IEEE Open Journal of
Intelligent Transportation
Systems

Received 2 June 2023; revised 13 September 2023; accepted 16 September 2023. Date of publication 2 October 2023; date of current version 19 October 2023.

Digital Object Identifier 10.1109/OJITS.2023.3321110

A Bayesian Optimization Approach for Calibrating Large-Scale Activity-Based Transport Models

SERIO AGRIESTI[1,2], VLADIMIR KUZMANOVSKI[2,3,4], JAAKKO HOLLMÉN[3,5] (Senior Member, IEEE), CLAUDIO RONCOLI[1], AND BAT-HEN NAHMIAS-BIRAN[6,7] (Member, IEEE)

[1]Department of Built Environment, School of Engineering, Aalto University, 02150 Espoo, Finland

[2]FinEst Centre for Smart Cities, Tallinn University of Technology, 19086 Tallinn, Estonia

[3]Department of Computer Science, Aalto University, 02150 Espoo, Finland

[4]Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

[5]Department of Computer and Systems Sciences, Stockholm University, 114 19 Stockholm, Sweden

[6]The Porter School of the Environment and Earth Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

[7]School of Social and Policy Studies, Tel Aviv University, Tel Aviv 6997801, Israel

CORRESPONDING AUTHOR: S. AGRIESTI (e-mail: serio.agriesti@aalto.fi)

This work was supported in part by the FINEST Twins Center of Excellence; in part by the H2020 European Union funding for Research and Innovation under Grant 856602; and in part by the Academy of Finland Project ALCOSTO under Grant 349327.

ABSTRACT Addressing complexity in transportation in cases such as disruptive trends or disaggregated management strategies has become increasingly important. This in turn is resulting in the rising adoption of Agent-Based and Activity-Based modeling. Still, a broad adoption is hindered by the high complexity and computational needs. For example, hundreds of parameters are involved in the calibration of Activity-Based models focused on behavioral theory, to properly frame the required detailed socio-economical characteristics. To address this challenge, this paper presents a novel Bayesian Optimization approach that incorporates a surrogate model defined as an improved Random Forest to automate the calibration process of the behavioral parameters. The presented solution calibrates the largest set of parameters yet, according to the literature, by combining state-of-the-art methods. To the best of the authors' knowledge, this is the first work in which such a high dimensionality is tackled in sequential model-based algorithm configuration theory. The proposed method is tested in the city of Tallinn, Estonia, for which the calibration of 477 behavioral parameters is carried out. The calibration process results in a satisfactory performance for all the major indicators, the OD matrix average mismatch is equal to 15.92 vehicles per day while the error for the overall number of trips is equal to 4%.

INDEX TERMS Activity-based transport modeling, model calibration, machine learning, Bayesian optimization, surrogate model.

I. INTRODUCTION

LARGE-SCALE transportation problems have always been prone to high complexity, approximations, and a lack of a univocal mathematical formulation [1]. This is especially the case for models involving human behavior and the numerous factors ruling over mobility choices, which have been applied to narrow scopes (e.g., modal choices) or have been designed as aggregated (e.g., four-step models). Still, current and future transportation challenges – e.g., urbanization, population growth, and congestion [2], [3] but also disruptive events such as pandemics [4] or the climate crisis – require solutions able to frame transport demand through the lenses of individual choices on a large scale. Future innovations on the transport supply spectrum [5], [6] are also foreseen to have disruptive effects on mobility demand. To evaluate said innovations, tools able to frame changed mobility choices and travel habits are needed. Currently, agent-based modeling (ABM) is

The review of this article was arranged by Associate Editor Jiaqi Ma.

the most promising solution due to the ability to frame both demand and supply at the agent level (the agent being either the individual or the single vehicle) in a disaggregate fashion, allowing the investigation of emergent behaviors [7]. Specifically, activity-based models are a particular kind of ABM, where the population is modeled in a disaggregate fashion, with each individual as an agent. To do so, allows one to frame each behavioral choice based on individual socio-economic features. Activity-based and ABM models have already successfully been used in policy analyses [8], accessibility studies [9], and forecasting experiments [10], [11], [12], [13] focusing on automated mobility. The disaggregated approach is particularly fit for scenarios involving automated vehicles because historical data cannot be exploited in these scenarios to forecast a realistic demand, which is necessary in turn to address problems such as fleet sizing or routing algorithms. Behavioral modelling offers an alternative to the lack of historical patterns. Another ABM application addressing the reported long-term challenges and strongly benefitting from a properly calibrated large-scale activity-based model is the modeling of remote and hybrid working patterns during public health crises [14].

Still, despite these needs, the state-of-the-art concerning activity-based models behind most ABM is limited, especially when it comes to the underlying calibration. The different structures of the available activity-based tools and the different magnitude of parameters to be calibrated each time are still a barrier against a wider adoption. A unified calibration approach, not dependent on specific software and able to include hundreds of parameters, is still needed to fill this research gap and foster the usage of behavioral activity-based models [15], [16]. The presented work tries to tackle the issue and thus to foster the adoption of activity-based models thanks to the global optimization (BO) and the resulting calibration. By doing so for a large-scale urban scenario, the presented work removes one of the main hurdles in the field, further advances the applicability of ABMs, and improves the performance of current calibration approaches.

BO [17], [18] exploits sequential sampling designing, a surrogate model and the resulting response surface to search for the global optimum. By exploiting a surrogate model, the computational time is strongly reduced and an extensive search process is made possible. By doing so, it becomes possible to balance the trade-off between exploration and exploitation. As such, the method is designed for optimization problems that feature "expensive" functions in terms of computational time, which are approximated through the surrogate surface. This study focuses on the development of a high-dimensional BO method that converges within a given computational budget and avoids the necessity to master the implementation of the underlying large-scale ABM. The proposed algorithm and the modifications to the BO approach are designed to be transferable to other large-scale problems involving dozens of parameters and fit to be analyzed through a surrogate model. The amount of calibrated parameters considered (up to 477) for BO has yet to be matched, as it will be shown in Section II. Besides, this work reports the first large-scale solution to the ABM calibration problem without the constraint of a specific tool or data structure. Automating the calibration process for an urban case study would remove one of the main obstacles to the wider adoption of ABM, as will be further reported in Section II.

In Section II, a review of the current literature is carried out, concerning the calibration of large-scale transport models and BO; Section III elaborates on the proposed methodology; Section IV describes the case study, while Section V reports the results obtained by applying the proposed methodology to the case study; Section VI discusses the results and highlights the main conclusions for this work.

The work presented in the paper aims to foster the adoption of both large-scale behavioral activity-based models and BO methods by reporting a methodological approach for the calibration of hundreds of parameters in the transportation domain. As it will be shown in Section II, no other work manages to calibrate 477 parameters at once through a surrogate model and BO techniques. Besides, the code and data used for all the experiments are made publicly available to foster replicability.

II. LITERATURE REVIEW

This work tackles two streams of literature in an interdisciplinary fashion: activity-based modeling for large transportation case studies and Bayesian Optimization techniques. Accordingly, this literature review is divided into two parts. The first subsection reports current calibration techniques focusing on activity-based transportation models and behavioral parameters. This means that the wider literature concerning the supply-side parameters is not addressed here, being the number and the nature of the calibrated parameters not comparable between behavioral activity-based models and traffic assignment ones. The subsection on Bayesian Optimization focuses instead on the current methodological approaches, their perks and limitations, as well as their scale and their applications.

A. CALIBRATION APPROACHES IN TRANSPORT MODELING

The calibration of behavioral activity-based models in transport has received far less attention than the calibration of other supply-focused ABMs, while existing approaches mainly rely on heuristics, which, in turn, is hindering the potentialities of these tools. Still, some notable works can be found in literature.

The work presented in [15] describes a gradient- and simulation-based optimization procedure designed to calibrate 28 parameters in a utility-based nested logit system. Similarly, [19] exploits the WSPSA algorithm to calibrate 94 behavioral parameters on the demand side, still, it does not exploit a surrogate model and thus needs multiple computationally expensive runs. Besides, the WSPSA requires

the definition of a weight matrix, which becomes difficult to define as the number of parameters increases. Different SPSA techniques have been more extensively designed and used to calibrate the aggregated demand in traffic assignment models rather than activity-based ones, examples of such applications are [20], [21], [22], [23], [24]. Not all works exploit an algorithm to calibrate the behavioral parameters in an activity-based model. In [25], for example, a trial-and-error approach is carried out to calibrate against the recorded modal share and departure times. The model is built for the city of Barcelona through the use of Call Detail Records.

