the paper.

4.1 Simultaneous Perturbation Stochastic Approximation (SPSA_I and SPSA_II) Algorithms

One optimization method that has attracted considerable attention is the simultaneous perturbation stochastic approximation (SPSA) method. It has been first developed by James Spall (see Spall, 1992, 1998, 2000, 2003, 2009 as references). It is a stochastic analogue of the Newton/Raphson algorithm of deterministic nonlinear programming and basically extends the Kiefer and Wolfowitz version of the stochastic approximation (SA) method (Kiefer and Wolfowitz, 1952) used to find extrema of functions in the presence of noisy measurements. The SA method numerically evaluates the gradient of the function using a standard finite difference gradient approximation (each step of the algorithm thus requiring $2p$ function evaluations, being p the number of variables of the function itself). SPSA improves the SA method since the gradient evaluation is based on only 2 measurements of the objective function (at each step of the procedure) in which all the variables of the function are simultaneously perturbated (and not one at time as in SA). As a consequence, SPSA has two features particularly relevant in the calibration of micro-simulation traffic models: i) it explicitly takes into account the presence of measurement errors in the objective function and ii) it is usually less expensive than many other algorithms. In addition, in the last decade, the performances of the algorithm have been mathematically demonstrated in the form of an extensive convergence theory, both for local optimization (Spall, 1992) and global optimization (Maryak and Chin, 2008). This latter case is particularly interesting for the calibration of microscopic traffic simulation models (in Ciuffo and Punzo, 2010 the authors have demonstrated that this calibration problem has to be seen in the framework of global optimization). More in details, in (Maryak and Chin, 2008) the authors have shown the strict convergence of the SPSA algorithm, using the method of the noise injection and a convergence in probability without the application of this method. This means that SPSA could be

suitable also in global optimization, but that this should be tested on the specific case under evaluation.

In the present paper two version of the SPSA have been tested, the standard first order SPSA (Spall, 1992, 1998) and the more recently developed second order SPSA (Spall, 2009). Since the SPSA_I has been already described in other transportation studies (see e.g. Ma et al., 2007, Vaze et al., 2009), we will concentrate our attention on SPSA_II (which however has a similar theoretical background).

The second order SPSA takes its motivation from the fact that in order to improve the quality of the solution found it would be desirable to have information of the Hessian matrix of the function. The evaluation of the Hessian matrix, in black-box optimization, however, is usually considered unfeasible, requiring several additional evaluation of the objective function. In the SPSA_II method the Hessian matrix evaluation follows the same logic used in SPSA_I for the definition of the gradient, i.e. it is numerically derived by the simultaneous perturbation of the elements of the Jacobian matrix. In practice, this means that at each step, SPSA_II requires four (instead of the two of the SPSA_I) noisy evaluations of the objective function which, however, should turn in a considerable improvement of the quality of the solution found. In practice, in SPSA_I, the variables of the function (the parameters of our simulation model) estimated at step *k+1*, $\widehat{\boldsymbol{\theta}}_{k+1}$ are given by

$$\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_k - a_k \widehat{\boldsymbol{g}}_k\left(\widehat{\boldsymbol{\theta}}_k\right) \tag{18}$$

being a_k a non-negative scalar gain coefficient (as defined in Spall, 1992, 1998) and $\widehat{\boldsymbol{g}}_k\left(\widehat{\boldsymbol{\theta}}_k\right)$ the estimation of the gradient of the function at step *k* obtained by simultaneously perturbing all the variables at once (as detailed in Spall, 1992, 1998). In SPSA_II equation 10 is substituted by

$$\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_k - a_k \bar{\bar{\boldsymbol{H}}}_k^{-1} \widehat{\boldsymbol{g}}_k\left(\widehat{\boldsymbol{\theta}}_k\right) \tag{19}$$

with

$$\begin{aligned} \bar{\bar{\boldsymbol{H}}}_k &= \boldsymbol{f}_k(\overline{\boldsymbol{H}}_k) \\ \overline{\boldsymbol{H}}_k &= (1 - w_k)\overline{\boldsymbol{H}}_{k-1} + w_k\left(\widehat{\boldsymbol{H}}_k - \widehat{\boldsymbol{\Psi}}_k\right) \end{aligned} \tag{20}$$

in which (Spall, 2009) $\boldsymbol{f}_k$ is an invertible *p*×*p* matrix (being *p* the number of variables of the function) designed to cope with possible non-invertibility of the (feedback-based) Hessian estimate $\overline{\boldsymbol{H}}_k$, $w_k \in [0,1]$ is a weight to apply to the new input to the recursion for $\overline{\boldsymbol{H}}_k$, $\widehat{\boldsymbol{H}}_k$ is the per-iteration estimate of the Hessian matrix and $\widehat{\boldsymbol{\Psi}}_k$ is the feedback-based term. In this light, equation (20) represents the second improvement of the SPSA_II methodology, since the Hessian matrix used in equation (19) is not that directly estimated by the simultaneous perturbation of the gradient components, but is its per-iteration (feedback-based) weighted average. Going into further details is beyond the aim of the paper (in particular because a detailed explanation is provided in Spall, 2009).

Here it is only important to underline that the complications in the implementation and use of the SPSA_II, introduced by the new formulation, are only apparent, since the parameters to be defined remain few as in the SPSA_I. The only obstacle is that, while for SPSA_I the implementation procedure is described in Spall (1998), this is not the case for SPSA_II and thus some initial fine-tuning is necessary. Finally, an issue which is important to mention is the definition of the algorithm's stopping rule. As already pointed out by Ma et al. (2007), for the calibration of traffic micro-simulation models the use of certain thresholds in the variation of the objective function between two successive steps of the algorithm is not straightforward, but there is also the additional problem of the global optimization and thus one cannot risk to stop the algorithm when it falls in a local minimum. For this reason, it is important to understand how many iterations the specific problem requires and then to define an appropriate limit on it. It is worth knowing that the MATLAB (Mathworks, 2010) codes for SPSA_I and for the identification of its parameters are available on the web site of the author at http://www.jhuapl.edu/spsa/Pages/MATLAB.htm.

For the calibration of traffic micro-simulation models SPSA_I has been adopted in (Balakrishna et al., 2007, Ma et al., 2007, Vaze et al., 2009, Lee and Ozbay, 2009), while SPSA_II has never been tested in this field.

4.2 Simulated Annealing (SA)

Simulated annealing is a well-known method for solving unconstrained and bound-constrained global optimization problems. It is based on the principle of annealing, meaning that the magnitudes of random perturbations (injected within the parameters' domain in a Monte Carlo fashion) are reduced in a controlled manner. This method is designed with the aim to increase the probability of avoiding local minima in path towards the global minimum of the function. The randomness injection (as already pointed out for the SPSA) helps prevent premature convergence, by providing a greater "jumpiness" to the algorithm (Spall, 2003). The term annealing comes from analogies to the controlled cooling of physical substances to achieve a type of optimal state (lower energy configuration) to the substance itself.

The theoretical basis of the simulated annealing algorithm can be found in Kirkpatrik et al. (1983) and in Spall (2003). For transportation application SA has been used in several cases (e.g. Chen et al., 2005 and Gardner et al., 2010). In the present work, the authors have used the general simulated annealing algorithm MATLAB code (Vandekerkhove, 2010), which implements the simulated annealing algorithm described in (Kirkpatrik et al., 1983).

4.3 Genetic Algorithm (GA)

Genetic algorithms are probably the most widely used algorithm type for the calibration of microscopic traffic simulation models. The reason is quite straightforward since no information on the objective function is required for their application (suitable for black-box optimization). Genetic algorithm is a stochastic global search method for solving both constrained and unconstrained optimization problems. It is based on the natural selection, the process that drives biological evolution. Genetic algorithms continuously modify a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population evolves toward an optimal solution. A fundamental difference between GAs and the other algorithms used so far is that they work with a population of potential solutions to the problem. This increases the probability for the algorithm to find the global solution.

The scientific literature on genetic optimization is quite wide and thus it is beyond the aim of this paper to repeat their theoretical fundamentals. For a review of the topic, please refer to Holland (1975) and to Spall (2003). For the calibration of microscopic traffic simulation model, they have been applied several times (see e.g. as guide for their application Ma and Abdulhai, 2001, Schultz and Rilett, 2004, Kim et al., 2005, Ma et al., 2007). In the present work, the authors have used the Genetic Algorithm Toolbox (for use with MATLAB) developed by the University of Sheffield (Chipperfield et al., 2010).

4.4 OptQuest/Multistart algorithm (OptQuest)

The OptQuest/Multistart heuristic (Ugray et al., 2002) is an optimization algorithm for solving both constrained and unconstrained global optimization problems. It has been recently used for the calibration of microscopic traffic simulation model in (Ciuffo et al., 2008).

Basically, the algorithm employs a Scatter Search meta-heuristic (Glover, 1998) to provide starting points for a Generalized Reduced Gradient NLP solver (Smith and Lasdon, 1992, Drud, 1994). In this way it tends to combine the seeking behaviour of gradient-based local NLP solvers with the global optimization abilities of a Scatter Search. In practice, the Scatter Search performs a preliminary exploration in the parameters' domain in order to locate different starting points for the gradient descent (which converges to the "nearest" local solution). Adopting a sufficient number of starting points, the probability to find the global solution of the optimization problem is quite high. The major shortcoming with this approach is in the high number of objective functions evaluations (i.e. traffic simulation) usually required (in particular as the number of parameters to be calibrated increases). In

the present work, the authors have used the OptQuest algorithm implemented in Lindo API (Lindo Systems, 2002).

4.5 Downhill Simplex

The Nelder–Mead method or downhill simplex method or amoeba method was proposed by John Nelder and Roger Mead (Nelder et al., 1965). The Nelder–Mead technique is a gradient-free optimization method, widely used in many traffic optimization studies since 2004 (Brockfeld et al., 2004; Brockfeld et al., 2005; Ossen et al., 2008; Ossen et al., 2009; Kim et al., 2011).

It is a common unconstrained nonlinear optimization technique, since it is a well-defined numerical method for twice differentiable problems. However, the Nelder–Mead technique is only a heuristic, since it can converge to non-stationary points (Powell, 1973; Lagarias et al., 1998; McKinnon, 1999) on problems that can be solved by alternative methods.

5 Kriging meta-modeling to test optimization algorithms performances

As a consequence of the methodology showed in Figure 1, in the present case, per each MoP to be tested, 35 calibration experiments would be required (all the combinations of 5 optimization algorithms with the 7 GoF measures that are resulted suitable in the analysis that will be described later on). In addition, each calibration experiment should be composed by different calibration attempts (randomly changing the starting point of the algorithm) in order to really verify the effectiveness of procedure adopted. Finally, each calibration requires several traffic simulations for the evaluation of the value attained by the objective in different points of the parameters' domain. If we optimistically consider that each simulation takes 5 minutes, that each calibration requires 100 simulations, and if we think to carry out 100 calibrations per each calibration experiment, the time required for evaluating all the combinations, per each MoP, would require more than three years to be performed. This figure represents the major obstacle for the issue of calibrating a traffic micro-simulation model (and of the simulation optimization in general) to be faced, and shows also why hitherto only partial studies have been carried out on it.

However, for the purpose of verifying the effectiveness of a calibration procedure, it could be sufficient to assess that the optimization algorithm is able to solve even a different problem if it maintains the "same" mathematical features than the calibration of the traffic simulation model. This means that if we are able to find, per each MoP/GoF combination, an analytical function which has the same mathematical features of the simulation model, then we could think to use it instead.

Owning the same mathematical properties means that this analytical function, both locally and globally, should have a shape very close to the function given by the simulation model. In other words, we need a function able to approximate the objective function of our black-box optimization problem. A kriging meta-model can be used for this aim. With respect to other meta-models, Kriging models are specifically applied in global applications. This is also attested by the fact that Kriging approximations are used in global optimization to find the minimum value of expensive black-box functions (the literature on this topic is becoming wider and wider, e.g. Jones et a., 1998, Jones, 2001, Huang et al., 2006, Villemonteix et al., 2009). Therefore, if they are successfully applied in the place of simulation models for their optimization, it is possible to claim here that they can be used to verify the capability of the optimization framework implemented (if an algorithm is able to find the global minimum on the Kriging surrogate, it is likely that it will be able to find it on the simulation model as well). In the next section, some details on Kriging meta-models are provided.

5.1 Kriging basics

In spite of its novelty in the transportation field (it has been used only in a couple of studies so far as a spatial predictor for traffic forecasting in Vichiensan et al., 2006, Wang and Kockelman, 2009, but never in simulation optimization) Kriging has become a popular mathematical method in several fields. For this reason, in the following we will provide only some elements to make the reader more familiar with the method, leaving the details to more specific articles and textbooks (the authors suggest to use the recent book of Kleijnen, Kleijnen, 2008, as a reference).

Kriging was first developed in geostatistics by Krige, even if the mathematical formulation was presented some years later by Matheron (1963). A thorough reference in this field can be considered (Stein, 1999). As a surrogate of complex simulation models they have been applied, so far, in particular, for deterministic simulation (Jones et al., 1998, Kleijnen, 2008). Only in the last years they have been applied to stochastic simulation models (Van Beers and Kleijnen, 2003). This is an important point, given the stochastic nature of common traffic simulation models.

The simplest type of Kriging, the Ordinary Kriging (which is considered in this work and which is usually sufficient in practice, see Kleijnen, 2008, Kleijnen, 2009) assumes that the output of a simulation model $w(\boldsymbol{d})$ (being $\boldsymbol{d}$ the vector of the models' variables) can be estimated by:

$$w(\boldsymbol{d}) = \mu + \delta(\boldsymbol{d}) \tag{21}$$

where μ is the simulation output averaged over the whole variables' domain (or at least over the available experimental points) and $\delta(\boldsymbol{d})$ is a zero mean stationary covariance process. It is worth knowing that, while in the Ordinary Kriging μ is a constant value, in the Universal Kriging it is a

regression model. In the case of a stochastic simulation model, it has been seen that equation (21) holds, being $w(\boldsymbol{d})$ the average simulation output over different replications (Kleijnen, 2009).

The output of a simulation model $y(\boldsymbol{d})$ predicted by Kriging for a new variables' combination $\boldsymbol{d}$ is provided by:

$$y(\boldsymbol{d}) = \boldsymbol{\lambda}(\boldsymbol{d}, \boldsymbol{D})'\boldsymbol{w}(\boldsymbol{D}) \tag{22}$$

in which $\boldsymbol{D}$ is matrix of the variables' combination for which the simulation output is known (the vector $\boldsymbol{w}(\boldsymbol{D})$) and $\boldsymbol{\lambda}(\boldsymbol{d}, \boldsymbol{D})$ is the matrix of weights for the new variables combination $\boldsymbol{d}$ estimated using the old ones $\boldsymbol{D}$. $\boldsymbol{\lambda}(\boldsymbol{d}, \boldsymbol{D})$ values are not constant but decrease as the distance between $\boldsymbol{d}$ and $\boldsymbol{D}$ increase (and this is one of the main peculiarities of Kriging with respect of other regression models). The selection of the optimal weights is made using the Best Linear Unbiased Predictor (BLUP) which minimizes the Mean Squared Error of the predictor in equation (22). The solution may be proven to be:

$$\boldsymbol{\lambda_0} = \boldsymbol{\Gamma}^{-1}\left[\boldsymbol{\gamma} + \mathbf{1}\frac{1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}}\right] \tag{23}$$

being $\mathbf{1}$ the n-dimensional identical vector (n is the number of the old variable's combinations in $\boldsymbol{D}$), $\boldsymbol{\Gamma}$ the n×n symmetric and positive semi-definite matrix with the covariances of the old outputs $\boldsymbol{w}(\boldsymbol{D})$ ($cov(w_i, w_j)$ with $i, j = 1, \dots, n$) and $\boldsymbol{\gamma}$ the n-dimensional vector with the covariances between the n old outputs and the output for the variables' combination to be predicted.

Finally, it can be proven that from equations (21), (22) and (23) it is possible to derive

$$y(\boldsymbol{d}^*) = \hat{\mu} + \boldsymbol{\gamma}(\boldsymbol{d}^*)'\boldsymbol{\Gamma}^{-1}(\boldsymbol{w} - \hat{\mu}\mathbf{1}) \tag{24}$$

with

$$\hat{\mu} = (\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{w} \tag{25}$$

In simulation applications, the elements of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ are estimated using a correlation function which is the product of *k* one-dimensional functions (being *k* the number of variables or parameters of the simulation model). In Kriging applications, a popular function is the Gaussian correlation function (which has been used also here in the application). Using it the covariances are calculated as follows:

$$cov(w_i, w_j) = \prod_{g=1}^{k} exp\left[-\theta_g\left(\left|d_{i,g} - d_{j,g}\right|\right)^2\right] \tag{26}$$

in which θ_g is a parameter of the correlation function for the variable *g*, denoting the importance of the variable itself (the higher θ_g is, the less effect the variable *g* has).

In order to find the best Kriging metamodel for a simulation model, it is therefore only necessary to estimate the k-dimensional vector of θ_g. This estimation is performed using a Maximum Likelihood Estimator. Unfortunately the constrained maximization required for this method is a hard problem for several reasons (see Martin and Simpson, 2005, Kleijnen, 2009). This problem can be overcome using the free of charge Matlab toolbox DACE (Lophanev et al., 2002a, 2002b).

The present section was aimed at proving some basics for the Kriging meta-modelling. Some other details will be provided in the description of the application performed. However, for a deeper comprehension of this kind of models the reader is suggested to refer to the sources previously mentioned and available in the references.

6 Case study 1: a freeway scenario in AIMSUN

Analyses were conducted on the same scenario presented in (Punzo and Ciuffo, 2009). The model used is AIMSUN 6.1 (TSS, 2008). The scenario consists of a 22-km network of the Napoli-Salerno E45 freeway (the schematic network layout is reported in Figure 2). The simulation period was in the morning from 7.00 to 10.00 a.m. OD demand input was segmented at one-minute intervals and consisted of two distinct vehicle classes: cars and heavy vehicles.

The freeway is monitored (both in real life and in the simulated scenario) by a number of detectors, located as shown in Figure 2. Along the freeway stretch different traffic conditions coexist. In particular:

- detectors 8561-2 and 15005-7 show congested to saturated conditions;
- detectors 8566-7 show free-flow conditions;
- detectors 8556-7 show vehicles escaping from congestion to free flow.

Time series shown in Figure 2 give a good representation of the existing traffic conditions. In the figure, however, the time series shown are the synthetic measurements that will be used for the verification of the calibration framework. They have been obtained with the AIMSUN model, using the real traffic demand and giving to the parameters of the model some values able to qualitatively reproduce the real conditions. In the figure, there is also shown the MoP that will be used for the AIMSUN calibration (see Punzo and Ciuffo, 2009).

According to the results presented in (Punzo and Ciuffo, 2009), four of the eight parameters explained more than 95% of output variance (as shown in the calibration methodology, the results obtained in the sensitivity analysis work as input to verify the calibration procedure). These parameters are all part of the car-following model used in AISMUN (namely the Gipps' car following

model, Gipps, 1981), and in particular vehicle reaction time (in the following RT), speed acceptance (SA, a proxy of maximum desired speed), maximum acceleration (MA) and the coefficient of variation (CV) of the distributions of all the car-following model parameters (for further details on the analysis, the results and the meaning of each parameter, please refer to Punzo and Ciuffo, 2009).

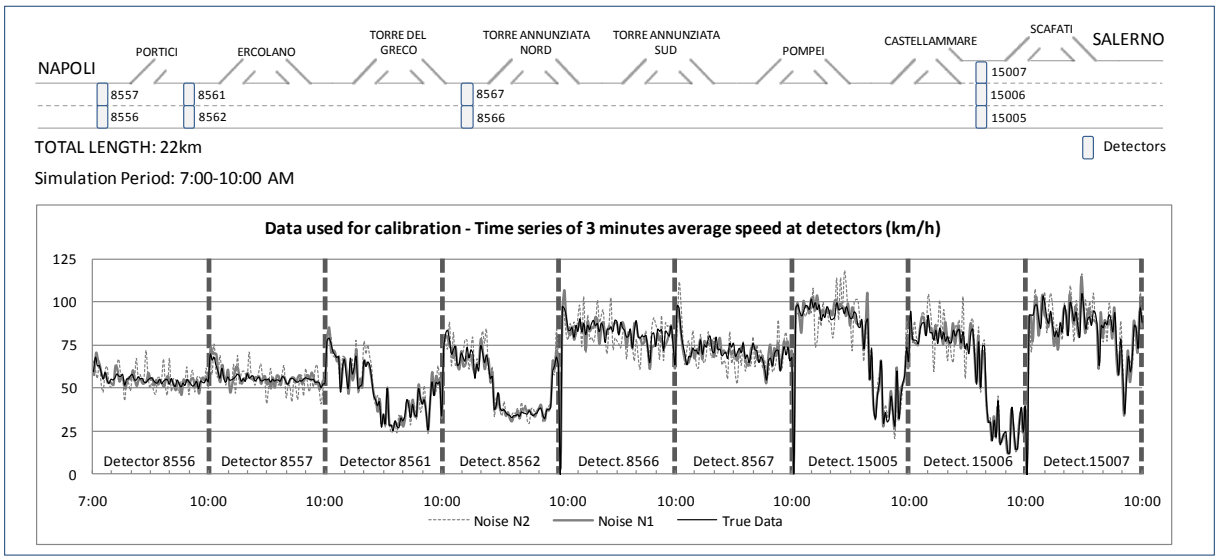


Figure 2. Test site layout and detected traffic conditions. In the graph, per each detector, time series of three minutes average speeds during the simulation period are reported (the black narrow line represents the synthetic measurements without measurement errors, while in the other two lines two different errors are introduced).

The share of the variance explained by the four parameters differed according to measurement location (Punzo and Ciuffo, 2009). This suggested the possibility of calibrating, at a first step, *SA* and *CV* on the measurement locations characterized by free-flow conditions (Pinetina-VMS and Portici on-ramp) and then to calibrate *RT* and *MA* on the data collected at the location characterized by a wide-moving jam (Castellammare off-ramp). Using data from these locations, the two pairs of parameters appeared sufficiently independent (for further details on the results, please refer again to Punzo and Ciuffo, 2009). This result gives the possibility of plotting the objective function of the calibration problem and suggested the idea to perform this preliminary graphical analysis.

6.1 Response surfaces for the case studies

The plots presented in this paper have been created with the following considerations:

- parameters used to create the objective functions were $\boldsymbol{\beta} = \{RT; MA; SA; CV\}$; for such parameter "true" values, lower and upper bounds were selected. In particular $\boldsymbol{\beta}_{true} = \{0.6; 3.5; 1.0; 0.1\}$, $\boldsymbol{l}_{\boldsymbol{\beta}} = \{0.1; 1.5; 0.45; 0.0\}$, $\boldsymbol{u}_{\boldsymbol{\beta}} = \{2.0; 5.5; 1.55; 0.4\}$;
- a "true" scenario was defined as a reference in the evaluation of the objective function

setting the four parameters to their "true" values; the results of the simulation were considered the "true" observations (M_{true}^{obs}) at detectors (i.e. average speed time series over three-minute time intervals);

- two "noised" scenarios were also defined by applying to the simulated outputs of the "true" scenario two random errors, normally distributed with zero mean and standard deviation equal to 5% and 10% of the observations ($M_{noiseN1}^{obs} = M_{true}^{obs} + \varepsilon \sim N\left(0, \left[0.05 \cdot M_{true}^{obs}\right]^2\right)$ and $M_{noiseN2}^{obs} = M_{true}^{obs} + \varepsilon \sim N\left(0, \left[0.1 \cdot M_{true}^{obs}\right]^2\right)$. The resulting time series are reported in FIGURE 2 (labelled as "Noise N1" and "Noise N2"). These scenarios were also created to understand the impact of measurement errors on calibration (i.e. depending on the errors the global minimum could also change);
- objective functions have been evaluated by simulating the model several times, changing the values of two out of the four parameters considered and leaving the others at the true values. In accordance with the results of the sensitivity analysis in Punzo and Ciuffo (2009), the objective functions were evaluated using only data collected at "free-flow" locations for SA and CV, while, with RT and MA using only those collected in one of the location with congested traffic conditions (i.e. detectors 15005-7).
- the results of a simulation are considered here as the average of the results from different replications (using different seeds in the random number generation process of the simulation). Six replications were performed for each simulation in order to reduce the impact of stochasticity with a certain confidence following the indications reported in (Law, 2007).

