

Multi-Objective Calibration of Microscopic Traffic Simulation for Highway Traffic Safety

Edgar Tamayo Cascan^{1,2}, Jordan Ivanchev^{1,4}, David Eckhoff^{1,4},
Alberto Sangiovanni-Vincentelli³, and Alois Knoll^{4,5}

Abstract—Microscopic traffic simulation has become an important tool to investigate traffic efficiency and road safety. In order to produce meaningful results, driver behaviour models need to be carefully calibrated to represent real world conditions. If this type of simulations are to be used to evaluate safety features of traffic, on top of macroscopic relationships such as the speed-density diagram, they should also adequately represent the average risk of accidents occurring on the road. In this paper, we present a two-stage computationally feasible multi-objective calibration process. The first stage performs a parameter sensitivity analysis to select only parameters with considerable effect on the respective objective functions. The second stage employs a multi-objective genetic algorithm utilizing only few influential parameters that produces a front of Pareto optimal solutions with respect to the conflicting objective functions. Compared to traditional methods which focus on only one objective while sacrificing the accuracy of the other, our method achieves a high degree of realism for both traffic flow and average risk.

I. INTRODUCTION

Microscopic traffic simulation offers a safe and cost-efficient environment to evaluate the impact on road safety of, e.g. a new street or intersection layout as well as upcoming technology such as autonomously driving vehicles. The main concern of using simulation for such a critical task is the trustworthiness of the results: the simulation can only show what the underlying behavioural models allow. The meaningfulness of a simulation therefore strongly depends on its capability to represent realistic movements of vehicles and their interaction with each other.

Ideally, a mixture of both simulation and field experiments should be applied to evaluate new traffic environments: the former to understand the ground truth of the to-be-modelled system and to capture effects that might have been overlooked (or are difficult to represent) in a simulation environment, and simulations to efficiently cover a larger system and parameter space. The key challenge here is *calibration*, i.e., to make the simulations as realistic as possible by tuning the parameters

of the models so the observed system behaviour reflects the recorded empirical data.

Calibration of complex simulation models is a non-trivial task. First, it has to be determined which input parameters of the model should be calibrated, which range of values is allowed and what is the desired output of the calibration, that is, which aspect of reality does the simulation experiment want to capture. In the context of road safety, this could be the number of accidents, injuries or fatalities – the latter two requiring detailed models of vehicle physics. Since these events are rare and most mobility models in microscopic traffic simulation are collision-free, usually the number of traffic conflicts (or near-misses) is taken into consideration instead. A traffic conflict occurs when two vehicles approach each other to such extent that there is risk of collision if their movements remain unchanged [1]. A traffic conflict occurs when two vehicles approach each other to such extent that there is risk of collision if their movements remain unchanged [1].

When solely calibrating underlying driver behaviour models to produce the same number of traffic conflicts in simulation as observed in real life, there is a risk that other important properties of the system will be overlooked. These properties can include the flow on the investigated road segment, the speed of vehicles or the number of lane changes. We therefore argue that choosing a single measure of performance (MOP) to calibrate the system is insufficient and that the underlying calibration algorithm should find a balance of multiple performance measures.

In this paper, we present a multiple-objective Genetic Algorithm (GA) calibration for car-following and lane-changing behaviour models. Our four contributions include (1) We introduce an extension of IDM and MOBIL to incorporate attention and aggression. This is necessary to be able to capture near-miss situations in otherwise too-safe traffic simulation. (2) We show that standard parameters found in the literature are not sufficient to generate a realistic simulation environment. (3) We furthermore illustrate that optimising only for one MOP will lead to undesired differences of other important traffic characteristics. (4) We show that our calibration approach is feasible using four different target measures to capture both traffic flow and traffic conflicts.

The remainder of this paper is organised as follows: In Section II, we discuss related work in the field of microscopic traffic simulation for traffic safety as well as the calibration of the used models. We present our system model and the

*This work was financially supported by the Singapore National Research Foundation under its Campus for Research Excellence And Technological Enterprise (CREATE) programme.

¹TUMCREATE, Singapore {edgar.cascan| jordan.ivanchev| david.eckhoff} @tum-create.edu.sg

²KTH Royal Institute of Technology, Sweden

³University of California, Berkeley, USA
alberto@berkeley.edu

⁴Technical University of Munich, knoll@in.tum.de

⁵Nanyang Technological University, Singapore

used simulator in Section III, followed by an explanation of the Traffic Conflict Technique (TCT) in Section IV. In Section V we introduce the output metrics used for comparison. Our calibration process is discussed in detail in Section VI. In Section VII we present our results. We conclude our work and discuss future steps in Section VIII.

II. RELATED WORK

The application of microscopic traffic simulation in the field of traffic safety was initially recognised by Darzentas et al. [2]. They simulated different traffic flows entering a T-Junction, and observed that conflicts increase linearly with flow on the main road, as expected from analytic studies. Later on, Archer and Kosonen tested a driver behaviour model created by Rumar [3] in their HUTSIM microsimulator for the study of conflicts [4]. Their study showed great promise and had significant implications for future traffic safety research. Nonetheless, criticism against simulated conflicts rose rapidly due to two main concerns: first, driver behaviour models follow specific crash avoidance rules, and therefore fail to explain the relation between high risk behaviour and crashes [5], [6]. Second, the effectiveness of simulation in safety studies lies in its ability to accurately model the interaction that occurs between vehicles, which was shown to be non-trivial [7], [8].