Somehow more attention has been dedicated to the calibration of the demand against supply-side parameters in ABMs. This is probably due to the availability of tools encompassing both dimensions and thus requiring a joint calibration. While this approach has many advantages (e.g., wider applicability), as it will be shown in the following the added complexity usually requires to compromise on some aspects of the calibration. In [26], [27], an iterative black box approach is adopted to calibrate an activity-based transport model, with the former applied to a small network (24 zones) and the latter calibrating only 9 behavioral parameters. Paper [28] succeeds in calibrating 25 behavioral parameters through a maximum-likelihood method exploiting link counts and/or plate scanning data. In [29], [30], [31], an activity-based model and a traffic assignment model are jointly calibrated. Papers [29], [30] calibrate both the MATSim software and CEMDAP [32], with MATSim receiving the final calibration based on traffic counts. In [29], for example, CEMDAP after the calibration produces a set of 5 eligible daily activity schedules but then it is up to MATSim (and the related module CaDyTS) to score them according to how well they reproduce the supply side measurements so that the supply performance does not result in a change of behavioral parameters. Besides, the modal share is manually calibrated at the end of the process. A similar approach for the demand is followed in [24] while the supply is calibrated through SPSA. A general description of the scoring system and the underlying behavioral models in MATSim is provided in [33], the work also crucially highlights some of the challenges currently related to integrating the demand and supply components. Besides, when different calibration steps are carried out separately for demand and supply, a complete traffic assignment model is needed, which increases the overall calibration effort by increasing the number of factors but also by somehow putting the two modules (activity-based and traffic assignment) "against" each other whereas a univocal optimization criterion for the two modules is not defined, with convergence being decided by supply-side metrics. This issue is tackled in [34], where MATSim is integrated with a multinomial discrete choice model considering 12 behavioral parameters. Results show that the choice of behavioral parameters becomes a key element in the simulation pipeline, without which it is not possible to reach a good integration with an ABM while avoiding convergence issues. To address this limitation, [35] decouples the traffic assignment from the behavioral components; however, the resulting agents' features and the related behavioral constants in a logit model remain not calibrated, which limits the transferability or even the usability of the calibrated model. Overall, MATSim applications, while generally more widespread, are more limited in their ability to predict new technologies and disruptive scenarios while, also, relying on a smaller number of parameters to be calibrated [36]. Based on the above, it is possible to state the following limitations concerning the state-of-the-art:

1) Existing literature tackling the calibration of behavioral parameters for activity-based models on large-scale scenarios is scarce and only a handful of works try to solve the problem without recurring to heuristics.
2) Many calibration methods aim to reproduce outputs of the activity-based models matching the desired supply-side measurements, rather than to calibrate the underlying behavioral parameters. Thus, the calibration of the supply overrules the calibration of the demand.
3) Even the works trying to formalize a rigorous methodology do not consider more than a few dozen parameters (the maximum number being 98 in [19]).

B. BAYESIAN OPTIMIZATION

BO finds applications in various scientific and industrial domains, e.g., machine learning for hyperparameter optimization [37], [38], modeling of population genetics [39], spreading of pathogens [40], atomic structure of materials [41], [42], as well as cosmology [43], and establishes as a state-of-the-art method in lower-dimensional problems [17], [44]. However, the BO performances and its computational efficiency decline as the dimensionality of a problem increases [37], [45], [46], [47], which is the case with the calibration of large-scale ABM that features a large number of behavioral parameters to be tuned.

In state-of-the-art applications of the BO, the Gaussian processes (GP) is usually exploited as a prior distribution to both model the surrogate surface and to approximate the posterior distribution of the parameters [17], [18], [48]. The advantages of GP are tied to its probabilistic nature, thanks to which it is possible to quantify the prediction uncertainty by assessing the distance in mathematical spaces between the known regions and the new samples. Such quantified uncertainty allows for an efficient trade-off that guides the search for better samples to be sampled, which helps the BO to achieve state-of-the-art performances. However, the GP comes with a computation bottleneck when applied to high-dimensional problems, which in turn hinders a wider adoption for complex parameter spaces [37], [45], [46], [47]. Therefore, the straightforward adoption of the BO method in the calibration of activity-based models is hampered by the large number of parameters to be tuned [15], [49].

For that purpose, various methods for dimensionality reduction are adopted [26], [27], [50], or transformations (including partitioning) of the parameter space are applied [51], [52], [53], [54], [55], [56], [57], but neither solve the issue since a higher number of runs is needed in

the former while two strong assumptions are required from the latter (i.e., low intrinsic dimensionality and compounding effects). References [26], [27] invest a significant amount of effort to introduce BO in the field of transportation modeling, but fails to do so without an expensive dimensionality reduction using a deep learning methodology, i.e., auto-encoders. The deep learning methodology has been shown in numerous applications to be a valuable approach in high-dimensional spaces, but it is known to be data-intensive [58], which in the context of transportation activity-based models means a greater number of executions of the models. Outside the activity-based applications, deep learning has nevertheless been applied in the transportation domain, e.g., to estimate the intersection's queue length and dissipation time [59] or to enhance prediction fairness in spatial-temporal demand forecasting of ride-hailing services [60].

Approaches based on transforming the parameter space and decomposing the optimization problem into subproblems – each mapped to a lower-dimensional space – depend on space properties, among which the intrinsic dimensionality is the most important. In that context, previous works explore a latent space where the function is decomposable, either by latent structures [52] or additive structures [61], [62]. Another approach to dimensionality reduction involves random projections into a latent space [51], [54], [55], [56] or low-rank matrix approximation [63]. Alternatively, a cylindrical transformation of the parameter space [53] and sequential optimization along with a subset of dimensions [64] are adopted in recent studies. However, the authors of these studies report the performance on benchmark optimization functions, showing that the methods perform well on problems with low intrinsic dimensionality, but fail to depart significantly from the initial points in optimization problems with high intrinsic dimensionality.

Without prior knowledge of the intrinsic dimensionality of the large-scale activity-based models and the corresponding calibration process, we consider a variation of the BO method that uses a dimensionality-wise more robust method to approximate the posterior of the parameters, i.e., Random forests [65]. Previously, the method of Random forests has been used in a study with low dimensional optimization problems, featuring discrete mathematical spaces [66].

III. METHODOLOGY

A. ACTIVITY-BASED TRANSPORT MODELS

An activity-based ABM aims at describing the behavior of potentially millions of agents, each representing a traveler (and/or a vehicle), each one capable of multiple choices through the simulation horizon. The decision process for each agent is typically modeled via a nested tree, where each node represents a choice, which is in turn defined via utility maximization, solved via evaluating multiple (utility) functions, each described by several parameters and, crucially, the corresponding weights. Utility functions are of the form

$$U_{\text{choice}} = f\left(V, \underline{\beta}_{\text{choice}}\right), \tag{1}$$

where V is a set of (known) parameters characterizing the agent, which could include, e.g., the socioeconomic features of each agent (defined while creating the synthetic population), and $\underline{\beta}_{\text{choice}}$ are the weights defining how much these parameters concur to the utility function. The hundreds of possible combinations among weights result in a problem for which it is impossible to calculate an analytical solution, while the stochastic nature of the utility-maximizing theory makes employing a numerical solution a necessity. In the following, all weights $\underline{\beta}_{\text{choice}}$ present in all utility functions defined in a decision tree are grouped in θ, while the observations are measurements resulting from the emergent behavior of the agents.

B. THE CONCEPTUAL DESIGN

BO [17], [18] as a methodology exploits sequential sampling design and a surrogate model over a surrogate (response) surface approximating a likelihood function. By doing so, BO seeks the global optima. The parameter space and the discrepancy between observations and simulation results both characterize the surrogate surface. The approach is iterative and each iteration produces new parameters' values (i.e., samples) maximizing the expected utility. Said utility corresponds to a capacity of the new parameters' values to minimize the optimizing quantity - cost and is estimated with an acquisition function, which balances the trade-off between exploration and exploitation of the search space and efficiently guides towards the global optima (Fig. 1). An iteration starts with fitting a surrogate function (data-driven model) based on previous evidence and simulations. Then, the BO generates a large number of candidate samples (parameters' values) that are attributed with a utility by using the surrogate function and an acquisition function. Finally, the candidate with the highest utility, i.e., the greatest estimated capacity to reduce the optimization cost, is selected and used in a run of a simulation model (e.g., activity-based ABM), a result of which updates the evidence for the next iterations.

Formally, a simulation model is a generative stochastic process and its calibration corresponds to a statistical inference of a finite number of parameters $\theta \in \mathbb{R}^d$ based on a set of observations Y_o:

$$p(\theta|Y_o) = \frac{p(Y_o|\theta) \cdot p(\theta)}{p(Y_o)}, \tag{2}$$

where $p(\theta)$ is the prior belief on the distribution of parameter values and $p(Y_o|\theta)$ is the likelihood of the observations, given the parameters, resulting from a known function $\mathcal{L}(\theta)$. $L(\theta)$ is used instead of $\mathcal{L}(\theta)$ because the analytical form of the latter is not known a priori. $L(\theta)$ is approximated over a set of N samples - $\tilde{L}^N(\theta)$. As the marginal distribution $p(Y_o)$ does not depend on θ, we omit it from the formulation,

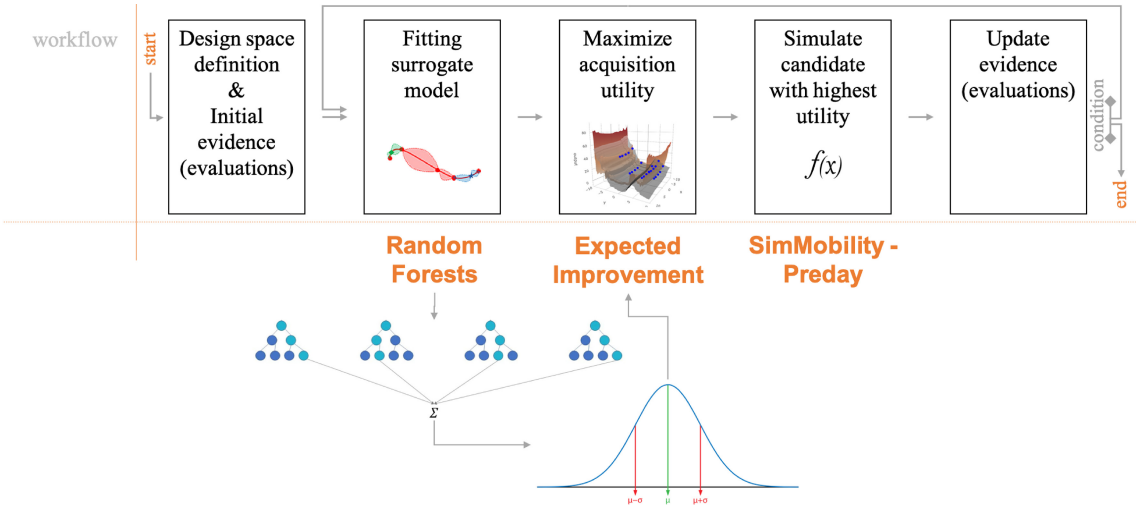


FIGURE 1. Conceptual design of the iterative Bayesian optimization method with Random Forest as a surrogate model and Expected Improvement (EI) as an acquisition function.

