Transportation Research Part C 98 (2019) 186–206

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Dynamic origin-destination matrix calibration for large-scale network simulators

Carolina Osorio

Civil and Environmental Engineering Department, Massachusetts Institute of Technology, Office 1-232, Cambridge, MA 02139, USA

ABSTRACT

This paper considers offline dynamic OD (origin-destination) calibration problems for large-scale simulation-based network models. We formulate the problem as a simulation-based optimization (SO) problem and propose a scalable and efficient metamodel SO algorithm. For a network with n links and a problem with T time intervals and Z OD pairs per time interval, the corresponding SO problem is a high-dimensional problem of dimension $T{\cdot}Z$. At each iteration of the SO algorithm, we solve a set of T independent analytical and differentiable metamodel optimization problems, each of dimension Z. The T analytical problems are constrained by n nonlinear equality constraints. Hence, they scale linearly with the number of links in the network and independently of other network attributes, such as the dimension of the route choice set or the link lengths. The temporal correlation, of the link performance metrics across time intervals, is approximately captured through the parameters of the metamodel. Since the T analytical optimization problems are decoupled and can be solved independently, the proposed approach scales independently of the number of calibration time intervals, making it suitable for the calibration of demand over numerous time periods.

The approach is efficient and scalable. It is suitable to address high-dimensional calibration problems for large-scale network models. It is benchmarked on both a toy network and a Singapore network versus two general-purpose algorithms: SPSA and a derivative-free pattern search algorithm. The validation experiments indicate that the proposed method identifies points with objective function estimates that outperform the benchmark methods by 1 to 2 orders of magnitude. As the problem dimension and the temporal correlation, across time intervals, increase, so does the magnitude by which the proposed method outperforms the benchmark methods. The Singapore case study considers a problem of dimension 16,200. This is 2 orders of magnitude higher than what is currently considered high-dimensional for continuous SO problems. The proposed method identifies solutions with an estimated objective function that is 1 order of magnitude better than those of the benchmark methods. It yields solutions that provide an average, across time intervals, improvement to link counts of 77%, compared to the benchmark methods. The case study also illustrates the robustness of the method to the quality of the initial points.

1. Introduction

There is an increased interest among both private and public urban transportation stakeholders to develop and use traffic models to inform the design of their urban mobility services. Additionally, as the resolution (i.e., granularity) of urban mobility data increases, so does that of the corresponding models. This leads to reduced computational efficiency. There is a pressing need for computationally efficient algorithms that enable the calibration of these higher resolution, yet inefficient, models. The calibration problems faced by practitioners are difficult optimization problems. They are high-dimensional, simulation-based and non-convex problems. Hence, the design of efficient calibration algorithms is challenging.

We consider the calibration of stochastic simulation-based traffic models. The general goal of a model calibration problem is to calibrate (i.e., fit) model inputs that minimize a distance function between estimates of network conditions (e.g., link flows, link speeds) provided by the simulator and those obtained from field measurements. We consider a specific type of calibration problem known as the dynamic OD calibration problem, which calibrates a set of high-dimensional time-dependent origin-destination (OD) matrices. The latter are model inputs that describe spatial and temporal travel demand patterns.

E-mail address: osorioc@mit.edu.

https://doi.org/10.1016/j.trc.2018.09.023
Received 21 April 2018; Received in revised form 24 September 2018; Accepted 24 September 2018
Available online 03 December 2018

We distinguish between two classes of problems. The first is the OD calibration problem, where the goal is to calibrate the demand inputs of a specific traffic model. The output is a calibrated model, yet the calibrated demand (i.e., set of OD matrices) is not intended to be used as stand-alone travel demand estimate. The second is the more general OD estimation problem, where the goal is to estimate travel demand for a given region and the outputs are the OD matrices themselves. These matrices are then used for a variety of transportation planning studies. We focus on the first class of problems: our goal is to calibrate the OD matrices of a specific traffic model.

This paper focuses on the design of dynamic OD calibration algorithms that are: (i) computationally efficient, such that they can address the needs of transportation practitioners by identifying solutions with good performance within tight computational budgets (e.g., few simulation evaluations), and (ii) scalable, such that they can address high-dimensional problems for large-scale networks.

