Bayesian pre-calibration of a large stochastic microsimulation model

Alexis Boukouvalas[1], Pete Sykes[2], Dan Cornford[1,4], Hugo Maruri-Aguilar[3],

[1]Aston University
[2]TORG Newcastle University
[3]Queen Mary University
[4]IGI Ltd

Calibration of stochastic traffic microsimulation models is a challenging task. This paper proposes a fast iterative probabilistic pre-calibration framework and demonstrates how it can be successfully applied to a real-world traffic simulation model of a section of the M40 motorway and surrounding area in the UK. The efficiency of the method stems from the use of emulators of the stochastic microsimulator which provide fast surrogates of the traffic model. The use of emulators minimizes the number of microsimulator runs required and the emulators probabilistic construction allows the consideration of the extra uncertainty introduced by the approximation. Visualisation methods are presented to help interpret the pre-calibration results. It is shown that automatic pre-calibration of this real-world microsimulator, using turn count observational data, is possible considering all parameters at once and that this pre-calibrated microsimulator improves on the fit to observations compared to the traditional expert tuned microsimulator.

I. Introduction

ROAD transport microsimulation models are fixed time step Monte Carlo simulations of individual vehicles moving on a road network and are intended to provide decision support to transport planners; [1] describes many contemporary commercial and academic software suites. Large models are often both computationally intensive to run and also labour intensive to calibrate with numerous interrelated parameters adjusted in an iterative process. The combination of long run times and the interactions between parameters makes calibration a challenging task on a large microsimulation model.

In this paper we demonstrate a pre-calibration, or 'history matching', approach to calibration of a traffic microsimulation model, building on the methods developed in [2]. History matching refers to the process of ruling out areas of the parameter space which are inconsistent with the available observations through the use of a probabilistic discrepancy criterion. [2] apply history matching to a deterministic galaxy formation simulator to determine plausible parameter sets. Galaxy formation simulators are amongst the most computationally intensive simulators in use today, and to address this issue, [2] utilise emulators, that is surrogate statistical models, which can be used to reduce the computational cost of calibration [3]. History matching is sometimes referred to as pre-calibration since it helps to identify regions of parameter space that cannot be ruled out given the observational evidence. Reducing the parameter space using history matching can then be followed by a more formal Bayesian calibration [3].

The workflow adopted in this paper is to: 1. elicit the critical inputs and outputs of the model; 2. produce emulators which reproduce the relationship between inputs and outputs; and 3. progressively refine the emulators in the area(s) of parameter space where calibration is most likely to be achieved. In the second and subsequent waves of refinement we are able to both limit the volume of parameter space to be investigated and also refine the emulators, moving from a simple to a more complex emulator. An additional challenge we face in calibrating microsimulation models is that they are stochastic simulators, thus when using emulators we need to predict not just the simulator outputs for a given input, but the *distribution* of the simulator outputs.

The paper begins reviewing the literature on calibration of stochastic traffic simulation (Section I-A) and then describes a general methodology to probabilistically calibrate a stochastic simulator (Section II). We present an extensive demonstration of the methodology on a model of the M40 motorway near Warwick, England in Sections III - IV. We conclude with a summary and suggestions for further work in Section V.

A. Literature

Calibration and sensitivity analysis are well studied problems [4], [5] that appear in a multitude of application domains. In the field of traffic simulation, operational calibration is often what might be referred to as hand tuning, in which model parameters are adjusted based on expertise and trial and error. We term this approach expert calibration and it is discussed in Section III-A within the context of the M40 model.

A wide range of statistical approaches to transport model calibration have been proposed. [6] discuss methodologies of calibration specific to road transport microsimulation models observing that the issue is addressed in some studies primarily through calibration based on comparisons of individual vehicle movements and in other studies by aggregated measures of flow rates and journey times. [6] also discuss the number of parameters involved in the calibration and the merits of using a pragmatic multi-stage process to reduce the scope of the calibration problem in each stage as advocated by [7] versus an approach that addresses all parameters simultaneously and is hence more likely to find a better calibration.

Manuscript received July 1, 2013; revised Nov 15, 2013. Corresponding author: Pete Sykes (email: Pete.Sykes@ps-ttrm.co.uk

[8] developed a microsimulation model for a large region of Des Moines, Iowa and used an automatic calibration method. However the complexity of the model, and in particular the computational time, meant that an iterative approach to calibration was undertaken, considering parameter groups sequentially. As noted in the paper, and in agreement with [6], it would be preferable to jointly tune all parameters; however, the driver behaviour is calibrated on a single road section only and then fixed while other parameters are tuned. The actual tuning of the parameters uses a generalised least squares method. [9] develop a general mathematical framework for the simultaneous calibration of the parameters and inputs to microscopic traffic simulation models using general traffic measurements while [10] used a neural network approach to the calibration of microsimulation models of roundabouts with mixed success depending on the measure chosen for calibration.

An example of Bayesian calibration, applied to stochastic biological simulators, is described in [11], which employs a relatively simple emulator that allows for input dependent variance in the outputs, but decouples the mean response from the variance response. Within the context of traffic simulation, [12] describes a Bayesian approach to the calibration of a traffic simulator on a small network, and exploits the relatively simple structure of the simulator to derive an efficient Markov Chain Monte Carlo sampling method. [13] stress the importance of uncertainty quantification and the application of Bayesian methods for calibration of traffic microsimulators.

Run time is a problem inherent in large transport models, [14] attempted to calibrate a very large microsimulation model of the Buffalo and Niagra region and found the model run time of 30 hrs even precluded heuristic based techniques and essentially required that the process be over-simplified. [15] developed an emulator of an agent based travel demand model utilising a multiple regression model to emulate the relationship of one key output (km travelled) with three socio-economic inputs including first order interactions. [15] argue the case for the use of emulators in applications where short run times are critical (in this case an investigative workshop environment) and extend this to the case for emulation in sensitivity analysis.

[16] used Gaussian process metamodels, or emulators, derived from a transport microsimulation model to examine the feasibility of undertaking parameter sensitivity analysis of the car following and gap acceptance algorithms. Their conclusion was that comparable results were achievable in parameter sensitivity analysis using both the original model and the emulator. [17] developed this work to examine the performance of different optimisation techniques and calibration criteria. Each experiment required many runs of a computationally expensive model and hence a kriging metamodel, equivalent to a Gaussian process emulator, was developed with the four vehicle behaviour parameters identified as critical in [16] as inputs.