which becomes:

$$p(\theta|Y_o) \propto L(\theta) \cdot p(\theta), \tag{3}$$

$L(\theta)$ is reconstructed as the number of samples increases:

$$\lim_{N\to\infty} \tilde{L}^N(\theta) = L(\theta). \tag{4}$$

C. THE SURROGATE FUNCTION AND THE SAMPLING DESIGN

The approximation ($\tilde{L}^N(\theta)$) of the likelihood function ($L(\theta)$) is the formal task we address with the BO methodology, using a surrogate function and the sequential sampling design [18]. To model the surrogate function, we use a regressor such as Random forests (RF) [65], which is used to estimate the acquisition utility of newly sampled parameter values through the *Expected Improvement* (EI) [77]:

$$EI\left(\theta|\mu,\sigma,f^*\right) = \sigma(\theta)[z\Phi(z) + \phi(z)] \tag{5}$$

$$z = \frac{f^* - \mu(\theta)}{\sigma(\theta)}, \tag{6}$$

where $\sigma(\theta)$ and $\mu(\theta)$ are the standard deviation and the mean of the inferred posterior distribution, f^* is the active optima discovered in the previous iterations, and Φ and ϕ are probability density and cumulative distribution function in terms of the standard normal distribution, respectively. The expected improvement $EI(\theta) = 0$ if $\sigma(\theta) = 0$. Eq. (5) represents the exploration-exploitation trade-off that favors higher uncertainties that are close to the latest discovered optimal region(s).

The RF [65] is an ensemble method composed of C regression trees, which follow the decision tree concept, with a structure of decision binary nodes built iteratively in a top-down fashion. Each regression tree is built from random subsets of the features and from bootstrap samples. Trees are conceived to explore a portion of the parameter space. Given a dataset, each regression tree outputs a prediction of the target for the specific region it is exploring within the search space. The prediction of the ensemble is instead an average of the outcomes of all C tree base predictors:

$$\mathcal{RF}(\theta|\Theta, Y) = \frac{1}{C}\sum_{i=1}^{C} \tau_i; \;\; \tau_i = T_i(\theta|\Theta_i, Y_i), \tag{7}$$

where C is the number of tree predictors, Θ_i and Y_i are training datasets of i-th regression tree T_i that provides a prediction τ_i, while Θ and Y are global training dataset and the corresponding label set, respectively. RF has been chosen for its robustness and scalability when compared to other machine learning models. Namely, RF performs well in cases of high-dimensional problems with limited dataset sizes, which is not the case with more advanced techniques like neural networks and deep learning [49], [65]. The latter requires much more data in order to generalize over a given problem domain. Additionally, RF de-prioritizes certain sections of the search space by handling conditional variables [17]. The outcome of the RF method can be strongly influenced by a limited number of hyper-parameters, usually, the ones undergoing a tuning process are: the number of tree components C, the minimum number of samples in a terminating node that controls the structure growth and over-fitting settings of the individual tree components, and the number of features to design a sub-space or partition. Additionally, the RF method is characterized by very high robustness over high-dimensional data, which results in a limited bias of the overall predictions due to the maximizing of the variance between predictors [67].

Still, RF models as surrogates for BO have the following limits: (a) they do not frame uncertainties while quantifying the predictions due to their non-probabilistic output and (b) the values outside the observed space are not predicted. Therefore, the efficiency of the probabilistic

acquisition function (EI) is greatly affected by the RF during the acquisition of new promising samples [17], [66].

D. THE IMPROVED RANDOM FOREST

In order to comply with the expected probabilistic output, the RF method is adapted so that it approximates a parametric (normal) probability distribution of the evaluated input parameter values (θ), through empirically derived mean (μ_θ) and standard deviation (σ_θ):

$$\mu_\theta = \mathcal{RF}(\theta|\Theta, Y) = \frac{1}{C}\sum_{i=1}^{C}\tau_i; \tag{8}$$

$$\sigma_\theta = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(\mu_\theta - \tau_i)^2}, \tag{9}$$

where τ_i corresponds to a prediction of a single decision tree model in the RF, as described in Eq. (7).

As we adapt the RF into a compatible method for modeling the surrogate function that enables estimation of the acquisition utility for each new sampled parameter value, a component of BO, yet to be formalized in our calibration framework, is the optimization of the acquisition utility at each iteration of the iterative optimization process. The optimization of the acquisition utility corresponds to finding a set of parameters values (θ^*) maximizing the utility:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, EI\left(\mu_\theta, \sigma_\theta, \theta^*_{\text{prev}}\right), \tag{10}$$

where θ^*_{prev} is the optimal set of parameter values obtained in previous iterations of the process.

Finding the set of parameters' values that maximizes the acquisition utility (estimated to perform best) can be performed in various ways, including Random search, Thompson sampling, gradient-based, or population-based (evolutionary) optimization methods [17], [18], [38]. In this study, we examined the performances of the Random search, gradient-based, and population-based methods, observing that the gradient-based outperforms the rest. Therefore, for our case study, we adopt the gradient-based Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B) [68], [69]. L-BFGS-B is an efficient optimization method for unconstrained and bounded-constrained optimization problems. Its efficiency is reflected through limited-memory approximation of the inverse Hessian matrix and gradient information, which are improved iteratively.

E. THE SELECTION OF THE BEST-PERFORMING SIMULATION

The stochastic nature of the proposed method requires that the optimization is performed multiple times, by which the possibility of finding a locally optimal solution is eliminated. However, depending on the complexity of the high-dimensional problem at a glance and the definition of the optimization objectives (cost), it may be not feasible to completely avoid locally optimal regions of the problem space, as there may be more of them. Therefore, *ex-post* Pareto analyses of simulated candidates according to a broader set of criteria is recommended. These shall represent some important benchmarks or margins of the considered case study and have to be defined accordingly. They differ from the optimization objectives (performance measure) that are exploited to guide the algorithm through the learning process, and they are used to filter out non-robust simulations in accordance with the Pareto objectives.

Pareto analysis [70] is a quantitative method that seeks to identify and prioritize the options that contribute most significantly to a particular outcome. It is based on the observation that a small subset of the options (located at the Pareto front) often accounts for a majority of the observed effects and dominates the rest of the options in terms of a quantifiable objective.

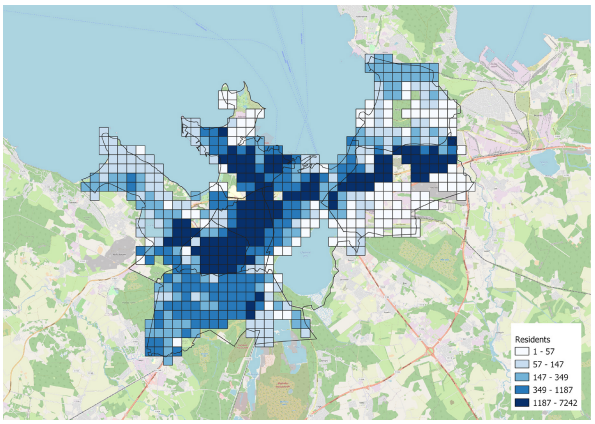


FIGURE 2. Spatial distributions of residents and grid corresponding to the 616 zones considered in the modeling.

IV. CASE STUDY

The proposed algorithm is tested by modeling Tallinn (the capital city of Estonia). The reference year is 2015 as the mobility survey used both for the generation of the database population and for the calibration was carried out in the said year (Fig. 2). SimMobility Preday [71] is employed as the activity-based model, which takes as input a Postgres database containing the following features:

1) A population of ∼400.000 synthetic individuals, matching the city's whole population, while each individual is characterized by socioeconomic features such as age, gender, income, workplace, etc.
2) A spatial map of 616 zones, each 500x500 meters wide (Fig. 2).
3) Skim tables detailing the costs of different travel modes, including waiting time, time on board, etc., among the different zones.

Four transport modes are considered: Public Transport, Private Vehicles, Walking, and Others (e.g., motorcycles). The spatial resolution has been set to 500 m to realistically

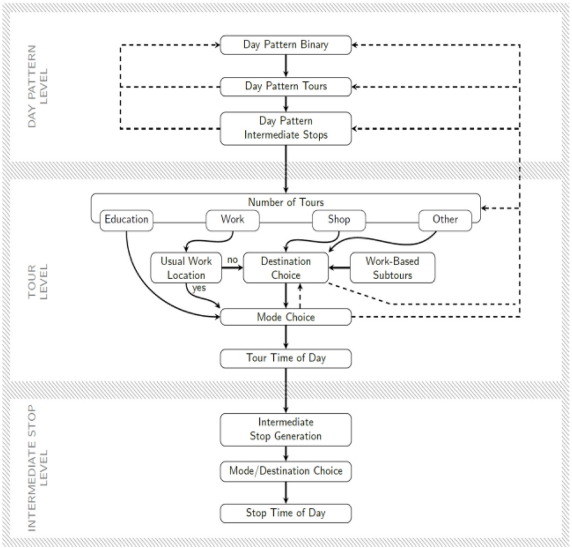


FIGURE 3. Nested logit in SimMobility - levels and choices included in the calibration [75].

capture the Walking mode of transport, while the skim matrix for public transport has been extracted through Open Trip Planner [72]. More details about the synthetic population, its generation, and assignment, may be found in [73].

As detailed in [71], SimMobility is an agent-based, fully econometric, activity-based demand model. Its structure is fully modular and allows us to focus independently on the demand side of the tool, called Preday. A run of SimMobility Preday results in a Daily Activity Schedule, a dataset including one entry item for each leg of each tour. An extract is provided in Table 1 for a randomly chosen individual (ID 107).