Accordingly, one of the most important stages in simulation is to ensure that model parameters are determined based on observational data, and that they generate conflict reports that can be verified from real-world observations [9]. Therefore, several studies have been carried out with a focus on calibration and validation of simulation models. Brockfeld, Kühne, and Wagner [10], for instance, calibrated ten microscopic models with test track data from Japan, minimising travel time and headway errors. Schultz and Rilett [8] determined the CORSIM parameters to match traffic volume and travel time data from Houston, Texas. Cheu et al. [11] used a Genetic Algorithm (GA) to calibrate FRESIM for Singapore expressway traffic flows, just to name a few.

As an attempt to further expand the Traffic Conflict Technique (TCT), FHWA sponsored a research project to investigate the potential of surrogate measures of safety for existing simulation models [12]. Therein, they developed a Surrogate Safety Assessment Model (SSAM). SSAM identifies conflict events by processing detailed vehicle trajectory data from traffic simulation software such as AIMSUN, Paramics, TEXAS, and VISSIM, all of which collaborated on the project. In SSAM's final report [13], Gettman et al. conducted a theoretical validation to assess the use of their tool to recognise the relative safety of pairs of design alternatives in eleven case studies that include intersections and interchanges. Results showed that SSAM could recognise statistically significant differences in the total number of conflicts for both design alternatives under equivalent traffic conditions.

Recently, Essa and Sayed [5] investigated the relationship between field-measured conflicts and simulated conflicts at a signalised intersection in Surrey, British Columbia, Canada. They proposed a simplified two-step calibration of VISSIM driving behaviour parameters. First, they balanced

the simulated desired speed and arrival type in order to match average delay time. Then, by implementing SSAM, a sensitivity analysis and a subsequent GA technique were applied to determine the optimal parameter configuration regarding the simulated rear-end conflicts. Similarly, Cunto and Saccomanno [9] proposed a calibration in which traffic flow and conflict level were matched to real conditions. However, these studies share a common limitation in that safety performance is calibrated by a single-objective GA that disregards its effect on the traffic flow. Ideally, the calibration procedure should include both traffic attributes and safety as objectives, since they are essentially linked [6]. Moreover, the driver behaviour models implemented in existing microscopic traffic simulators follow specific rules aimed at avoiding collisions. Therefore, it is challenging to represent unsafe vehicle interactions and near misses.

Overall, research has recognised great potential in both TCT and microscopic traffic simulation for safety evaluations. Several studies have indicated that there is a strong relation between real and simulated conflicts, however, the effectiveness of these studies is limited by their calibration approach. This paper presents a holistic calibration procedure for safety studies that takes into account both traffic flow and safety performances. Our goal in doing so is to improve the applicability of microscopic traffic simulation to road safety.

III. SYSTEM MODEL

The traffic simulator utilised in this study is the Behaviour Evaluation of Human and Autonomous VEHICLES (BEHAVE) [14] tool based on CityMoS [15]. Our approach is not limited to the chosen simulator but can be applied to any other microscopic traffic simulator, such as SUMO, VISSIM or AIMSUN. In this paper, we will focus on the calibration for the human driver behaviour models and make use of real vehicle traffic data from the Next Generation SIMulation (NGSIM) project [16]. This data set contains high resolution real world data from a highway section in California under changing traffic conditions, i.e., from almost free flow to a congested state.

A. Simulated Data

We configure BEHAVE to simulate a dynamic 1600 meter-long stretch of highway with five lanes. According to the speed limit of the real world data, we set the desired speed of all agents, i.e., vehicles, to 112 km/h (70 miles per hour). The simulated highway section is part of an endless highway and moves dynamically with one pre-selected vehicle. To study traffic characteristics, we have to gather data from different traffic density and speed levels. Therefore, the number of simulated agents is linearly increased with time for a total of 5,000 seconds, describing a congestion evolution.

B. Behaviour models

The BEHAVE simulator contains implementations of known car-following and lane-changing models which can be customised before and during the simulation. In this study, we focus on the enhanced Intelligent Driver Model (IDM) from

Treiber et al. [17], and on the Minimizing Overall Braking Induced by Lane Changes (MOBIL) model, also from Kesting, Treiber, and Helbing [18].

The enhanced IDM describes the dynamics of the positions and velocities of single vehicles through two ordinary differential equations. Within them, there are six configurable model parameters: *desired velocity* v_0 is the velocity at which the vehicle would drive during free flow traffic; *minimum spacing* s_0 is the minimum desired distance to the vehicle in front, particularly important when vehicles are stopped; *desired time headway* T is the preferred time gap that is kept between the vehicle and its predecessor; *acceleration* a is the maximum preferred acceleration of the vehicle; *comfortable braking deceleration* b is the maximum preferred deceleration; *coolness factor* c describes how reliant a driver would be on the vehicle in front continuing to drive without major changes in acceleration. However, enhanced IDM incorporates collision-free behaviour which is sub-optimal for safety studies. We have extended this version of the IDM by adding some human-typical factors such as *aggression* and *attention* to the already existing IDM, so crashes are possible in the simulation environment and near-misses are more realistic.

The lane-changing model MOBIL contains three additional parameters: *politeness* p is used as a weight for how much the change in comfort of other vehicles is considered relative to your own; *acceleration threshold* Δa_{th} is the potential acceleration gain required to motivate a lane change; and *bias* adds a 'keep-right' tendency.

The aggression of a driver $G \in [0, 1]$ modulates the preferred speed and time headway parameters of IDM, the politeness factor, and the acceleration threshold factors of MOBIL such that: $\tilde{v}_0 = v_0(1 + 0.5G)$, $\tilde{T} = T(1 - 0.5G)$, $\tilde{p} = p(1 - 0.9G)$, $\tilde{a}_{th} = a_{th}(1 - 0.9G)$.