II. Probabilistic Calibration via History Matching

We first describe the methodology we propose to calibrate the stochastic microsimulator via history matching. The process may be summarised by the following iterative scheme:

1) *Elicitation* (Section II-A): initially an elicitation exercise [18] is undertaken with the model stakeholders to identify the key input parameters in the simulator and the key outputs, and their associated plausible ranges. The elicitation exercise also considers the simulator structural error, intrinsic variance and observational uncertainties.
2) *Initial experimental design*: an experimental design to vary the elicited inputs is created. The maximin Latin Hypercube design is widely used in the computer experiment literature [3] as it provides good coverage of the input space and is fast and straight-forward to generate. Due to the stochastic nature of the simulator, multiple runs are undertaken for each design point to obtain estimates of the simulator variance.
3) *Simulator evaluation*: The simulator is executed at the design points to obtain the corresponding outputs and the design is split into non-overlapping training and validation sets.
4) *Emulator fitting and validation* (Section II-B): emulators are trained to approximate the simulator using the previously obtained training set of simulator evaluations. The performance of the emulators is checked on the validation set to ensure the probabilistic description is correct [19].
5) *Implausibility* (Section II-C): emulators are used with observations to calculate the implausibility criterion for each proposed parameter set, taking into account all sources of uncertainty (Section II-A). This allows us to rule out areas of implausible parameters, creating a new denser experimental design in the 'not ruled out' space.
6) *Iterate*: return to step 3 by evaluating the simulator at the design locations still deemed *non*-implausible. The emulators constructed at the next stage are defined only in the non-implausible region. Continue until either the computational budget is exhausted or the emulator uncertainty is dominated by other sources of uncertainty.

We now describe each stage in more detail.

A. Expert Elicitation

In addition to eliciting the ranges for the simulator inputs, estimates of simulator and observational uncertainties are needed. Specifically we elicit:

- Model Discrepancy (MD): the difference of the simulator output to reality. This is also known as structural or model error and is due to simplifications and other approximations incorporated in the construction of the simulator, and can be very challenging to quantify.
- Observation Error (OE): the expected error of the observations, and is generally better understood.

Additionally we need to consider the Intrinsic Stochastic Simulator (ISS) variance that arises from running the simulator

repeatedly at a single parameter setting. This is usually input-dependent and we propose to estimate the ISS variance by using replicated experimental designs and constructing emulators that provide a smoothed estimate for this quantity coupled with 'emulator uncertainty'. The latter is defined as the additional uncertainty due to the use of the emulator rather than the simulator.

B. Emulation

Although calibration can be performed without the use of emulation, emulation allows for considerable savings in the number of simulator runs required and is thus commonly employed when dealing with computationally demanding simulators. In this exercise, the following additive model is assumed for each simulator output:

$$t(x) = f(x) + \epsilon(x),$$

where x denotes the simulator inputs (parameters), $f(x)$ is the logarithm of the unknown mean of the simulated traffic count, $\epsilon(x)$ is an input dependent, zero mean, additive Gaussian random variable representing the intrinsic simulator variability and $t(x)$ represents the stochastic computed simulator outputs.

A common approach to emulation for deterministic simulators, where $\epsilon(x) = 0$, is to place a Gaussian Process (GP) prior on $f(x)$ [3]. For stochastic simulators, the GP emulator can be extended to incorporate the stochastic nature of the output by including a range of additional variance models. A GP is defined as 'a collection of random variables, any finite number of which have a joint Gaussian distribution' [20]. GPs are an example of a non-parametric method as they characterise a prior over functions directly instead of requiring an explicit parameterisation of the unknown function f [21].

A GP is defined by a *mean* and a *covariance* function, the specification of which allows the incorporation of prior knowledge in the emulation construction such as the smoothness and differentiability of the approximated function. Formally

$$f(x) \sim \mathcal{GP}\left(m(x), c(x, x')\right),$$

where $x \in \mathbb{R}^p$ the vector of inputs. The mean function $m(x)$ and covariance function $c(x, x')$ are defined as:

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)], \\ c(x, x') &= \mathrm{cov}\left[f(x), f(x')\right]. \end{aligned}$$

Any finite collection of samples from a GP has a joint Gaussian distribution $\{f(x_1), f(x_2), \ldots, f(x_N)\} \sim \mathcal{N}(\mu, C)$, where C has entries $C_{ij} = c(x_i, x_j)$ and the mean μ has entries $\mu(x_i)$. The Squared Exponential and Exponential covariance functions [20] are considered:

$$\begin{aligned} k_\theta^{SE}(r) &= \sigma_p^2 \exp\left(-\frac{r^2}{2\lambda^2}\right), \\ k_\theta^{Exp}(r) &= \sigma_p^2 \exp\left(-\frac{r}{\lambda}\right), \end{aligned}$$

where $r = ||x_i - x_j||$ the Euclidean distance between input points. The process-variance parameter σ_p^2 controls the amplitude of the kernel response. The correlation length-scale parameter λ rescales the inputs. The GP parameters may be estimated by maximum likelihood or integrated out by using sampling. For computational efficiency we use the maximum likelihood approach.

Assuming Gaussian noise and conditioning on the maximum likelihood estimates of the parameters, the posterior predictive distribution for a new point x_* can be analytically calculated [20]. This allows for fast prediction at new points x_* where the simulator has not been evaluated:

$$\begin{aligned} E[t_*|x_*, x, t] &= C_*(C + R)^{-1} t, \\ Var[t_*|x_*, x, t] &= C_{**} + R_* - C_*(C + R)^{-1} C_*^T, \end{aligned}$$

where $\{x, t\}$ is the training set, $C_* = [c(x, x_*)]$ and $C_{**} = [c(x_*, x_*)]$ the train-prediction and prediction only covariance matrices respectively [20]. R and R_* refer to the variance model $\epsilon(x)$ and $\epsilon(x_*)$ respectively.

Different models for the simulator variability $\epsilon(x)$ are considered:

- *Homoscedastic*: input-independent Gaussian noise $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. In this case $R = R_* = \sigma^2 I$ where I the identify matrix.
- *Heteroscedastic*: polynomial input-dependent Gaussian noise $\epsilon(x) \sim \mathcal{N}\left(0, \exp\left(H(x)^T\beta\right)\right)$, where $H(x)$ is the set of fixed basis functions with known parameters [22]. A simple example in 2D space is a linear variance model $\exp\left(\beta_0 + x_1\beta_1 + x_2\beta_2\right)$. In this study we use polynomial variance models up to degree 3. In terms of the variance matrices, R and R_* are diagonal with entries $\exp\left(H(x)^T\beta\right)$ and $\exp\left(H(x_*)^T\beta\right)$ respectively.