The main strength of SimMobility Preday lies in the behavioral models used, i.e., a series of nested logit functions allowing to simulate the travel demand based on an established methodology [9], [71], [74]. This feature of the model allows simulating future scenarios for which the ground data (e.g., traffic counts) are not yet available, which is done via computing utilities at different levels (binary choice to leave the residence, type and number of tours, modes and destinations, time of the day, and stops). These choices are interrelated, an aspect that is accounted for through the computation of logsums (defined in [9] as the log of the denominator of a logit choice probability).

A representation of all the levels of the choice tree is provided in Figure 3. Each level is characterized by its own utility functions and a specific set of β s. Figure 3 shows the model components and process flow of the Preday model. The overall system can be viewed as a hierarchical (or nested) series of choice models. The solid arrows indicate that models from lower levels are conditioned on decisions made with models from higher levels. There are three different hierarchies in the Preday model: *day pattern level*, *tour level*, and *intermediate stop level*. Each level consists of several models. The first of the three levels (day pattern) characterizes the choice of each agent in participating in one of the available activities (i.e., education, work, shops, or other). It results in a list of tours and intermediate stop availability for the agents. The tour level on the other hand includes the choices made within a tour, such as the mode choice, the destination, eventual sub-tours, etc. Finally, the intermediate stop level includes two types of discrete choices, mode/destination (for the intermediate stop), and time of day. Preday consists of 22 behavioral models overall, which are described in detail in the *Supplementary Information*[1] and in [19].

Besides, while the choice process is carried out individually for each agent, SimMobility Preday considers household data through β s which are dedicated to framing the impacts of the household structure on the choices [71]. Examples are β malenone$_{\text{edu}}$ which weights the number of education trips of a single man with no children or β femaleage515$_{\text{edu}}$, weighting the number of education trips in a family with a female child (5 to 15 years old). The structure of SimMobility [71] guarantees that when drafting the activity schedule, these interactions are accounted for through different probability distributions, allowing to frame trips to school in which an adult on their way to work drives their child to school first. It does so by extracting the household structure from the population database and then using the relevant β s (e.g., β malenone and β femaleage515) at each level to define how many chained trips are carried out. It should be explicitly stated, though, that SimMobility Preday considers household features through probabilities and not constraints, so it does not completely frame intra-household constraints. For this specific case study, a synthetic population was designed in a previous work [73], in which relevant information (e.g., the number of private vehicles in a household, age and gender distribution) was used to build a realistic database for SimMobility. Finally, it should be highlighted how a specific structure for the nested logit is not by itself a fundamental requirement of the developed methodology, so the method may be easily transferable to similar activity-based models (e.g., CT-RAMP [76]). Still, to guarantee this transferability, further work should be carried out by applying the proposed methodology to other tools, to ensure the lack of unforeseen barriers.

In the following, the formulation of utility for the binary choice to perform an activity or not (top level of the nested logit tree) and the mode choice (bottom level of the nested logit tree) are reported (11), (12) to highlight the high number of behavioral parameters involved (and, crucially, to be calibrated).

$$U_{\text{binary}} = f\Big(V_{\text{case\_study}}, \underline{\beta}_{\text{female\_travel}}, \underline{\beta}_{\text{age\_category}}, \underline{\beta}_{\text{children\_in\_household}}, \underline{\beta}_{\text{income}}, \underline{\beta}_{\text{missing\_income}}, \underline{\beta}_{\text{work\_at\_home}}, \underline{\beta}_{\text{number\_of\_cars\_in\_household}}, \underline{\beta}_{\text{dptour\_logsum}}, \underline{\beta}_{\text{employment\_status}}\Big) \tag{11}$$

1. https://github.com/smart-fm/simmobility-prod/wiki/Mid-Term-Parameters

TABLE 1. The daily activity schedule for individual 107.

person id	tour no	tourType	stop no	stop type	stop location	stop mode	primary stop	arrival time	departure time	prev stop location	prev stop departure time
107-1	1	Education	1	Education	166	BusTravel	True	7.25	13.75	569	7.25
107-1	1	Education	2	Home	569	BusTravel	False	13.75	15.75	166	13.75
107-1	2	Work	1	Work	415	BusTravel	True	15.75	18.25	569	15.75
107-1	2	Work	2	Home	569	BusTravel	False	18.25	19.25	415	18.25
107-1	3	Shop	1	Shop	496	Walk	True	19.75	23.25	569	19.25
107-1	3	Shop	2	Home	569	Walk	False	23.75	26.75	496	23.25

Note that each variable in (11) is a vector of β s, including as many behavioral variables as categories considered. For example, $\underline{\beta}$ for age category is a vector with 5 β s, since 5 are the age categories considered. Overall, binary alone includes 25 β s to be calibrated.

On the other side of the tree (Mode/Destination choice), the utility related to the bus mode is computed as:

$$U_{\text{bus}} = f\Big(V_{\text{case\_study}}, \underline{\beta}_{\text{cons\_bus}}, \underline{\beta}_{\text{tt}}, \underline{\beta}_{\text{walk\_time}}, \underline{\beta}_{\text{wait\_time}}, \underline{\beta}_{\text{cost}}, \underline{\beta}_{\text{cost\_over\_income}}, \underline{\beta}_{\text{central\_district}}, \underline{\beta}_{\text{transfer}}, \underline{\beta}_{\text{female\_num\_of\_cars}}, \underline{\beta}_{\text{number\_of\_cars\_in\_hh}}, \underline{\beta}_{\text{agecat\_num\_of\_cars}}\Big) \tag{12}$$

Also in this case, the above is a simplified version for presentation purposes and the number of β s required to compute U_{bus} is 18. When all the modes and levels of the nested logit tree are considered (Figure 3), 477 is the total number of behavioral parameters (β s) characterizing the calibration problem. The whole set of β s represents the θ parameters described in Section III, they constitute the search space explored by the algorithm. The full list of β s is publicly available.[2]

The calibration is carried out against the following baseline data:

1) OD matrix at subdistrict level. Tallinn has 82 subdistricts and the movements across them at each time of the day are extracted and upscaled from a mobility survey obtained from Taltech University.
2) Statistical margins concerning workplace distributions and totals at the cell level (500x500m). The methodology behind these is detailed in [73]. A similar method, albeit simplified, was applied for the margins concerning school institutions.
3) Modal share, as detailed in [?].

These represent the observations Y_o introduced in (2) and the achieved match is detailed in the calibration and results section.

V. CALIBRATION AND RESULTS

A. THE OBJECTIVE FUNCTION

The design of the calibration process with the proposed methodology features a) a custom objective function, b) an iteration process for the calibration runs, and c) the definition of hyperparameters for both optimization methods, i.e., the global objective and the inner acquisition function.

Formally, the calibration of a simulation model is an optimization task that aims to minimize the discrepancy between a set of simulated and observed outputs. The discrepancy is measured via an objective function and, within this study, we compile a custom function that comprises three components: (i) Origin-Destination (OD) matrix; (ii) share of transportation modes; and (iii) coverage of the employed individuals with scheduled work tours. All three components are adjusted to result in similar ranges and as such share equal contributions to the final objective function:

$$\epsilon_{\text{global}} = \epsilon_{\text{od}} \cdot \epsilon_{\mathcal{M}} \cdot \epsilon_{\text{workers}} \tag{13}$$

$$\epsilon_{\text{od}} = \frac{1}{100}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(od_{ij} - \tilde{od}_{ij}\right)^2} \tag{14}$$

$$\epsilon_{\mathcal{M}} = 1 + \sqrt{\sum_{i\in\mathcal{M}}\left(\mathcal{M}_i - \tilde{\mathcal{M}}_i\right)^2} \tag{15}$$

$$\mathcal{M} = \{\text{public, car, walk, other}\} \tag{16}$$

$$\epsilon_{\text{workers}} = 2 - w_{\text{assign}}/w_{\text{total}}, \tag{17}$$

where n is the number of districts covered within the OD matrix, od_{ij} and $\tilde{od}_{ij}$ are numbers of observed and simulated tour legs for a given element of the OD matrix, respectively, where i is the origin and j is the destination. $\mathcal{M}_i$ and $\tilde{\mathcal{M}}_i$ are shares of observed and simulated transportation modes, respectively, whereby, for both, holds $\sum_{i\in\mathcal{M}}\mathcal{M}_i = 1$ and $\sum_{i\in\mathcal{M}}\tilde{\mathcal{M}}_i = 1$. Regarding the workers component in Eq. (17) (i.e., the ratio between the number of assigned workers in the daily activity schedule and the number of workers within the whole dataset), w_{assign} corresponds to the number of employed individuals with scheduled at least one work-based tour, and w_{total} is the total number of employed adults in the population. The choice of the elements to include in the discrepancy function was guided by the calibration objectives. The function guides the algorithm towards a solution with a small discrepancy, which means that a modeler may want to prioritize the error elements that are considered to be most important to their model/analysis. In this work, the first two items in Equation (17) were chosen because of the importance of the represented measure to the calibration quality, while the last item was chosen after noticing a higher than acceptable error in the number of work trips, also in sets of simulations converging to small errors. The authors' hypothesis is that this bias arose from a conflict between the general OD and the specific spatial distribution

2. https://github.com/Angelo3452/Tallinn-Synthetic-Population/tree/main/SimMobility%20MT%20Database/Postgres%20database

of work-related trips. Without an explicit mention in the discrepancy function, the developed algorithm was fulfilling its objective in the most efficient way but failed to properly frame the work category of trips. Also in other case studies, the discrepancy function should be monitored in the initial runs and adapted accordingly to properly frame all the important dimensions of the problem at hand. Finally, it should be highlighted that the presented methodology would still be applicable in cases where a modeler may prefer prioritizing other measurements to assess the calibration performance (e.g., the spatial distribution of leisure trips rather than work ones).