In conclusion, twelve categories of figures were created (the symbols used here are those explained in the second section):

i. objective functions against *SA* and *CV* using "true" outputs and setting *RT* and *MA* to their true values
$\boldsymbol{GoF_k^{true,true}} = \left\{GoF_{k,i,j}\left(\boldsymbol{M_{true}^{obs}}, \boldsymbol{M_{true,i,j}^{sim}}\right)\right\} \forall\, k = 1, \dots, 16;\ i = 1, \dots, 23;\ j = 1, \dots 17$
$\boldsymbol{M_{true,i,j}^{sim}} = S\left(\boldsymbol{\beta_{true,i,j}}\right)$
$\boldsymbol{\beta_{true,i,j}} = \left\{0.6; 3.5; SA_i; CV_j\right\}$
$SA_i = 0.45 + 0.05 \cdot i \qquad \forall\, i = 1, \dots, 23$
$CV_j = 0 + 0.025 \cdot j \qquad \forall\, j = 1, \dots, 17$

ii. objective functions against *RT* and *MA* using "true" outputs and setting *SA* and *CV* to their true values
$\boldsymbol{GoF_k^{true,true}} = \left\{GoF_{k,i,j}\left(\boldsymbol{M_{true}^{obs}}, \boldsymbol{M_{i,j,true}^{sim}}\right)\right\} \forall\, k = 1, \dots, 16;\ i = 1, \dots, 20;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}} = S(\boldsymbol{\beta}_{\boldsymbol{i,j,true}})$

$\boldsymbol{\beta}_{\boldsymbol{i,j,true}} = \{RT_i; MA_j; 1.0; 0.1\}$

$RT_i = 0.1 + 0.1 \cdot i \qquad \forall\, i = 1, \dots, 20$

$MA_j = 1.5 + 0.25 \cdot j \qquad \forall\, j = 1, \dots, 17$

iii. objective functions against *SA* and *CV* using "noisedN1" outputs and setting *RT* and *MA* to their true values

$\boldsymbol{GoF}^{\boldsymbol{noise1,true}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN1}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 23;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}} = S(\boldsymbol{\beta}_{\boldsymbol{true,i,j}})$

$\boldsymbol{\beta}_{\boldsymbol{true,i,j}} = \{0.6; 3.5; SA_i; CV_j\}$

$SA_i = 0.45 + 0.05 \cdot i \qquad \forall\, i = 1, \dots, 23$

$CV_j = 0 + 0.025 \cdot j \qquad \forall\, j = 1, \dots, 17$

iv. objective functions against *SA* and *CV* using "noisedN2" outputs and setting *RT* and *MA* to their true values

$\boldsymbol{GoF}^{\boldsymbol{noise2,true}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN2}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 23;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}} = S(\boldsymbol{\beta}_{\boldsymbol{true,i,j}})$

$\boldsymbol{\beta}_{\boldsymbol{true,i,j}} = \{0.6; 3.5; SA_i; CV_j\}$

$SA_i = 0.45 + 0.05 \cdot i \qquad \forall\, i = 1, \dots, 23$

$CV_j = 0 + 0.025 \cdot j \qquad \forall\, j = 1, \dots, 17$

v. objective functions against *RT* and *MA* using "noisedN1" outputs and setting *SA* and *CV* to their true values

$\boldsymbol{GoF}^{\boldsymbol{noise1,true}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN1}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 20;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}} = S(\boldsymbol{\beta}_{\boldsymbol{i,j,true}})$

$\boldsymbol{\beta}_{\boldsymbol{i,j,true}} = \{RT_i; MA_j; 1.0; 0.1\}$

$RT_i = 0.1 + 0.1 \cdot i \qquad \forall\, i = 1, \dots, 20$

$MA_j = 1.5 + 0.25 \cdot j \qquad \forall\, j = 1, \dots, 17$

vi. objective functions against *RT* and *MA* using "noisedN2" outputs and setting *SA* and *CV* to their true values

$\boldsymbol{GoF}^{\boldsymbol{noise2,true}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN2}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 20;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}} = S(\boldsymbol{\beta}_{\boldsymbol{i,j,true}})$

$\boldsymbol{\beta}_{\boldsymbol{i,j,true}} = \{RT_i; MA_j; 1.0; 0.1\}$

$RT_i = 0.1 + 0.1 \cdot i \qquad \forall\, i = 1, \dots, 20$

$MA_j = 1.5 + 0.25 \cdot j \qquad \forall\, j = 1, \dots, 17$

vii. objective functions against *SA* and *CV* using "true" outputs and setting *RT* and *MA* to very wrong values

$\boldsymbol{GoF}^{\boldsymbol{true,noise}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{true}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{noise,i,j}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 23;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{noise,i,j}} = S(\boldsymbol{\beta}_{\boldsymbol{noise,i,j}})$

$\boldsymbol{\beta}_{\boldsymbol{noise,i,j}} = \{0.9; 2.5; SA_i; CV_j\}$

$SA_i = 0.45 + 0.05 \cdot i \qquad \forall\, i = 1, \dots, 23$

$CV_j = 0 + 0.025 \cdot j \qquad \forall\, j = 1, \dots, 17$

viii. objective functions against *RT* and *MA* using "true" outputs and setting *SA* and *CV* to very wrong values

$\boldsymbol{GoF}^{\boldsymbol{true,noise}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{true}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,noise}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 20;\ j = 1, \dots 17$

$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}} = S(\boldsymbol{\beta}_{\boldsymbol{i,j,noisee}})$
$\boldsymbol{\beta}_{\boldsymbol{i,j,true}} = \{RT_i; MA_j; 1.3; 0.15\}$
$RT_i = 0.1 + 0.1 \cdot i \qquad \forall\, i = 1, \dots, 20$
$MA_j = 1.5 + 0.25 \cdot j \qquad \forall\, j = 1, \dots, 17$

ix. objective functions against *SA* and *CV* using "noisedN1" outputs and setting *RT* and *MA* to very wrong values

$\boldsymbol{GoF}^{\boldsymbol{noise1,noise}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN1}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 23;\ j = 1, \dots 17$
$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}} = S(\boldsymbol{\beta}_{\boldsymbol{true,i,j}})$
$\boldsymbol{\beta}_{\boldsymbol{true,i,j}} = \{0.6; 3.5; SA_i; CV_j\}$
$SA_i = 0.45 + 0.05 \cdot i \qquad \forall\, i = 1, \dots, 23$
$CV_j = 0 + 0.025 \cdot j \qquad \forall\, j = 1, \dots, 17$

x. objective functions against *SA* and *CV* using "noisedN2" outputs and setting *RT* and *MA* to very wrong values

$\boldsymbol{GoF}^{\boldsymbol{noise2,noise}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN2}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 23;\ j = 1, \dots 17$
$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{true,i,j}} = S(\boldsymbol{\beta}_{\boldsymbol{true,i,j}})$
$\boldsymbol{\beta}_{\boldsymbol{true,i,j}} = \{0.6; 3.5; SA_i; CV_j\}$
$SA_i = 0.45 + 0.05 \cdot i \qquad \forall\, i = 1, \dots, 23$
$CV_j = 0 + 0.025 \cdot j \qquad \forall\, j = 1, \dots, 17$

xi. objective functions against *RT* and *MA* using "noisedN1" outputs and setting *SA* and *CV* to very wrong values

$\boldsymbol{GoF}^{\boldsymbol{noise1,noise}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN1}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,noise}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 20;\ j = 1, \dots 17$
$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}} = S(\boldsymbol{\beta}_{\boldsymbol{i,j,true}})$
$\boldsymbol{\beta}_{\boldsymbol{i,j,true}} = \{RT_i; MA_j; 1.0; 0.1\}$
$RT_i = 0.1 + 0.1 \cdot i \qquad \forall\, i = 1, \dots, 20$
$MA_j = 1.5 + 0.25 \cdot j \qquad \forall\, j = 1, \dots, 17$

xii. objective functions against *RT* and *MA* using "noisedN2" outputs and setting *SA* and *CV* to very wrong values

$\boldsymbol{GoF}^{\boldsymbol{noise2,noise}}_{\boldsymbol{k}} = \{GoF_{k,i,j}(\boldsymbol{M}^{\boldsymbol{obs}}_{\boldsymbol{noiseN2}}, \boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,noise}})\}\ \forall\, k = 1, \dots, 16;\ i = 1, \dots, 20;\ j = 1, \dots 17$
$\boldsymbol{M}^{\boldsymbol{sim}}_{\boldsymbol{i,j,true}} = S(\boldsymbol{\beta}_{\boldsymbol{i,j,true}})$
$\boldsymbol{\beta}_{\boldsymbol{i,j,true}} = \{RT_i; MA_j; 1.0; 0.1\}$
$RT_i = 0.1 + 0.1 \cdot i \qquad \forall\, i = 1, \dots, 20$
$MA_j = 1.5 + 0.25 \cdot j \qquad \forall\, j = 1, \dots, 17$

However, since the level of noise did not affect the results of this first analysis, in the following section we have considered only the level of noise N2.

6.2 Results of the response surface analysis

The following figures and tables summarize the results achieved. In particular, in Table 1 and in Table 2, for each of the eight diagram categories, the global minimum (minima if more than one) of the 16 functions is reported. By contrast, in Figure 3-10 the response surfaces are shown. In this study, such response surfaces have not been obtained by fitting a specific meta-model (such as a quadratic

surface or a kriging meta-model, see Jones, 2001) with the results of the model simulation on a set of point and then plotting the meta-model obtained. Instead, the visual output was obtained directly using the MATLAB software (Mathworks, 2010). Indeed, at this stage, we are not interested in model calibration but at gaining insights into the problem. In particular, it was chosen to display the 3-D diagrams as surface plots in order to allow the reader to capture as much information as possible (the value of the objective function is shown in the color-bar attached to each diagram). In the labels of both tables and figures, some GoFs were taken with the opposite sign. This was done in order to always compare objective functions referring to minimization problems. In addition, the GoF indicated as GEH refer to the ordinary definition as provided in Ma et al. (2007), while GEH1, GEH3 and GEH5 refer to the alternative formulation proposed here (with threshold, 1, 3 and 5).

Table 1 Global minimum (minima) found in each response surface for SA and CV parameters and for each GoF (Plot categories derive their name from the description provided in the previous section).

GoF_k	Parameter	Plot category					
		True RT-MA True Data	True RT-MA Noise Data	Noise RT-MA True Data		Noise RT-MA Noise Data	
-GEH1	SA	1	1	1	1.05	1.1	1.3
	CV	0.1	0.1	0.07	0.07	0.1	0.15
-GEH3	SA	more than ten minima	more than ten minima	more than ten minima		more than ten minima	
	CV						
-GEH5	SA	more than ten minima	more than ten minima	more than ten minima		more than ten minima	
	CV						
GEH	SA	1	1	1.05	1.1	1.05	1.1
	CV	0.1	0.1	0.07	0.07	0.07	0.1
MAE	SA	1	1	1.05		1.05	1.1
	CV	0.1	0.1	0.07		0.07	0.1
MANE	SA	1	1	1.05	1.1	1.05	1.1
	CV	0.1	0.1	0.07	0.07	0.07	0.1
RMSE	SA	1	1	1		1.1	
	CV	0.1	0.1	0.07		0.1	
RMSNE	SA	1	1	1.1		1.05	1.1
	CV	0.1	0.1	0.07		0.07	0.1
SE	SA	1	1	1		1.1	
	CV	0.1	0.1	0.07		0.1	
-r	SA	1	1	more than ten minima		6 minima	
	CV	0.1	0.1				
U	SA	1	1	1		1.1	
	CV	0.1	0.1	0.07		0.1	
U_m	SA	1.5	0.95	1.15		1.15	
	CV	0.25	0.05	0		0.07	
U_s	SA	0.95	1.25	0.65		1.05	
	CV	0.2	0.13	0.03		0.1	
$-U_c$	SA	7 minima	4 minima	3 minima		more than ten minima	
	CV						
Green = exact solution; Yellow = acceptable solution; Red = wrong solution							

A number of considerations stem from the diagrams obtained:

a) the response surfaces confirm that is argued in the introduction about the complexity of the calibration problem and the need for using a global optimization approach (e.g. this poses serious questions on the choice made in Balakrishna et al. (2007) and Ma et al. (2007) of using the SPSA algorithm for dealing with this kind of problems). Indeed most of the

response surface show the existence of several local minima as well as of wide areas wide approximately constant values. In addition it is evident how different choices in setting up a calibration problem generate different response surfaces (e.g. compare –GEH1, RMSE, RMSNE, SE, -Um, Us and -r in Figure 3 or –GEH1, RMSE, SE, Um, Us, -r in Figure 9 for the influence of the GoF). In addition comparing the pictures of FIGURE 3-6 with the respective ones of Figure 7-10 it is clear how significant is the selection of the sensitive parameters on the calibration problem.

b) the possibility of splitting the calibration problem into two sub-problems has to be further proven. That said, the analyses carried out here show that, as also hypothesized in [4], calibrating first the SA and CV parameters, and then RT and MA, is likely to provide positive results. In particular, the results in Table 1, using SE or RMSE as the GoF measure, show that even by introducing significant errors in both the observed measurement and in the RT and MA parameters, the problem solution does not change considerably (less than 10%). This also highlights the necessity of performing sensitivity analysis of models used. Indeed, as happens for the response surfaces of RT and MA, in the case that significant parameters are disregarded from the calibration, the minimum of the objective function may change considerably (see Table 2) and thus the value retrieved at the end of the process for the parameters calibrated may also be different from the real one, achieving also physically unrealistic values.

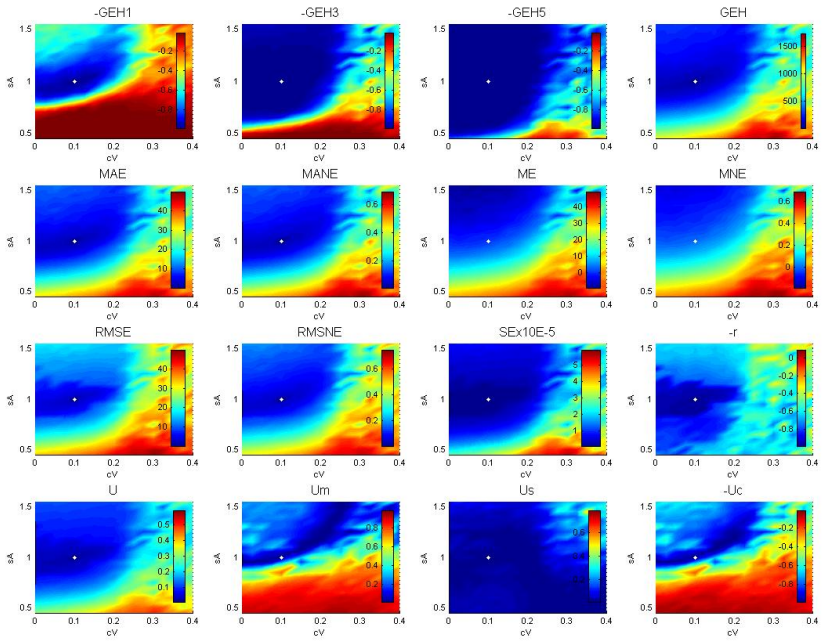


Figure 3. Plots of the objective function against the SA and CV parameters, obtained using true data and true values for the RT and MA parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

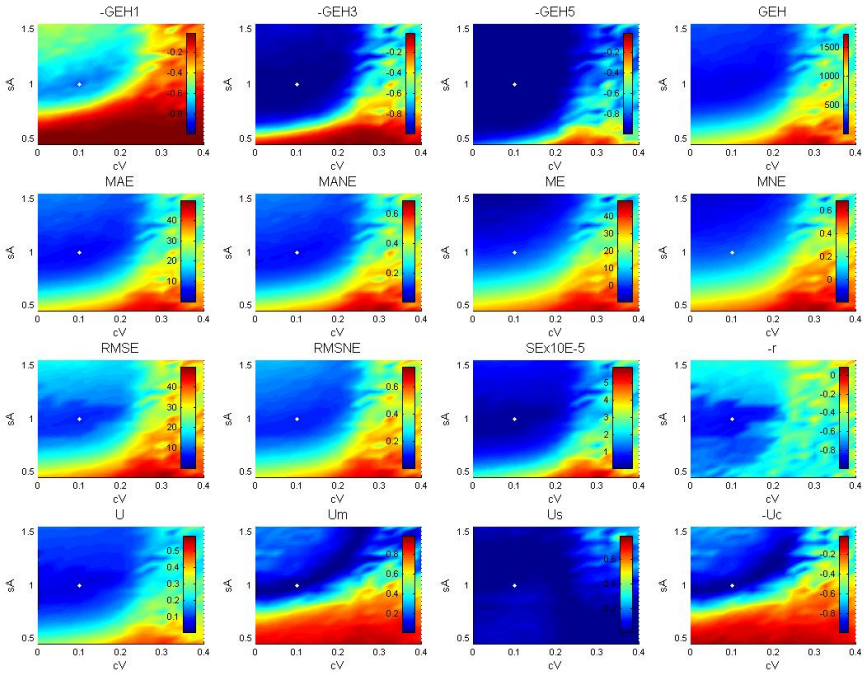


Figure 4. Plots of the objective function against the SA and CV parameters, obtained using noised data and true values for the RT and MA parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

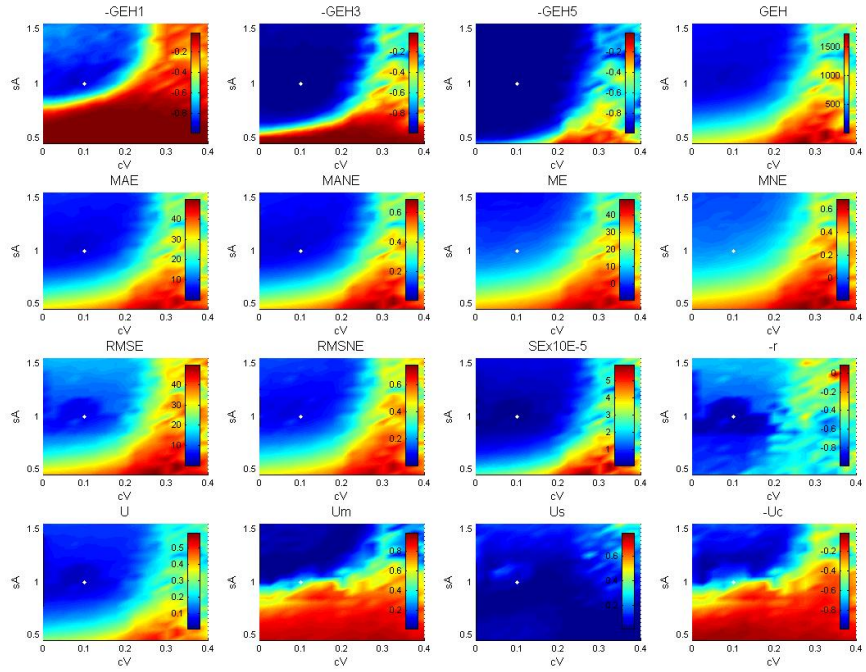


Figure 5. Plots of the objective function against the SA and CV parameters, obtained using true data and noised values for the RT and MA parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

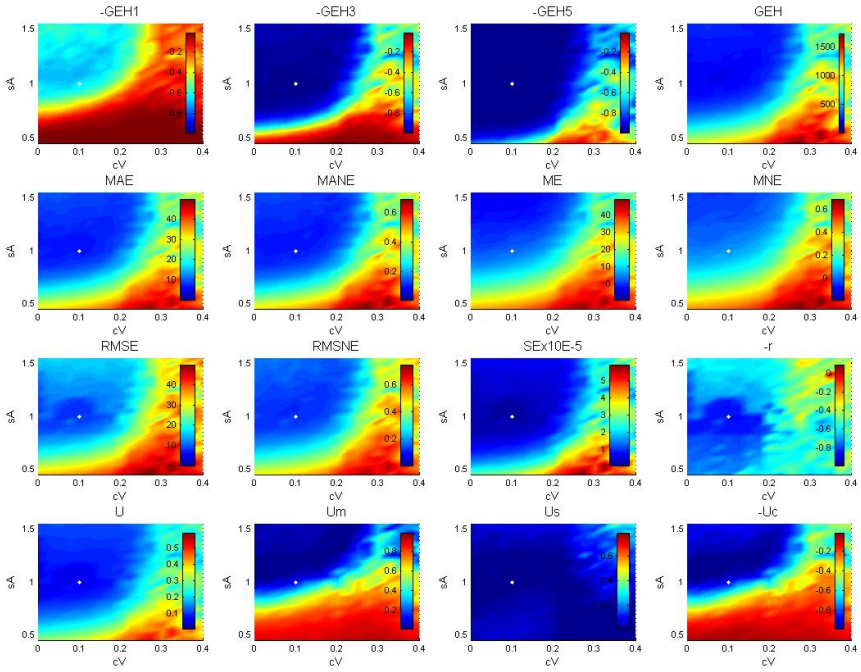


Figure 6. Plots of the objective function against the SA and CV parameters, obtained using noised data and noised values for the RT and MA parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

c) on the other hand, plots regarding RT and MA parameters (Figure 7-10) seriously focus attention on whether or not Maximum Acceleration should be included in the calibration phase. The value of the objective functions seems to be very little influenced by MA at least in the neighborhood of the "true" solution. In addition, neither does reaction time seem to affect the objective functions very much for values less than 1. This is explained by the fact that under certain values of the RT, even increasing the road capacity, the speed of the traffic flow is only regulated by the speed that vehicles wish to attain. This is perfectly highlighted by the –GEH1 plots in Figure 9-10, in which, the simulations being performed with a higher value of Speed Acceptance, for lower values of RT the objective function starts to increase again. This is another reason why it is necessary to first calibrate SA and CV.

d) as already pointed out the results obviously confirmed the impossibility of using ME and MNE as GoF in the objective function (they were also removed from Table 1 and Table 2);

e) Um, Us, -Uc and –r proved less suitable than the others to be used in the objective function of the calibration problem. In particular, Us, -Uc and –r were always more irregular, showing several different minima in all the plots; in addition, the values reported in Table 1 and Table 2 show that also without errors in the data and in the value of the other parameters, for Um, Us, and -Uc the minimum was quite different from the "true" one;

Table 2 . Global minimum (minima) found in each response surface for RT and MA parameters and for each GoF (Plot categories derive their name from the description provided in the previous section).

GoF_k	Parameter	Plot category			
		True SA-CV True Data	True SA-CV Noise Data	Noise SA-CV True Data	Noise SA-CV Noise Data
-GEH1	RT	0.6	0.6	1 1.1	1.1 1.1
	MA	3.5	3.5	3 4.5	4 4.5
-GEH3	RT	more than ten minima	more than ten minima	more than ten minima	more than ten minima
	MA				
-GEH5	RT	more than ten minima	more than ten minima	more than ten minima	more than ten minima
	MA				
GEH	RT	0.6	0.6	1	1
	MA	3.5	3.5	2.75	3
MAE	RT	0.6	0.6	1	1
	MA	3.5	3.5	3	3
MANE	RT	0.6	0.6	1	1
	MA	3.5	3.5	2.5	3
RMSE	RT	0.6	0.6	1 1	1
	MA	3.5	3.5	2.75 3	3
RMSNE	RT	0.6	0.6	1	1
	MA	3.5	3.5	2.5	3
SE	RT	0.6	0.6	1	1
	MA	3.5	3.5	2.75	3
-r	RT	0.6	0.6	0.9	0.4 0.9
	MA	3.5	3.5	5.5	3.25 5.5
U	RT	0.6	0.6	1	1
	MA	3.5	3.5	2.75	3
U_m	RT	0.4	0.6	1.3	1.3
	MA	5	2	4.25	4.25
U_s	RT	0.6	0.4	1.9	0.7
	MA	3.5	3.5	1.75	2.75
$-U_c$	RT	more than ten minima	more than ten minima	1	1
	MA			2.5	1.75

Green = exact solution; Yellow = acceptable solution; Red = wrong solution

f) on the other hand, plots regarding RT and MA parameters (Figure 7-10) seriously focus attention on whether or not Maximum Acceleration should be included in the calibration phase. The value of the objective functions seems to be very little influenced by MA at least in the neighborhood of the "true" solution. In addition, neither does reaction time seem to affect the objective functions very much for values less than 1. This is explained by the fact that under certain values of the RT, even increasing the road capacity, the speed of the traffic flow is only regulated by the speed that vehicles wish to attain. This is perfectly highlighted by the –GEH1 plots in Figure 9-10, in which, the simulations being performed with a higher value of Speed Acceptance, for lower values of RT the objective function starts to increase again. This is another reason why it is necessary to first calibrate SA and CV.