The attention of the driver $A \in [0, 1]$ is a process, which if unaltered will asymptotically return to 1 modulated by λ set to 0.99. Alternatively, with probability f (distraction intensity) the value of the attention term will be reduced by a random number uniformly distributed between 0 and the current attention value:

$$A[i + 1] = \begin{cases} \lambda(A[i] - 1) + 1 & \text{w.p. } 1-f \\ A[i] - x & \text{w.p. } f \end{cases}, \quad (1)$$

where $X \sim U[0, A[i]]$.

At every time step i the car-following model will simply not be executed with probability $1 - A[i]$. Similarly, the vehicles at neighbouring lanes, which should be considered by the lane changing model will not be detected with the same probability. We will refer to these new models as EIDM-E and MOBIL-E.

Vehicles should show diverse behaviours to emulate reality. Research has shown, for instance, that headway times generally follow a normal distribution among drivers [19]. In this paper, we assume that all other driver characteristics, such as aggression, maximum acceleration, or politeness, follow as well normal distributions. Therefore, in order to achieve variety between vehicles, all parameters are set by indicating means and standard deviations, rather than single

values. However, for simplicity reasons, standard deviations σ are not independent variables but relative to the mean μ of the respective parameter p such that $\sigma_p = \frac{\mu_p}{5}$.

IV. TRAFFIC CONFLICT TECHNIQUE (TCT)

The aim of this study is to propose a multi-objective calibration for all mentioned parameters of the EIDM-E and MOBIL-E models to improve the feasibility of traffic safety studies in simulation. Since the trajectory data from NGSIM do not include accidents, the models' safety performance cannot be calibrated by comparing collisions. Consequently we focus on assessing risks (or dangerous situations) by making use of the TCT. In order to apply this concept in practice, many surrogate safety indicators have been proposed. In this study, the widely used Time to Collision (TTC) will be described and adopted as a safety performance parameter.

TTC is defined as '*the time that remains until a collision between two vehicles would have occurred if the collision course and speed difference are maintained*' [1]. The minimum TTC (TTC_{min}) during the approach of two vehicles on a collision course is taken as an indicator for the severity of a conflict [20].

TTC can be expressed as,

$$TTC_i[n] = \frac{x_{i-1}[n] - x_i[n] - l_{i-1}}{v_i[n] - v_{i-1}[n]}, \quad (2)$$

where n is the discrete instant of study, $x_i[n]$ is the position of the vehicle i , $v_i[n]$ its speed, and $x_{i-1}[n]$, $v_{i-1}[n]$, and l_{i-1} are the position, speed, and length of the preceding vehicle, respectively.

A threshold value (TTC^*) should be chosen to distinguish between relatively safe and critical encounters [20]. It is crucial to choose an adequate threshold value for each case study. When analysing intersections, for instance, low TTC values are more frequent than on freeways, therefore the limit for TTC has to be modified accordingly. Research has shown that a desirable TTC threshold of 1.5s for intersections is adequate [21], [22], [23], [24], while for rural roads or highways, this increases to around 3 seconds [25], [26], [27], [28]. We therefore apply a threshold of 3 seconds in the present study.

Some of the safety indicator values, such as those from TTC, are inversely proportional to the risk that they aim to represent. That is to say, the lower the TTC is, the higher the risk of a collision will be [29]. In order to transform TTC outputs into indicators of risk, these can be compared to the chosen thresholds. Therefore,

$$IR_i[n] = \begin{cases} TTC^* - TTC_i[n], & TTC^* > TTC_i[n] \\ 0, & otherwise \end{cases}, \quad (3)$$

where $IR_i[n]$ represents the Individual Risk (IR) of a conflict scenario i at the discrete instant of study n , $TTC_i[n]$ describes the indicator's value, and TTC^* the selected TTC threshold.

Since TTC's technique assumes that consecutive vehicles will keep constant speed, conflicts related to acceleration or

deceleration variations will not be considered. Moreover, TTC can provide the number of crashes but not their severity [20]. However, this indicator is suitable for rear-end, turning, and crossing conflicts analysis. It is not only the most common indicator in traffic safety research, but is also used in many automobile collision avoidance or driver assistance systems as a warning criterion [30].

V. OUTPUT METRICS: FLOW AND AVERAGE RISK

In order to calibrate our behaviour models, it is first required to study the relation of speed to traffic density and conflicts to density from real traffic flows.

We make use of the detailed vehicle trajectory data that was collected on eastbound I-80 in Emeryville, CA, on 13th April 2005. As stated in [16], the study area was approximately 500 metres in length and consisted of five freeway lanes, including a high-occupancy vehicle (HOV) lane, with speed limit of 112 km/h (70 miles per hour). This vehicle trajectory data provided the precise location of each vehicle within the study area every one-tenth of a second, for a total of 45 minutes segmented into three 15-minute periods. These periods were: 4:00pm to 4:15pm, 5:00pm to 5:15pm, and 5:15pm to 5:30 pm. They represent the transition between uncongested and congested conditions, and full congestion during the peak period.

In order to compare the traffic flow response of simulation to reality, traffic characteristics such as density, speed, and conflict level need to be clearly defined beforehand. Consequently, density is calculated at each time step τ (i.e. 0.1 s) and for each lane, as the amount of vehicles within the area of study, divided by the length of the road extension considered. Speed is defined as the average of all vehicles at each time step, and similarly, the conflict level is the average of the IR (see Eq. 3) of all automobiles on the studied lane. Observations are averaged over a 60-second time period which is ultimately defined by its density ρ_i , speed v_i , and conflict level IR_i , where i represents the studied period.

As we will show later, ρ_i and v_i data are essential for the study of traffic flow. For the relation of conflicts and traffic density, the individual risk will serve as the main metric. However, numerous time periods involve close-to-zero IR_i , while others are unusually high. Therefore, a procedure described by Kuang, Qu, and Yan [31] is followed in order to avoid large variations of data, which could impair its readability.