B. THE ITERATIVE ALGORITHM

To consider the stochasticity of the Bayesian optimization and produce stable results, the calibration process is repeated multiple times with the initial dataset (five in our experiments, each with a different random seed and random initialization of the parameters). The starting values of the behavioral parameters have been set within a realistic range but this preadjustment does not amount to a pre-calibration, as the results from the first simulation reported in Figs. 4 and 5 show. The reported results are summarized across all runs, with the selected optimal solution (run no. 2) outperforming all other runs, where the difference in the results is due to the varying random seeds and initial parameters across the different runs.

Each independent run is performed with the same hyperparameters for the optimization methods. A summary of all hyperparameters used in this study is presented in Table 2. The selection of the hyperparameter values is done via trial-and-error and their transferability depends on the size of the modeled area. This, in particular, is important for the termination conditions for both Bayesian optimization and L-BFGS-B gradient-based optimization methods, where the convergence is affected by the size and the underlying complexity of the modeled domain, i.e., urban area. However, the given values are good starting point for further adjustment. The hyperparameters that reflect the choices over particular methods for initial sampling and acquisition optimization are transferable as-is to new studies.

In Fig. 4, the progression of the performance measure for 5 sets of iterations is reported.

In order to investigate broader aspects of the solutions and confirm their robustness, we examine the Pareto front of all potential solutions through six additional criteria: i) the modal share ($\mathcal{M}_i - \tilde{\mathcal{M}}_i < 10\%$), the spatial distribution of ii) work and iii) education trips, iv) the absolute number of workers, the v) total legs, and vi) the spatially distributed legs (note that the last two items are directly derived from the OD matrix). Hence, the defined cost function guides the algorithm in its optimization process, but the final results are analyzed and a selection is made against a broader set of criteria. The post analysis confirms that run number 2 outperforms the others after slightly more than 150 iterations and reaches a performance value $\epsilon_{\text{global}} \approx 1.06$.

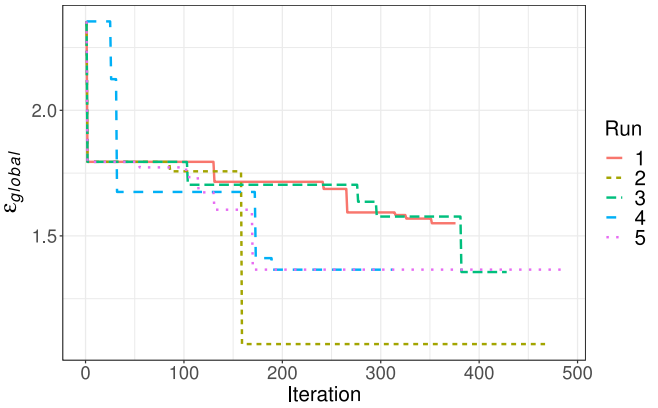


FIGURE 4. Performance measure progression across 500 iterations for 5 different runs.

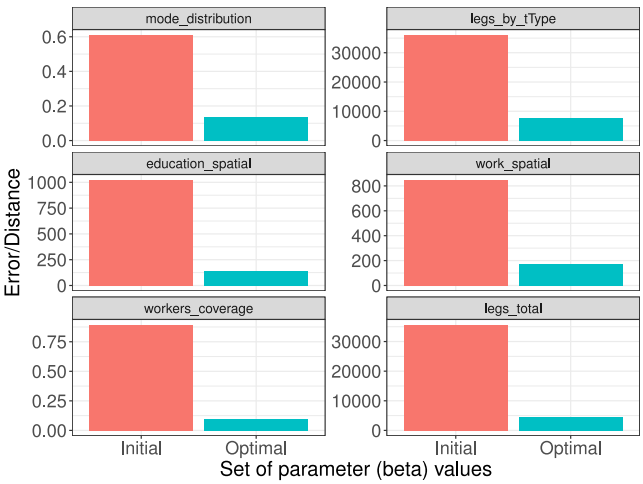


FIGURE 5. Comparison between the starting point (1) and the best simulation in run 2 (182) – benchmarks against the baseline and residual error; the y-axis represents the percentage for worker_coverage and mode distribution, and the number of legs for all the other benchmarks.

C. CALIBRATION AGAINST THE BASELINE

The results reported in the following are the ones arising from run 2. The overall improvement is evident in Figs. 4 and 5. Fig. 5 shows the best-performing simulation in run 2, namely 182. It is compared with the initial simulation which results instead in 0 satisfactory benchmarks and high errors across all the measurements. The second and third benchmarks (legs_by_rType and educational spatial, not perfectly met in 182) reflect instead absolute quantities and, while the latter is quite small and can be considered a match, the former will be further commented on in the following. We would like to stress that these six criteria differ from the metrics selected for the performance measure. In fact, while the latter guides the algorithm and the learning process, the former ones are exploited to filter the best simulations after each iteration.

While Fig. 5 and Table 3 report only the error in the total number of legs, their distribution across the 82 subdistricts of Tallinn has also been compared against the baseline (i.e., the mobility survey). The algorithm is remarkably able to reproduce correct spatial distributions as in Fig. 6.

TABLE 2. Hyperparameters used for the algorithms utilized in this study. For the hyperparameters not included in this table, default values are used.

Algorithm	Hyperparameter	Value(s)
Bayesian optimization	Termination conditions	500 iterations or 500 hours
Bayesian optimization	Initial sampling	Latin Hypercube Sampling
Bayesian optimization	Retrain surrogate model	every 5 iterations
Bayesian optimization	Acquisition optimization	L-BFGS-B [68], [69]
Random forests	Number of trees	1000, 3000, 5000
L-BFGS-B	Termination conditions	1000 iterations
L-BFGS-B	Step sizes for approximation to gradient	0.5

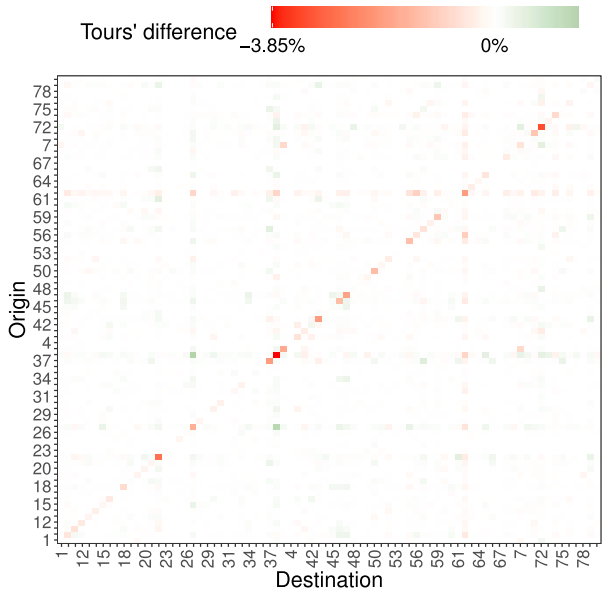


FIGURE 6. Difference between the number of trips for each OD pair, calculated between the simulated ones and the total obtained instead by upscaling the mobility survey to the whole population. The axis labels include every third district.

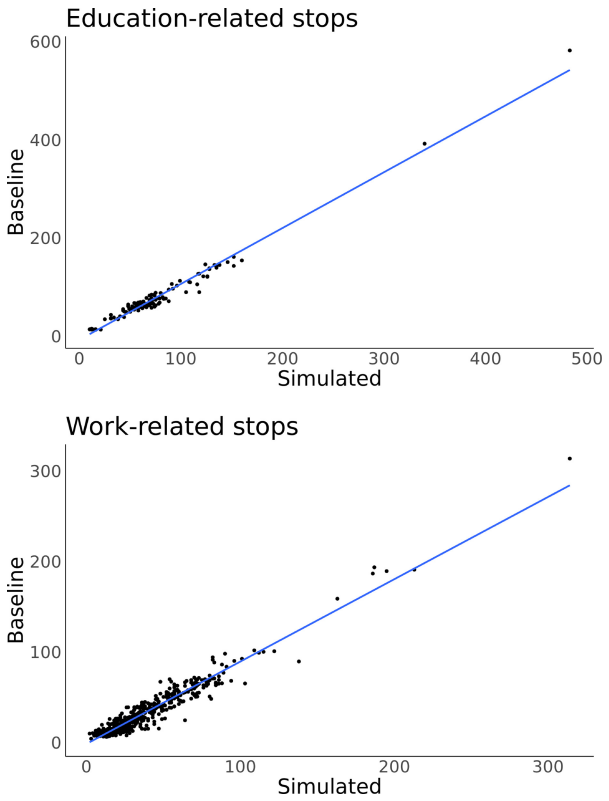


FIGURE 7. Comparison of attendances at anchor points for education-related and work-related reasons. x-axis: number of trips to the anchor point in the calibrated simulation; y-axis: baseline number of trips to the anchor point from the baseline.

It should be stressed that these matrices cover a period of 24 hours. This means that, as in Fig. 6, the absolute error of private vehicle trips outside the diagonal averages 15.92 vehicles. If one considers that the magnitude of simulated trips by private vehicles settles around 50,000 trips (each simulation runs over 10% of the population), the match between the two matrices is impressively close. The diagonal does instead show a somewhat higher variance (albeit still within a reasonable margin, an absolute average of 267.82 trips) and all the cells appear to be underestimated in the number of intrazonal trips. Still, intrazonal trips are commonly more difficult to frame, so a higher error was expected.

The algorithm succeeds also in framing the tour types and spatial distribution of work and education trips, somehow trickier because mandatory, thus subject to stricter correspondence against the benchmark. Table 3 reports a comparison of the totals while Fig. 7 reports the spatial distributions of school and work trips.

As can be seen, the algorithm faithfully frames both education and totals, while slightly overestimating work tours. The overestimation has been considered less of a problem than an underestimation since it was considered mandatory that all the work trips would be framed (every employed person goes to the workplace). An overestimation of work trips does instead mistake only the aim of a tour, this is probably caused by the "emphasis" that the performance measure puts on simulating all the employees going to their workplace.