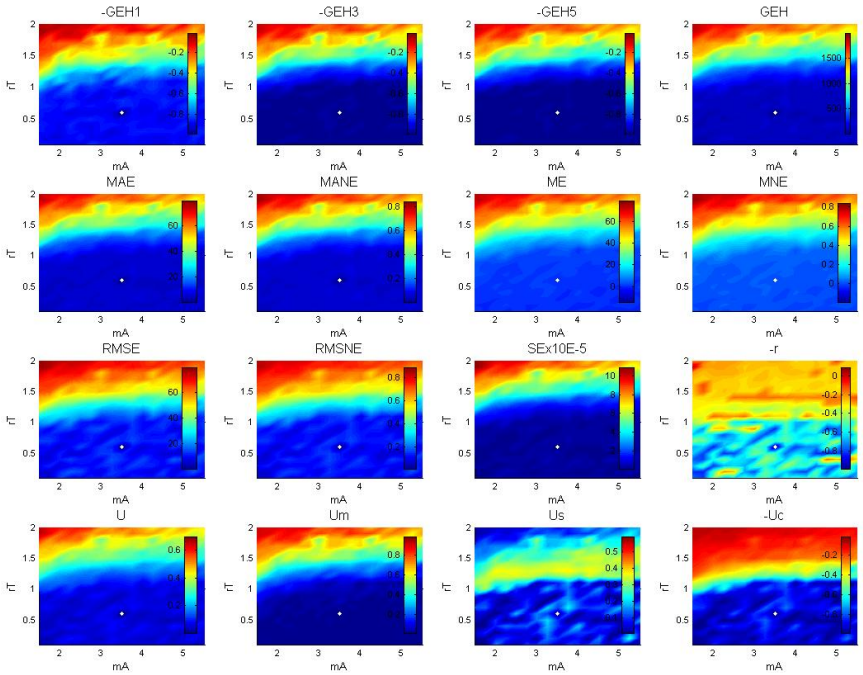


Figure 7. Plots of the objective function against the RT and MA parameters, obtained using true data and true values for the SA and CV parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

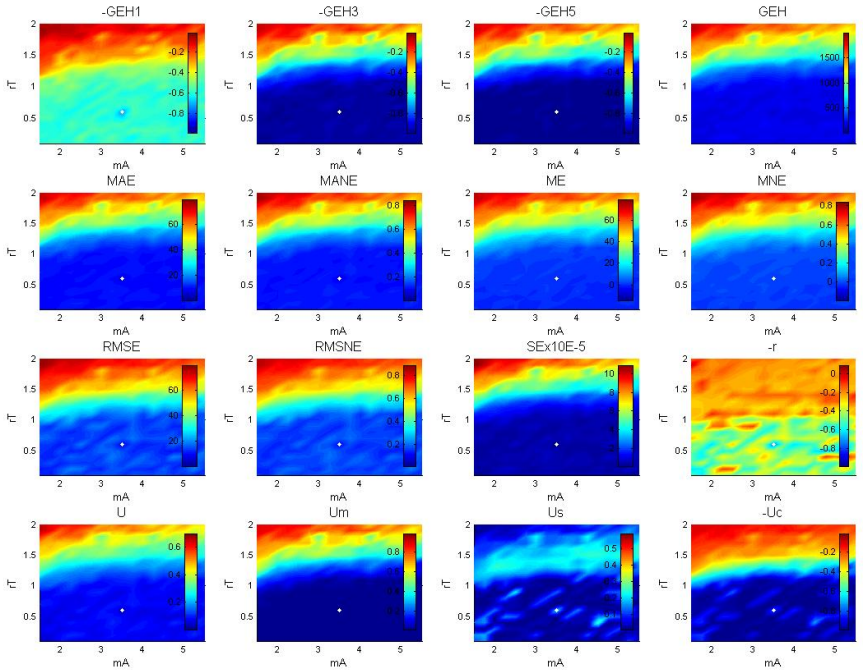


Figure 8. Plots of the objective function against the RT and MA parameters, obtained using noised data and true values for the SA and CV parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

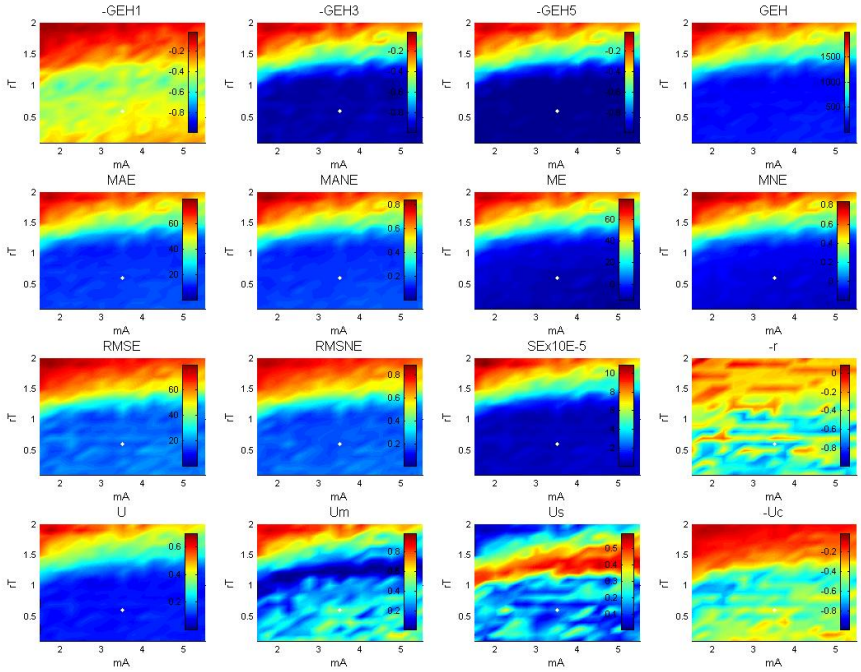


Figure 9. Plots of the objective function against the RT and MA parameters, obtained using true data and noised values for the SA and CV parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

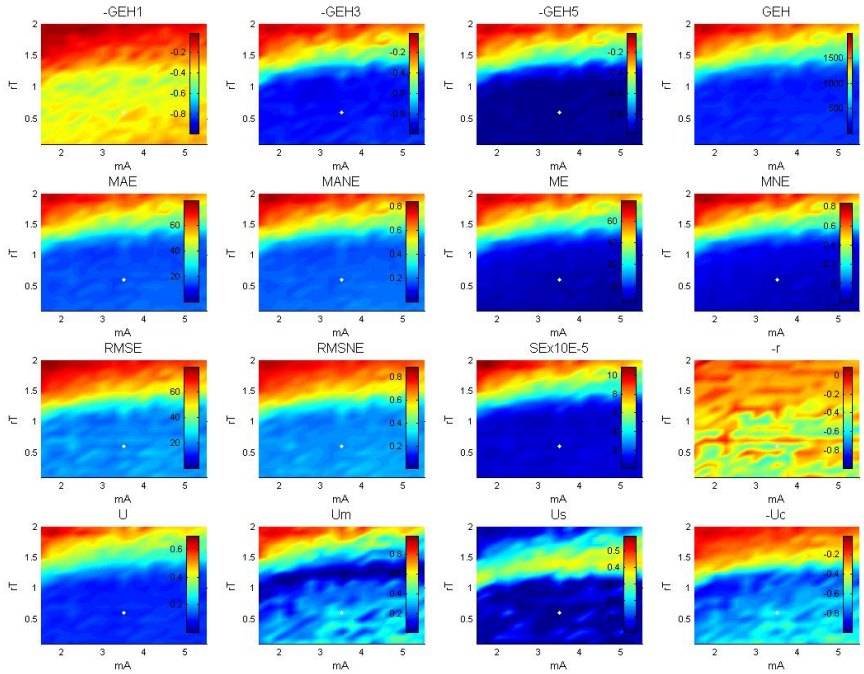


Figure 10. . Plots of the objective function against the RT and MA parameters, obtained using noised data and noised values for the SA and CV parameters with 16 different goodness-of-fit measures. The white point is the "true" solution.

g) as already pointed out the results obviously confirmed the impossibility of using ME and MNE as GoF in the objective function (they were also removed from Table 1 and Table 2);

h) Um, Us, -Uc and –r proved less suitable than the others to be used in the objective function of the calibration problem. In particular, Us, -Uc and –r were always more irregular, showing several different minima in all the plots; in addition, the values reported in Table 1 and Table 2 show that also without errors in the data and in the value of the other parameters, for Um, Us, and -Uc the minimum was quite different from the "true" one;

i) the values of 3 and 5 as thresholds in GEH3 and GEH5 evaluations proved very high, and consequently a wide area in all their plots resulted with a constant value of the objective function. This suggests that, at least in the transportation field, 5 is probably too high a threshold to assess that two series of data show a good fit, as instead proposed in Highway Agency (1996).

j) all the other GoFs show on the whole a similar behavior, even if:
   - SE appears to be the least sensitive GoF (with the widest deep area around the "true" solution), but also the most regular around the minimum value;
   - RMSE seems to offer higher irregularity than the other GoFs around the optimum value;
   - -GEH1 probably shows the best capacity in highlighting the position of the minimum (in this case the threshold used seems to have a good impact), even if it is the GoF which most suffers from the presence of noise in the data (compare for example FIGURE 5-6;
   - MAE and MANE show the highest flatness of the objective function around the global solution. This could represent a problem identifying an effective stopping rule for any optimization algorithm;
   - all these GoFs prove robust with respect to the introduction of noise to the data (in the category of plots with noise data and true parameters, in all of them the minimum value is the true one) even with large errors (which is the case of that introduced here);
   - SE turns out to be the GoF with the highest stability, with respect to errors in the value of the parameters, (also without showing multiple minimum values, see TABLE 1 and TABLE 2).

As a conclusion, as already stated, we decided to test the methodology proposed in section 1.1 on seven GoF measures. We decided to test the five algorithms on the surfaces categories referred to in the points *i-vi* in section 6.1 and thus on 42 different objective functions.

6.3 Kriging meta-models

From the objective function evaluations which have led to the response surfaces definition, the parameters of 42 different Kriging models have been defined. With this aim, the Matlab toolbox DACE (Lophanev et al., 2002a, 2002b) has been used.

The DACE toolbox has essentially two functions, "*dacefit*" and "*predictor*". The former is used to find the parameters of the Kriging which better fit the experimental points provided as input. Once the parameters of the model are known, the latter is used to calculate the value of the Kriging function. For this reason, the first function is used here in order to define the 42 Kriging models, while the second is used in order to put the Kriging models in the place of the traffic simulation model for the calibration experiments.

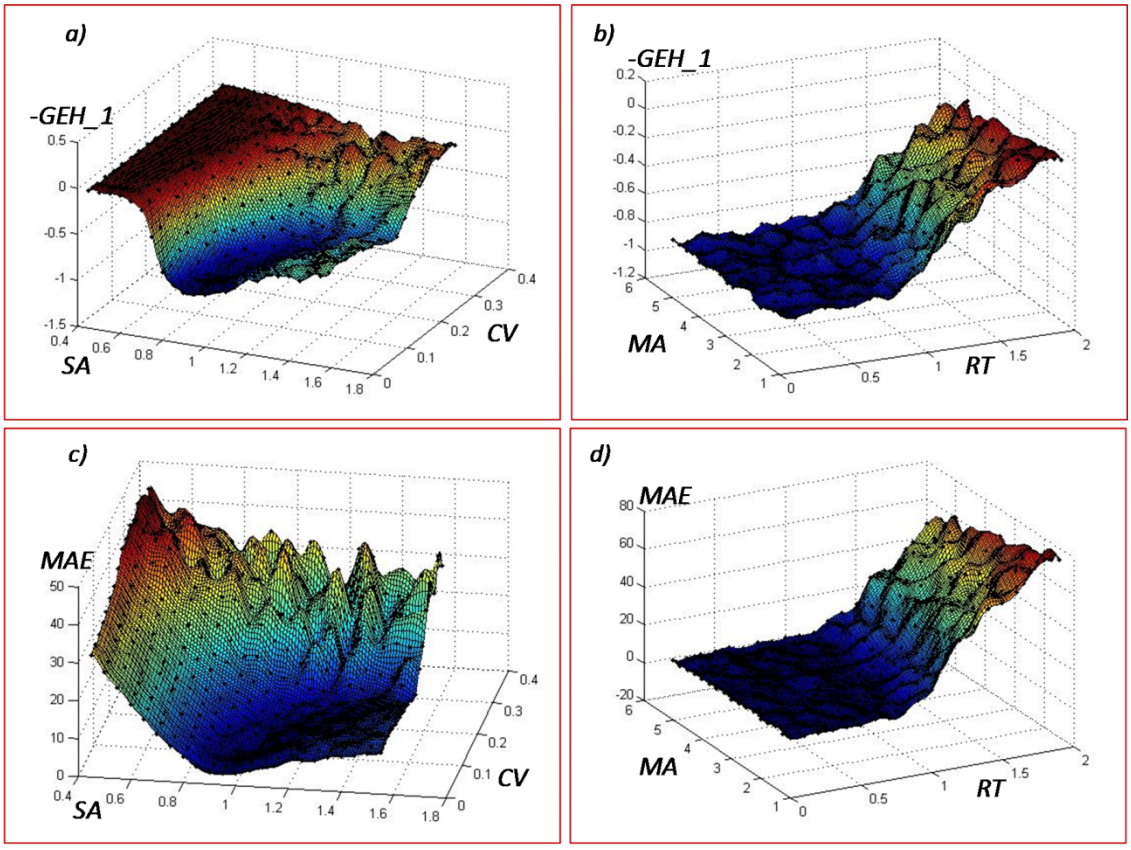


Figure 11. Comparison between simulation outputs (the black dots) and Kriging prediction in four different objective functions (all the cases here reported refer to objective functions evaluated using the time series without noise injection).

The only potential problem with the function "dacefit" is the necessity to provide as input vectors of starting values, upper bounds and lower bounds for the parameters of the correlation function (ϑ_g). Indeed the final values of these parameters, as provided by the tool, will depend upon the input. In order to overcome the problem, after some analyses, upper and lower bounds have been set to

define a quite large domain (with respect to the common applications, Lophanev, 2002a), while for the starting values an iterative procedure has been created, which uses the final values provided by "dacefit" at one step as the starting values for the next one. The iteration ends when the starting values for the parameters ϑ_g equal the results obtained. In this way, it has been noted a substantial improvement of the Kriging capability to reproduce the simulation model.

In Figure 11, four examples of the capability of the Kriging surrogates to reproduce the experimental points (represented by the black dots) are shown for 4 out of the 42 objective functions considered. The graphs further confirm what claimed in the first section of the paper concerning the dependence of the simulation optimization problem from several factors (here the dependence from the parameters to calibrate and from the GoF measure considered is clear, as well as the importance of performing a preliminary model sensitivity analysis in order to remove parameters like MA, which may dramatically affect the calibration efforts).

6.4 Optimization algorithms set-up

In this case study we applied the algorithms described in sections 4.1-4.4. One of the main improvements led by the use of Kriging surrogates of the objective function in the verification of simulation optimization procedures is the possibility, otherwise unfeasible, to fine-tune the parameters of the optimization algorithms. The performance of whatever algorithm in the solution of an optimization problem depends also on the value of its parameters. This issue is often neglected in simulation optimization for the unfeasible efforts that it would need. On the contrary, in our case it was possible (even if time consuming), being the Kriging surrogates very fast to be evaluated.

The following steps have been followed:

- definition of the algorithms' maximum number of iterations. We decided not to consider, in this phase, the convergence speed of the different algorithms. Therefore, we searched a maximum number of iterations allowing all the algorithms to work in a satisfactory manner. We found this number to be around 1000.
- per each objective function we have designed a full factorial experimental plan on the parameters of each algorithm, in order find the parameters combination ensuring the best algorithm performance. The number of levels considered per each parameter was 4. This required around 10.000 preliminary calibrations experiments.

6.5 Methodology of appraisal

Two indicators have been used in order to assess the performance of each algorithm on each objective function:

- the percentage of times, over 1000 attempts (in which the starting point of the algorithm is modified), the solution found by the optimization algorithm is in the surroundings of the global solution (in the present study an error of the 5% in the parameters' estimation has been accepted);
- the value of the Optimization Performance Indicator (OPI) defined as follows:

$$OPI = \sum_{j=1}^{N} \left(\sqrt{\left(\frac{X_{1j} - X_{1,min}}{ub_1 - lb_1}\right)^2 + \left(\frac{X_{2j} - X_{2,min}}{ub_2 - lb_2}\right)^2} \cdot \left(exp\left(\frac{Y_j - Y_{min}}{Y_{max} - Y_{min}}\right)\right)\right) \quad (27)$$

  which measures the sum, over all the calibration attempts, of the distance between the solution found by the algorithm and the global solution of the problem, normalized with respect to the parameters range of variability, and weighted with the exponential of the normalized value attained by the objective function in correspondence of the solution found. It allows understanding how wide is the cloud of the solutions found but also which algorithms suffer the most for the presence of local minima whose objective function value is higher than the true one.

6.6 Results

The 42 calibration experiments carried out have required 42 millions evaluation of the objective function. Given that each simulation of the scenario described requires approximately 10 minutes, the full experiment would have required around 800 years to be completed on a single computer (without considering the phase of fine-tuning of the algorithms). This means that, without losing generality, the procedure adopted is likely to provide a substantial step forward in the evaluation of simulation optimization procedures.

For the particular case study presented, the results are presented in Figure 12 and 13 and in Table 3 and 4. The following considerations can be made:

- the performance of the different algorithms over the 42 experiments considerably differ. This definitely prove the dependence of the performance from the parameters to be calibrated, from the GoF measure and from the quality of the data;
- SE has resulted the worst GoF measure, affecting the optimization problem solution as

shown in Figure 12a) and Figure 13a). For this reason it has been removed from the successive analyses (42 experiments have become 38);

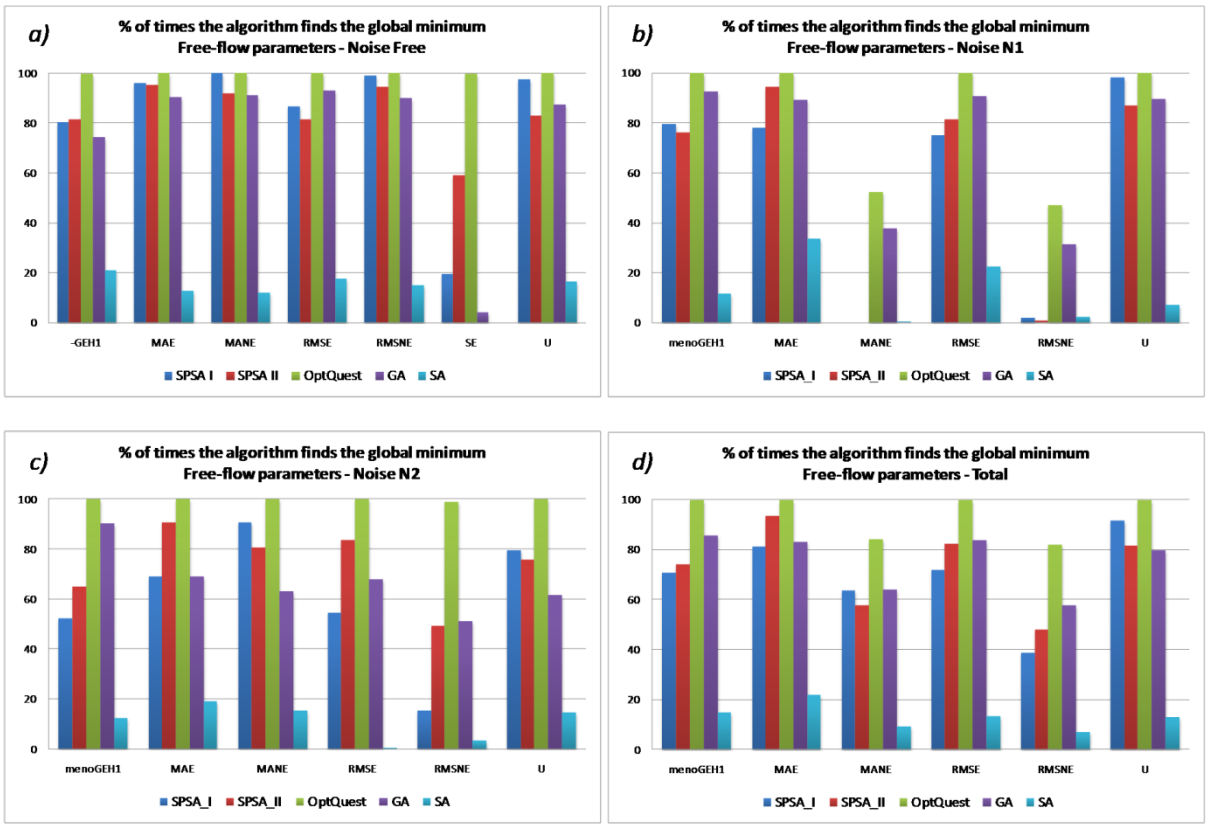


Figure 12. Percentage of times the algorithm finds the global solution for the different combination of GoF measures/optimization algorithms on the objective function for the SA-CV couple of parameters (a, b and c present the results respectively without and with the injection of noise, while d presents the global figure).

- OptQuest resulted the best algorithm. This is clearly shown by the percentage of times it is able to find the global minimum in all the cases and from the values of the OPI. In general only the overall performance of the GA with –GEH_1 as GoF has overcome the performance of the OptQuest (but only with MANE and RMSNE as GoF as showed in Table 4). This is also an indication of the fact that the presence of several "similar" local minima in the objective functions requires an exploration of the objective function to be performed. This also means that, increasing the number of parameters, the efforts for the model calibration would considerably increase;
- SA resulted the worst algorithm. Its capability to escape local minima seemed less effective than that of the others. This is clear from both Figure 12-13 and Table 3-4
- the difference of performance among the other algorithms is less evident and in general depends on the particular combination parameters/GoF/noise. This is clearly shown by

Figure 12-13. From Table 3 it seems that globally GA outperforms the others, but looking at Table 4 it is clear how the performances are unstable. Furthermore, the improvements brought by the SPSA_II algorithm are not evident in the present applications;

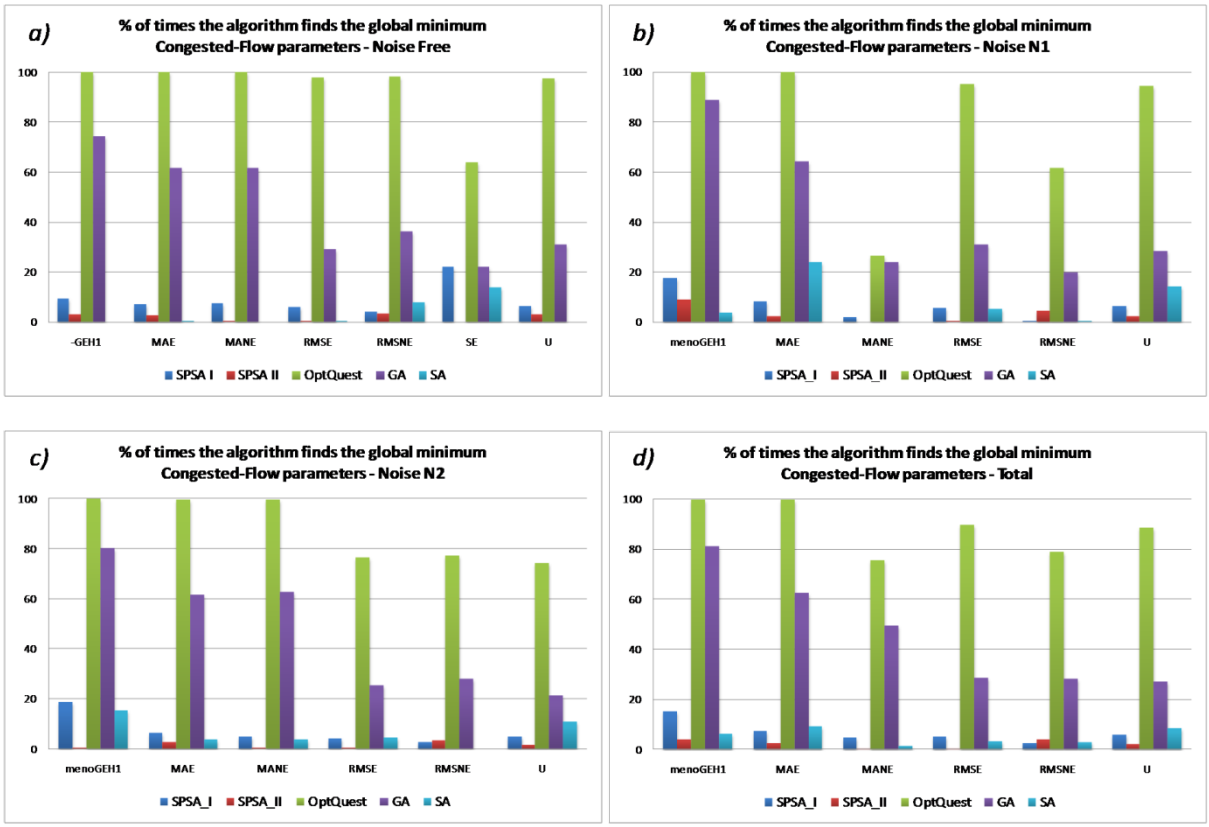


Figure 13. Percentage of times the algorithm finds the global solution for the different combination of GoF measures/optimization algorithms on the objective function for the RT-MA couple of parameters (a, b and c present the results respectively without and with the injection of noise, while d presents the global figure).

Table 3. Optimization Performance Index (OPI) aggregated over the different levels of noise and alternatively over the different GoF measures and the different algorithms for SA-CV objective functions, RT-MA objective functions and for all the cases.

OPI	Tot Free-Flow parameters	Tot Cong-Flow parameters	Tot All Cases
-GEH1	4006.0	4228.8	8234.8
MAE	**1803.0**	**3803.3**	**5606.4**
MANE	2707.9	5962.3	8670.2
RMSE	2020.3	5456.2	7476.5
RMSNE	3490.4	5203.1	8693.6
U	2288.5	4784.7	7073.2
SPSA_I	1836.7	6125.8	7962.5
SPSA_II	2660.0	7731.7	10391.6
OptQuest	**345.5**	**1075.8**	**1421.3**
GA	1261.4	5178.2	6439.7
SA	10212.5	9327.1	19539.6

Table 4. Optimization Performance Index (OPI) per each combination GoF measure/optimization algorithm aggregated over the different level of noise for SA-CV objective functions, RT-MA objective functions and for all the case.

IID	Tot Free-Flow param			Tot Congested-Flow param			Tot All cases		
	Algorithm	GoF	OPI	Algorithm	GoF	OPI	Algorithm	GoF	OPI
1	**OptQuest**	**-GEH1**	**8.5**	**OptQuest**	**MAE**	**9.8**	**OptQuest**	**-GEH1**	**22.5**
2	**OptQuest**	**RMSE**	**14.7**	**OptQuest**	**-GEH1**	**14.0**	**OptQuest**	**MAE**	**25.2**
3	**OptQuest**	**U**	**15.0**	**OptQuest**	**RMSE**	**154.0**	**OptQuest**	**RMSE**	**168.7**
4	OptQuest	MAE	15.3	GA	-GEH1	163.9	OptQuest	U	198.7
5	OptQuest	MANE	109.7	OptQuest	U	183.7	GA	-GEH1	289.4
6	GA	-GEH1	125.5	OptQuest	RMSNE	289.7	OptQuest	RMSNE	472.0
7	SPSA_I	U	138.2	OptQuest	MANE	424.5	OptQuest	MANE	534.2
8	GA	RMSE	143.6	GA	MAE	571.7	GA	MAE	722.4
9	SPSA_II	MAE	150.6	SPSA_I	-GEH1	902.4	SPSA_I	MAE	1066.1
10	GA	MAE	150.8	SPSA_I	MAE	905.7	SPSA_I	U	1077.5
11	SPSA_I	MAE	160.4	SPSA_I	U	939.3	SPSA_I	RMSE	1175.2
12	GA	U	173.9	SPSA_I	RMSE	988.9	GA	MANE	1249.3
13	OptQuest	RMSNE	182.3	GA	MANE	991.3	SPSA_II	MAE	1252.2
14	SPSA_I	RMSE	186.3	SPSA_II	MAE	1101.6	GA	RMSE	1269.3
15	SPSA_I	MANE	236.1	GA	RMSE	1125.7	SPSA_I	-GEH1	1319.1
16	GA	MANE	258.0	SPSA_I	MANE	1155.1	GA	U	1342.9
17	SPSA_II	MANE	295.8	GA	RMSNE	1156.7	SPSA_I	MANE	1391.3
18	SPSA_II	U	326.5	SPSA_II	U	1157.5	SPSA_II	U	1484.0
19	GA	RMSNE	409.7	GA	U	1169.0	GA	RMSNE	1566.4
20	SPSA_I	-GEH1	416.7	SA	RMSNE	1213.0	SPSA_II	MANE	1777.0
21	SPSA_II	RMSNE	503.9	SA	MAE	1214.5	SPSA_II	RMSNE	1813.4
22	SPSA_II	RMSE	554.9	SPSA_I	RMSNE	1234.3	SPSA_I	RMSNE	1933.3
23	SPSA_I	RMSNE	699.0	SPSA_II	-GEH1	1294.3	SPSA_II	RMSE	1942.4
24	SPSA_II	-GEH1	828.3	SPSA_II	RMSNE	1309.5	SPSA_II	-GEH1	2122.6
25	SA	RMSE	1120.8	SA	U	1335.2	SA	MAE	2540.5
26	SA	MAE	1325.9	SPSA_II	RMSE	1387.6	SA	RMSNE	2908.5
27	SA	U	1634.9	SPSA_II	MANE	1481.2	SA	RMSE	2920.9
28	SA	RMSNE	1695.5	SA	RMSE	1800.1	SA	U	2970.1
29	SA	MANE	1808.3	SA	-GEH1	1854.2	SA	MANE	3718.4
30	SA	-GEH1	2627.0	SA	MANE	1910.2	SA	-GEH1	4481.1

- Table 3 shows that, globally, the most suitable GoF is MAE. However, in Table 4 the situation seems not as clear. In particular looking at the performances of the OptQuest, on one hand, on the SA-CV objective functions, -GEH1 is the best GoF with RMSE and U being 2nd and 3rd (but very close each other and very close to MAE). On the other hand, on the RT-MA objective functions, MAE is the best GoF with –GEH1 being 2nd. Anyway, considering all the

algorithms (Table 3), -GEH1 is one of the worst GoF, resulting very unsuitable to be used with the SA algorithm.