The homoscedastic model is the simplest but likely to be inappropriate for stochastic simulators where the variance of the simulator is expected to depend on the inputs. However, it is useful as a baseline measure from which to judge the more complex heteroscedastic models. [22] reviews approaches to heteroscedastic GP emulation in more detail.

Lastly, we note that more complex emulator structures are available and can be employed at any iteration of the calibration process. Rather than using independent emulators for each output other methods that model the correlation between output emulator uncertainties are possible. Dynamic emulators, as proposed in [23], explicitly model the time dependency in the response and would be useful to model time series data such as that produced in the M40 Model for a single location (Section III). More general types of correlations between different responses could also be modeled using a multivariate response GP [24]. Such methods would also require a multivariate elicitation of model discrepancy and observations error that are challenging to elicit in practice [2].

C. Implausibility

The essence of the history matching method employed in this paper lies in the probabilistic criterion known as implausibility [2]. A large value suggests a large difference between the simulator output and reality, considering all sources of uncertainty. For input vector x, the implausibility for output i is defined as:

$$I_i(x) = \left(E_i[t] - z_i\right)^2 / \left(V_i[t] + V_i^O + V_i^{MD}\right),$$

where z_i is the log of the observed data (turn count data in our case), $E_i[t]$, $V_i[t]$ is the mean and variance of the emulator, V_i^O is the observation variance and V_i^{MD} the model discrepancy variance. Large values of $I_i(x)$ suggest if the simulator was evaluated for the input vector x, it is very unlikely the response would be an acceptable match to the observed data, accounting for all the sources of uncertainty [2].

The implausibility across all outputs is summarised by making the simplifying assumption of the independence of output errors. This leads to a multivariate version of $I(x)$ with diagonal matrices:

$$I(x) = (E[t] - z)^T \left(V[t] + V^O + V^{MD}\right)^{-1} (E[t] - z), \quad (1)$$

where $E[t]$ the vector of $E_i[t]$ and z the vector of all observations. $V[t]$ is a diagonal covariance matrix with emulator variances $V_i[t]$ on the diagonal. V^O and V^{MD} are similarly diagonal matrices of the observation errors and model discrepancy respectively. As [2] argue, non-independent constructions of the implausibility measure are possible and lead to more efficient reduction in the parameter space, however they require elicitation of error correlations which is challenging.

The theoretical distribution of (1) assuming z is sampled from a multivariate Gaussian distribution $\mathcal{N}\left(E[t], V[t] + V^O + V^{MD}\right)$ is a Chi-squared distribution with m degrees of freedom where m is the number outputs. This allows the calculation of a cut-off as a suitable percentile of the Chi-square distribution. The cut-off is set to $c = \mathcal{F}^{-1}\left(\mathcal{X}_m^2 < 99.5\right)$ where $\mathcal{F}^{-1}$ the inverse cumulative distribution function. If $I(x) > c$ it is very unlikely the simulator will produce output that will be close to the observations, even taking into account all the uncertainties.

III. M40 Microsimulation Model

The subject for this exercise is an operational model of a section of the M40 Motorway in England between junctions 12 and 14 (Figure 1). It lies south of Warwick and covers 58km of road with 44 zones acting as sources and sinks for traffic. It models over 50,000 trips in the four hour AM peak period. The model, commissioned in 2011 by Warwick County Council, uses the S-Paramics microsimulation software suite [25] and is intended to test options for traffic management and road improvements with a goal of reducing chronic congestion, and improving journey times and journey time reliability.

This model was chosen because it shows many of the more complex facets of a transport microsimulation model. It has route choice and uses dynamic routing in which knowledge of congestion is imparted to some, but not all, vehicles in the simulation. The area is overcongested, exhibits flow breakdown on motorways, and has significant queueing at two key junctions. The execution time of the model is $\sim$ 10–30 minutes depending on the parameter settings, which implies that, while it is not unmanageable, running it many times to investigate parameter options during model calibration is non-trivial.

The model calibration process is described in the Local Model Validation Report (LMVR) [26]. The LMVR describes a structured approach to calibration following the process

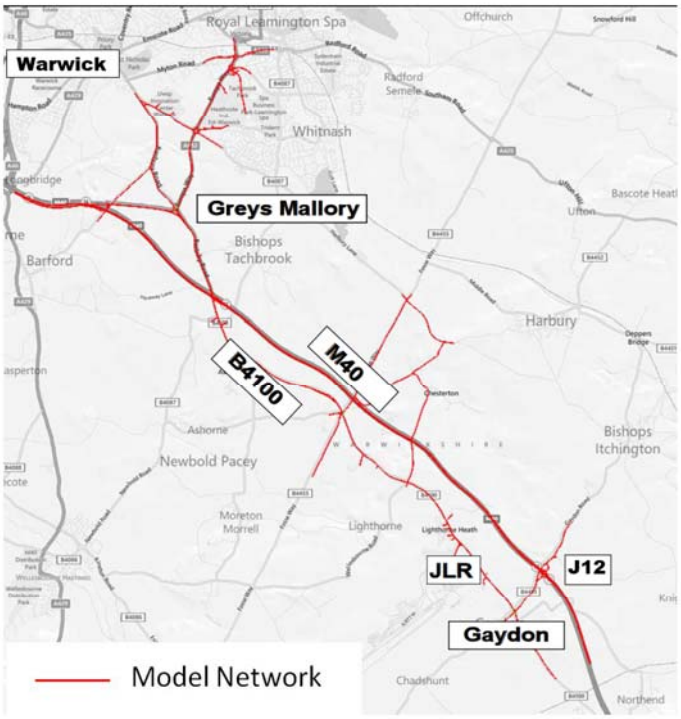


Fig. 1. M40 Modelled Area.

recommended by the software supplier [27]. The key junctions in the model are cordoned and calibrated independently by adjusting locus points, junction visibility parameters and by making local vehicle behaviour adjustments. The next stage is to examine the road network as a whole; assigning route cost multipliers to particular links or classes of links, and by making behavioural changes to reflect observations and by defining the merge and weave areas on motorways prior to exit ramps. Calibration of the distribution of trips in the model, the detailed time profile of trip generation, and the route knowledge of the drivers making those trips forms the third stage. The relationships between route choice, junction congestion, vehicle release rates and timing, and driver knowledge form a network of many interelated parameters controlling different aspects of the model and all contributing to its calibration. Changing one may require complementary changes in others and the inherently linear process of hand calibration struggles to manage the complexity of the task for a large model.