The distribution of work and education destinations modeled has been checked against the workplaces and education anchor points previously assigned to each eligible individual in the synthetic population [73]. Fig. 7 shows a good match between the two, which implies that the calibrated utility parameters do indeed result in the right number of trips and spatial distribution for these two categories. Modal share also reaches a reasonable level of precision against the share recorded in [77], as shown in Fig. 8.

TABLE 3. Numerical comparison between the best-performing simulation 182 in run 2 and the baseline.

Variable	Category	Simulated	Baseline	Difference	Percentage Error
legs [abs]	Education	19259	19563	-304	1.55
legs [abs]	Total	117093	112481	4612	4.10
legs [abs]	Work	44521	36788	7733	21.02
mode [%]	car	0.420	0.488	-0.069	-
mode [%]	other	0.014	0.012	0.003	-
mode [%]	PT	0.366	0.256	0.110	-
mode [%]	walk	0.2	0.239	-0.039	-

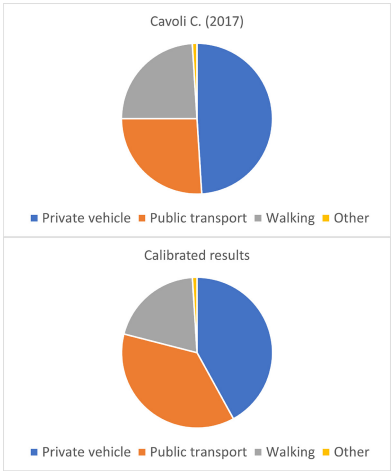


FIGURE 8. Baseline modal share from [77] (top) and modal share arising from the calibrated simulation (bottom).

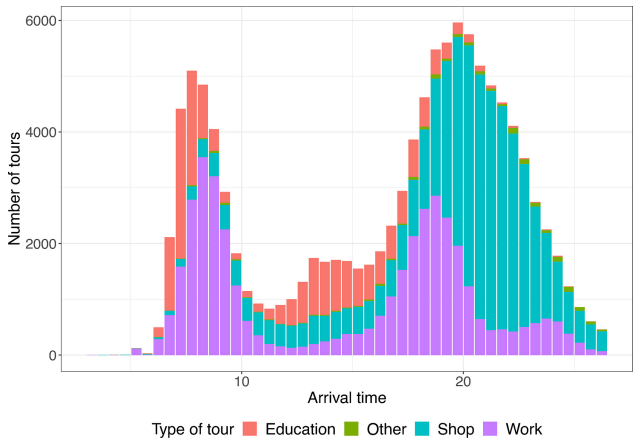


FIGURE 9. Temporal distribution of tours throughout the day for the best-performing simulation 182 in run 2.

Another important result of the calibration is the time distribution of the stops in each tour throughout the day. This is reported in Fig. 9. As mentioned, the case study has 2015 as the reference year (ante COVID-19 pandemic). This means that the usual travel pattern showing two peaks (one in the morning and one in the afternoon) was to be expected and is coherent with the case study. Besides, the model clearly captures the different dynamics such as an education peak in the early afternoon or a spike in leisure trips in the evening (after work).

While most of the tours do fall in a realistic pattern, the calibrated behavioral parameters result in a small percentage of tours (4% percent) allocated in the last available time slot. This is bound to how SimMobility Preday models the time of the day, allocating at the very end all the tours that could not be fitted through the day.

To assess the effectiveness of the proposed method, the only possible comparison is against an achieved manual calibration of the same case study since, as shown in Section II, no other method can currently be adapted to account for such a large number of parameters. Nevertheless, the average absolute error in the manual calibration is equal to 8.5% across non-null OD pairs outside the diagonal (against a 6.7% for the BO method). Yet, the overall number of trips falls well short of the baseline one with only 984108 trips simulated (and thus an error of 12% against the 4% achieved through BO). It appears how, during the manual calibration, the OD matrix distribution was prioritized to the detriment of the total trips, as not all the metrics could be satisfied at once. Most importantly, the manual calibration did not address such a large number of parameters, considering not more than 25 $\sim$ 30 β s. This means that the manual calibration fails to capture all the behavioral factors contributing to each choice (a subset of which is reported in Fig. 10 for the BO). The Daily Activity Schedule resulting from the manual calibration has been uploaded in the available repository[2]. The comparison has been done in % terms, as the manual calibration was carried out on 33% of the population rather than 10%.

D. ANALYSIS OF THE CALIBRATED PARAMETERS

The set of results provided in this section and their comparison with baseline values should clarify how the algorithm reaches an acceptable solution in a completely automated way. It is important to stress how this calibration process differs from the traditional one for SimMobility Preday, which is carried out manually by tuning the various parameters,[3] improving its results. Besides, in the literature review reported above, it was highlighted how other, more complex methods, do not encompass as many β s as the proposed

3. https://github.com/smart-fm/simmobility-prod/wiki/Mid-term-calibration

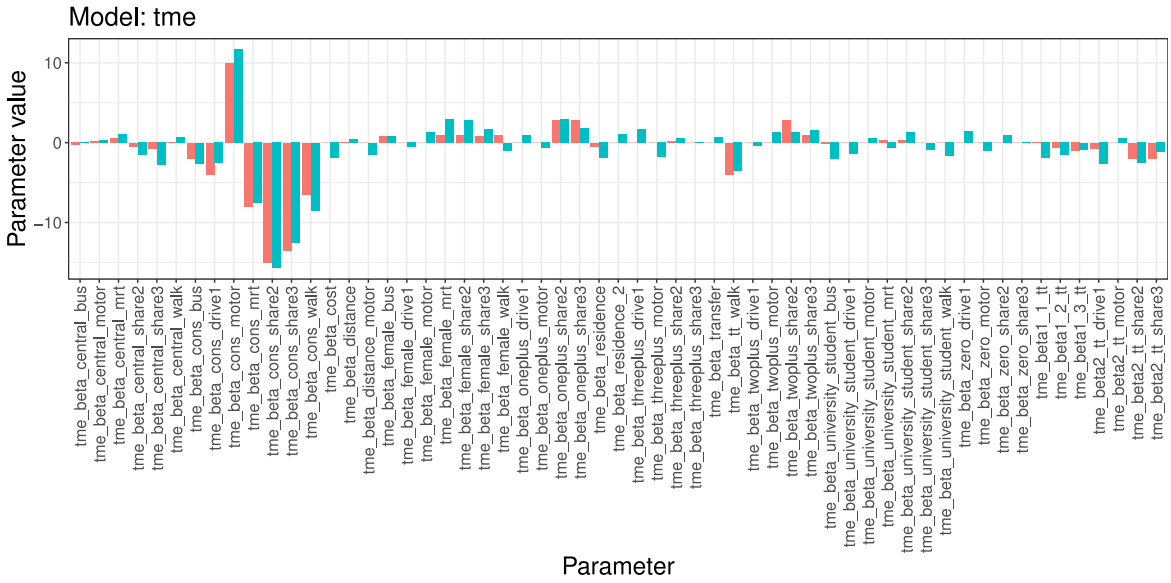


FIGURE 10. Comparison of the parameters – starting (red) and final (green) values – in the logit branch modeling the mode choice for trips to school or other educational institutions.

algorithm. Fig. 10 provides a snapshot of the behavioral parameters and their original and final values for one level of the nested logit tree. For the complete list of β s, please refer to the publicly available repository[2]. One fundamental aspect should be highlighted: the algorithm allows to expand the set of parameters that are calibrated, as it appears in the tme branch of the logit tree (the branch addressing modal choice for education) plot in Fig. 10, where many parameters have non-null values only for the calibrated results. This is because the starting values (manually defined) could not encompass such a large set of behavioral parameters that were then set to 0.

VI. DISCUSSION AND CONCLUSION

The paper presented a new algorithm to calibrate a large number of parameters by exploiting a surrogate model and BO techniques, applied to a real-life case study to prove the effectiveness of the proposed method in calibrating hundreds of behavioral parameters for an activity-based model. The result shows a satisfactory match between the modeled outputs and the baseline, built from an available mobility survey and aggregate data. By calibrating the model through the presented algorithm, it was possible to tune a wider set of behavioral parameters than it would have been manually or through heuristics. As shown in the literature review, no other work succeeded in calibrating as many as 477 parameters, although avoiding doing so would strongly reduce the effectiveness of a nested logit model detailed enough to consider different socio-demographic features. The algorithm searches for the best-performing solution by perturbing all 477 behavioral parameters (β s through the paper). This is a task that could hardly be performed by hand.

By automating the process and exploiting a surrogate model, the algorithm bypasses the need to set up and run the activity-based model to test each plausible combination of parameters. The proposed approach requires instead one run of SimMobility at each iteration, followed by multiple runs of the surrogate model testing the candidate combinations of parameters. Running one iteration of SimMobility and surrogate model for all the candidate combinations took approximately 20-30 min on a Triton high-performance computing cluster part of the Finnish Grid and Cloud Infrastructure.[4] The whole set of 500 iterations lasted around 6 days and was run in parallel for the 5 runs reported in Figure 4. The proposed algorithm may be used in other scenarios and case studies, improving the feasibility of large-scale activity-based modeling rooted in behavioral science, thus fostering the number of similar studies/tools. The database, software, and the resulting behavioral parameters are available as open-source[2]. Furthermore, the applicability of the proposed algorithm is not limited to activity-based modeling and transportation problems, greatly increasing its applicability.

Finally, it is still worth noting that the study has some limitations that may be addressed in future research directions. Since the calibration has been carried out against aggregate benchmarks (e.g., the modal share of the whole population), future developments may strive to apply the calibration algorithm to a more disaggregate set of data (calibrating, for example, modal share for the type of tour or type of individual) reducing local discrepancies in the modal share. Moreover, the empirically quantified uncertainty, i.e., standard deviation, of the Random forests base predictions tends to collapse to 0 in the regions of the space that are distant from the observed points. This implies similar predictions by all base models and hence inaccurately estimated uncertainty.