- MANE and RMSNE show their limits in particular when in the time series there are low values different from zero (as it happens in the case of noise N1).

- the presence of a parameter with a low, even if not negligible, influence on the model outputs, heavily affects the possibility for the model to be calibrated. Apart from OptQuest, the performances of all the other algorithms are largely unsatisfactory, in particular as the level of noise increases. However, whatever the algorithm, one could question the meaning of the solution found for an objective function like those reported in Figure 11b-d). For this reason, again, it is fundamental that whatever simulation optimization has to be preceded by a sensitivity analysis of the model for individuating the parameters that should be calibrated and those which can be fixed a priori to some common sense values.

- the presence of random noise (of not negligible magnitude), had an influence on the optimization procedures. This influence has revealed to be particularly important with the normalized GoF measures (MANE and RSMNE), and less relevant for the other combinations of parameters/algorithms/GoF measures.

7 Case study 2: the Gipps car-following model

The Gipps model (Gipps, 1981) is a safety-based model. It provides two different transfer functions according to the two different driving regimes assumed. In the free-flow regime, the following vehicle *n* plans its speed ($v_{a,n}$) for the successive instant so that (a) it will not exceed the driver's desired speed and (b) its free acceleration should first increase with speed, as engine torque increases, then decrease to zero, as the vehicle approaches the desired speed. In the car-following regime, instead, the driver adopts a speed ($v_{b,n}$) that allows him a safely stop in case of a sudden braking of the leading vehicle.

The switch between the two driving regimes is not explicitly dealt with and it is driven by a simple rule: at each simulation instant (t) the following vehicle *n* adopts a speed for the instant $\left(t+\tau\right)$ that is the minimum between the values given by the two above-mentioned models:

$$v_{a,n}(t+\tau) = v_n(t) + 2.5 \cdot a_n^{Max} \cdot \tau \cdot \left(1 - \frac{v_n(t)}{V_n^{Max}}\right) \cdot \sqrt{0.025 + \frac{v_n(t)}{V_n^{Max}}} \quad (28)$$

$$v_{b,n}(t+\tau) = -b_n \cdot \left(\frac{\tau}{2} + \theta\right) + \sqrt{b_n^{\,2} \cdot \left(\frac{\tau}{2} + \theta\right)^2 + b_n \cdot \left[2 \cdot \left(x_{n-1}(t) - x_n(t) - S_{n-1}\right) - \tau \cdot v_n(t) + \frac{v_{n-1}(t)^2}{\hat{b}_{n-1}}\right]} \quad (29)$$

$$v_n(t+\tau) = Max\left\{0, \min\left\{v_{a,n}(t+\tau), v_{b,n}(t+\tau)\right\}\right\} \quad (30)$$

$$x_n(t+\tau) = x_n(t) + \tau \cdot \left[\frac{v_n(t) + v_n(t+\tau)}{2}\right] \quad (31)$$

where:

- $v_n(t)$ and $v_{n-1}(t)$ are, respectively, the follower's and leader's speed at time *t* [m/s];
- a_n^{Max} is the follower's maximum acceleration rate [m/s^2];
- τ is "the apparent reaction time, a constant for all vehicles" (Gipps, 1981) [s];
- V_n^{Max} is the follower's maximum desired speed, that is "the speed at which the driver of vehicle *n* wishes to travel" (Gipps, 1981) [m/s];
- b_n is "the most severe braking that the driver of vehicle *n* (i.e. the follower) wishes to undertake" (Gipps, 1981) [m/s^2];
- $\theta = \tau/2$ is an additional "comfort" time lag that allows the follower not to brake always at his or her maximum desired rate [s];
- $x_n(t)$ and $x_{n-1}(t)$ are, respectively, the follower's and leader's position at time *t*, measured at the front bumper [m];
- $S_{n-1} = L_{n-1} + Safety$ is the effective size of the leader's vehicle, that is "the physical length plus a margin into which the following vehicle is not willing to intrude, even when at rest" (Gipps, 1981) [m];
- L_{n-1} is the physical length of the leader's vehicle of the leader [m];

- $Safety$ is the safety margin "into which the following vehicle is not willing to intrude, even at rest" (Gipps, 1981) [m];

- $\hat{b}_{n-1}$ is the follower's estimate of the leader's maximum deceleration rate [m/s^2];

Please note that throughout the paper the deceleration rates, b_n and $\hat{b}_{n-1}$, must be intended as absolute values.

7.1 Simulation setup

In the following, a detailed report on the simulation settings is presented. We will first outline the integration scheme adopted, followed by the description of the methodology to use the leaders' and followers' characteristics in the model. The procedure for model initialization ends the section.

7.1.1 The integration scheme

The Gipps' car-following model is a delayed differential equation (being τ the delay). In his original paper (Gipps, 1981), Gipps found the solution of Eq. (3) by adopting an integration step just equal to the delay.

The integration scheme of the Gipps' car-following model is presented in Figure 1. At the instant (t), the model calculates the follower's speed at the instant $(t+\tau)$. The reaction time τ is assumed to be a multiple of the data resolution (i.e. 0.1 s) and, thus, it is treated as a discrete variable. The follower's speed function between the instants (t) and $(t+\tau)$ is assumed linear.

Finally, a forward Euler method on acceleration (i.e. a trapezoidal integration scheme on speed, see Figure 1) is there adopted for calculations. The same approach is applied here.

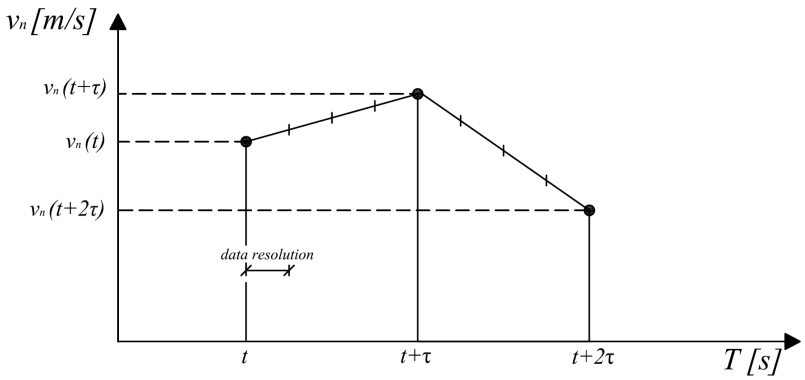


Figure 14. The integration scheme of the Gipps' car-following model.

7.1.2 Leader's and follower's characteristics

According to the structure presented in section 7, the Gipps car-following model is supplied with the leader speed profile. Leader's positions are calculated by the integration of speeds using the scheme presented in Figure 1, that is:

$$x_{n-1}(t+\tau) = x_{n-1}(t) + \tau \cdot \left[\frac{v_{n-1}(t) + v_{n-1}(t+\tau)}{2} \right] \tag{32}$$

If the resolution of the acquired raw speed data does not allow their use for car-following studies, while local vehicles' positions do, the leader speed profile can be derived reversing the trapezoidal integration scheme:

$$v_{n-1}(t+\tau) = 2 \cdot \left[\frac{x_{n-1}(t+\tau) - x_{n-1}(t)}{\tau} \right] - v_{n-1}(t) \tag{33}$$

7.1.3 Model initialization

Initial conditions relate to the leader's and follower's positions at $t = 0$.

7.2 Data description

According to the objective of the paper, the data used for this study were synthetic. It means that the follower trajectory has been generated through the simulation, by fixing the model parameters to a set of "known" values. The leader's trajectory, used to fed the Gipps' car-following model, was, instead, taken from a series of experiments carried out in the area surrounding Naples (Italy), under real traffic conditions, between October 2002 and July 2003.

These experiments were performed by driving four vehicles in a platoon along urban and extra-urban roads under different traffic conditions. Data were acquired through instrumented vehicles. The vehicles were equipped with kinematic differential GPS receivers that recorded the position of each vehicle at each 0.1 seconds. More details on data, including the description of collection experiments and the filter designed to process raw data, can be found in (Punzo et al, 2005).

For the current study, the leader's trajectory is taken from the experiment 30B (Punzo et al., 2005) carried out on a two-lane extra-urban highway. The complete set of trajectory data is available on the MULTITUDE website (COST Action TU0903 – Multitude, 2011) for the forum members.

The values of the parameters that were used to generate the synthetic follower trajectory were the following: τ = 1.0 s, V_n^{Max} = 30 m/s, a_n^{Max} = 2 m/s^2, $Safety$ = 2 m, b_n = 2 m/s^2, $\hat{b}_{n-1}$ = 2 m/s^2.

The leader's and the (synthetic) follower's speed profiles are shown in Figure15(a), while the spacing profile between the leader and the (synthetic) follower is presented in Figure 15(b).

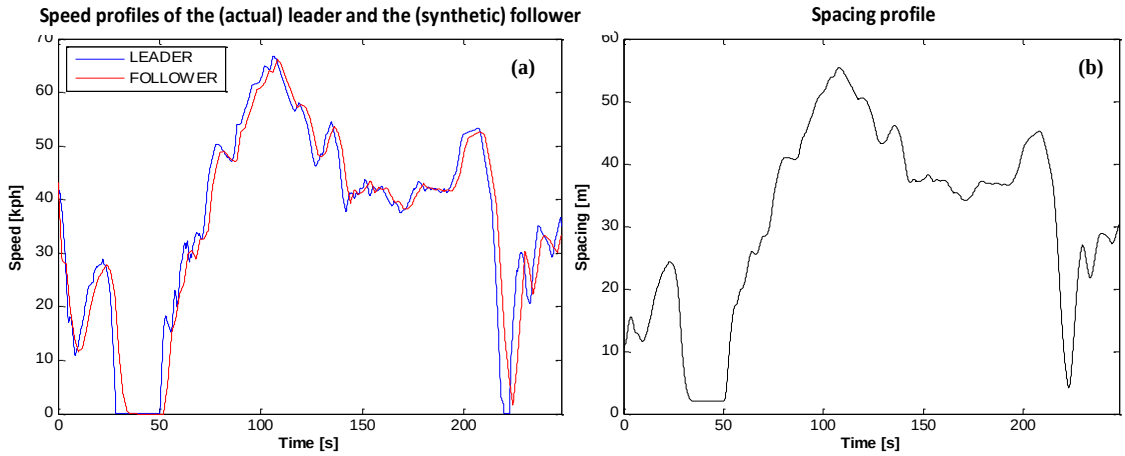


Figure 15. Leader's and (synthetic) follower's speed (a) and spacing (b) profiles.

7.3 The optimization setup

In the following, the description of the optimization settings is presented. First, parameters' upper and lower bounds are discussed. Then, insights in the non-linear constraints are given.

7.3.1 Parameters upper and lower bounds

The calibration of the Gipps' car-following model is performed on the following 6 parameters: τ, V_n, a_n, $Safety$, b_n and $\hat{b}_{n-1}$. The upper and lower bounds of the parameters were fixed at the values reported in Table 5.

As the calibration experiment was performed on synthetic data, we expected that the response surface of the model would have been very steep in the neighborhood of the well-defined global minimum point (i.e. the "known" values of the parameters). Therefore, in such a case, if the combination of the optimization algorithm, the Measure of Performance (MoP) and the Goodness of Fit function (GOF) was effective in finding the unique global minimum, the width of the range of variability of the parameters values should not influence the finding procedure. Thus, the range of variability for the parameters was set wide enough.

Table 5. Parameters upper and lower bounds.

Parameters	Lower bound	Upper bound
τ [s]	0.1	3.0
V_n^{Max} [m/s]	10	40
a_n^{Max} [m/s^2]	0.1	8
$Safety$ [m]	0.1	10
b_n [m/s^2]	0.1	8
$\hat{b}_{n-1}$ [m/s^2]	0.1	8

7.3.2 Non-linear constraints

In order to preserve the simulation from crashing (i.e. to obtain, at a certain time step, imaginary follower's speed values), the domain of the parameters was constrained. Two non-linear constraints were required.

The first one relates to the initial state of the simulation where, for given initial values of spacing and speeds, the values of the reaction time τ, safety margin $Safety$ and deceleration rates (b_n and $\hat{b}_{n-1}$) could not allowed the follower vehicle to stop before intruding the effective size of the leader. This condition happens when the root argument in equation (29) becomes negative, thus, it needs to be at t=0:

$$b_n^{\,2} \cdot \left(\frac{\tau}{2} + \theta\right) + b_n \cdot \left[2 \cdot \left(h_n(0) - S_{n-1}\right) - \tau \cdot v_n(0) + \frac{v_{n-1}(0)^2}{\hat{b}_{n-1}}\right] \geq 0 \qquad (34)$$

The second constraint preserve the consistency of the speed-headway function. In (Wilson, 2001), the author presented the steady state solutions of the Gipps' car-following model in uniform flow. In such conditions, all vehicles travel at the same speed (v^*) and thus their spacing (h^*) is constant and time-independent. The relationship which arises between speed and spacing in steady state is the so-called *speed-headway* function $v^* = V(h^*)$. Since vehicles will drive more slowly (for safety reasons) as the spacing decreases, V is expected to be an increasing function. Following the mathematical derivations, the speed-headway function results in the following quadratic equation in v^*:

$$\left(\frac{b_n}{\hat{b}_{n-1}}-1\right)\cdot v^{*2}-2\cdot b_n\cdot(\tau+\theta)\cdot v^{*}+2\cdot b_n\cdot(h^{*}-S_{n-1})=0 \tag{35}$$

Graphically, it results in the parabolic curve shown in Figure 16 (taken from Wilson, 2001).

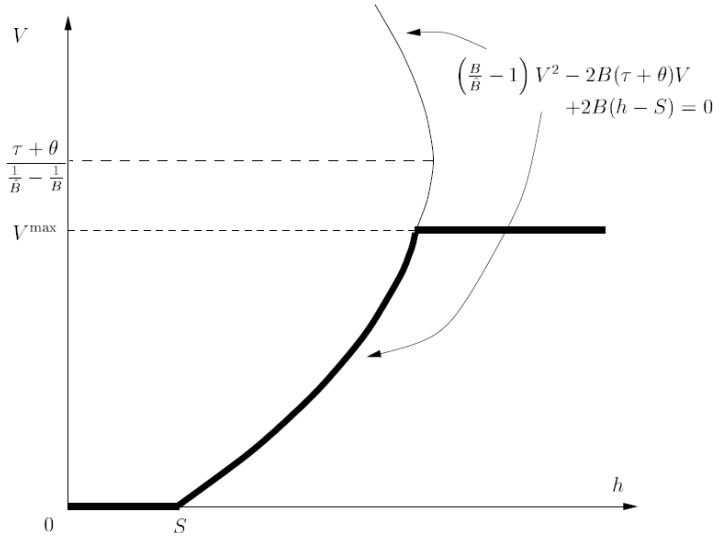


Figure 16. Sketch of the *speed-headway* function in uniform flow condition (taken from Wilson, 2001).

Figure 16 shows that the *speed-headway* function could be multi-valued at some points, for specific sets of the parameters. Since it is widely accepted in the traffic engineering community that this function should be a single-valued non-decreasing function, a constraint needs to be set.

The rightmost point of the parabolic curve (see Figure 16) is obtained at:

$$V=\frac{\tau+\theta}{\frac{1}{\hat{b}_{n-1}}-\frac{1}{b_n}} \tag{36}$$

Hence problems occur if:

$$V_n^{Max}>\frac{\tau+\theta}{\frac{1}{\hat{b}_{n-1}}-\frac{1}{b_n}} \tag{37}$$

Thus, the *speed-headway* function, obtained in uniform conditions, is well-defined if:

$$V_n^{Max} \le \frac{\tau + \theta}{\frac{1}{\hat{b}_{n-1}} - \frac{1}{b_n}} \tag{38}$$

It is worth noting that condition in equation (36) is valid under the assumption of uniform flow, which can be never reached in real traffic conditions. Nevertheless, if we assume that the follower's speed and the headway at each simulation step are representative of a (possible) steady state solution in uniform flow, the non-linear constraint in equation (36) still holds. Moreover, such constraint does not preserve the model parameters from generating a global instable car-following regime, in uniform flow. Indeed, the relation which must hold at the onset of linear instability is the following (see Wilson, 2001 for the details):

$$v^* > \frac{\theta}{\frac{1}{\hat{b}_{n-1}} - \frac{1}{b_n}} \tag{39}$$

Since $v^* \le V_n^{Max}$, the region of the instable parameters, for a well defined *speed-headway* function, can be derived from the following inequalities:

$$\frac{\theta}{\frac{1}{\hat{b}_{n-1}} - \frac{1}{b_n}} < v^* \le V_n^{Max} \le \frac{\tau + \theta}{\frac{1}{\hat{b}_{n-1}} - \frac{1}{b_n}} \tag{40}$$

This identifies the range of useful parameter values, where a uniform flow is unstable, as there is at least one unstable mode, yet the *speed-headway* function is properly defined at all headway arguments.

7.4 Tested algorithms

The algorithms used in this case study are the following:

- Downhill Simplex
- Genetic Algorithm
- OptQuest Multistart

Since none of the cited algorithms is considered to be a global optimization tool, each calibration experiment (i.e. a problem with a defined combination of Algorithm/MOP/GOF) was repeated 64

times, by using different starting points (in the case of gradient-based algorithms) or different random seeds (in the case of search-based algorithms). This approach allowed us to perform an analysis of local minima, in order to evaluate the power of the heuristic towards the finding of the (existing) global solution.

The different starting conditions were sampled using the Sobol' LPτ low-discrepancy quasi-random sequence, coded in MATLAB (Levitan et al., 1992), which is often used to explore the parameters' domain when conducting Global Sensitivity Analysis.

7.4.1 Downhill Simplex setup

In our calibration experiments, we adopted the algorithm code embedded in MatlabR2009b (Mathworks Inc., 2009). Since the algorithm does not allowed the setting neither of parameters bounds nor of constrained, we applied the following penalty function:

$$ObjFuncValue = \begin{cases} 100{,}000 & if\ x \notin D \\ ObjFuncValue & elsewhere \end{cases}$$

where:

- $ObjFuncValue$ is the value of the objective function related to the calibration experiment;
- x is the set of parameters value chosen by the algorithm at each functional evaluations;
- D is the domain of feasibility of the parameters, constrained by the upper and lower bounds (see Section 7.1), and by the non-linear constraints (see Section 7.2);

Stringent termination criteria were set in order to try avoiding to get stuck in local minima. Here are the defined stopping rules:

- Maximum number of function evaluations allowed is 100,000;
- Maximum number of iterations allowed (i.e. the maximum number of non-stationary points that can be found) is 100,000;
- Termination tolerance on the function value is 1e-30;
- Termination tolerance on the parameters values is 1e-30.

Since the Nelder-Mead technique do not assure the convergence towards a global solution, each calibration experiment was repeated starting from 64 different initial points.

7.4.2 Genetic Algorithm setup

As for the downhill simplex, stringent termination criteria were set in order to try avoiding to get stuck in local minima. Here are the defined stopping rules:

- Maximum number of generations (i.e. the maximum number of iterations allowed) is 100,000;
- Maximum number of stalling generations (i.e. with no improvements in the objective function) is 1,000;
- Cumulative change in the fitness function value over the maximum number of stalling generations is less than 1e-6.

Since it is not a global optimizer, the genetic algorithm could face difficulties in finding a stationary global solution. However, the genetic algorithm can sometimes overcome this deficiency with the right settings. Indeed, with a large population size, the genetic algorithm searches the solution space more thoroughly, thereby reducing the chance that the algorithm will return a local minimum that is not a global minimum (Powell, 1973). Concurrently, a large population size also causes the algorithm to run more slowly. Thus, a compromise was found and the number of individuals in each generation was set equal to 20.

Moreover, the genetic algorithm is not a gradient-based optimizer but a search-based technique. Since it is stochastic - that is, it makes random choices - you get (even very) slightly different results each time you run the algorithm. The implemented code uses the default MATLAB pseudorandom number stream and, each time the algorithm runs, the state (i.e. the seed) of the random stream state changes. In this view, the sensitivity of the choice of the random seed in the finding of the (possible) global solution was also tested by running the same calibration experiment with 64 different random seeds.

7.5 Tested Measures of Performance (MoPs)

In the case of calibration of car-following models the measures of performances should capture the dynamics of the phenomenon as it develops (Punzo et al., 2005). In the studies reviewed in the literature, the MOPs used to this aim were the following:

- Time series of the follower's speeds (V);
- Time series of the intervehicle spacing between leader and follower (S);

7.6 Tested Goodness of Fit functions (GoFs)

Considering the previously mentioned experiences, in this application we used the following GoF measures:

- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- GEH Statistic (GEH)
- Theil's Inequality Coefficient (U)

7.7 Summary of the experiments

According to the setting presented in this section, each calibration experiment was defined as an optimization problem given the solving heuristics and the response surface, which is univocally defined by the choice of the measure of performance and the functional form of the objective function.

Combining the 3 tested optimization algorithms (Downhill Simplex, Genetic Algorithm) and the 9 different response surfaces (RMSE(V), RMSE(S), MAE(V), MAE(S), GEH1(V), GEH1(S), U(V), U(S) and U(V)+U(S)), it resulted into 27 experiments. Moreover, each calibration problem was solved 64 times (i.e. 64 replications), in order to investigate the stability of the solution, thus resulting in a total number of 1728 calibrations of the parameters of the model.

7.8 The analysis of local minima

In this section, the study of the solutions of the calibration experiments is presented.

7.9 Cobweb plots

We adopted a graphical representation of the map of the solutions of each different calibration experiment. The so-called *Cobweb* plots were used for this purpose. Basically, they are line charts that display information as a series of data points (vertexes) connected by straight line segments. Unlike the time series, the horizontal axis is made of different categories and the vertexes of the plotted line are the values associated to each category. Since the range of values associated to different categories (for example, the model parameters) can be even very wide, a normalization of those values was required, limiting the range of variability between 0 and 1 for each category.

The *Cobweb* plots were constructed as follows. The categories were:

- The number of evaluations of the objective function (Nr_of_Iter) when the stopping conditions were reached;

- The value (Validation_Score) of a defined function, which was adopted to compare the solutions of the different calibration experiments. The chosen validation function was the sum of the Theil's Inequality coefficients related to speed and spacing (U(V)+U(S)), and its value was computed using the resulting calibrated parameters;
- The optimal value of the objective function (Obj_Funct), resulting from the current calibration experiment;
- The values of the calibrated model parameters (Tau=τ, Max_Vel=V_n, Max_Acc=a_n, Safe_Dist=$Safety$, Est_Dec=$\hat{b}_{n-1}$, Max_Dec=b_n);

In order to compare the results of all the calibration experiments with a certain optimization algorithm (i.e. 64x9 calibrations), the Validation_Score was used for this purpose. Therefore, the values of this category were normalized between the minimum and the maximum Validation_Score among all the calibrations with a specific optimizer. Further, to give visual information of the best overall solutions (i.e. those associated to the minimum Validation_Score, as defined before) among all these calibrations, a color bar was added.

Regarding the number of evaluations of the objective function, they were normalized between 1 and the maximum number among all the calibrations with a specific optimizer (i.e. 64x9 calibrations).

Conversely, the optimal value of the objective function, resulting from a single replication of the same calibration experiment, was normalized between 0 and the maximum value among the results of all the 64 replications with a specific GOF and optimization algorithm. Moreover, in the Cobweb plot, the bold line is associated to the results of the replication where the objective function was the minimum.

Concerning the optimal values of the parameters resulting from each replication (i.e. independently from the optimization algorithm and the objective function), they were normalized between the lower and upper bounds of the parameters.

An example of the Cobweb plots described above is shown in Figure 4.

It is worth noting that, according to the normalization methodology used, when the algorithm finds the "known" global solution of the optimization problem, both the validation score and the objective functional value will be equal to 0.