As reported in [31], the processed data must be divided into traffic state intervals, sorted by density, with uniform span. Accordingly, all 60-second observations are ranked regarding their density, from smallest to largest,

$$(\rho_1, v_1, IR_1), \dots, (\rho_i, v_i, IR_i), \dots, (\rho_m, v_m, IR_m), \quad (4)$$

where $\rho_1 \leq \dots \leq \rho_i \leq \dots \leq \rho_m$, and v_i and IR_i are, respectively, the corresponding speed and conflict level for the period i . In order to classify the data in clusters, a constant span δ of 0.0015 veh/m is specified. The number of intervals is determined as follows:

$$n_{total} = \lceil \frac{k_{max} - k_{min}}{\delta} \rceil, \quad (5)$$

where n_{total} is the total number of intervals and k_{max} and k_{min} are the maximum and minimum density among the 60-second observations, respectively. Subsequently, the density range R_n for all clusters can be computed as:

$$R_n = [k_{min} + \delta \cdot (n - 1), k_{min} + \delta \cdot n], \quad (6)$$

with $n \in (1, 2, \dots, n_{total})$. Next, the number of observations N_n that fall into each of these n_{total} ranges are counted and stored. Then the Cumulative Risk (CR) value is calculated for each interval n as follows,

$$CR_n = \sum_{i=M_{n-1}}^{M_{n-1}+N_n} IR_i, \quad (7)$$

$$M_{n-1} = \sum_{n=1}^{n-1} N_n, \quad (8)$$

where M_{n-1} denotes the lower bound of the n 'th interval and IR_i is the IR value of the observation i . Finally, the Average Risk (AR) value is determined for each interval n by dividing the CR by its respective number of observations N_n ,

$$AR_n = \frac{CR_n}{N_n} \quad (9)$$

In the next sections, we will use the average speed and the average risk to evaluate the quality of our calibration process.

VI. ANALYSIS AND MODEL CALIBRATION

In order to calibrate the models' parameters, suitable Measures of Performance (MOP) need to be selected to characterise both simulation and reality outputs. Usually, this selection depends on the aim of the study. For safety performance studies, TTC is one of several indicators of crash potential [9]. When the focus is traffic flow, these measures are related to average speed or traffic volume [32]. However, as stated by Duong et al., traffic is an input for safety performance, and therefore the two objectives are linked [6]. Therefore, we choose different MOPs to both traffic flow and conflicts.

A. Measures of Performance (MOP)

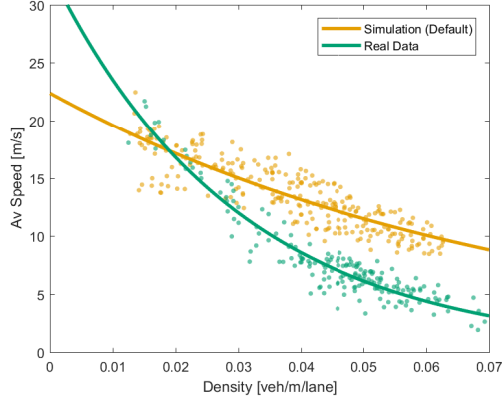
From other studies, it has been observed that both average speed and conflicts depend on the vehicle density. If these relations follow clear trends, the data points will be modelled by functions such that,

$$y = f(x) + \epsilon, \quad (10)$$

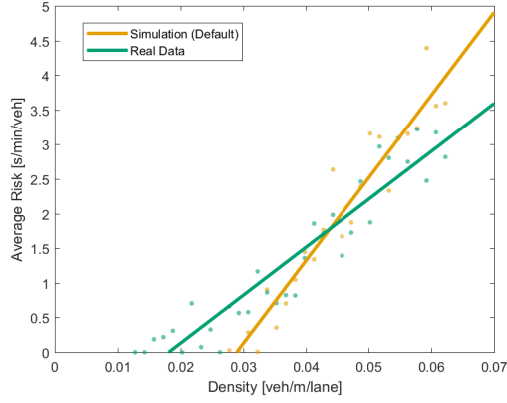
where x and y are the studied variables, f is the selected function, and ϵ is a random error. For instance, after analysing the traffic characteristics from the I-80 real world trace, it was observed that the relation speed-density can be appropriately fitted by an Underwood model [33]. In this model, average speed v relates to density k by using an exponential model:

$$v = v_f \cdot e^{-\frac{k}{k_o}} + \epsilon_v, \quad (11)$$

where free flow speed v_f and optimal density k_o are the parameters that characterise how the average speed decreases when the density on a given road increases. The random speed



(a) Traffic flow data and fitted Underwood models



(b) Safety performance data and fitted linear equations

Fig. 1: Comparison between real data and simulation with default parameters

error ϵ_v represents the discrepancies between the model and the actual measurements. We fit these parameters to the I-80 freeway data set using nonlinear least squares, as shown in Figure 1a. Similarly, for every simulation run, trajectory data is collected and processed, and Underwood parameters are derived from it. To motivate our work further, we also plot the Underwood model derived with the default parameters of the EIDM and MOBIL models [17], [18]. We find that without calibration, the results are considerably different and the simulation does not represent the real traffic situation. The calibration process has to alter the behaviour of the vehicles to more closely resemble the recorded trajectories. Therefore, regarding traffic flow, we select two MOPs: the squares of the relative difference between free flow speed and optimal density from simulation and the real world trace.