A. Expert Calibration

The model is calibrated using criteria are set out in the UK WebTAG Section 3.19 [28] which requires that 85% of the count locations must have a GEH value < 5 defined as $GEH = \sqrt{\frac{2(m-o)^2}{(m+o)}}$, where m is the modelled count and o is the observed count in vehicles per hour. This criteria was designed for traditional transport models which typically use trip matrices specified at 1 hour intervals. For the microsimulation model, a finer time profile of release of vehicles into the model is possible giving a much improved ability to reproduce the turn counts using the time resolution of the observed data (15 minutes). The elicitation stage revealed that the build up and dispersal of the queues occurred in under one hour implying

that calibration using 1 hour aggregated data was sub-optimal. Hence the decision was taken to modify the calibration criteria to use the equivalent GEH comparison for 15 minute data. [6] comment that most studies use aggregate data such as link flow and journey time in calibration relying on the micro-simulation software supplier to have calibrated the underlying behaviour model for generic applications with only local modifications based on observed deviation from normal behaviour required. This approach was adopted in this study not least due to the constraints of the available traffic flow data.

B. Elicitation

The goal of the elicitation exercise was to identify the critical inputs, outputs and uncertainties in the model.

1) Model Inputs

The first stage is to identify those inputs that, in the opinion of the stakeholders, had the most influence on the simulator calibration or had the greatest uncertainty as to the most appropriate value.

a) Junction Calibration: The model stakeholders were confident that the congested junctions had been adequately calibrated in isolation and showed representative throughput. However, some critical parameters were discussed and, on the recommendation of the software expert, these were included in the inputs to the emulators. These are described in Table II.

b) Route Behaviour: The description of the road network also affects routing; links are described as major or minor depending on whether they form the main signposted routes or not. While the M40 motorway is obviously a major road, the designation for the parallel B4100, a secondary road, but a primary route to the Jaguar-Landrover plant is more subjective. Dynamic route choice gives those drivers labelled *familiar* knowledge of the current achieved speeds in the model, in effect mimicking a process of longitudinal learning by regular commuters. Drivers with higher awareness and aggressiveness attributes will react more strongly to this information but all will have an uncertainty in their perception of congestion delay. As the model was disaggregated to the extent that the users of the main commuter routes could be identified by vehicle type as well as by Origin - Destination (OD) trip, it was possible to assign different route choice parameters to these vehicles. These are described in Table III.

c) Vehicle Release: The number of vehicles released into the model on a particular trip is described by OD matrices which may be disaggregated by vehicle type (i.e. car, HGV) or by trip purpose (i.e. commuter, leisure) and more closely controlled by a time profile with five minute granularity. In the simple case where one profile serves an entire zone, the profile can be derived directly from the observed flow rates from that zone. In the more complex case where multiple profiles are assigned to one zone, it is necessary to examine observed data at junctions further into the model. However, when elements of the route are shared, generating profiles is a non-trivial task and is often subject to the modeller's professional judgement.

The elicitation process revealed that the project team were confident in the OD matrix, cordoned from a wider area model, but were less confident about the departure time profiles. Four key profiles were identified and are described in Table IV.

Parameterising the release profiles was undertaken by defining a spread and shift variable for each profile. The spread variable reduced the peak of a profile by distributing it across the shoulder of the profile using an exponential transformation with a bound of 20% deviation, the shift variable moved the peak forwards or back in time by up to 3 steps (15 min).

d) Vehicle Dynamics: Vehicle dynamics may be globally specified for the model or may be overridden in chosen locations to model observed behaviour. On the B4100, the secondary road running parallel to the M40 motorway, the headway between vehicles is observed to extend beyond the typical range. Subjective opinion expressed in the LMVR is that, with experience, drivers have learned that making smooth progress on this road, where the majority of vehicles share the same destination, is better than stop/start queuing. This observation was coded into the model altering driver behaviour on this link by coding it as either *urban* or *rural* and by extending the headway between individuals. The sensitivity of the model to these changes and the relationship of these parameters to the route choice options was of interest to the model stakeholders. The elicited parameters controlling vehicle behaviour are described in Table V.

2) Model Outputs and Uncertainties

The next stages of elicitation were to determine which of the model outputs should be emulated and what was the stakeholders understanding of the inherent uncertainty in the model. As the calibration focus in this model was on the aggregate level of matching turn counts [6] and, with 255 count locations, it is inevitable that many are interdependent. Hence the elicitation task was to identify which are the main points where driver choice is made and where routes diverge, and which are the main points which allow the characteristics of the flow in the adjacent area to best be inferred. After consultation we narrowed the focus of the calibration to four timesteps and nine locations that were deemed critical.

In terms of the Model Discrepancy (MD) and Observation Error (OE) variances (Section II-A), they were treated similarly and were assumed mutually independent and input independent. For the MD error a value of 15% of the simulator turn count output was elicited. The elicited value for the OE was only 2% of the observed count, the low errors being due to the nature of the data considered. In all cases the model experts felt that proportional errors were more likely, with the quoted percentages referring to two times the standard deviation (giving 95% confidence intervals) of the Gaussian noise in the log transformed space (see Section III-D).

C. Experimental Design

The simulator was evaluated in three waves. For wave 1, the simulator was evaluated on a maximin Latin Hypercube design of 484 points with 5 replicates at each point. This design provided an initial estimate of the mean and variance response of the model across the entire domain of elicited parameter values. An independent validation set of 250 runs with 5 replicates at each point was used for emulator validation.

Wave 2 was performed on the subset of the original space deemed as warranting further investigation by evaluating the

implausibility criterion (Section II-C). The wave 1 emulators were evaluated on a 1100 point maximin Latin Hypercube on the original space. Evaluating the implausibility criterion at each point resulted in 233 of the points classified as non-implausible, i.e. we could not conclude with high confidence that the simulator would result in unrealistic predictions at those settings (Section II-C). The simulator was then evaluated at each of the 233 points using 20 replicate runs to obtain more accurate estimates of the simulator variance. The data set was split into a training and small validation set with 213 training points and 20 validation points.

A final, wave 3, set of simulator runs was created using another 1100 point Latin hypercube over the complete input space and evaluating the implausibility using the emulators from wave 1 and wave 2. This resulted in 32 more simulator runs, further refining the emulators in the area of minimum implausibility.

D. Emulation construction and validation

In the analysis an independent emulator is utilised for each output (location $\times$ time). As the analysis is focussed on nine locations and four time points at each, this results in a total of 36 emulators being fitted in each wave. The emulator structure is simplified by using a log transformation of the outputs, which also serves to ensure predicted traffic counts remain positive. The emulator output can then be treated as a continuous real number and the observation and structural errors become additive rather than multiplicative in the transformed space (Section III-B2).