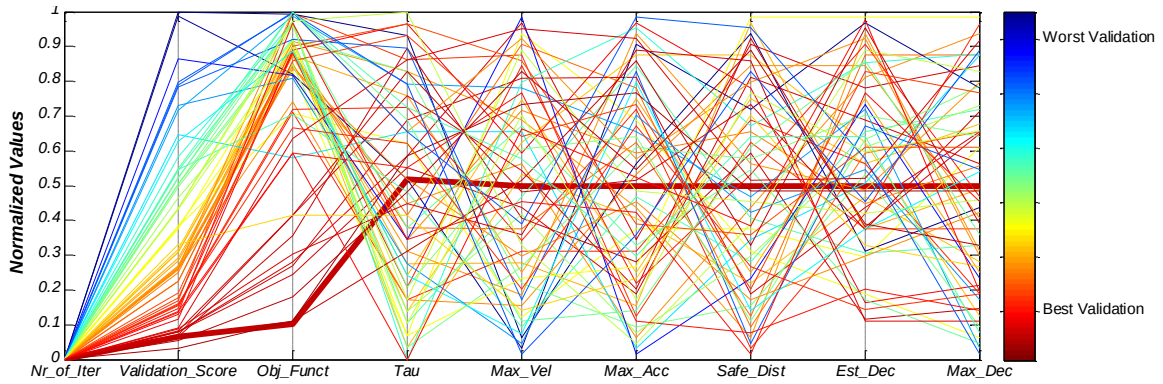


Figure 17. Example of a *Cobweb* plot for a calibration experiment with a specific optimization algorithm. Nr_of_Iter is normalized between 1 and the maximum among all the calibrations with a specific optimizer (i.e. 64x9). Validation_Score is normalized between the minimum and the maximum among all the calibration with a specific optimizer (i.e. 64x9). Obj_Funct is normalized between 0 and the maximum among the 64 replications of a calibration experiment with a specific optimizer. Parameters' values are normalized between their upper and lower bounds. The different colors reflects the rank of the validation score. The bold line is associated to the results of the calibrations with the minimum value of the objective function.

7.10 Results

In the present section, we are reporting the results of the experimental plan. They are reported in the following table and figures. In particular, the following figures show the *Cobweb* plots related to the calibration experiments performed on the Gipps' car-following model. The OPI presented in Table 6 refers to the description provided in section 6.5 (equation 19).

The main outcomes can be summarized as follows.

1. The analytical formulation of the GEH Statistics, even with a strict threshold value set at 1 (in the place of 5, which is considered to be a good match between the observed and the model simulated outputs (Ma et al., 2007)), does not allow any algorithm to find the global solution, that is the "true" value of the parameters of the synthetic data. All the algorithms converge (more or less frequently) to points where the value of the objective function is 0, but none of them corresponds to the real ("known") global minimizer.

2. The Downhill Simplex algorithm is unable to find the global solution with none of the tested GoF functions. Further, the algorithm is not even robust towards the starting conditions, since it converges in almost all the replications to different optimal solutions and the minimum optimal one is found just once. As a consequence, the algorithm gets always stuck in local minima.

Table 6. Analysis of the performances of each calibration procedure

ALGORITHM	GOF	*OPI* at the minimum Obj_Funct point	Frequency of the Minimum *Obj_Funct value* (%)	Frequency of the *true* set of parameters ± 5% error (%)	*Total OPI*
Downhill Simplex with penalty function	RMSE(V)	9.01E-02	2	0	42.21
	RMSE(S)	5.54E-01	2	0	51.16
	MAE(V)	3.18E-01	2	0	49.62
	MAE(S)	3.57E-01	2	0	52.00
	GEH1(V)	More than one value	14	0	57.29
	GEH1(S)	More than one value	6	0	55.38
	U(V)	1.73E-01	2	0	41.55
	U(S)	4.99E-01	2	0	52.23
	U(V)+U(S)	3.63E-01	2	0	49.07
Genetic Algorithm with penalty function	RMSE(V)	1.76E-02	2	5	11.54
	RMSE(S)	1.03E-01	2	0	21.60
	MAE(V)	2.05E-02	2	3	13.21
	MAE(S)	1.36E-01	2	0	24.30
	GEH1(V)	More than one value	77	0	21.27
	GEH1(S)	More than one value	28	0	20.27
	U(V)	9.72E-03	2	3	12.90
	U(S)	1.44E-01	2	2	20.99
	U(V)+U(S)	4.65E-02	2	2	19.30
OptQuest Multistart	RMSE(V)	2.56E-05	75	75	5.62
	RMSE(S)	9.92E-04	34	34	17.36
	MAE(V)	5.66E-05	58	58	11.80
	MAE(S)	1.38E-03	2	44	16.44
	GEH1(V)	More than one value	100	0	33.87
	GEH1(S)	More than one value	61	0	31.37
	U(V)	2.56E-05	58	58	9.07
	U(S)	9.92E-04	25	25	19.59
	U(V)+U(S)	5.11E-04	23	23	18.82

3. The OptQuest Multistart is able to rediscover the “true” values of the parameters with very high frequencies and converging repeatedly to the global solution. It is the less “sensible” to the initial condition.

4. The Genetic Algorithm and the OptQuest Multistart performed very differently when optimizing on speed or on spacing. According to the findings, it emerges that the Genetic Algorithm was able to better reproduce the spacing profile when calibrating on speeds than when it works on the spacing itself. On the contrary, the OptQuest Multistart performs better when optimizing the spacing between the leader and the follower vehicle, without suffering underperformances in the speed profile.

5. The use of mixed GoF functions that combine both the MoPs (speed and spacing), such as the sum of Theil's Inequality coefficients, performed worse than when calibrating separately on speed or on spacing. Further, the use of absolute measures of the distance between observed and (model) simulated outputs, such as the MAE, entails very low efficiency in the optimization, as they require a high number of evaluations of the objective functions to satisfy the same stopping rules adopted with the other GoF functions. Moreover, the improvements in finding the global minimizer are negligible.

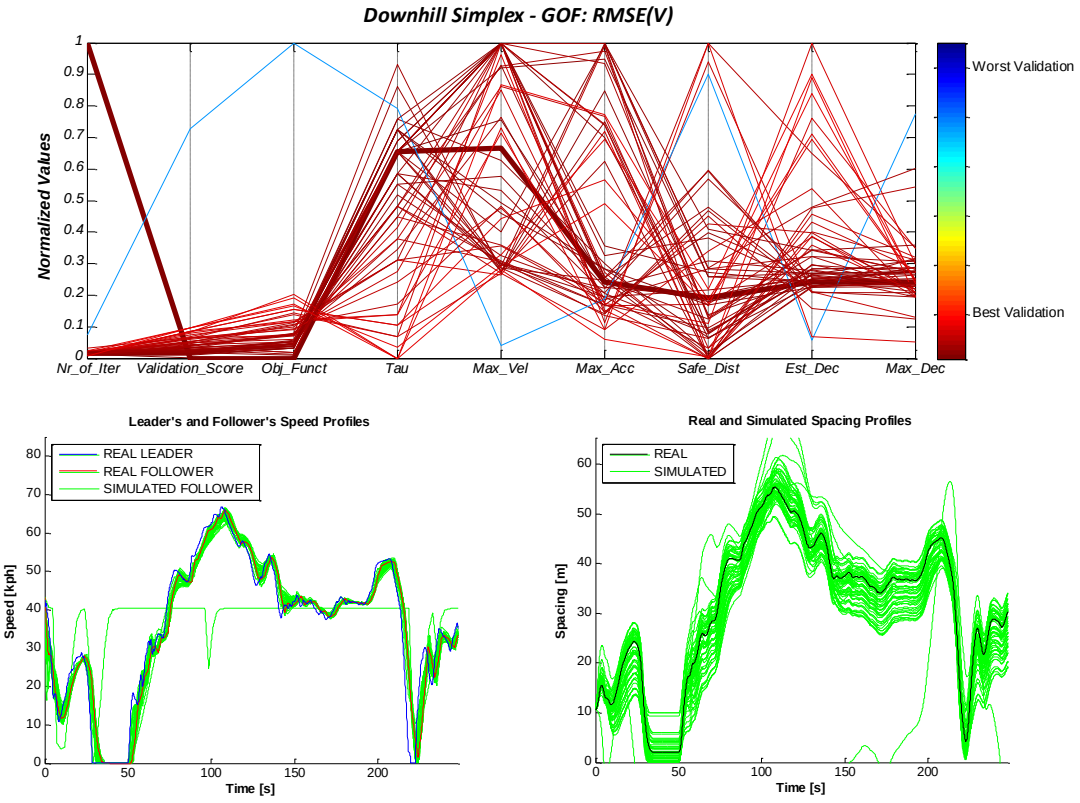


Figure 18. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with RMSE(V)

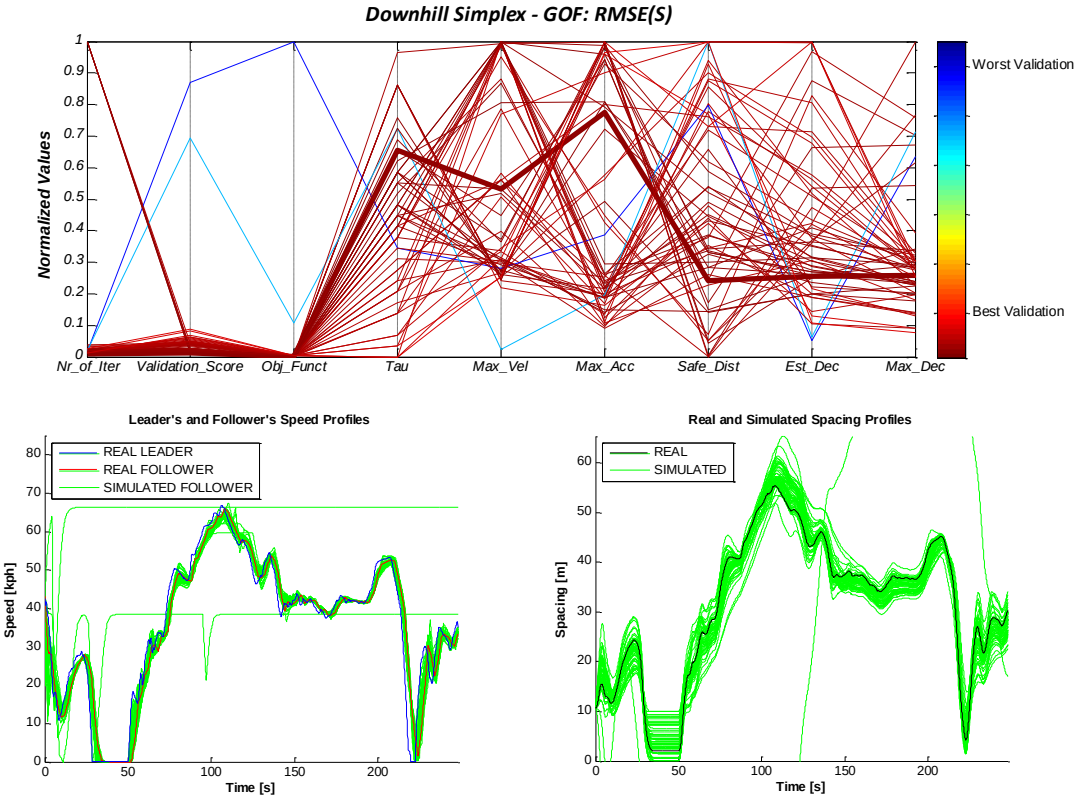


Figure 19. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with RMSE(S)

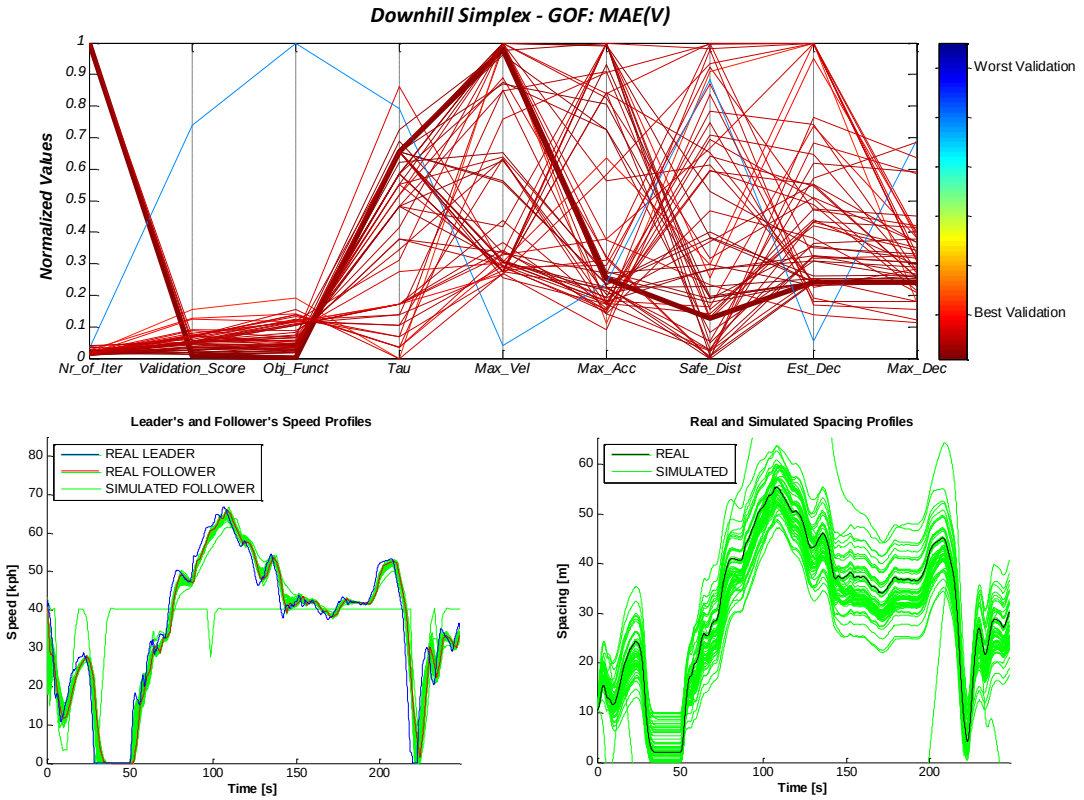


Figure 20. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with MAE(V)

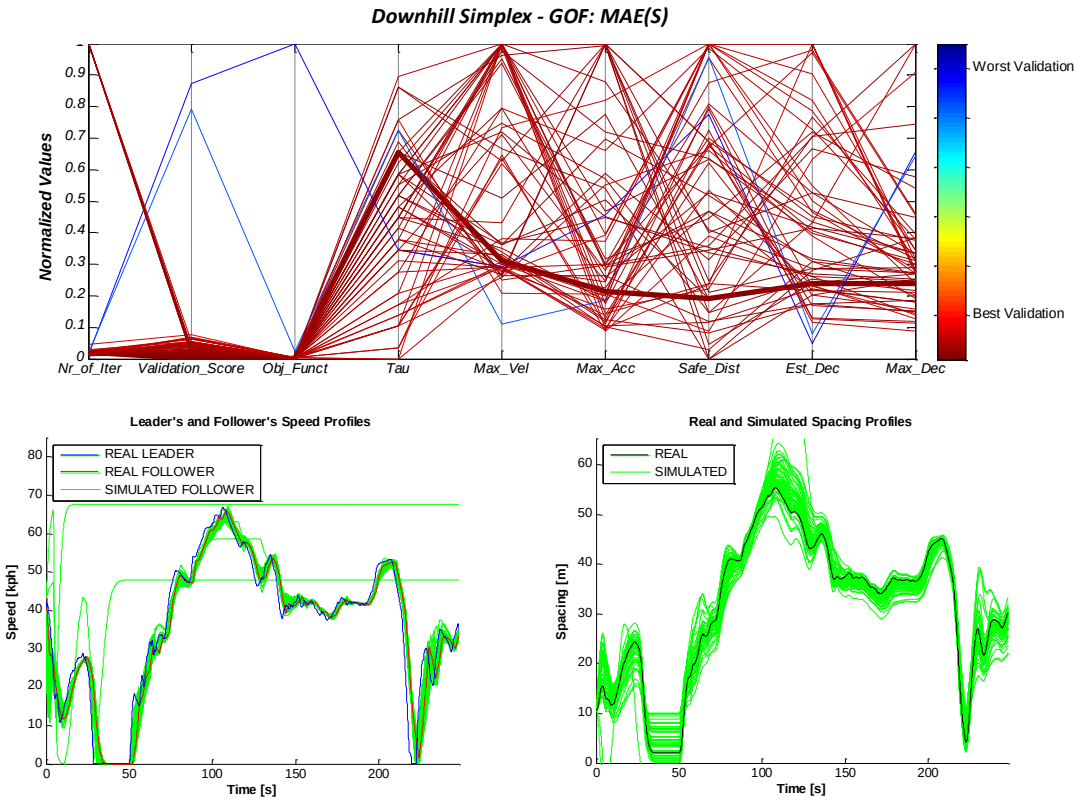


Figure 21. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with MAE(S)

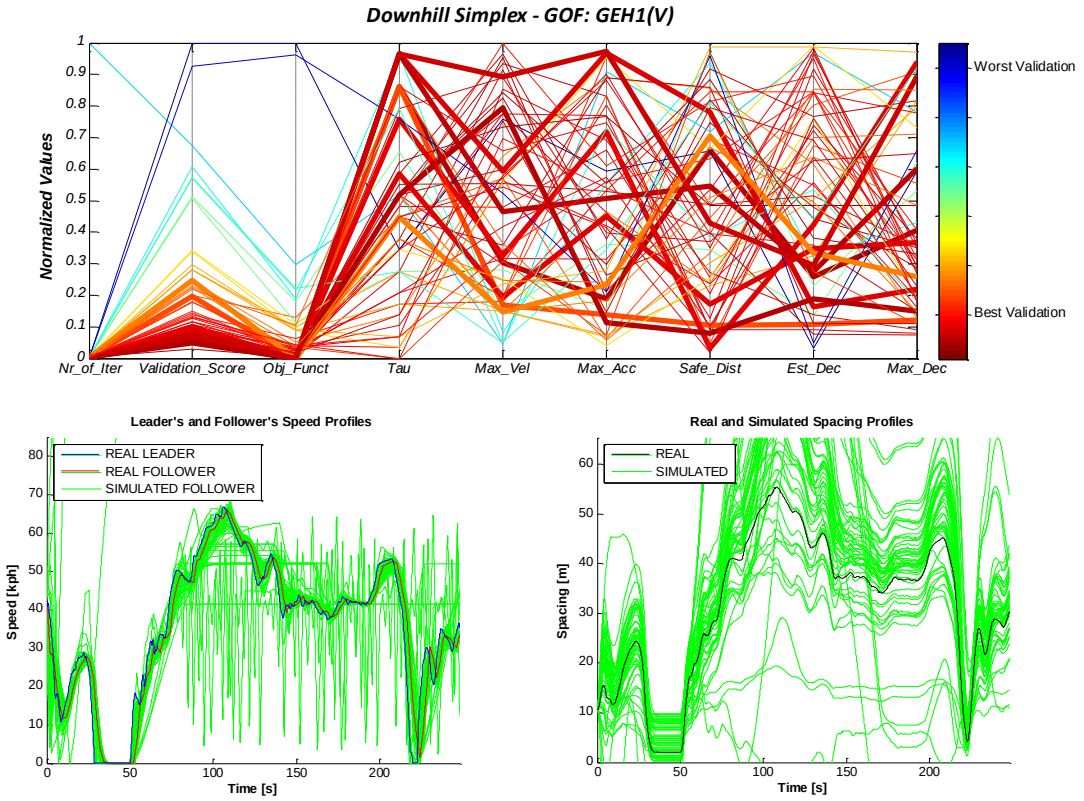


Figure 22. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with GEH1(V)

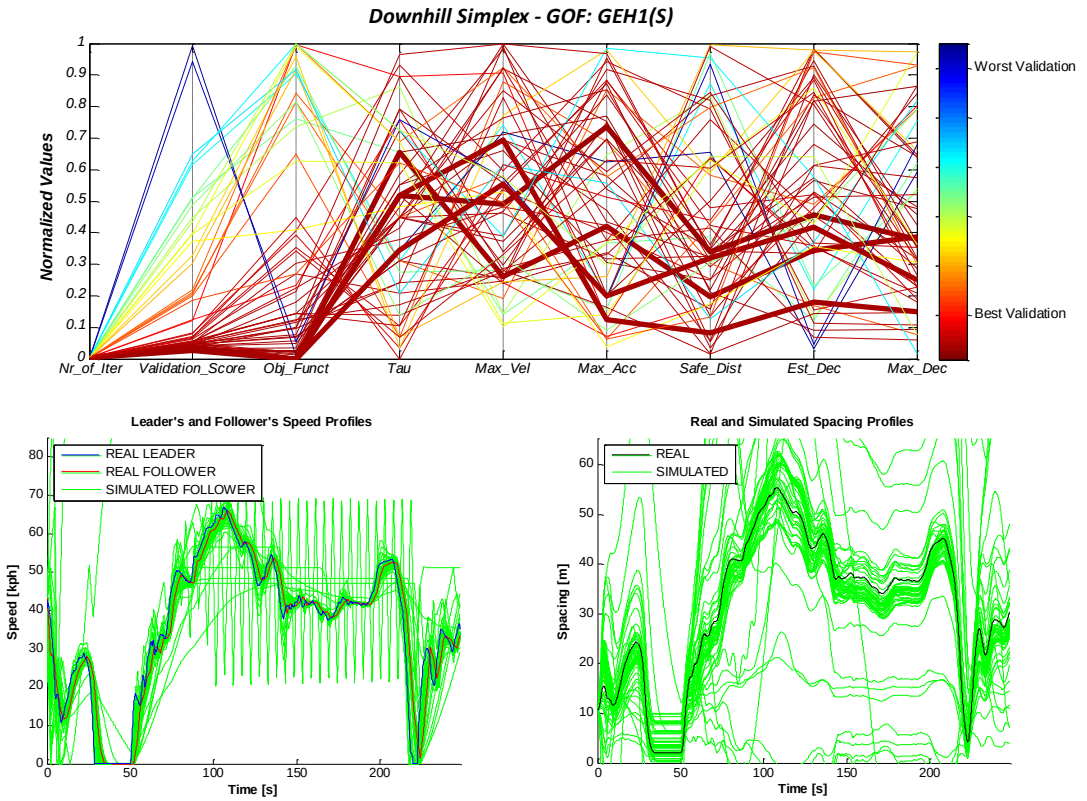


Figure 23. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with GEH1(S)

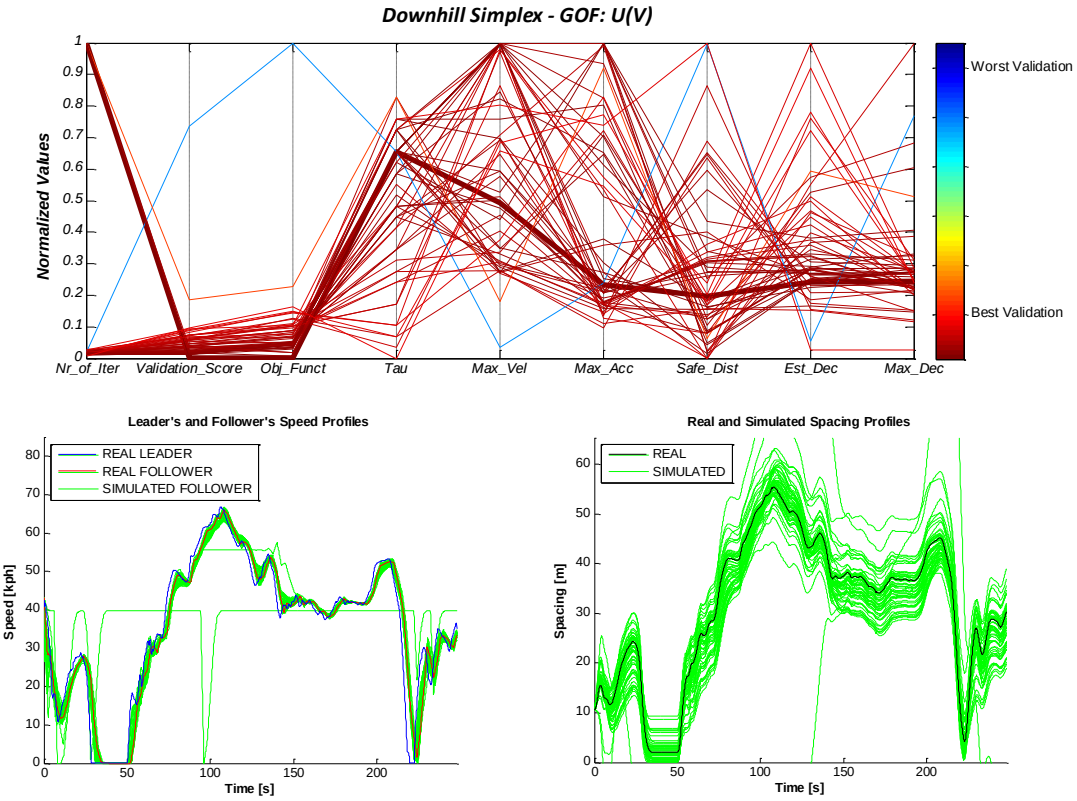


Figure 24. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with U(V)