$$\text{MOP}_{\widehat{v}_f} = \left(\frac{\widehat{v}_{f\text{sim}} - \widehat{v}_{f\text{real}}}{\widehat{v}_{f\text{real}}} \right)^2, \quad (12)$$

$$\text{MOP}_{\widehat{k}_o} = \left(\frac{\widehat{k}_{o\text{sim}} - \widehat{k}_{o\text{real}}}{\widehat{k}_{o\text{real}}} \right)^2 \quad (13)$$

Regarding the conflicts observed by TTC in the real world trace, we observe that the number of dangerous situations increases linearly with traffic density. Therefore, we make use of a linear function $AR_{\text{TTC}} = a \cdot k + b + \epsilon_{AR}$ to the data points

extracted from the I-80 trace, as shown in Figure 1b. AR_{TTC} is the average risk explained by the indicator TTC, k is the linear density, and a and b are the parameters that characterise the equation, i.e., slope and vertical intercept, respectively. Again, the statistical method used to fit the data and estimate the parameters is non-linear least squares. Compared to a default parametrisation from [17], [18], we observe that neither the intercept nor the slope of this function could be captured.

Therefore, regarding safety performance, the two MOPs selected are the square of the relative difference between the AR's slope \widehat{a} and the vertical intercept \widehat{b} from simulation and the real world trace.

$$\text{MOP}_{\widehat{a}} = \left(\frac{\widehat{a}_{\text{sim}} - \widehat{a}_{\text{real}}}{\widehat{a}_{\text{real}}} \right)^2, \quad (14)$$

$$\text{MOP}_{\widehat{b}} = \left(\frac{\widehat{b}_{\text{sim}} - \widehat{b}_{\text{real}}}{\widehat{b}_{\text{real}}} \right)^2 \quad (15)$$

B. Parameter Selection

The aim of this methodology is to minimise the differences between real and simulated traffic flow outputs by calibrating BEHAVE's parameters. As explained before, E-IDME includes seven adjustable parameters: *Maximum acceleration*, *maximum deceleration*, *minimum gap*, *headway*, *aggression*, *coolness*, and *distraction*. MOBIL-E uses three parameters: *Politeness*, *acceleration threshold*, and *bias*.

Altogether, the calibration methodology must produce a set of 10 parameters that yield satisfactory outcomes. This high number of variables gives a significant level of freedom, which becomes a challenge for a genetic algorithm (GA).

In order to narrow down the search, we decided to select the parameters with highest effect in relation to conflicts and traffic flow, and find their best estimates by utilising a GA.

The first step in the calibration process was to understand the parameters' impact on the behaviour of the vehicles. Subsequently, a few grid searches were carried out to select the best families of parameters which yield to desired \widehat{v}_f , \widehat{k}_o , \widehat{a} and \widehat{b} outputs. It is worth mentioning that, while matching traffic flow characteristics such as free flow speed \widehat{v}_f and optimal density \widehat{k}_o was relatively straightforward, accomplishing a matching slope \widehat{a} and vertical intercept \widehat{b} for the conflicts was much more demanding. This indicates the importance of this study, showing that while the simulated traffic flow might be realistic, safety relevant characteristics may not.

For all parameters, we selected a high and a low level. The best model configuration found during the grid search was used as high level for all parameters. The low level consisted in a decrease of 10% from the high level. Due to the high number of factors, it was decided to implement a fractional factorial design of experiment (DOE). In statistics, fractional factorial designs are experimental designs consisting of a carefully chosen subset of the experimental runs of a full factorial design [34].

The decided 2^{10-3} design is 1/8 of a two level, ten factor factorial design. While the full 2^{10} factorial would require 1024 experiments, this approach requires only 128. After all experiments, effects of the parameters were calculated as the

difference between the responses average with high and low levels:

$$E_{p,j} = \frac{1}{128} \sum_{i=1}^{128} R_{j,p} \cdot a_{i,p} \quad (16)$$

where $E_{p,j}$ is the effect or impact of the parameter p on the objective j , $R_{j,i}$ is the result of the objective j in the i 'th simulation, and $a_{i,p}$ is the element for row i and column p of the indicator matrix A used for the fractional factorial design, which describes the position of parameter p during the experiment i , -1 for low level and 1 for high level.

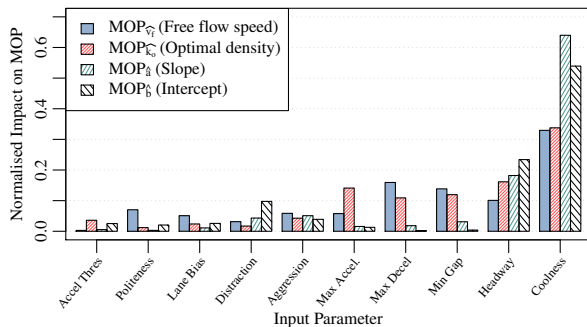


Fig. 2: Sensitivity analysis of input parameter on the measures of performance. Bars represent the normalised impact on the output value caused by a 10% change of the input parameter.

Figure 2 shows the effect of input parameters on the selected MOP. The four parameters with the largest effect on the free flow speed v_f are the *coolness*, *maximum deceleration*, *headway* and the *minimum gap*. For the optimal density k_o , the *maximum acceleration* replaced the *minimum gap*. In terms of conflicts, we observe that *coolness* and *headway* affect the slope and intercept of the average risk AR the most. As expected, *aggression* and *distraction* both have a considerable influence on the observed conflicts, as both of them are crucial to create safety-critical traffic situations. According to Kesting et al. [17], *minimum gap* is only relevant for very low velocities, and therefore it is not considered for further analysis. We sorted the input parameters by their accumulated impact (ascending from left to right in the figure) and decided to select the three most influential (excluding *minimum gap*) for the genetic algorithm. This keeps the complexity and required computational power at a manageable level while at the same time allowing to significantly affect the desired simulation output.