4. https://scicomp.aalto.fi/triton/overview/

Such phenomena may exhibit greater visibility at the initial stage of the BO, when the space is very scarcely sampled, thus future works may allow limiting the exploration to the distant regions of the parameter space.

ACKNOWLEDGMENT

The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project. The authors would like to thank Dago Antov from TalTech for sharing the travel survey exploited in this work.

REFERENCES

[1] D. Sha, K. Ozbay, and Y. Ding, "Applying Bayesian optimization for calibration of transportation simulation models," *Transp. Res. Rec.*, vol. 2674, no. 10, pp. 215–228, Aug. 2020. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0361198120936252

[2] United nations department of economic and social affairs. "The world's cities in 2018." 2018. [Online]. Available: https://digitallibrary.un.org/record/3799524

[3] D. Schrank, B. Eisele, and T. Lomax. "2019 urban mobility report." 2019. [Online]. Available: https://static.tti.tamu.edu/tti.tamu.edu/documents/umr/archive/mobility-report-2019.pdf

[4] D. Schrank, L. Albert, B. Eisele and T. Lomax. "2021 urban mobility report." 2021. [Online]. Available: https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-report-2021.pdf

[5] S. A. Agriesti, R. Soe, and M. A. Saif, "Framework for connecting the mobility challenges in low density areas to smart mobility solutions: The case study of Estonian municipalities," *Eur. Transp. Res. Rev.*, vol. 14, p. 32, Jul. 2022.

[6] E. Thonhofer et al., "Infrastructure-based digital twins for cooperative, connected, automated driving and smart road services," *IEEE Open J. Intell. Transp. Syst.*, vol.4, pp. 311–324, 2023.

[7] G. O. Kagho, M. Balac, and K. W. Axhausen, "Agent-based models in transport planning: Current state, issues, and expectations," *Procedia Comput. Sci.*, vol. 170, pp. 726–732, Jan. 2020.

[8] P. M. Bösch, F. Ciari, and K. W. Axhausen, "Transport policy optimization with autonomous vehicles," *Transp. Res. Rec.*, vol. 2672, no. 8, pp. 698–707, 2018. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0361198118791391

[9] B. Nahmias-Biran, J. B. Oke, N. Kumar, C. L. Azevedo, and M. Ben-Akiva, "Evaluating the impacts of shared automated mobility on-demand services: An activity-based accessibility approach," *Transportation*, vol. 48, pp. 1613–1638, Aug. 2021.

[10] K. A. Marczuk et al., "Autonomous mobility on demand in SimMobility: Case study of the central business district in Singapore," in *Proc. IEEE 7th Int. Conf. Cybern. Intell. Syst. (CIS) IEEE Conf. Robot. Autom. Mechatron. (RAM)*, 2015, pp. 167–172.

[11] A. Araldo et al., "Implementation & policy applications of AMOD in multi-modal activity-driven agent-based urban simulator SimMobility," in *Proc. Annu. Meeting Transp. Res. Board*, Feb. 2018, pp. 1–20, [Online]. Available: https://www.researchgate.net/profile/Arun-Akkinepally/publication/323150735_Implementation_Policy_Applications_Of_AMOD_In_Multi-_Activity-Driven_Agent-Based_Urban_Simulator_Simmobility/links/5a82fb5545851504fb37d1eb/Implementation-Policy-Applications-Of-AMOD-In-Multi-Modal-Activity-Driven-Agent-Based-Urban-Simulator-Simmobility.pdf

[12] C. Llorca, A. Moren, and R. Moeckel, "Effects of shared autonomous vehicles on the level of service in the greater Munich metropolitan area," in *Proc. Int. Conf. Intell. Transp. Syst. Theory Pract.*, 2017. [Online]. Available: https://scholar.google.com/scholar?hl=it&as_sdt=0%2C5&q=EFFECTS+of+shared+autonomous+vehicles+on+the+level+of+service+in+the+greater+Munich+metropolitan+area&btnG=

[13] B.-H. Nahmias-Biran, G. Dadashev, and Y. Levi, "Demand exploration of automated mobility on-demand services using an innovative simulation tool," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 580–591, 2022.

[14] V. A. Cruz. "An activity-based modeling approach to assess the effects of activity-travel behavior changes and in-home activities on mobility." 2021. [Online]. Available: https://repository.tudelft.nl/islandora/object/uuid:414ca108–d79b-485b-87d1–75386742a43b

[15] S. Chen, A. A. Prakash, C. L. Azevedo, and M. Ben-Akiva, "Formulation and solution approach for calibrating activity-based travel demand model-system via microsimulation," *Transp. Res. Part C*, vol. 119, Oct. 2020, Art. no. 102650.

[16] H. J. Miller, "Activity-based analysis," in *Handbook of Regional Science*. Berlin, Germany: Springer, 2021, pp. 187–207. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-3-642-36203-3_106-1#citeas

[17] B. Shahriari, K. Swersky, Z. Wang, P. R. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[18] M. U. Gutmann and J. Corander, "Bayesian optimization for likelihood-free inference of simulator-based statistical models," *J. Mach. Learn. Res.*, vol. 17, no. 125, pp. 1–47, 2016.

[19] L. Siyu. "Activity-based travel demand model: Application and innovation." 2015. [Online]. Available: https://scholarbank.nus.edu.sg/handle/10635/121998

[20] S. Oh, R. Seshadri, C. L. Azevedo, and M. E. Ben-Akiva, "Demand calibration of multimodal microscopic traffic simulation using weighted discrete SPSA," *Transp. Res. Rec.*, vol. 2673, no. 5, Apr. 2019, Art. no. 36119811984210. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0361198119842107

[21] L. Lu, Y. Xu, C. Antoniou, and M. E. Ben-Akiva, "An enhanced SPSA algorithm for the calibration of dynamic traffic assignment models," *Transp. Res. Part C Emerg. Technol.*, vol. 51, pp. 149–166, Feb. 2015.

[22] M. Qurashi, T. Ma, E. Chaniotakis, and C. Antoniou, "PC–SPSA: Employing dimensionality reduction to limit SPSA search noise in DTA model calibration," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1635–1645, Apr. 2020.

[23] B. Kostic, G. Gentile, and C. Antoniou, "Techniques for improving the effectiveness of the SPSA algorithm in dynamic demand calibration," in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Naples, Italy, 2017, pp. 368–373.

[24] B. Y. He et al., "A validated multi-agent simulation test bed to evaluate congestion pricing policies on population segments by time-of-day in New York City," *Transp. Policy*, vol. 101, pp. 145–161, Feb. 2021.

[25] A. Bassolas, J. J. Ramasco, R. Herranz, and O. G. Cantú-Ros, "Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona," *Transp. Res. Part A Policy Pract.*, vol. 121, pp. 56–74, Mar. 2019.

[26] L. Schultz and V. Sokolov, "Bayesian optimization for transportation simulators," *Procedia Comput. Sci.*, vol. 130, pp. 973–978, Jan. 2018.

[27] L. Schultz and V. Sokolov. "Practical bayesian optimization for transportation simulators." 2018. [Online]. Available: arXiv:1810.03688.

[28] T. Siripirote, A. Sumalee, H. W. Ho, and W. H. K. Lam, "Statistical approach for activity-based model calibration based on plate scanning and traffic counts data," *Transp. Res. Part B*, vol. 78, pp. 280–300, Aug. 2015.

[29] D. Ziemke, I. Kaddoura, and K. Nagel, "The MATSim open Berlin scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data," *Procedia Comput. Sci.*, vol. 151, pp. 870–877, Jan. 2019.

[30] D. Ziemke, K. Nagel, and C. Bhat, "Integrating CEMDAP and MATSIM to increase the transferability of transport demand models," *Transp. Res. Rec.*, vol. 2493, no. 1, pp. 117–125, Apr. 2019. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3141/2493-13

[31] M. Cools, E. Moons, and G. Wets, "Calibrating activity-based models with external origin-destination information," *Transp. Res. Rec.*, vol. 2175, no. 1, pp. 98–110, Jan. 2010. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3141/2175-12

[32] C. R. Bhat, J. Guo, S. Srinivasan, and A. Sivakumar. "CEMDAP user's manual." 2003. [Online]. Available: https://www.caee.utexas.edu/prof/bhat/REPORTS/CEMDAPUserManual_4080.pdf

[33] S. Hörl, M. Balac, and K. W. Axhausen, "A first look at bridging discrete choice modeling and agent-based microsimulation in MATSim," in *Proc. 7th Int. Workshop Agent-based Mobil. Traffic Transp. Models Methodol. Appl. (ABMTrans)*, 2018, pp. 900–907.

[34] S. Hörl, M. Balać, and K. W. Axhausen, "Pairing discrete mode choice models and agent-based transport simulation with MATSim," in *Proc. 98th Annu. Meeting Transp. Res. Board (TRB)*, 2019, pp. 1–20.

[35] G. Flötteröd, Y. Chen, and K. Nagel, "Behavioral calibration and analysis of a large-scale travel microsimulation," *Netw. Spat. Econ.*, vol. 12, pp. 481–502, Dec. 2012.

[36] A. Horni, K. Nagel, and K. W. Axhausen, *The Multi-Agent Transport Simulation MATSim*. London, U.K.: Ubiquity Press, 2016.

[37] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and efficient hyperparameter optimization at scale," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1437–1446.

[38] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated Machine Learning*. Cham, Switzerland: Springer, 2019, pp. 3–33. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-05318-5_1#citeas

[39] M. Järvenpää, M. U. Michael, A. Vehtari, and P. Marttinen, "Gaussian process modelling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria," *Ann. Appl. Statist.*, vol. 12, no. 4, pp. 2228–2251, 2018.