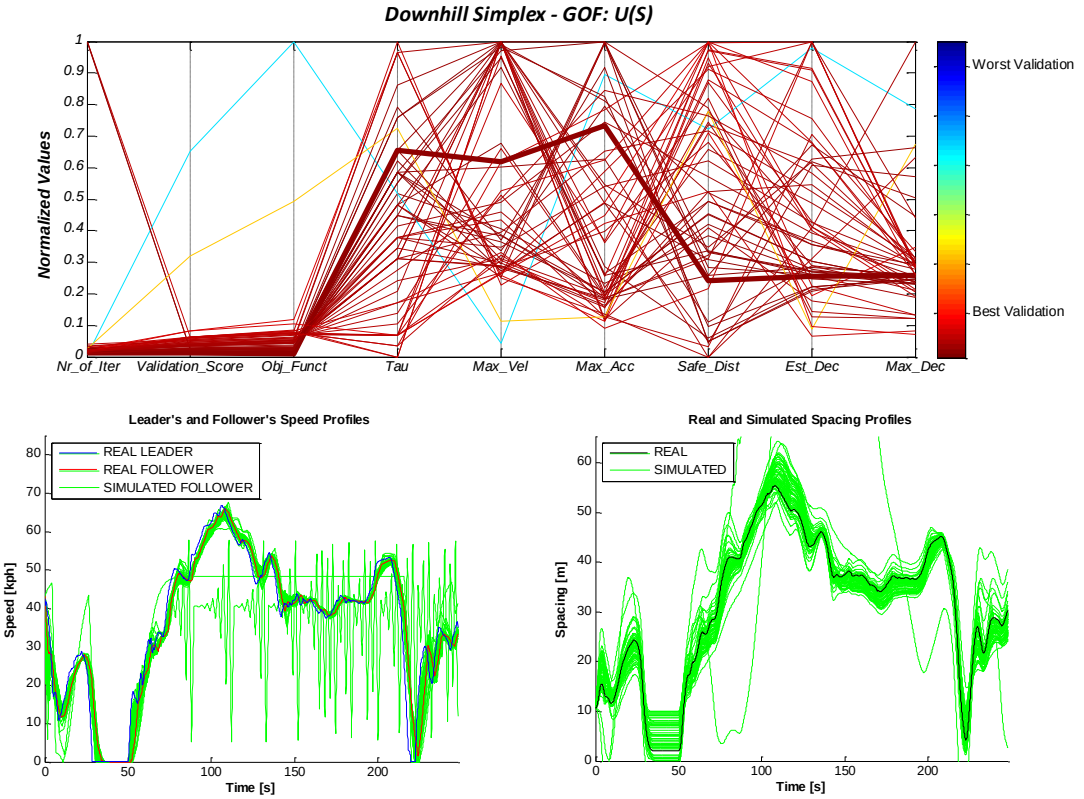


Figure 25. *Cobweb* plot, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with U(V)

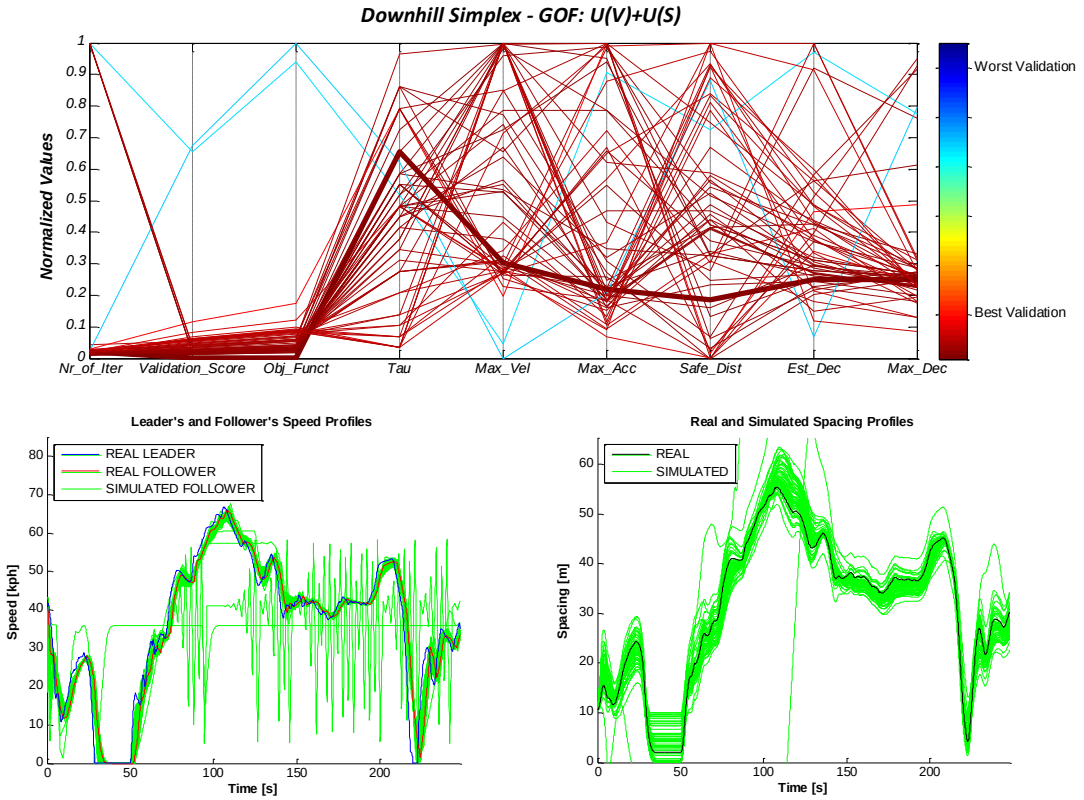


Figure 26. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Downhill Simplex with U(V)+U(S)

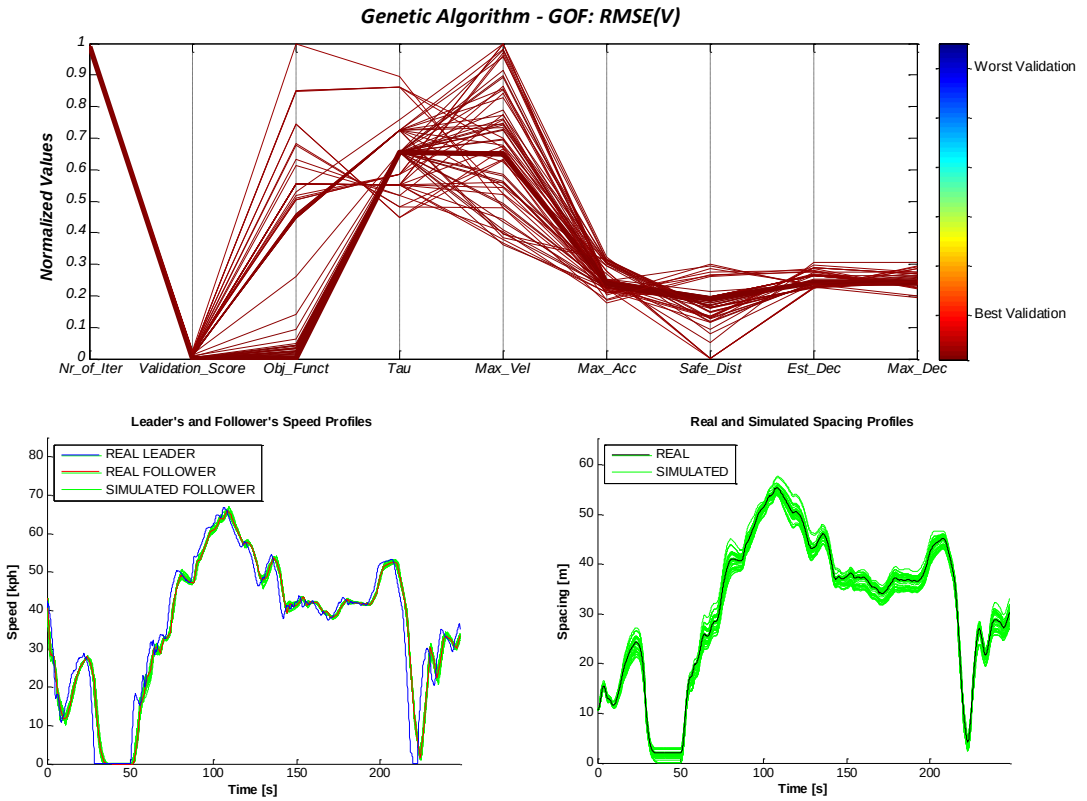


Figure 27. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with RMSE(V)

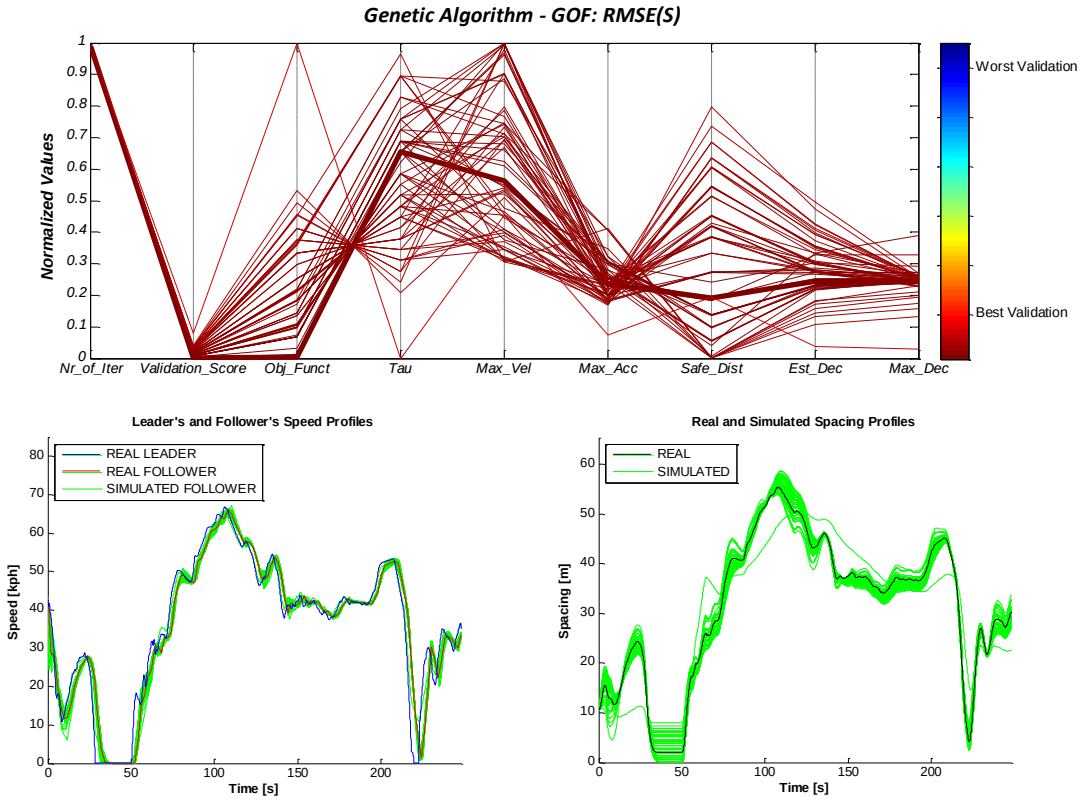


Figure 28. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with RMSE(S)

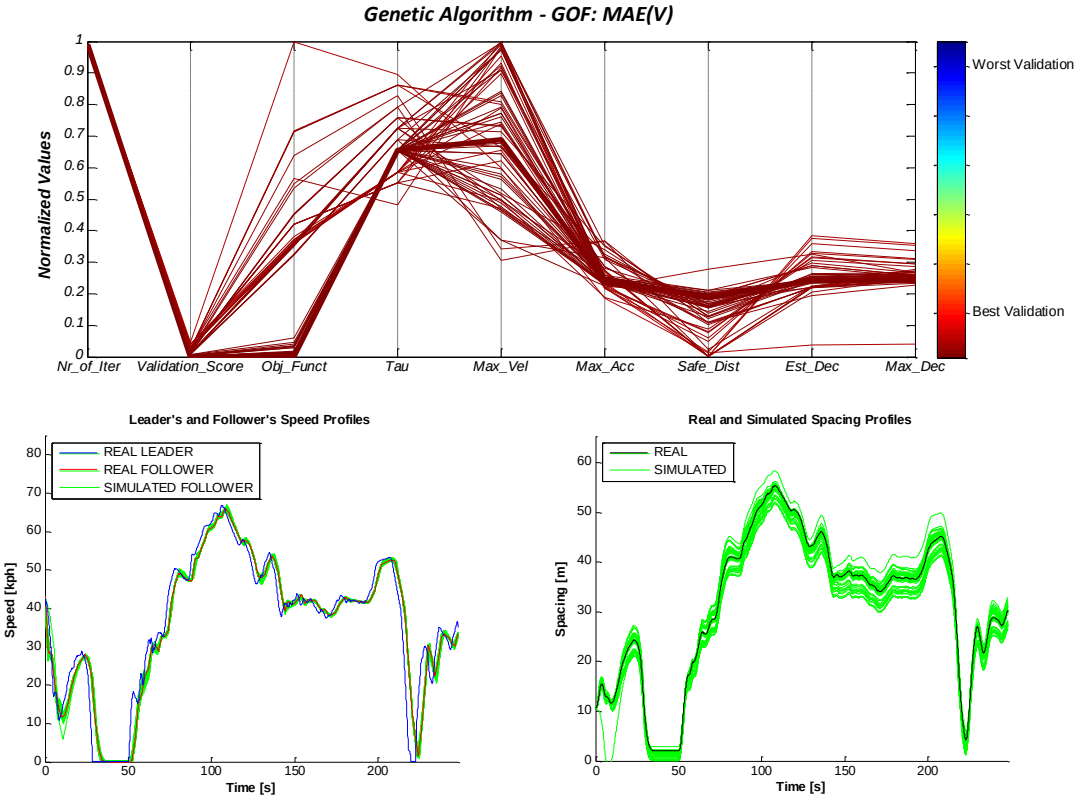


Figure 29. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with MAE(V)

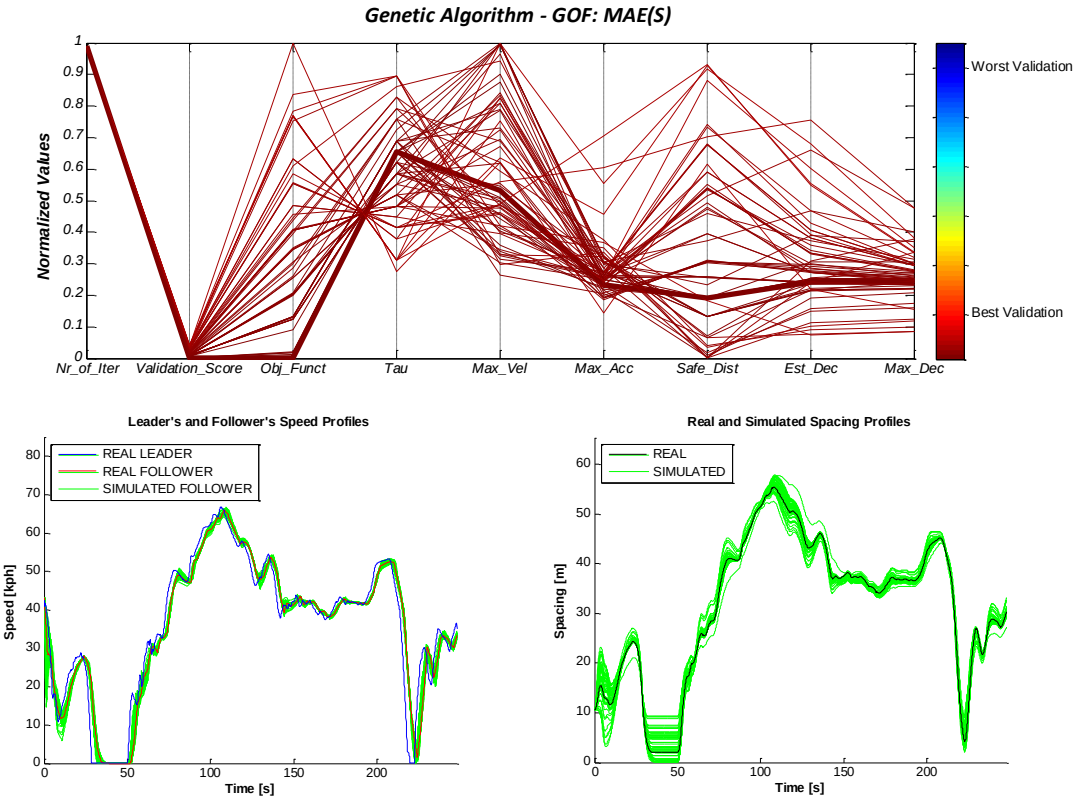


Figure 30. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with MAE(S)

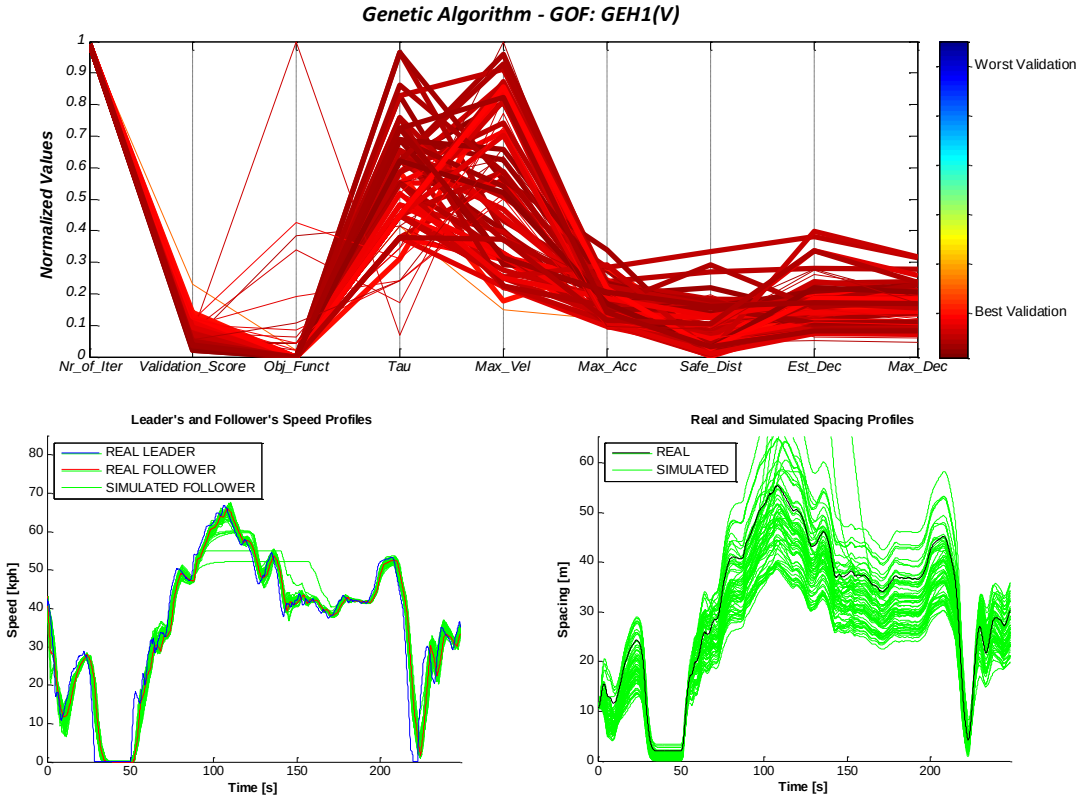


Figure 31. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with GEH1(V)

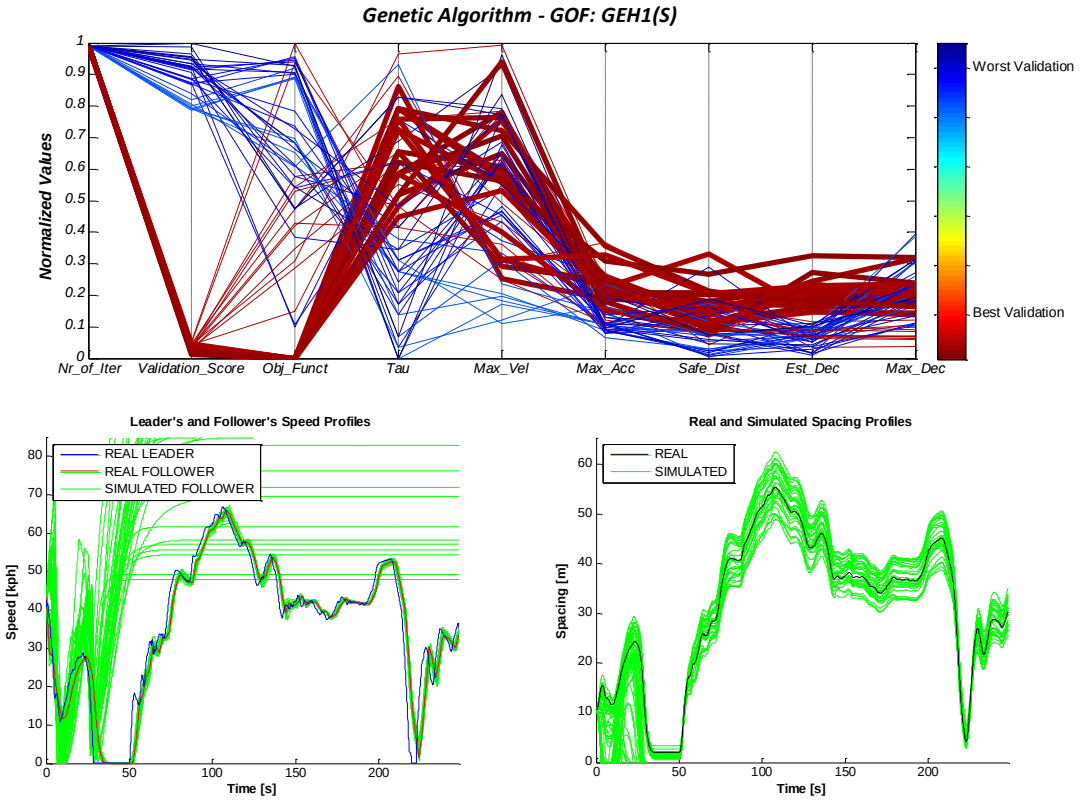


Figure 32. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with GEH1(S)

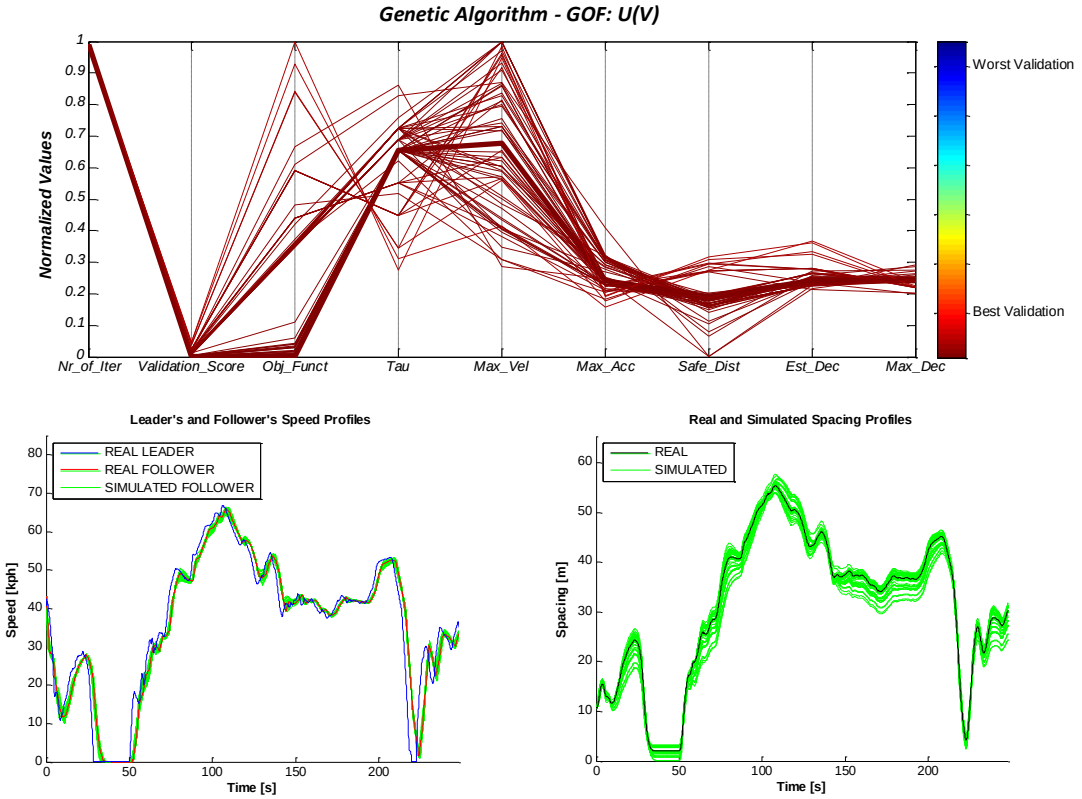


Figure 33. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with U(V)

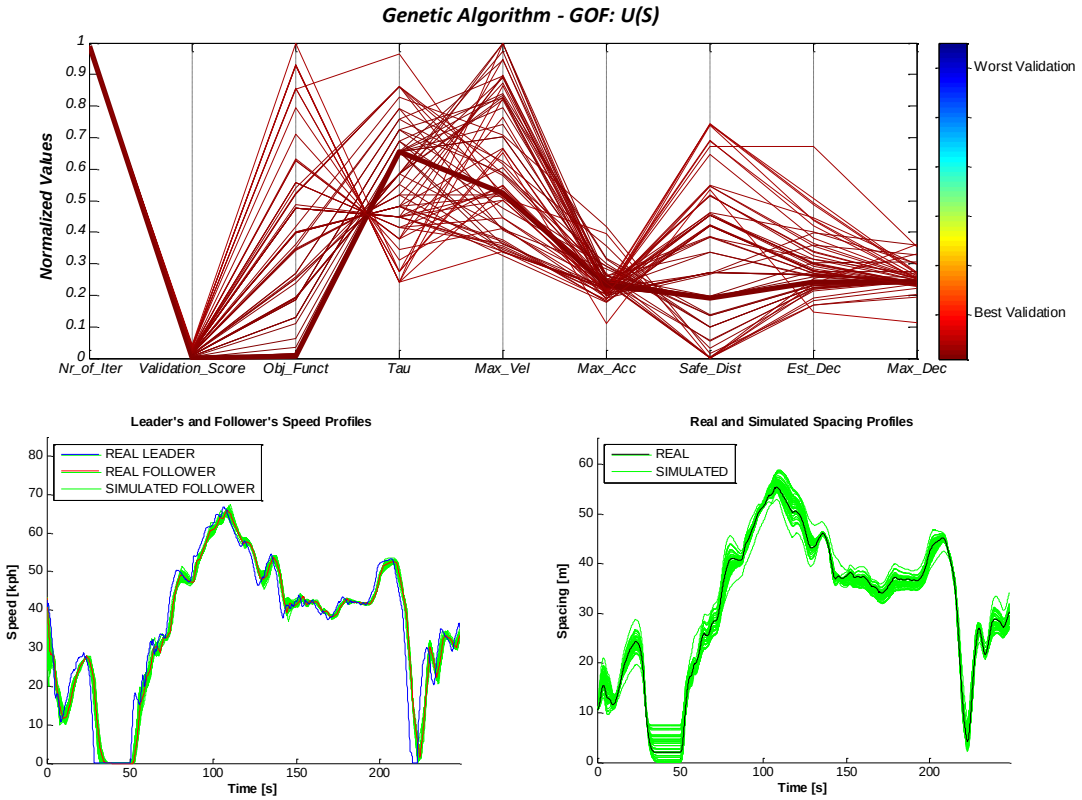


Figure 34. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with U(S)