C. Genetic Algorithm (GA)

The GA is an optimisation technique based on the mechanics of natural selection and evolution [35]. The three main operations executed within the GA are crossover, mutation, and reproduction. Although the GA technique has already been used for calibration of models regarding safety performance, these calibrations did not take into account their effect on the traffic flow response [5], [9], [24]. This study aims to calibrate the direct interaction between vehicles, and therefore, we include both conflict and traffic flow performance.

In order to initialise the optimisation, we generate a random (within an allowed range) population of 16 chromosomes or individuals. By combining the three main GA operations, this population is modified throughout 100 generations, resulting in 1600 simulation runs. Every simulation is post-processed following the methodology described in Section V, and evaluated over a fitness function.

We formalise the optimisation problem as the multi-objective minimisation of vector F wrt. the chosen model parameters:

$$\arg \min_{\Omega} F \left(\text{MOP}_{\hat{v}_f}, \text{MOP}_{\hat{k}_o}, \text{MOP}_{\hat{a}}, \text{MOP}_{\hat{b}} \right) \quad (17)$$

where $\text{MOP}_{\hat{v}_f}$, $\text{MOP}_{\hat{k}_o}$, $\text{MOP}_{\hat{a}}$, and $\text{MOP}_{\hat{b}}$ are the objectives for free flow speed, optimal density, conflicts slope, and conflicts vertical intercept, as introduced in Eq. 12 to 15, and $\Omega = \{\text{max. deceleration}, \text{headway}, \text{coolness}\}$ is the set of parameters to calibrate.

In this paper, we consider traffic flow and conflicts equally important and therefore choose the set of parameter values with the lowest sum of MOPs from the Pareto front.

VII. RESULTS

During the initial grid search and the sensitivity analysis, we observed that calibrating the traffic flow was considerably easier than matching the safety performance outputs. Nonetheless, finding a set of parameter values that produce similar conflicts as the real data was well possible; the challenge for a calibration procedure is to achieve both objectives at the same time.

In this study, the driver behaviour models composed by E-IDME and MOBIL-E have been calibrated simultaneously for safety performance and traffic flow. For comparison purposes, two additional Single-Objective (SO) GA, representing the state of the art, have been implemented as well. The first one intends to solely match the traffic flow observed in the real data, the second one focuses on reproducing the average risk.

Figure 3 shows the overall results from our comparison as well as the performance of the multi-objective calibration. As expected, the SO GAs (blue and purple curves) perform well on their respective optimisation goals, closely reproducing the real world results. However, since they do not consider both objectives, the conflicts-optimised configuration produces a much better flow than real data (around $4m/s$ faster). Similarly, the average risk obtained from the flow-optimised GA strongly deviates from the real world data in terms of both slope and y-intercept. Our approach (multi-objective genetic algorithm, yellow line) produces good results for both traffic flow and safety performance, outperforming the single-objective genetic algorithms when considering the overall picture.

We note that it is not trivial to improve the flow response without weakening the safety performance outputs. Especially for high traffic densities, the simulated vehicles move faster than desired. Slowing them down has a direct effect on the conflict counts, effectively lowering them too much. As a result, as shown in Figure 3a, for a density of 0.06

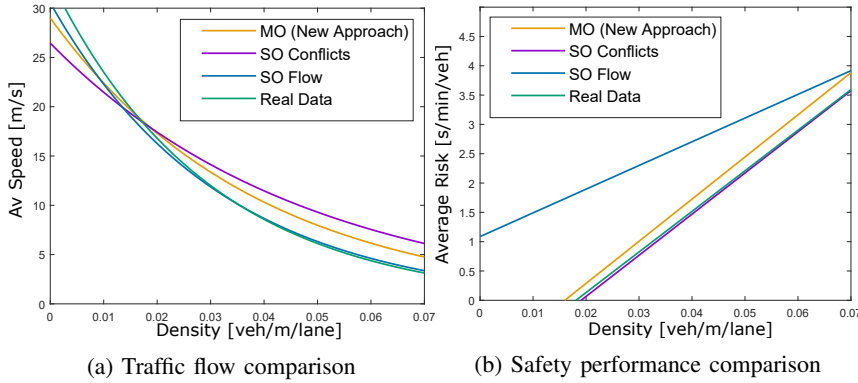


Fig. 3: Comparison between real data and single-objective (SO) and multiple-objective (MO) calibrated simulation.

veh/m/lane, vehicles on the I-80 drive at 4.5 m/s while their simulated counterparts move at 6 m/s. Nonetheless, for medium and low densities, the calibrated models were able to accurately capture the flow from the real world trace. In terms of average risk, the MO GA produced an almost identical slope with only a minor deviation of y-intercept. Altogether, when using the calibrated parameters found during the MO GA, our calibration accomplished $MOP_{\hat{v}_f} = 0.0019$, $MOP_{\hat{k}_o} = 0.0230$, $MOP_{\hat{a}} = 0.0051$, and $MOP_{\hat{b}} = 0.0055$.

Figure 4 displays the fitness values for some of the first rank solutions found by the GA (NSGA-II [36]). The axes x and y represent the $MOP_{\hat{b}}$ and $MOP_{\hat{a}}$, respectively. The combination of the axis z and the gradual colour legend represent the $MOP_{\hat{k}_o}$. Information on the solutions for $MOP_{\hat{v}_f}$ are omitted in this graph. We observe in that there is no individual solution which accomplishes to minimise all of them. Instead, a trade-off in MOPs is observed: On the one hand, solutions that closely replicate \hat{a} and \hat{b} , such as those represented by yellowish and greenish colours, fail to mimic \hat{k}_o . On the other hand, dark blue solutions reproduce adequately \hat{k}_o , but are unsuccessful at minimising $MOP_{\hat{b}}$ and $MOP_{\hat{a}}$ simultaneously.