The emulators were validated at each wave using the Negative Logarithmic Predictive Density (NLPD) score which weights the errors on the mean prediction by the predictive variance, therefore penalising incorrect mean and variance estimates [20]: $NLPD = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log(2\pi \mathrm{Var}[t]_i) + \frac{(E[t]_i - t_i)^2}{\mathrm{Var}[t]_i}\right)$, where N the number of validation points, $E[t]_i$ and $\mathrm{Var}[t]_i$ the emulator predictive mean and variance and t_i the simulator response at validation point i. Smaller values indicate a more accurate prediction.

In wave 1 a common emulator structure was specified for all outputs; a zero mean GP emulator with a squared exponential kernel and a homoscedastic noise model. For the first iteration only a rather coarse emulator is needed since the aim is to quickly exclude areas of the parameter space where the output is predicted as implausible with high confidence. The emulators were validated using the NLPD score evaluated at the wave 1 validation set. The median NLPD for all outputs was 0.12 as compared to 1.49 for an ordinary least squares linear regression model.

In wave 2 a wider range of emulators were considered. For each of the 36 outputs, five independent models were fit:

- A linear model $t(x) = x^T b + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ and σ^2 a input-independent variance.
- Linear mean GP models with an exponential kernel for $f(x)$ and polynomial variance models up to degree 3.

The GP specification was changed from wave 1 as exploratory analysis of the wave 1 non-implausible simulator runs indicated a linear mean process with a less smooth response was more appropriate. The model which minimised the NLPD score evaluated on the wave 2 validation set, was selected for that output. The linear model was selected for most outputs (58.3%) whilst the homoscedastic GP was never selected for any outputs. The heteroscedastic model with log linear variance was selected 25% of the time, with log quadratic variance 8.3% of the time and with log cubic variance also 8.3% of the time. The median NLPD for all outputs was -0.74. The selection of the linear model for most outputs is due to a variety of reasons. Whereas the wave 1 emulator structure was imposed on all outputs, the model selection performed in wave 2 allowed for simple mappings to be described concisely by a linear model. A simple mapping is also more likely in the reduced space of wave 2 compared to wave 1. The small training data sets used also make more likely a better fit with a linear model whereas the additional flexibility of a GP model would necessitate larger training sizes.

IV. Results

This section presents the results of the emulation and pre-calibration of the M40 microsimulation model. Section IV-A discusses how variable selection was used to reduce the number of inputs considered in the pre-calibration. Visualisation of the implausibility space is discussed in Section IV-B. The pre-calibrated model fit is discussed in Section IV-C.

A. Wave 1 Variable Selection

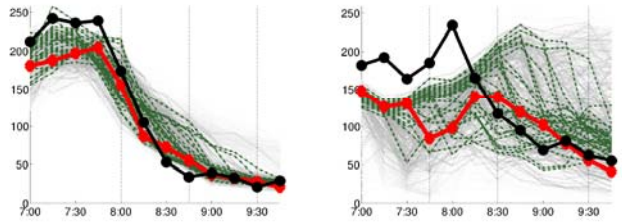
Through discussion with the domain experts at the elicitation stage, the initial set of 37 simulator parameters was reduced to a subset of 25 for the study. Using stepwise polynomial regression, a further 5 parameters were removed that were never selected as inputs for any of the 36 outputs.

B. Simulator parameter investigation

In summary 3 iterations (known as waves) were performed. The first wave included 484×5 simulator runs, the second wave 233×20 runs and the third wave 32×20 runs. In Figure 2 the mean value across the replicate runs for all three waves are shown for one typical location representative of the majority of the output locations (Figure 2(a)) and one atypical location (Figure 2(b)).

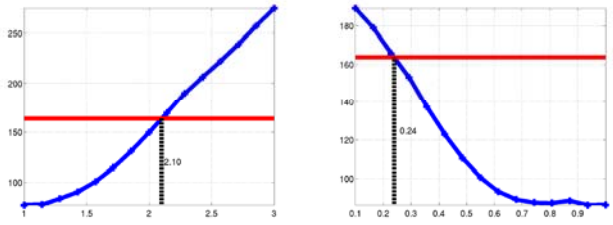
Wave 1 reduced the parameter space to 20% of the original volume considered. Wave 2 further reduced the volume of the parameter space to 3% of the original. This reduction in the parameter space is an important aspect of the pre-calibration and allows us to develop increasingly accurate emulators on an increasingly compact input space, where the emulator response is likely to be simpler and more straightforward to model.

Of note in Figure 2 is that for some locations the simulator is unable to give responses close to the observations under any parameter settings considered. For example for location M40 J12 onoff (Figure 2(b)), the observed turn counts between 7-8am are higher than all simulator runs. While errors such as this do not preclude the approval of the model using the



(a) SE Entrance From B4100 S (b) M40 J12 on-off slip

Fig. 2. Turn counts for runs generated using wave 1 (dotted light grey), wave 2 (dotted medium grey) and wave 3 (dash dark green) parameterizations. The expert calibrated model is shown in connected red diamonds and the observed data in connected black circles.



(a) Headway (b) Motorway Cost

Fig. 3. Wave 2 minimum implausibility profile plots for three of the simulator parameters. The implausibility threshold is shown as a solid horizontal red line.

GEH criteria described in [28], it would be worthwhile to investigate this structural error and if possible to reduce it by including other parameters in the analysis or enlarging the ranges of the parameters considered. Overall we note that for most locations wave 2 and 3 runs are more concentrated around the observations compared to the wave 1 runs.

In Figure 2 the output produced by the expertly calibrated model described in Section III-A is also shown. We term this the default run. These parameter settings were arrived at by an empirical calibration exercise performed prior to history matching. For most locations, the default run is reasonably close to the observations but we also note significant improvements in the wave 3 runs in several locations. For example in Figure 2(a) the default run consistently underestimates the turn counts between 7:00 and 7:45 while wave 3 includes runs achieving significantly smaller error.

The effect on parameter space can be explored in a variety of ways by visualization of the implausibility space. The simplest is to look at the implausibility a single parameter at a time. In Figure 3 we show these plots for two parameters, Headway, and Motorway Cost. These plots were constructed by evaluating the implausibility using wave 2 emulators (for all locations and times examined) on a large set of designs. Specifically, for each parameter a grid set of values was generated. For each value in the grid, a large set of designs (200) for the other 19 parameters was generated. The implausibility for this large set of designs was then calculated. The minimum implausibility value is then shown for each grid point in the plot. The calculation of implausibility threshold shown in the implausibility plots is discussed in Section II-C. The interpretation of these plots is as follows: for any parameter values over the threshold (red line), the simulator will very likely produce unrealistic outputs across the entire range of all other parameters. This can be more simply stated by saying that implausibility values over the threshold represent with near certainty regions in which the simulator will produce unrealistic outputs whilst values under the threshold represent a state of ignorance, that is we cannot say with any certainty that for such parameter values the simulator output will be unrealistic. For the parameters shown in Figure 3, we can state that setting Headway greater than 2.1 or the Motorway cost less than 0.24 will result in unrealistic simulator runs.