[40] J. Lintusaari, M. Gutmann, R. Dutta, S. Kaski, and J. Corander, "Fundamentals and recent developments in approximate Bayesian computation," *Syst Biol*, vol. 66, pp. e66–e82, Jan. 2017.

[41] M. Todorović, M. Gutmann, J. Corander, and P. Rinke, "Bayesian inference of atomistic structure in functional materials," *NPJ Comput. Mater.*, vol. 5, p. 35, Mar. 2019.

[42] Y. Zhang, D. W. Apley, and W. Chen, "Bayesian optimization for materials design with mixed quantitative and qualitative variables," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020.

[43] F. Leclercq, "Bayesian optimization for likelihood-free cosmological inference," *Phys. Rev. D*, vol. 98, no. 6, 2018, Art. no. 63511.

[44] A. Aushev, H. Pesonen, M. Heinonen, J. Corander, and S. Kaski. "Likelihood-free inference with deep Gaussian processes." 2020. [Online]. Available: arXiv:2006.10571.

[45] D. Nott, Y. Fan, L. Marshall, and S. A. Sisson, "Approximate Bayesian computation and Bayes' linear analysis: Toward high-dimensional ABC," *J. Comput. Graph. Statist.*, vol. 23, no. 1, pp. 65–86, 2014.

[46] R. Izbicki, A. B. Lee, and T. Pospisil, "ABC–CDE: Toward approximate Bayesian computation with complex high-dimensional data and limited simulations," *J. Comput. Graph. Statist.*, vol. 28, no. 3, pp. 481–492, 2019.

[47] L. Raynal, J. Marin, P. Pudlo, M. Ribatet, C. Robert, and A. Estoup, "ABC random forests for Bayesian parameter inference," *Bioinformatics*, vol. 35, no. 10, pp. 1720–1728, 2019.

[48] Z. Cheng, X. Wang, X. Chen, M. Trepanier, and L. Sun, "Bayesian calibration of traffic flow fundamental diagrams using Gaussian processes," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 763–771, 2022.

[49] V. Kuzmanovski and J. Hollmén, "Composite surrogate for likelihood-free Bayesian optimisation in high-dimensional settings of activity-based transportation models," in *Proc. Int. Symp. Intell. Data Anal.*, 2021, pp. 171–183.

[50] M. Blum, M. Nunes, D. Prangle, and S. A. Sisson, "Comparative review of dimension reduction methods in approximate Bayesian computation," *Statist. Sci*, vol. 28, no. 2, pp. 189–208, 2013.

[51] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. De Freitas, "Bayesian optimization in high dimensions via random embeddings," in *Proc. IJCAI*, 2013, pp. 1778–1784.

[52] Z. Wang, C. Gehring, P. Kohli, and S. Jegelka, "Batched large-scale Bayesian optimization in high-dimensional spaces," in *Proc. Int. Conf. Artif. Intell.d Statist.*, 2018, pp. 745–754.

[53] C. Oh, E. Gavves, and M. Welling, "BOCK: Bayesian optimization with cylindrical kernels," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3868–3877.

[54] M. Mutny and A. Krause, "Efficient high dimensional Bayesian optimization with additivity and quadrature fourier features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9019–9030.

[55] A. Nayebi, A. Munteanu, and M. Poloczek, "A framework for Bayesian optimization in embedded subspaces," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4752–4761.

[56] J. Kirschner, M. Mutny, N. Hiller, R. Ischebeck, and A. Krause, "Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3429–3438.

[57] R. B. Gramacy, *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Boca Raton, FL, USA: CRC Press, 2020.

[58] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[59] H.-H. Chen, Y.-B. Lin, I.-H. Yeh, H.-J. Cho, and Y.-J. Wu, "Prediction of queue dissipation time for mixed traffic flows with deep learning," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 267–277, 2022.

[60] Y. Zheng, Q. Wang, D. Zhuang, S. Wang, and J. Zhao. "Fairness-enhancing deep learning for ride-hailing demand prediction," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 551–569, 2023.

[61] C. Li, K. Kandasamy, B. Póczos, and J. Schneider, "High dimensional Bayesian optimization via restricted projection pursuit models," in *Proc. Artif. Intell. Statist.*, 2016, pp. 884–892.

[62] J. Gardner, C. Guo, K. Weinberger, R. Garnett, and R. Grosse, "Discovering and exploiting additive structure for Bayesian optimization," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1311–1319.

[63] J. Djolonga, A. Krause, and V. Cevher, "High-dimensional Gaussian process bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1025–1033.

[64] C. Li, S. Gupta, S. Rana, V. Nguyen, S. Venkatesh, and A. Shilton. "High dimensional Bayesian optimization using dropout." 2018. [Online]. Available: arXiv:1802.05400.

[65] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[66] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proc. Int. Conf. Learn. Intell. Optim.*, 2011, pp. 507–523.

[67] J. H. Friedman and P. Hall, "On bagging and nonlinear estimation," *J. Statist. Plan. Infer.*, vol. 137, no. 3, pp. 669–683, 2007.

[68] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[69] J. Nocedal and J. Stephen, *Numerical Optimization*. New York, NY, USA: Springer, 1999. [Online]. Available: https://link.springer.com/book/10.1007/b98874

[70] J. E. Stiglitz, "Pareto optimality and competition," *J. Finan.*, vol. 36, no. 2, pp. 235–251, 1981.

[71] Y. Lu et al., "SimMobility mid-term simulator: A state of the art integrated agent based demand and supply model," in *Proc. 94th Annu. Meeting Transp. Res. Board (TRB)*, 2015, pp. 1–17.

[72] M. Young. "OpenTripPlanner–creating and querying your own multi-modal route planner." 2021. [Online]. Available: https://github.com/marcusyoung/otp-tutorial

[73] S. Agriesti, C. Roncoli, and B. Nahmias-Biran, "Assignment of a synthetic population for activity-based modeling employing publicly available data," *Int. J. Geo-Inf.*, vol. 11, no. 2, p. 148, 2022.

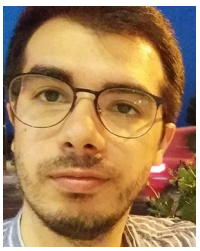
[74] T. Viegas de Lima, M. Danaf, A. Akkinepally, C. De Azevedo, and M. Ben-Akiva, "Modeling framework and implementation of activity- and agent-based simulation: An application to the greater Boston Area," *Transp. Res. Rec.*, vol. 2672, no. 49, pp. 146–157, 2018. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0361198118798970

[75] J. B. Oke et al., "A novel global urban typology framework for sustainable mobility futures," *Environ. Res. Lett.*, vol. 14, p. 9, Sep. 2019.

[76] P. Vovsha, J. Freedman, V. Livshits, and W. Sun, "Design features of activity-based models in practice," *Transp. Res. Rec.*, vol. 2254, no. 1, pp. 19–27, 2011. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3141/2254-03

[77] C. Cavoli. "CREATE–city report tallinn, estonia." 2017. [Online]. Available: http://www.create-mobility.eu/create/resources/general/download/CITY-REPORT-Tallinn-WSWE-AV3MMA

[78] J. Močkus, "On Bayesian methods for seeking the extremum," in *Proc. Optim. Techn. IFIP Tech. Conf.*, 1975, pp. 400–404.



SERIO AGRIESTI is currently pursuing the Ph.D. degree with Aalto University. Before starting his doctoral studies, he was a Research Fellow with the Politecnico di Milano, where he focused on the impact assessment of innovative transport systems, such as connected and automated vehicles and truck platooning. He has been involved in multiple European research projects and is part of the EIT Urban Mobility Doctoral Training Network. His current research activities focus on agent-based modeling, performance evaluation, and connected and automated driving.

VLADIMIR KUZMANOVSKI received the Ph.D. degree from International School Jožef Stefan, Slovenia, in 2016. He is a Postdoctoral Researcher of Machine Learning and Artificial Intelligence with Aalto University, Finland, and a Visiting Research Assistant with Jožef Stefan Institute, Slovenia. His research interests include global optimization in high-dimensional spaces, surrogate modeling, probabilistic methods, and modeling of complex systems.

JAAKKO HOLLMÉN (Senior Member, IEEE) is a Senior Lecturer with the Department of Computer and Systems Sciences, Stockholm University, Sweden, and a Senior University Lecturer with the Department of Computer Science, Aalto University, Finland. His research interests include machine learning and data mining, and applications within health and medicine as well as environmental informatics.

CLAUDIO RONCOLI received the Ph.D. degree from the University of Genova, Italy, in 2013. He is an Associate Professor of Transportation Engineering with Aalto University, Finland. Before joining Aalto University, he was a Research Assistant with the University of Genova, a Visiting Research Assistant with Imperial College London, U.K., and a Postdoctoral Researcher with the Technical University of Crete, Greece. He has been involved in several national and international research projects as a principal investigator. His research interests include real-time traffic management; modeling, optimization, and control of traffic systems with connected and automated vehicles; as well as smart mobility and intelligent transportation systems.

BAT-HEN NAHMIAS-BIRAN (Member, IEEE) received the B.Sc. degree in civil and environmental engineering, the M.Sc. degree in transportation engineering, and the Ph.D. degree in transportation systems from Technion in 2008, 2011, and 2016, respectively. She is a Senior Lecturer of Transportation Engineering with Ariel University, Israel, and the Head of Future Mobility Lab. She is also a Research Affiliate with the Intelligent Transportation Systems Lab, Massachusetts Institute of Technology. Prior to joining Ariel University, she was a Postdoctoral Associate with the SMART, Future Urban Mobility Lab, MIT. Her main research interests are the modelling and evaluation of new technologies in transportation, simulation of shared and automated mobility, activity-based modelling, machine learning capabilities for transportation, and transport equity.