The best solution found by our approach was *maximum acceleration* = 0.8, *maximum deceleration* = 1.355, *minimum gap* = 2.6, *headway* = 2.13, *coolness* = 0.934, *aggression* = 0.63, *distraction* = 0.00045, *politeness* = 0.45, *acceleration threshold* = 0.675, and *bias* = 0.22.

One limitation of the presented approach is that it was only feasible to efficiently explore a reduced parameter space of three input parameters. Therefore, the genetic algorithm was only able to find only close approximations of traffic flow and safety metrics compared to the almost identical curves achieved using the respective single-objective approach. One possible solution is to employ a two step calibration process, where in a first step the algorithm weighs traffic flow slightly higher than the safety metrics and then in a second step calibrates parameters that have been shown to have a stronger impact on the safety aspects, e.g., the aggression and distraction. Alternatively, a straight forward approach is to utilise more computing power.

To fully simulate *realistic* traffic, the two chosen typical metrics may not be enough. In fact, it might even be possible to simulate evidently (visually) unrealistic traffic while

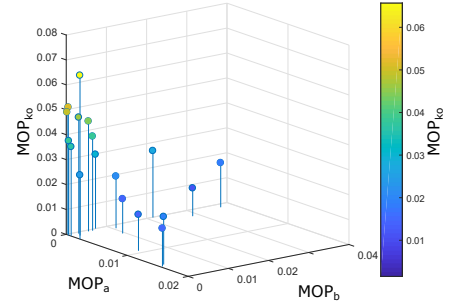


Fig. 4: Pareto-Front formed by first rank solutions with $MOP_{\hat{v}_f} < 10^{-4}$.

still meeting the selected objective functions. Incorporating additional measures such as the lane-change count or driving profile-related metrics has great potential to further improve the degree of realism of microscopic traffic simulations.

A. Discussion on Validation

We applied the identified parameter values to a different scenario (NGSIM US-101 [16]). This scenario differs in terms of lane-count and the presence of a high-occupancy vehicle lane. While the US-101 scenario included one on-ramp and one off-ramp, the used I-80 dataset only included one on-ramp.

Our results showed that the validation with this second dataset did not yield satisfactory results. The observed traffic flow (also conflicts, not shown) is significantly different (see Figure 5). As the baseline for both scenarios is already different, we conclude that the calibration for one scenario will not work for both. Reasons for this can be manifold, ranging from road geometry to traffic mix, to time of day.

Splitting the I-80 dataset in training and validation subsets is not helpful as the data includes the build-up of congestion, transitioning from uncongested to fully congested. The data contains a total 45 minutes of traffic; there is simply not enough data to adequately validate the model on the same dataset.

We conclude that in order to work on a calibrated model, the calibration has to be carried out for each scenario separately. While this is already the case for current calibration methods,

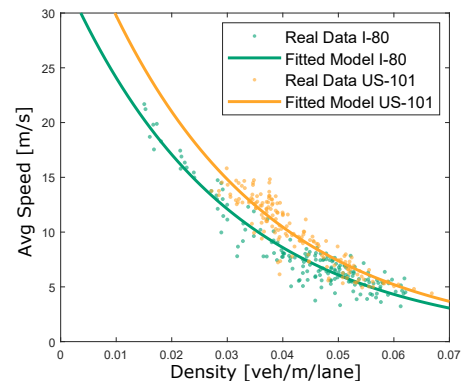


Fig. 5: Comparison between two real data-sets recorded on Interstate 80 and Highway 101, respectively.

our approach can provide a simulation environment producing realistic results for more than one metric.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce an approach that incorporates a multiple objective genetic algorithm to calibrate driver behaviour models to yield realistic traffic flow and road safety metrics. To allow driver behaviour to match both metrics at the same time, we extend the well-established EIDM and MOBIL models to incorporate attention and aggression. The used models incorporate a total number of ten parameters – too many to calibrate in a computationally feasible manner. Followed by an initial grid search, we therefore conducted a sensitivity analysis and identified the three most important parameters (coolness, headway, maximum deceleration) to affect traffic flow and average risk.

For the calibration process, we define a total of four Measures of Performance (MOP) which are used by the genetic algorithm to tune the identified set of parameters. We show that while conventional single-objective calibration methods fail to capture both output metrics adequately, our multi-objective approach is able to produce sufficiently accurate results. The necessity of such an approach is furthermore emphasised by the fact that using the incorporated behaviour models with their suggested default parameters did not match the real world data.

Future research includes a closer examination of the limitation caused by the trade-off between traffic flow and conflicts. Secondly, we plan to apply the methods introduced in this paper in other scenarios including intersections or urban roads. Finally, integrating other surrogate safety indicators for the calculation of conflicts and new traffic characteristics, such as lane-changes counts can give valuable insights in assessing the realism of a traffic simulation.