The univariate plots do not reveal interactions and so we include examples of two dimensional plots where such interactions are apparent. This is generated in the same manner as previously outlined, the only difference being a two dimensional grid is now used. In Figure 4 we show the interaction of the Headway parameter with the CarPerturb, GVFam, CarFam and MwayCost parameters. For the first two, the headway parameter dominates the implausibility criterion, with the threshold boundary being nearly linear at a value of Headway of approximately 2. With the other two parameters however, the threshold boundary is non-linear with interactions apparent between parameters. For instance, the simulator output is implausible for lower values of headway than 2 and MwayCost higher than 0.2 which is not evident from the one dimensional plots. Generating these plots without the use of emulation would require $15^2 \times 200 = 45000$ simulator runs, and with 6 replications of a Monte Carlo simulation this would require at least 1800 days of computer time (assuming the current best case 10 minute simulator runtime), demonstrating the utility of employing emulation for visualisation and interpretation of the pre-calibration process results.

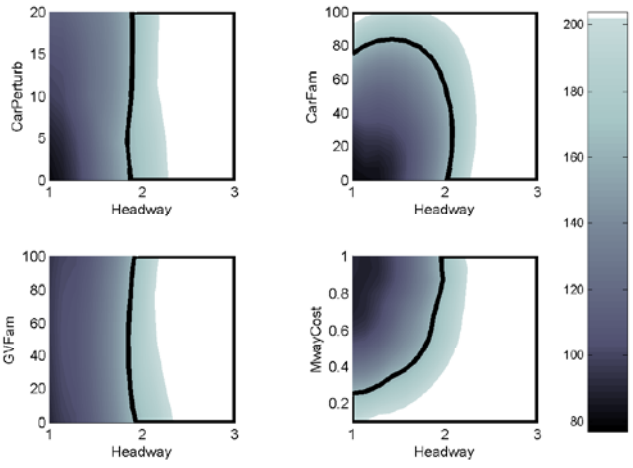


Fig. 4. Wave 2 2-D minimum implausibility parameter profile plots. The decision threshold in the implausibility plot is shown as a solid black line.

To better understand the structure of the wave 3 non-implausible space we used hierarchical cluster analysis. Variables `B4100-Behaviour` and `B4100Routeclass` are categorical so the data set is of mixed type and standard

clustering is not appropriate. We applied a k-medoids clustering method adapted to mixed type variables [29]. Pearson's adaptation of Hubert's Γ metric (PH metric) was calculated, which gives larger values when the parameters are in different clusters. Numbers of clusters ranging from $k = 2$ to $k = 8$ were explored; the PH metric achieved higher values when the number of clusters were 4, 5 or 6, thus pointing to a complex structure in the non-implausible region of parameter space:

No. clusters k	2	3	4	5	6	7	8
PH ($\times$100)	11	32	39	41	38	36	38

To better understand the clustering a technique called persistent homology analysis [30] was applied. Persistent homology enables us to describe the topological structure of the non-implausible parameter values, giving insight to the connectivity of the parameter values at different scales of variation. The most topologically complex cluster corresponds to `B4100-Behaviour`=Highway and `B4100Routeclass`=minor clusters, identified in the hierarchical cluster analysis. The complex geometry of this cluster may reflect non-linearities in the simulator, due to the formation of queues, which creates a fragmented non-implausible parameter region.

C. Calibrated Model Fit

We can also compare the performance of the model under the default expert calibrated parametrisation versus the best parametrisation obtained via the iterative probabilistic calibration methodology. While this is not a formal Bayesian calibration it allows us to compare the 'best' model we found automatically with the best model found by hand tuning. In Table I we compare the GEH of the model under the two parameter settings using two different perspectives. Firstly we show the percentage of the time series where $GEH < 2.5$. The threshold value of 2.5 is used since the error is taken with respect to the 15 minute observational data used in the calibration. For all locations, the probabilistically calibrated model achieves a higher score except for the M40 J12 location where the model prediction consistently underestimates the observations under all parametrisation considered (see Section IV-B and Figure 2(b)). Examining the median GEH, we note large reductions in error for all locations reflecting the closer proximity of the model prediction to the observations.

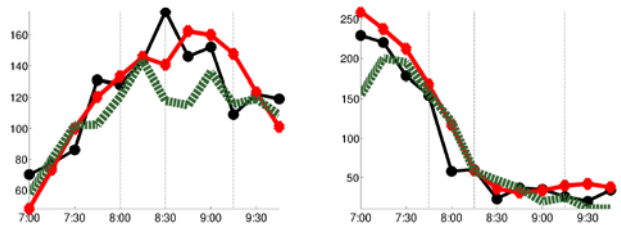
Figure 5 shows two examples of the automatically calibrated simulator output versus the default run and the observations. As expected the calibrated simulator output is closer than the default run on most of the observations used in the calibration. The error outside the observations calibrated against however can be significantly higher than the default run since it is not considered in the calibration exercise. Taking into account more time steps in future waves of calibration would address this issue at the cost of increased computational complexity.

V. Conclusions

This study considers the pre-calibration of a real traffic microsimulation model in a moderate complexity, operational scenario. The work has shown that using Gaussian processes

TABLE I
PERCENTAGE OF GEH < 2.5 AND MEDIAN GEH FOR THE HAND-CALIBRATED SIMULATOR RUN (DEFAULT) AND THE BEST SIMULATOR RUN FOUND USING PROBABILISTIC CALIBRATION (BEST).