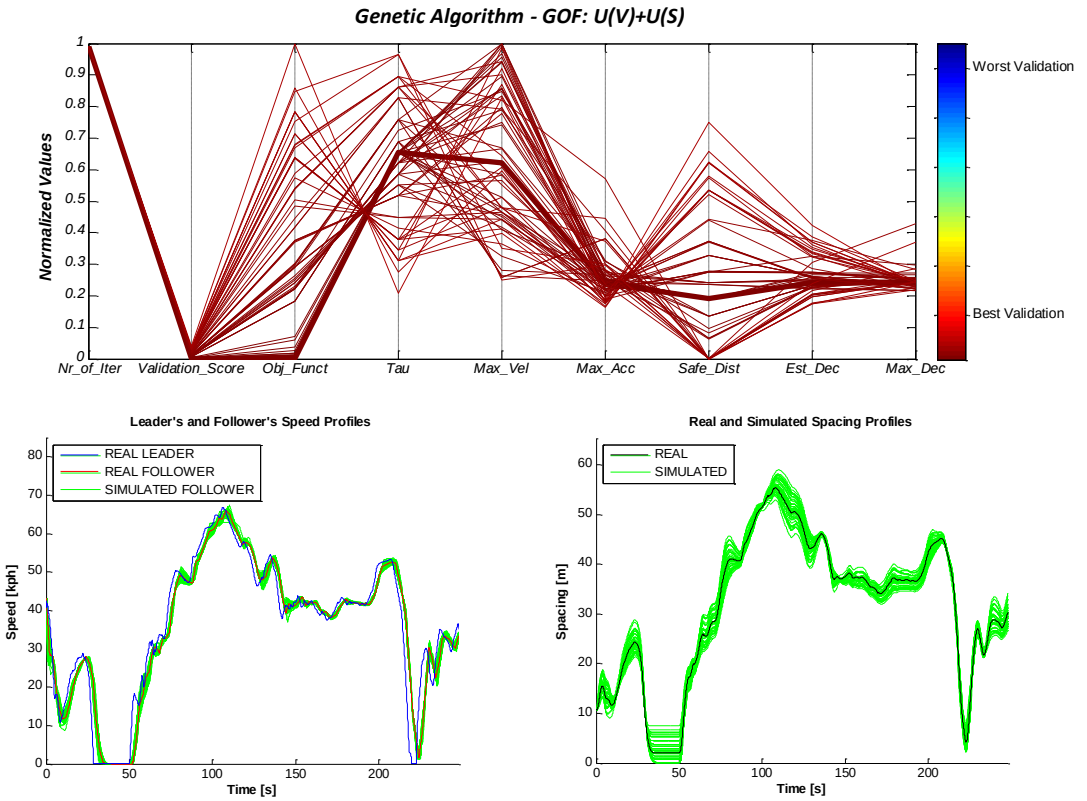


Figure 35. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the Genetic Algorithm with U(V)+U(S)

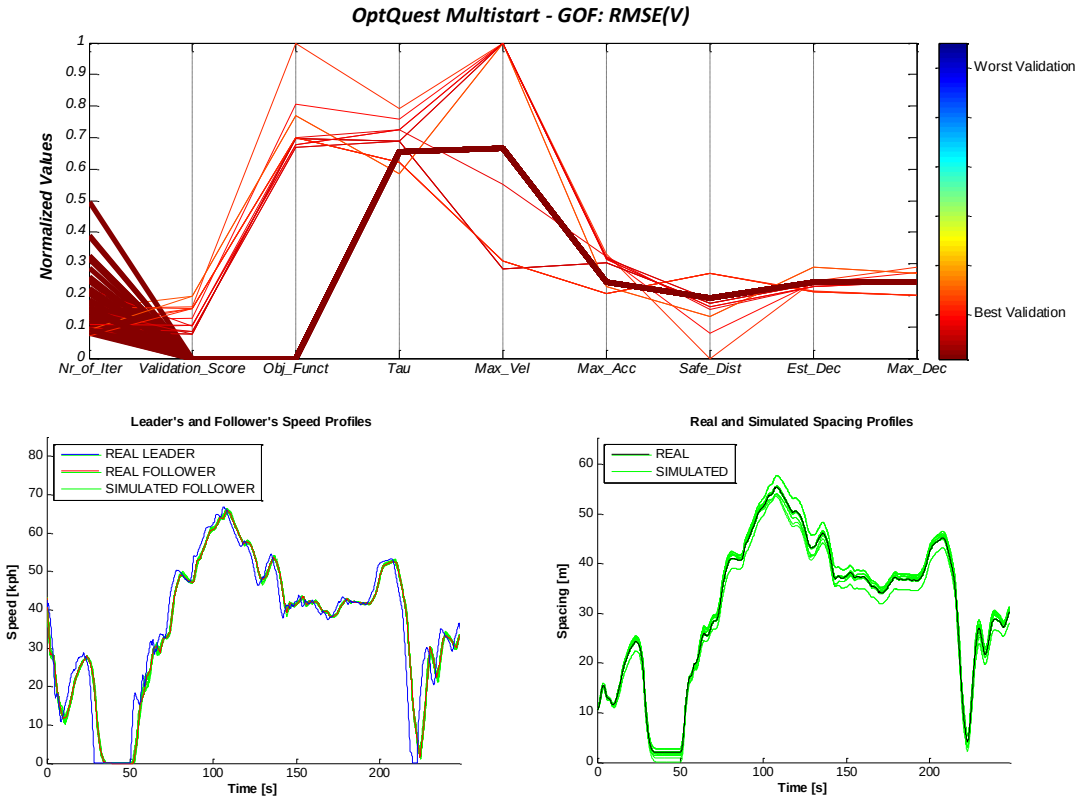


Figure 36. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with RMSE(V)

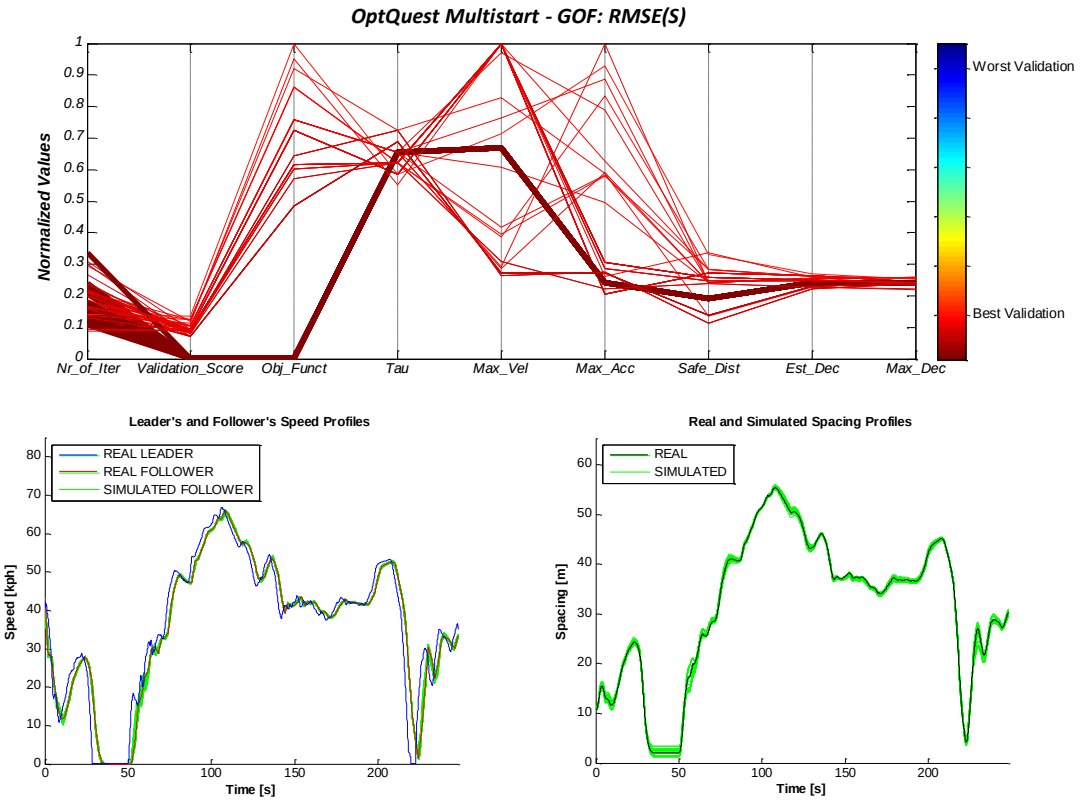


Figure 37. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with RMSE(S)

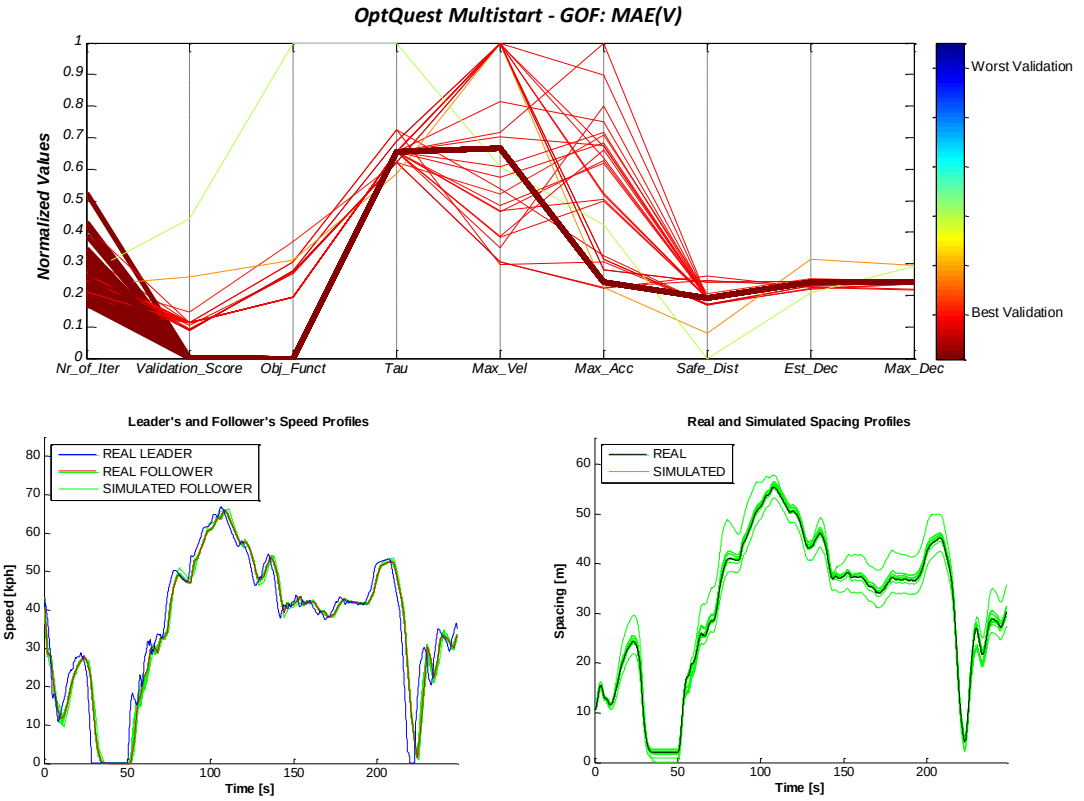


Figure 38. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with MAE(V)

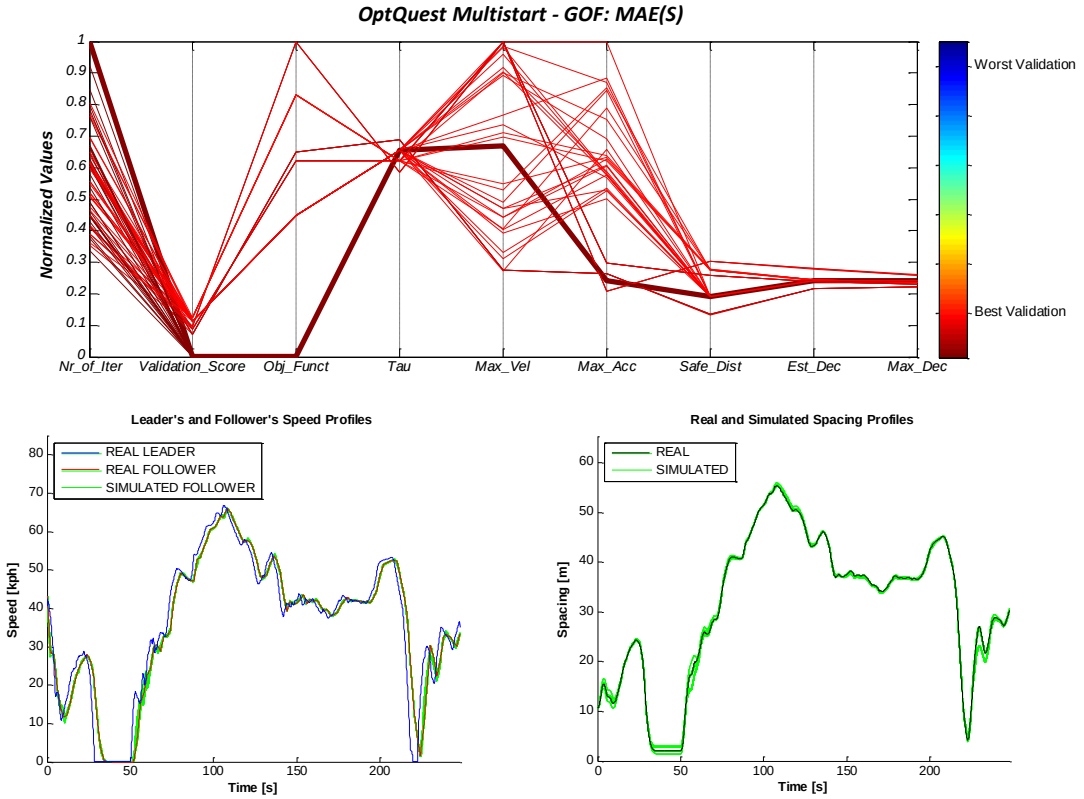


Figure 39. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with MAE(S)

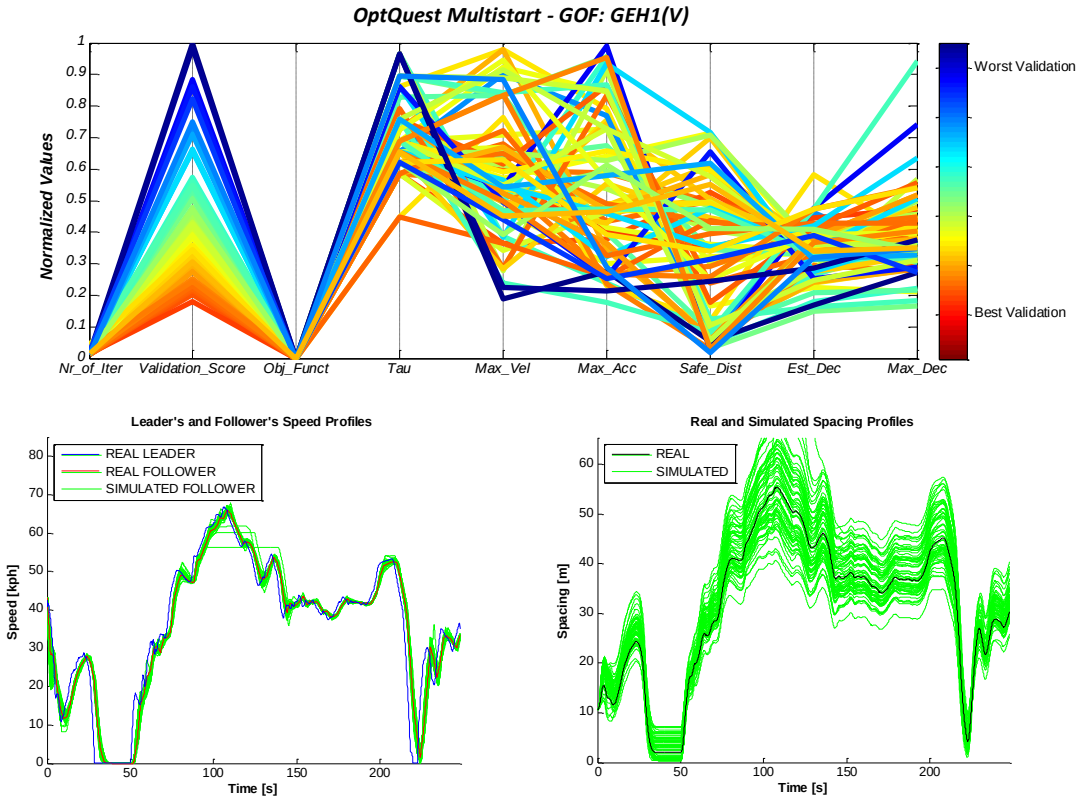


Figure 40. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with GEH1(V)

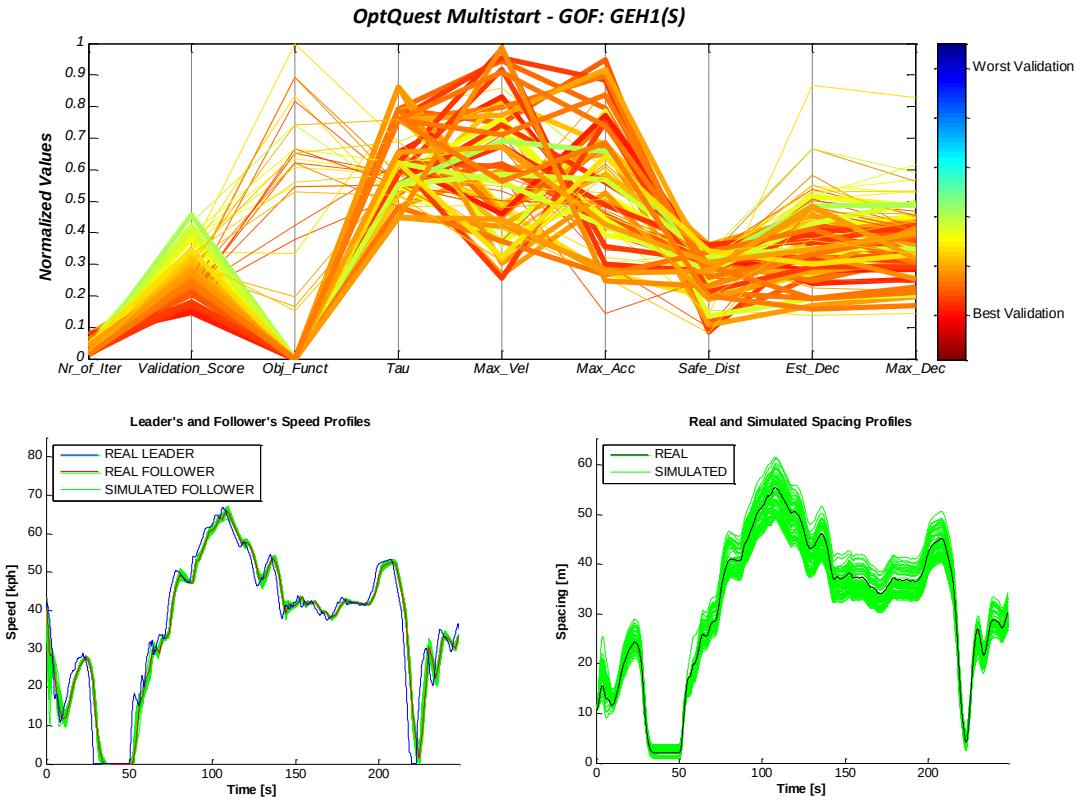


Figure 41. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with GEH1(S)

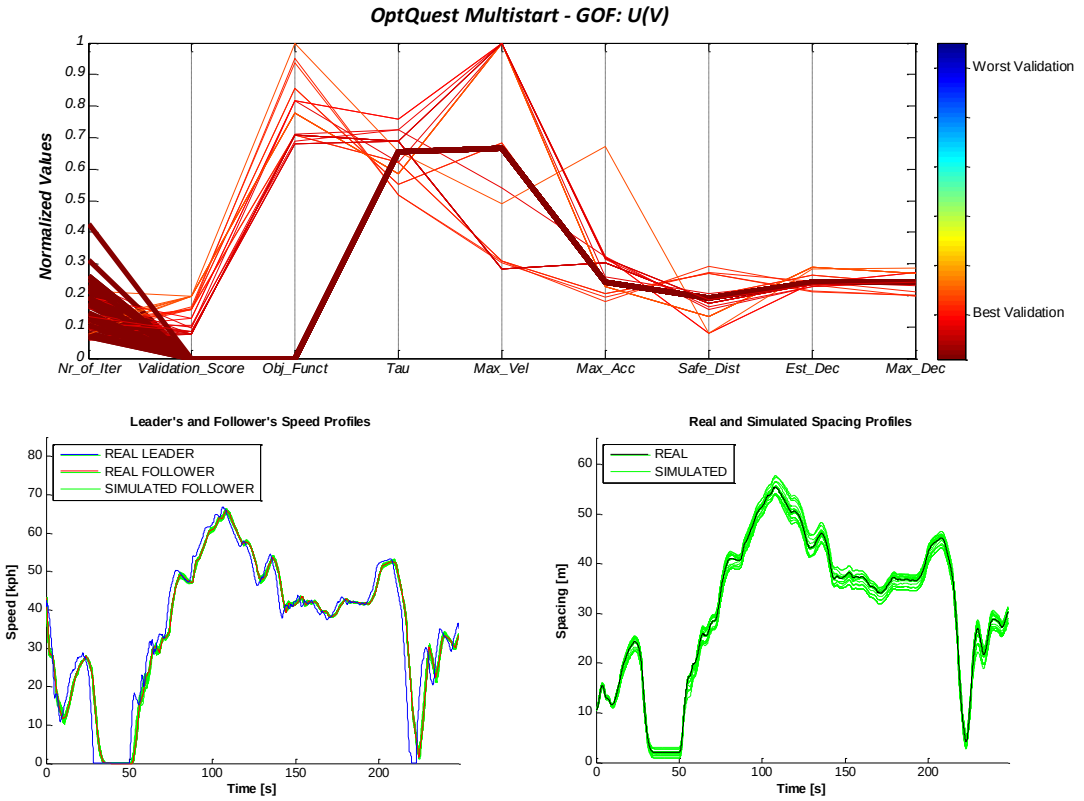


Figure 42. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with U(V)

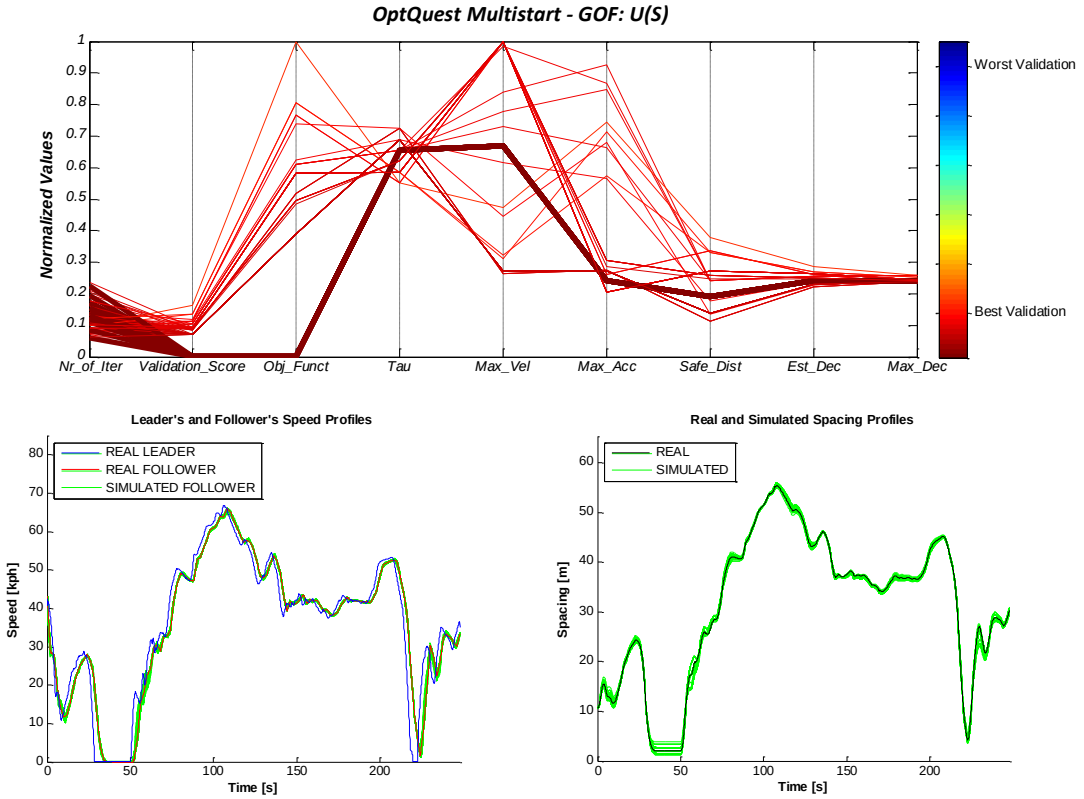


Figure 43. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with U(S)

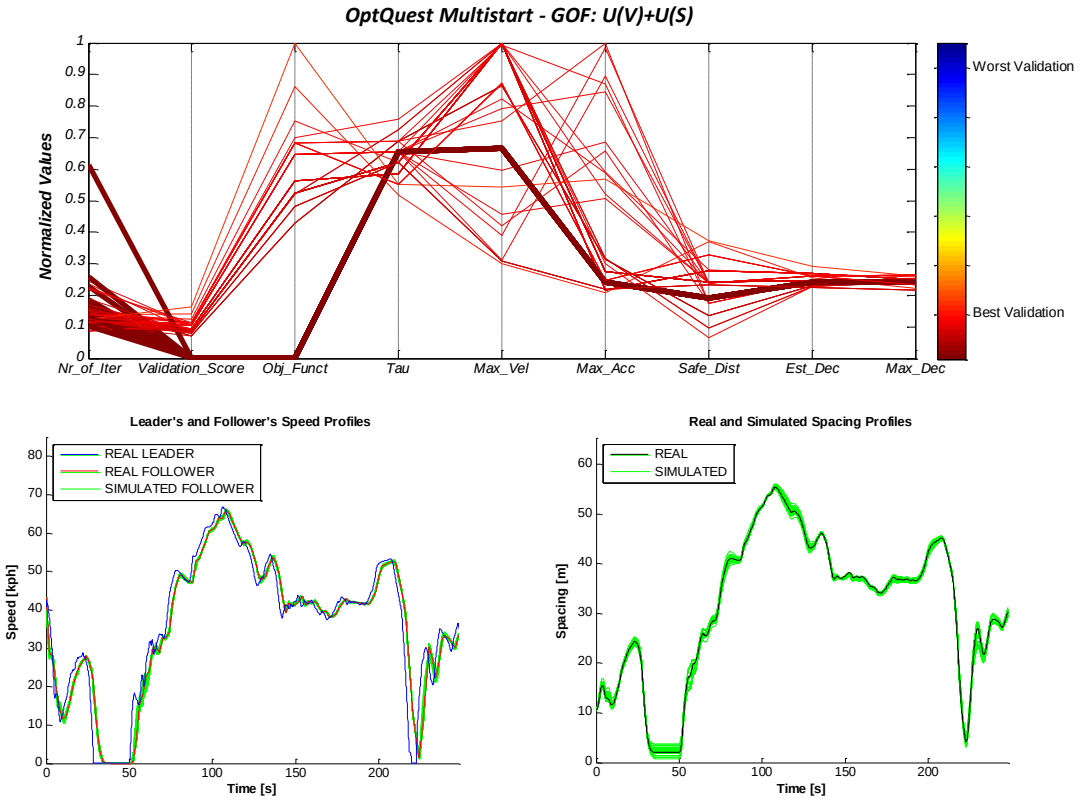


Figure 44. *Cobweb* plots, together with speed and spacing profiles, related to the calibration experiment using the OptQuest Multistart with U(S)

8 Conclusions and future research

In the last decades simulation optimization has received considerable attention from both researchers and practitioners. Simulation optimization consists in the use of simulation models to maximize the efficiency of the real system which is simulated. In simulation optimization a simulation model is then jointly used with a mathematical programming model.