REFERENCES

- [1] C. Hydén, “Traffic conflicts technique: state-of-the-art,” *Traffic safety work with video processing*, no. 37, pp. 3–14, 1996.
- [2] J. Darzentas, D. F. Cooper, P. Storr, and M. McDowell, “Simulation of road traffic conflicts at t-junctions,” *Simulation*, vol. 34, no. 5, pp. 155–164, 1980.
- [3] K. Rumar, “The role of perceptual and cognitive filters in observed behavior,” in *Human behavior and traffic safety*, 1985, pp. 151–170.
- [4] J. Archer and I. Kosonen, “The potential of micro-simulation modelling in relation to traffic safety assessment,” in *Proceedings of the ESS Conference*. Citeseer, 2000, pp. 427–443.
- [5] M. Essa and T. Sayed, “Simulated traffic conflicts: do they accurately represent field-measured conflicts?” *Transportation Research Record: Journal of the TRB*, no. 2514, pp. 48–57, 2015.
- [6] D. Duong, F. Saccomanno, and B. Hellinga, “Calibration of microscopic traffic model for simulating safety performance,” in *89th An. Meeting of the Transportation Research Board*, Washington, DC, 2010.
- [7] J. Hourdakis, P. Michalopoulos, and J. Kottommannil, “Practical procedure for calibrating microscopic traffic simulation models,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1852, pp. 130–139, 2003.
- [8] G. Schultz and L. Rilett, “Analysis of distribution and calibration of car-following sensitivity parameters in microscopic traffic simulation models,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1876, pp. 41–51, 2004.
- [9] F. Cunto and F. F. Saccomanno, “Calibration and validation of simulated vehicle safety performance at signalized intersections,” *Accident analysis & prevention*, vol. 40, no. 3, pp. 1171–1179, 2008.
- [10] E. Brockfeld, R. Kühne, and P. Wagner, “Calibration and validation of microscopic traffic flow models,” *Transportation Research Record: Journal of the TRB*, no. 1876, pp. 62–70, 2004.
- [11] R.-L. Cheu, X. Jin, K.-C. Ng, Y.-L. Ng, and D. Srinivasan, “Calibration of fresim for singapore expressway using genetic algorithm,” *Journal of Transportation Engineering*, vol. 124, no. 6, pp. 526–535, 1998.
- [12] D. Gettman and L. Head, “Surrogate safety measures from traffic simulation models,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1840, pp. 104–115, 2003.
- [13] —, *Surrogate safety measures from traffic simulation models, final report*. The Center, 2008.
- [14] J. Ivanchev, T. Braud, D. Eckhoff, D. Zehe, A. Sangiovanni-Vincentelli, and A. Knoll, “On the Need for Novel Tools and Models for Mixed Traffic Analysis,” in *26th ITS World Congress*, Singapore, October 2019.
- [15] D. Zehe, S. Nair, A. Knoll, and D. Eckhoff, “Towards CityMoS: A Coupled City-Scale Mobility Simulation Framework,” in *5th GI/ITG KuVS FG-IVC 2017*, Erlangen, Germany, April 2017.
- [16] J. Colyar and J. Halkias, “Us highway 101 dataset. federal highway administration research and technology fact sheet. publication number: Fhwa-hrt-07-030,” Tech. Rep., 2007.
- [17] A. Kesting, M. Treiber, and D. Helbing, “Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity,” *Philosophical Trans. of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1928, pp. 4585–4605, 2010.
- [18] —, “General lane-changing model mobil for car-following models,” *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.
- [19] A. D. May, *Traffic flow fundamentals*, 1990.
- [20] S. S. Mahmud, L. Ferreira, M. S. Hoque, and A. Tavassoli, “Application of proximal surrogate indicators for safety evaluation: a review of recent developments and research needs,” *IATSS research*, 2017.
- [21] A. R. A. Van der Horst, *A time based analysis of road user behaviour in normal and critical encounters*, 1990, no. HS-041 255.
- [22] T. Sayed, G. Brown, and F. Navin, “Simulation of traffic conflicts at unsignalized intersections with tsc-sim,” *Accident Analysis & Prevention*, vol. 26, no. 5, pp. 593–607, 1994.
- [23] K. Vogel, “A comparison of headway and time to collision as safety indicators,” *Accident analysis & prevention*, vol. 35, no. 3, pp. 427–433, 2003.
- [24] F. Huang, P. Liu, H. Yu, and W. Wang, “Identifying if vissim simulation model and ssam provide reasonable estimates for field measured traffic conflicts at signalized intersections,” *Accident Analysis & Prevention*, vol. 50, pp. 1014–1024, 2013.
- [25] J. Hogema and W. Janssen, “Effects of intelligent cruise control on driving behaviour: a simulator study,” 1996.
- [26] G. Hegeman, *Assisted overtaking: An assessment of overtaking on two-lane rural roads*. TRAIL Research School Delft, the Netherlands, 2008, no. T2008/4.
- [27] T. Officials, *A Policy on Geometric Design of Highways and Streets, 2011*. AASHTO, 2011.
- [28] H. Farah, S. Bekhor, A. Polus, and T. Toledo, “A model for passing decisions on two-lane rural highways,” 2008.
- [29] A. Van Der Horst, “A time-based analysis of road user behaviour at intersections,” in *ICTCT Conference Proceedings (Krakow)*, 1990, pp. 91–104.
- [30] R. van der Horst and J. Hogema, *Time-to-collision and collision avoidance systems*. TNO Institute for Human Factors, 1993.
- [31] Y. Kuang, X. Qu, and Y. Yan, “Will higher traffic flow lead to more traffic conflicts? a crash surrogate metric based analysis,” *PLoS one*, vol. 12, no. 8, p. e0182458, 2017.
- [32] T. Ma and B. Abdulhai, “Genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1800, pp. 6–15, 2002.
- [33] R. T. Underwood, “Speed, volume, and density relationships.”
- [34] G. E. Box, J. S. Hunter, and W. G. Hunter, *Statistics for experimenters: design, innovation, and discovery*. Wiley-Interscience New York, 2005, vol. 2.
- [35] D. E. Goldberg and J. H. Holland, “Genetic algorithms and machine learning,” *Machine learning*, vol. 3, no. 2, pp. 95–99, 1988.
- [36] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: Nsga-ii,” *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.