Location	Default %	Best %	Default Median	Best Median
M40 J12 NB onoff slip From M40 SB slip to B4451 south	50	50	2.50	1.87
SE Entrance FROM B4100 S TO JLR SE Entrance	100	100	2.28	0.82
NW Entrance Aston Martin Rbout From B4100 north to Aston Martin	75	100	0.81	0.43
Banbury Rd RB From Banbury Road north to Banbury Road south	100	100	1.49	0.59
Gallows hill (Greys mallory) FROM WARWICK BY-PASS TO BANBURY ROAD (S)	100	100	1.38	0.51
Gallows hill (Greys mallory) FROM WARWICK BY-PASS TO BANBURY ROAD (N)	75	100	1.94	0.98
Gallows hill (Greys mallory) FROM WARWICK BY-PASS TO EUROPA WAY	50	100	2.15	0.43
Europa Way Rbout From J3 Europa Way West to J3 Eurpoa Way North	75	100	0.88	0.11
Europa Way Rbout From J3 Europa Way West to J3 Queensway	100	100	1.41	0.42



(a) Europa Way Roundabout (b) Banbury Road Roundabout

Fig. 5. Observations (solid black with circles), default (red diamonds connected with solid line) and best calibrated model run (dashed green) in terms of GEH. The dashed vertical lines denote the time points used in the probabilistic calibration.

emulators with the microsimulation model it is possible to jointly constrain many parameters in the simulator, signficantly reducing the volume of parameter space that need to be considered in the formal calibration exercise. Indeed the pre-calibrated (non-implausible) simulator runs are shown to improve on the expert calibration of the simulator. The emulation approach offers even greater benefits when using stochastic simulators compared with deterministic models, since once trained the emulators are able to capture the intrinsic variability in the simulator without the requirement for replicated simulator runs.

The ability to describe the topology and structure of clusters in the pre-calibration space provides a means to try and understand whether the calibration processes finds many competing solutions across parameter space, or a single global maximum in the implausibility measure. It also helps to describe the complexity of the non-implausible space, and

when used in conjunction with the implausibility surfaces can help improve our understanding of the simulator and its associated parameters. This is critical to both operational use of the microsimulator, and identifying how to improve the microsimulator. It must be emphasised that emulation works with the microsimulator, and does not replace it.

In this study the observational data used to constrain the simulator consisted of only 9 locations each extending over 4 time steps. This has simplified the subsequent analysis but at the cost of efficiency, i.e. some runs that produce clearly implausible output are not detected as such because that behaviour occurs outside the subset of locations and times we have looked at. Extending our approach to more time points and locations is straight-forward but requires the development of many more independent emulators. Lighter-weight emulators as used in [2], which are essentially linear in parameter models, can be utilised to reduce the computational requirement, although the computational cost of training an emulator is signficantly less than a single simulator evaluation. Another option that could be explored is creating multivariate emulators for the functional output, or potentially emulating the misfit (implausibility) as was done in [31].

In future, we plan to extend this work to consider a formal Bayesian calibration of such a microsimulation model, using these methods to explore road development and traffic management option sensitivity. This will require further consideration of the various sources of uncertainty and in particular model discrepancy in microsimulation models.

APPENDIX A
ELICITED VARIABLES

Tables II-V show the ranges and values of the elicited input variables.

TABLE II
ELICITED VARIABLES: JUNCTION CALIBRATION

Group	Variable	Range	Default	Units
M40 J12	J12EastBound	500-4000	3000	m
		Where vehicles start to get in lane before Junction 12		
Greys Mallory	EastBound	0-60	60	m
	SouthBound	0-60	20	m
		Junction visibility parameters at Greys Mallory Roundabout		
Gaydon	4415SWB	0-15	15	m
Gaydon	4415NEB	0-60	20	m
Gaydon	4100SEB	0-18	18	m
Gaydon	4100NWB	0-17	17	m
		Junction visibility parameters at Gaydon Junction		

TABLE III
ELICITED VARIABLES: ROUTE CHOICE

Group	Variable	Range	Default	Units
Feedback	FeedbackTime	60 - 300	120	Seconds
	FeedbackCoeff	0 - 100	50	%
		Dynamic routing parameters		
Perception	RoutingScale	50-150	100	%
	RoutingShift	0	0	
	RoutingDrift	0	0	
		Controls the effect of aggression and awareness in route choice		
Fuzziness	CarCostPerturbation	0-20	5	%
	CarCostPerturbation-Mainline	0-20	5	%
		Route cost perturbation for commuters and seperately for motorway through routes		
Familiarity	Car-Familiarity	0-100	5	%
	M40-Car-Familiarity	0-100	85	%
	Goods-V-Familiarity	0-100	40	%
		% vehicles with congestion awareness		
Route cost	Cost-A	0-1	1	
	Cost-B	0-1	0.65	
		*Route Cost = A*time + B*distance*		
	M40CostBiasFactor	0.1 - 1	0.8	scalar
	B4100RouteClass	Maj—Min	Minor	enum
		Route weighting		

TABLE IV
ELICITED VARIABLES: PROFILES

Group	Variable	Range	Default
2	Shift	± 3	0
	Spread	$[-0.07, 0.06]$	0
		Controls traffic to Warwick and Royal Leamington Spa	
3	Shift	± 3	0
	Spread	$[-0.095, 0.105]$	0
		Controls though traffic on the M40	
40	Shift	± 3	0
	Spread	$[-0.038, 0.041]$	0
		Controls traffic to JLR W Entrance	
41	Shift	± 3	0
	Spread	$[-0.038, 0.038$	0
		Controls traffic to JLR E Entrance	

TABLE V
ELICITED VARIABLES: VEHICLE BEHAVIOUR

Group	Variable	Range	Default	Units
Following	Headway	1-3	1	Sec
	MinGap	0-3	2	Sec
		Headway between vehicles		
	CrawlSpeed	5-15	10	mile/hr
	HWCrawlMod	1-2	1.8	
		Change from time to distance based car following, with modifier for highway driving		
B4100	B4100-Behaviour	Urban—H'way	Urban	Categorical
	B4100-Headway	0.8-3.0	1.5	
		Behaviour specific to the B4100		

ACKNOWLEDGEMENTS

The authors would like to thank SIAS Ltd. for the loan of an S-Paramics simulation software licence and James Edwards of ARUP Ltd and Alan Law of Warwickshire County Council for their assistance during the elicitation stage and the provision of the simulation model. This work was funded as part of the Managing Uncertainty in Complex Models project (EPSRC grant D048893/1) and the Aston Research Centre for Healthy Ageing (ARCHA). We also wish to thank the anonymous reviewers for their comments that helped improve the presentation of the paper.

REFERENCES

[1] J. Barcelo, *Fundamentals of Traffic Simulation*, ser. International Series in Operations Research and Management Science. Springer Verlag, 2010.

[2] I. Vernon, M. Goldstein, and R. G. Bower, "Galaxy formation: a Bayesian uncertainty analysis," *Bayesian Analysis*, vol. 5, pp. 619–670, 2010.

[3] A. O'Hagan, "Bayesian analysis of computer code outputs: a tutorial," *Reliability Engineering and System Safety*, vol. 91, pp. 1290–1300, 2006.