A specific example of simulation optimization is the calibration of a simulation model. In traffic applications, more and more this issue is attracting the interest of researchers on traffic modelling. It represents an indispensable preliminary phase for using traffic models to evaluate investments on transportation systems. Difficulties hidden behind this phase have however been fairly underestimated by scientists and practitioners so far.

The calibration of a microscopic traffic simulation models is an optimization problem in which the objective function is an analytical function (called goodness-of-fit measure) able to estimate the likeliness between the traffic data collected in the real transportation system and the same data as obtained by the simulation model. The parameters of the model are the variables of this optimization problem.

Since microscopic traffic models are simulation models, very few elements are available on the properties of the objective function. In particular, one does not know anything about the smoothness of the function, its linearity, its convexity, the existence of different local minima as well as the existence of a wide area with fairly constant value. Analyzing the properties of the objective function is even more urgent in this field in which different models are available and in which different scenarios may be simulated, producing significant change in the objective function. For this reason, in this paper it is claimed that the problem has to be seen in the framework of the “no free lunch” theory.

In this light, in the present paper, a general methodology to verify the effectiveness of the optimization framework implemented for simulation optimization is described and applied to the calibration of a microscopic traffic simulation model.

The methodology presented and specified is based on a phase of optimization “verification” implemented by means of tests with synthetic data. This procedure allows all the steps of a calibration process to be controlled and inferences to be drawn about the effect that several choices have on the calibration process.

Then we applied the methodology on two different case studies.

In the first one, the application of the proposed methodology allowed drawing inference about the effect on the calibration problem of different combinations of five optimization algorithms, seven measures of goodness-of-fit, two couples of model's parameters and three levels of noise in the data.

In order to complete the study in reasonable time, a Kriging surrogate of the simulation model has been used in its place. This is totally admissible since the Kriging meta-model owns both locally and globally the same mathematical properties of the real objective function (i.e. Kriging is not used to solve the optimization problem but to check if an optimization algorithm is able to solve another problem which has the same features of the original one). As an additional major outcome, the use of Kriging surrogates allowed for the algorithms' parameters to be fine-tuned (a practice which is usually impossible to be performed using simulation models and directly calibrating it with real data).

The results showed the usefulness of the procedure and the indicators used in order to understand the key points of the problem. Results give some preliminary indications concerning the necessity to perform the sensitivity analysis of the simulation model before its calibration, the importance to use an optimization algorithm able to deal with problems of global optimization and the preference of non-normalized measures of Goodness of Fit.

In the second case study the proposed methodology is applied to the Gipps' car-following model. The use of a synthetic follower trajectory was an essential requirement to assess the goodness of the specification of the optimization problem, since, in this case, the values of the model parameters' used to generate the trajectory were "known" (i.e. fixed by the authors), and thus the global solution of the optimization problem was well-defined.

In order to generate such a trajectory, the Gipps' car-following model was fed with the data of the leader taken from one of the experiments carried out on a two-lane extra-urban highway, in the area surrounding Naples. Data were acquired through instrumented vehicles, equipped with kinematic differential GPS receivers that recorded the position of the vehicle at each 0.1 seconds.

In the field of the calibration of car-following models, to the best of the authors' knowledge there is not a consolidate approach for the specification of this class of optimization problem, that is to define the combination of algorithm, MoP and GoF function to be used to calibrate the model. Further, the lack of evidence about the capability of each problem specification in finding the global minimum, led us to test all the combinations among the most used algorithms, MoPs and GoF functions to investigate the performances of the calibration procedure.

In this view, we tested three optimization algorithms, four measure of goodness of fit and two measures of performance. Both to inquire on the capability of each problem specification to rediscover the "true" values of the parameters and to evaluate the robustness towards the starting condition, a novel graphic method based on *Cobweb* plots was used. Further, this helped us to explore the existence and the nature of the local minima found by the algorithms, as well as to give insights into the MoPs and GoF functions used in the calibration experiments.

The approach adopted allowed identifying unsuitable measures of goodness of fit and optimization algorithms for the car following model calibration. In addition it was very useful to get novel insights into the topic.

Further research should be performed in both the optimization side and the traffic flow side of the issue. On one hand, further efforts are needed in order to understand the possibility to apply Kriging meta-models also in the place of the simulation model for simulation optimization, and the minimum number of simulations required to achieve a good surrogate of the simulation model. On the other hand, it is worth repeating the same analysis with different simulation models and different traffic scenarios, in order to highlight the different problems arising in the different cases and to provide the users with practical recommendations to be followed during studies of simulation optimization in transportation/traffic applications.

Again, the present study also confirmed the complexity of the problem of calibrating car-following models against real trajectory data. As a matter of facts, none of the tested settings gave completely satisfactory results, and future research shall necessary aim at finding more robust settings. Therefore, in the light of the previous findings, the following research lines have been envisaged by the authors, and already object of ongoing investigations:

1. to limit the calibration process to the most sensitive parameters, via e.g. sensitivity analysis of model outputs, in order to reduce both the number of dimensions of the input space and the flatness of the response surface. This would drastically decrease the complexity of the optimization problem;
2. to seek for "'global"' GOFs which were able to capture the inner structure/driving behaviour/driving style contained in the trajectory data, as expressed/interpreted by the specific model in use. This is also in the course of the recent studies performed by Chiabaut et al. (2010). Local GOF indeed are sensitive to errors in the data, and especially least square ones tend to compensate errors over the whole length of the trajectory;
3. to appropriately bound the space of the admissible inputs in order to preserve the well

established macroscopic characteristics of the traffic flow.

Previous points would contribute at the end to address the problem of model over-fitting (mostly relevant for car-following models given their manifest inadequacy) and to increase the transferability of calibration results.

Acknowledgements

Authors are grateful to James Spall, for his precious suggestions during the implementation of both the SPSA algorithms and to Egidio Quaglietta for his contribution to derive some of the results presented in this report.

Research contained within this paper benefited from the participation in EU COST Action TU0903 MULTITUDE – Methods and tools for supporting the use calibration and validation of traffic simulation models.

References

Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., Wen, Y., 2007. Calibration of Microscopic Traffic Simulation Models: Methods and Application. Transportation Research Records: Journal of the Transportation Research Board, No. 1999, TRB, National Research Council, Washington, D.C.

Ben Akiva M., Darda, D., Jha, M., Koutsopoulos, H.N., Toledo, T., 2004 Calibration of Microscopic Traffic Simulation Models with Aggregate Data. Transportation Research Records: Journal of the Transportation Research Board, No. 1876, TRB, National Research Council, Washington, D.C.

Brockfeld E., Kuhne, R.D., Wagner, P., 2005. Calibration and Validation of Microscopic Models of Traffic Flow. In Transportation Research Records: Journal of the Transportation Research Board, No. 1934, TRB, National Research Council, Washington, D. C., pp. 179-187.

Chang, T.-H., Li, Z.-Y., 2002. Optimization of mainline traffic via an adaptive co-ordinated ramp-metering control model with dynamic OD estimation. Transportation Research Part C: Emerging Technologies. Vol. 10 (2), pp. 99-120

Chen, C.-H., Schonfeld, P., Paracha, J., 2005. Work zone optimization for two-lane highway resurfacing projects with an alternate route. Transportation Research Record: Journal of the Transportation Research Board, No. 1911, pp. 51-66

Chiabaut, N., L. Leclercq and C. Buisson. From heterogeneous drivers to macroscopic patterns in congestion. In Transportation Research Part B, Vol. 44, Issue 2, 2010, 299-308.

Chipperfield, A., Fleming, P., Pohlheim, H., Fonseca, C., 2010. Genetic Algorithm Toolbox User's Guide. Available at http://www.shef.ac.uk/acse/research/ecrg/gat.html. Last accessed July, 21st, 2010

Ciuffo, B., Punzo, V., Torrieri, V., 2007. A framework for calibrating a microscopic simulation model. Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington, D.C.

Ciuffo, B., Punzo, V., Torrieri, V., 2008. A comparison between simulation-based and model-based calibrations of traffic flow micro-simulation models. Transportation Research Records: Journal of Transportation Research Board, No. 2088, TRB, National Research Council, Washington D.C., pp. 36-44.

Ciuffo, B., and V. Punzo, 2010. Verification of traffic micro-simulation model calibration procedures: analysis of Goodness-of-Fit measures. *Presented at 89th Annual Meeting of the Transportation Research Board*, Washington, D.C..

COST Action TU0903 - Multitude, 2011. www.multitude-project.eu. Last accessed July, 30, 2011.

Dowling, R., Skabardonis, A., Halkias, J., McHaie, G., Zammit, G., 2004. Guidelines for calibration of Microsimulation Models: Framework and Application. In Transportation Research Records: Journal of the Transportation Research Board, No. 1876, TRB, National Research Council, Washington, D. C.

Drud, A. S., 1994. CONOPT – A large-Scale GRG Code. In ORSA Journal of Computing, Vol. 6, No. 2. Operation Research Society of America, pp. 207-216.

Fu, M.C., 2002. Optimization for simulation: Theory vs. practice. INFORMS Journal of Computing, vol. 14 (3), pp. 192-215

Fu, M.C., Chen, C.H., Shi, L., 2008. Some topics for simulation optimization. Proceedings of the 2008 Winter Simulation Conference, vol. 1, pp. 27-38

Gardner, L.M., Unnikrishnan, A., Waller, S.T., 2010. Solution methods for robust pricing of transportation networks under uncertain demand. Transportation Research Part C: Emerging Technologies, vol. 18 (5), pp. 656-667

Gartner, N.H., Assman, S.F., Lasaga, F., Hou, D.L., 1991. A multi-band approach to arterial traffic signal optimization. Transportation Research Part B, vol. 25 (1), pp. 55-74

Gipps P.G., 1981. A behavioural car-following model for computer simulation. Transportation Research - B Vol. 15B, No. 2, pp. 105-111.

Glover, F., 1998. A Template for Scatter Search and Path Relinking in Artificial Evolution, Lecture Notes in Computer Science 1363, J.-KRelinking. J.-K. Hao, E. Lutton, E. Ronald, M. Schoenauer and D. Snyers, eds., Springer Verlag, pp. 13-54.

Highway Agency, 1996. Design Manual for Roads and Bridges, Volume 12, Traffic Appraisal of Road Schemes, Section 2, Part I, Traffic Appraisal in Urban Areas. The Stationery Office, London.

Holland, J. H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Hollander, Y., Liu, R., 2008a. The principles of calibrating microsimulation models. Transportation, No. 35, Springer Science, pp. 347-362

Hollander, Y., Liu, R., 2008b. Estimation of the distribution of travel times by repeated simulation. Transportation Research Part C, No. 16. pp. 212-231.

Hourdakis, J., Michalopoulos, P.G., Kottommannil, J., 2003. Practical Procedure for Calibrating Microscopic Traffic Simulation Models. Transportation Research Records: Journal of the Transportation Research Board, No. 1852, TRB, National Research Council, Washington, D.C., pp. 130-139.

Huang, D., Allen, T.T., Notz, W.I., Miller, R.A., 2006 Sequential kriging optimization using multiple-fidelity evaluations. Structural and Multidisciplinary Optimization, vol. 32, pp. 369-382

Jayakrishnan, R., Oh,, J.S., Sahraoui, A., 2001. Calibration and Path Dynamics Issues in Microscopic Simulation for Advanced Traffic Management and Information Systems. In, Transportation Research Record: Journal of the Transportation Research Board, No. 1771, TRB, National Research Council, Washington, D.C.

Jones, D.R., 2001. A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, vol. 21, pp. 345-383

Jones, D.R., Schonlau, M., Welch, W.J., 1998 Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization, vol. 13, pp. 455-492

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. Annals of Mathematical Statistics, vol. 23, pp. 462-466

Kim, S.-J., Kim, W., Rilett, L.R., 2005. Calibration of Microsimulation Models Using Nonparametric Statistical Techniques. In Transportation Research Records: Journal of the Transportation Research Board, No. 1935, TRB, National Research Council, Washington, D. C., pp. 111-119.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by Simulated Annealing. Science, vol. 220, pp. 671-680

Kleijnen, J.P.C., 2008 Design and Analysis of Simulation Experiment. Springer, International Series in Operations Research and Management Science. Stanford, USA.

Kleijnen, J.P.C., 2009 Kriging metamodeling in simulation: A review. European Journal of Operational Research, vol. 192, pp. 707-716.

Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright, 1998. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, Vol. 9, No. 1. pp. 112–147.

Levitan, Y. L., B. V. Shukman, I. M. Sobol, and V. I. Turchaninov, 1992. *Quasirandom sequence generators*. IPM Zak., n. 30.

Law, A.M., 2007. Simulation Modeling and Analysis. Fourth Edition. McGraw Hill, New York

Lee, J.-B., Ozbay, K., 2009. A New Calibration Methodology for Microscopic Traffic Simulation Using Enhanced Simultaneous Perturbation Stochastic Approximation (E-SPSA) Approach. In Transportation Research Records: Journal of the Transportation Research Board, No. 2124, TRB, National Research Council, Washington, D. C., pp. 233-240.

Lindo Systems, 2002. LINDO API User Manual 2.0

Lophanev, S.N., Nielsen, H.B., Sondergaard, J., 2002a. DACE: A Matlab Kriging Toolbox, version 2.0. IMM Technical Report IMM-REP-2002-12. Technical University of Denmark

Lophanev, S.N., Nielsen, H.B., Sondergaard, J., 2002b. Aspects of the Matlab Toolbox DACE. IMM Technical Report IMM-REP-2002-13. Technical University of Denmark

Ma, J., Dong, H., Zhang, H. M., 2007. Calibration of Micro Simulation with Heuristic Optimization Methods. Transportation Research Records: Journal of the Transportation Research Board, No. 1999, TRB, National Research Council, Washington, D. C.

Ma, T., Abdlulhai, B., 2002. Genetic Algorithm-Based Optimization Approach and Generic Tool for Calibrating Traffic Microscopic Simulation Parameters. Transportation Research

Records: Journal of the Transportation Research Board, No. 1800, TRB, National Research Council, Washington, D.C., pp. 6-15.

Martin, J.D., Simpson, T.W., 2005 Use of Kriging models to approximate deterministic computer models. AIAA Journal, vol. 43 (4), pp. 853-863

Maryak, J.L., Chin, D.C., 2008 Global Random Optimization by Simultaneous Perturbation Stochastic Approximation. IEEE Transactions on Automatic Control, vol. 53(3), pp. 780-783

Matheron, G., 1963. Principles of Geostatistics. Economic Geology, vol. 58 (8), pp. 1246-1266

Mathworks Inc., www.mathworks.com. 2010

McKinnon, K.I.M. Convergence of the Nelder–Mead simplex method to a non-stationary point. *SIAM Journal of Optimization*, Vol. 9, 1999. pp. 148–158.

Menneni, S., Sun, C., Vortisch, P., 2008. Microsimulation Calibration Using Speed Flow Relationships. Transportation Research Records: Journal of Transportation Research Board, No. 2088, TRB, National Research Council, Washington D.C.

Nelder, J. A. and R. Mead, 1965. A simplex method for function minimization. *Computer Journal*, Vol. 7. pp. 308–313.

Olafsson, S., Kim, J., 2002. Simulation Optimization. Proceedings of the 2002 Winter Simulation Conference, vol. 1, pp. 79-84

Ossen, S.J.L., Hoogendoorn, S.P., 2008. Validity of Trajectory-Based Calibration Approach of Car-Following Models in the Presence of Measurement Errors. Transportation Research Records: Journal of the Transportation Research Board, No. 2088, TRB, National Research Council, Washington, D.C.

Plackett, R.L., 1950. Some Theorems in Least Squares. Biometrika, No. 37.. pp. 149-157

Punzo, V., Ciuffo, B. 2010. Sensitivity analysis of car-following models. *Presented at 90th Annual Meeting of the Transportation Research Board*, Washington, D.C..

Punzo V., Ciuffo, B., 2009. How parameters of microscopic traffic flow models relate to traffic conditions and implications on model calibration. Forthcoming on Transportation Research Records: Journal of Transportation Research Board, TRB, National Research Council, Washington D.C., 2009.

Punzo, V., Tripodi, A., 2007. Steady-state solutions and multi-class calibration of Gipps' microscopic traffic flow model. Transportation Research Records: Journal of the Transportation Research Board, No. 1999, TRB, National Research Council, Washington D.C., 2007, pp. 104-114.

Punzo, V., Formisano, D.J., Torrieri, V. 2005. Non-Stationary Kalman Filter for the Estimation of Accurate and Consistent Car-Following Data. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1934, Transportation Research Board of the National Academies, Washington, D.C.. pp. 3-12.

Punzo, V., and F. Simonelli. 2005. Analysis and comparison of microscopic traffic flow models with real traffic microscopic data. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1934, Transportation Research Board of the National Academies, Washington, D.C.. pp. 53-63.

Saltelli, A., Ratto, T., Andres, T., Campolongo, F., Cariboni. J., Gatelli, D., Saisana, M., and S. Tarantola, 2008. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd., West Sussex.

Sánchez-Medina, J.J., Galán-Moreno, M.J., Rubio-Royo, E., 2010. Traffic signal optimization in la Almozara District in Saragossa under congestion conditions, using genetic algorithms, traffic microsimulation, and cluster computing. IEEE Transactions on Intelligent Transportation Systems, vol. 11 (1), pp. 132-141.

Schultz, G.G., Rilett, L.R., 2004. Analysis of distribution and calibration of car-following sensitivity parameters in microscopic traffic simulation models. In Transportation Research Record: Journal of the Transportation Research Board, No. 1876, TRB, National Research Council, Washington, D.C., pp.41-51.

Schultz, G.G., Rilett, L.R., 2005. Calibration of Distributions of Commercial Motor Vehicles in CORSIM. In, Transportation Research Record: Journal of the Transportation Research Board, No. 1934, TRB, National Research Council, Washington, D.C., pp.246-255.

Smith, S., Lasdon, L., 1992. Solving Large Sparse Nonlinear Programs Using GRG. ORSA Journal on Computing 4 1, pp.3-15.

Spall, J.C., Hill, S., Stark, D.R., 2006. Theoretical Framework for Comparing Several Stochastic Optimization Approaches. In Probabilistic and Randomized Methods for Design under Uncertainty (G. Calafiore and F. Dabbene, eds.), Springer-Verlag, London.

Spall, J.C., 1992. Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. IEEE Transactions on Automatic Control, vol. 37, pp. 332-341

Spall, J.C., 1998. Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization. IEEE Transactions on Aerospace and Electronic Systems, vol. 34, pp. 817-823.

Spall, J.C., 2000. Adaptive Stochastic Approximation by the Simultaneous Perturbation Method. IEEE Transactions on Automatic Control, vol. 45, pp. 1839-1853.

Spall, J.C., 2003. Introduction to Stochastic Search and Optimization. Estimation, Simulation and Control. Wiley-Interscience. Hoboken, New Jersey

Spall, J.C., 2009. Feedback and Weighting Mechanisms for Improving Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm. IEEE Transactions on Automatic Control, vol. 54 (6), pp. 1216-1229.

Stein, M.L., 1999 Interpolation of spatial data: some theory for Kriging. Springer, New York.

Stevanovic, J., Stevanovic, A., Martin, P.T., Bauer, T., 2008 Stochastic optimization of traffic control and transit priority settings in VISSIM. Transportation Research Part C: Emerging Technologies. Vol. 16 (3), pp. 332-349.

Toledo T., Koutsopoulos, H.N., Davol, A., Ben-Akiva, M., Burghout, W., Andréasson, I., Johansson, T., Lundin, C., 2003. Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study. Transportation Research Records: Journal of the Transportation Research Board, No. 1831, TRB, National Research Council, Washington, D.C.

TSS-Transport Simulation Systems, 2008. AIMSUN 6 Microsimulator User's Manual, TSS, Barcelona, Spain.

Ugray, Z., Lasdon, L., Plummer, J., Glover F., Kelly, J., Marti, R., 2002. A Multistart Scatter Search Heuristic for Smooth NLP and MINLP Problems. In INFORMS Journal of Computing. July 25.

Van Beers, W.C.M., Kleijnen, J.P.C., 2003. Kriging for interpolation in random simulation. Journal of the Operational Research Society. Vol. 54, pp. 255-262.

Vandekerckhove, J., 2010. General simulated annealing algorithm. Available at the following link http://www.mathworks.it/matlabcentral/fileexchange/10548-general-simulated-annealing-algorithm. Last accessed July, 21st, 2010

Wilson, E.R., 2001. An analysis of Gipps' car-following model of highway traffic. *IMA Journal of Applied Mathematics*, Vol. 66. pp. 509-537.

Vaze, V., Antoniou, C., Wen, Y., Ben Akiva, M., 2009. Calibration of Dynamic Traffic Assignment Models with Point-to-Point Traffic Surveillance. In Proceedings of the 88th TRB Annual Meeting, Washington D.C.

Vichiensan, V., Páez, A., Kawai, K., Miyamoto, K., 2006. Nonstationary spatial interpolation method for urban model development. Transportation Research Record: Journal of the Transportation Research Board, No. 1977, pp. 103-111

Villemonteix, J., Vazquez, E., Sidorkiewicz, M., Walter, E., 2009. Global optimization of expensive-to-evaluate functions: An empirical comparison of two sampling criteria. Journal of Global Optimization, vol. 43, pp. 373-389, 2009

Wang, X., Kockelman, K.M., 2009. Forecasting network data spatial interpolation of traffic counts from Texas data. Transportation Research Record: Journal of the Transportation Research Board, No. 2105, pp. 100-108

Wolpert, D.H., MacReady, W.G., 1997. No free lunch theorems for optimization. IEEE Transaction on Evolutionary Computation, 1, pp. 67-82.

European Commission

EUR 25188 EN – Joint Research Centre – Institute for Environment and Sustainability

Title: The Calibration of Traffic Simulation Models - Report on the assessment of different Goodness of Fit measures and Optimization Algorithms. MULTITUDE Project – COST Action TU0903

Author(s): Ciuffo, B., Punzo, V., Montanino, M.

Luxembourg: Publication Office of the European Union

2012 – 92 pp. – 21 x 29.7 cm

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

ISBN 978-92-79-22812-4

doi:10.2788/7975

Abstract

In the last decades, simulation optimization has received considerable attention from both researchers and practitioners. Simulation optimization is the process of finding the best values of some decision variables for a system whose performance is evaluated using the output of a simulation model.

A possible example of simulation optimization is the model calibration. In traffic modelling this topic is particularly relevant since the solutions to the methodological issues arising when setting up a calibration study cannot be posed independently. This calls for methodologies able to check the robustness of a calibration framework as well as further investigations of the issue, in order to identify possible "classes" of problems to be treated in a similar way. Therefore in the present work, first a general method for verifying a traffic micro-simulation calibration procedure (suitable in general for simulation optimization) is described, based on a test with synthetic data. Then it is applied, my means of two different case studies, to draw inferences on the effect that different combinations of parameters, optimization algorithms, measures of Goodness of Fit and levels of noise in the data may have on the optimization problem. Results showed the importance of verifying the calibration procedure with synthetic data. In addition they ascertained the need for global optimization solutions, giving new insights into the topic.

Research contained within this paper benefited from the participation in EU COST Action TU0903 MULTITUDE

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

LB-NA-25188-EN-N

JRC
EUROPEAN COMMISSION

Publications Office

ISBN 978-92-79-22812-4

9 789279 228124