[4] A. Saltelli, K. Chan, and E. Scott, *Sensitivity Analysis*. Wiley, 2009.

[5] M. Hill and C. Tiedeman, *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*. Wiley, 2006.

[6] Y. Hollander and R. Lui, "The principles of calibrating traffic microsimulation models," *Transportation*, vol. 35, pp. 347–362, 2008.

[7] FHWA, "Traffic analysis toolbox volume iii: Guidelines for applying traffic microsimulation modeling software," FHWA, Tech. Rep., 2004.

[8] M. Jha, G. Gopalan, A. Garms, B. Mahanti, T. Toledo, and M. Ben-Akiva, "Development and calibration of a large-scale microscopic traffic simulation model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1876, pp. 121–131, 2004.

[9] R. Balakrishna, C. Antoniou, M. Ben-Akiva, H. N. Koutsopoulos, and Y. Wen, "Calibration of microscopic traffic simulation models: Methods and application," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1999, pp. 198–207, 2007.

[10] I. Otkovic, T. Tollazzi, and M. Sraml, "Calibration of microsimulation traffic model using neural network approach," *Expert Systems with Applications*, vol. 40, p. 5965 5974, 2013.

[11] D. A. Henderson, R. J. Boys, K. J. Krishnan, C. Lawless, and D. J. Wilkinson, "Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons," *Journal of the American Statistical Association*, vol. 104, pp. 76–87, 2009.

[12] G. Molina, M. J. Bayarri, and J. O. Berger, "Statistical inverse analysis for a network microsimulator," *Technometrics*, vol. 47, pp. 388–398, 2005.

[13] M. J. Bayarri, J. O. Berger, G. Molina, N. , and J. Sacks, "Assessing uncertainties in traffic simulation: A key component in model calibration and validation," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1876, pp. 32–40, 2004.

[14] Y. Zhaoa and W. Sadeka, A, "Large-scale agent-based traffic microsimulation: Experiences with model refinement, calibration, validation and application," *Procedia Computer Science*, vol. 10, p. 815 820, 2012.

[15] A. Rasouli and H. Timmermans, "Using emulators to approximate predicted performance indicators in complex micro-simulation and multi-agent models of travel demand," in *Proceedings of the 4th Conference on Innovations in Travel Modelling Conference*, 2012.

[16] B. Ciuffo, J. Casas, M. Montanino, J. Perarnau, and V. Punzo, "From theory to practice: Gaussian process metamodels for the sensitivity analysis of traffic simulation models. a case study of the Aimsun mesoscopic model." in *Proceedings of the Transportation Research Board 92nd Annual Meeting*. TRB, 2013.

[17] B. Ciuffo and V. Punzo, "'No free lunch' theorems applied to the calibration of traffic simulation models," *IEEE Transactions on ITS*, vol. Accepted for publication, 2013.

[18] A. O'Hagan, "Eliciting expert beliefs in substantial practical applications," *The Statistician*, vol. 47, pp. 21–35, 1998.

[19] L. S. Bastos and A. O'Hagan, "Diagnostics for Gaussian process emulators," *Technometrics*, vol. 51, pp. 425–438, 2008.

[20] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[21] D. J. C. Mackay, "Introduction to Gaussian processes," *Neural Networks and Machine Learning*, 1998.

[22] A. Boukouvalas, "Emulation of random output simulators," Ph.D. dissertation, Aston University, 2011. [Online]. Available: wiki.aston.ac.uk/foswiki/pub/AlexisBoukouvalas/WebHome/thesis.pdf

[23] S. Conti, J. P. Gosling, J. E. Oakley, and A. O'Hagan, "Gaussian process emulation of dynamic computer codes," *Biometrika*, vol. 96, pp. 663–676, 2009.

[24] S. Conti and A. O'Hagan, "Bayesian emulation of complex multi-output and dynamic computer models," *Journal of Statistical Planning and Inference*, vol. 140, pp. 640 – 651, 2010.

[25] SIAS., *S-Paramics V2011.1*, Edinburgh UK, 2011. [Online]. Available: www.paramics.co.uk

[26] J. Edwards, "M40 junction 12 to 14 paramics modelling: M40 paramics model development report," Arup, Tech. Rep. 211439-18/R001, 1 May 2012 2012.

[27] SIAS., *The Microsimulation Consultancy Good Practice Guide*, 2006.

[28] DfT, "Highway assignment modelling," 01/08/2012 2012. [Online]. Available: http://www.dft.gov.uk/webtag/documents/expert/unit3.19.php

[29] C. Hennig and T. Liao, "How to find an appropriate clustering for mixed type variables with application to socio-economic stratification (with discussion)," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 62, pp. 309–369, 2013.

[30] A. Zomorodian, *Topology for Computing*. Cambirdge University Press, 2005.

[31] M. T. Pratola, S. R. Sain, D. Bingham, M. Wiltberger, and E. J. Rigler, "Fast Sequential Computer Model Calibration of Large Nonstationary Spatial-Temporal Processes," *Technometrics*, vol. 55, pp. 232–242, 2013.

Alexis Boukouvalas received his PhD from Aston University, UK, in July 2011. He has been part of the Managing Uncertainty in Complex Models project as both a research student and a research fellow. He is currently a Research Fellow at the Aston Research Center for Healthy Ageing. His research interests lie in the area of non-parameteric Bayesian modelling for the analysis of computer experiments as well as longitudinal studies.

Pete Sykes was the Director of the Software Division of SIAS Ltd from 1998 to 2011 with responsibility for the Paramics Microsimulation Software suite. He is now a part time PhD student at Newcastle University researching methods of eliciting and managing the drivers of uncertainty in transport planning projects. He also maintains an interest in the EU COST project MULTITUDE describing the content and use of guidelines for microsimulation modellers.

Dan Cornford is currently a Reader in Computer Science at Aston University and a director of IGI Ltd. He obtained his PhD from Birmingham University in 1996 and has since worked on various aspects of probabilistic modelling of geophysical systems including work on data assimilation and how to best combine our knowledge of physical laws with partial information on unmodelled processes and incomplete observation. He maintains an interest in machine learning and statistics including visualisation, learning from data, and communication of uncertainty.

Hugo Maruri-Aguilar received his PhD from Warwick University, UK in July 2007. He worked as research fellow in London School of Economics as part of the project 'Managing Uncertainty in Complex Models' in the period 2006-2009. He is currently Lecturer in Statistics in the School of Mathematical Sciences in Queen Mary, University of London, UK. His research interests lie in design and analysis of computer experiments, and the use of algebraic techniques for modelling data in statistics.