Reviews of OD calibration literature are included in Balakrishna (2006), Djukic (2014), Zhang and Osorio (2017). The most common approach to simulation-based OD calibration has been the use of general-purpose algorithms, such as Simultaneous Perturbation Stochastic Approximation (SPSA) (Balakrishna et al., 2007; Vaze et al., 2009; Lee and Ozbay, 2009; Ben-Akiva et al., 2012) and metaheuristics, such as the genetic algorithm (GA) (Kim et al., 2001; Stathopoulos and Tsekeris, 2004; Kattan and Abdulhai, 2006; Vaze et al., 2009).

These general-purpose algorithms are designed for a broad class of optimization problems. In particular, their use is not limited to calibration problems, let alone to transportation optimization problems. They are designed based on asymptotic properties (e.g., asymptotic convergence guarantees). Little is theoretically known about their short-term performance (i.e., performance under tight computational budgets). Yet there is recent interest in the theoretical and empirical analysis of their short-term performance (Dong et al., 2017).

These general-purpose algorithms lack computational efficiency because they are not designed to identify good quality solutions within tight, or small, computational budgets. Nonetheless, in the transportation calibration community, they are most commonly used under tight budgets. Djukic (2014, page 33) discusses the lack of computational efficiency of both SPSA and of metaheuristics for OD calibration problems. Discussions of the advantages and limitations of SPSA for transportation problems are also presented in Antoniou et al. (2015), Tympakianaki et al. (2015), Tympakianaki (2018). Extending these general-purpose algorithms such as to enhance their efficiency is an active area of research. Recent extensions to SPSA include Cipriani et al. (2011), Lu et al. (2015), Antoniou et al. (2015), Tympakianaki et al. (2015). The design of efficient calibration algorithms has also been driven by the need for real-time feasible algorithms (e.g., Ashok and Ben-Akiva, 2002; Bierlaire and Crittin, 2004; Barceló and Montero, 2015; Zhou and Mahmassani, 2007).

The design of simulation-based optimization (SO) algorithms suitable for high-dimensional problems is also a challenge. In the field of continuous SO, problems with dimension in the order of 200 are considered high-dimensional (Wang et al., 2016), while OD calibration problems have a dimension in the order of thousands or tens of thousands of variables. Reviews and discussions of the design of scalable OD calibration methods include Djukic (2014), Prakash et al. (2018). A spatial network decomposition approach for offline calibration was proposed by Frederix et al. (2014). Research on scalable techniques has been mostly driven by online calibration problems, where the most common approach is the use of dimensionality reduction techniques (Djukic et al., 2012; Prakash et al., 2017, 2018).

General-purpose algorithms exploit limited, or no, problem-specific structural information. There is an opportunity to improve their scalability as well as their short-term performance by embedding problem-specific information and tailoring them to specific classes of transportation problems. This line of thinking has led to some of the extensions of SPSA (Lu et al., 2015; Tympakianaki et al., 2015). In this paper, we pursue this line of thinking.

We propose a computationally efficient and scalable algorithm for dynamic OD calibration problems. To achieve efficiency, we formulate and embed analytical problem-specific information within a general-purpose SO algorithm. More specifically, we use an analytical approximation of the simulation-based objective function. This approximation is known as a metamodel. The general concept of metamodel SO is detailed in Section 2.2. The metamodel we use embeds information from an analytical network model. The latter provides a problem-specific approximation of how the inputs (OD matrices) relate to the simulation-based functions (expected link flows). The proposed approach is scalable. The formulation scales linearly with the number of links in the network, making it suitable for large-scale networks. It scales independently of the number of calibration time intervals, making it suitable for the calibration of demand over numerous time periods.

The efficiency and scalability of the proposed approach are evaluated by applying it to a Singapore case study (Section 4). The network considers over 1000 links and over 18,000 routes. The calibration problem is of dimension 16,200, this is 2 orders of magnitude larger than what is currently considered high-dimensional for continuous SO problems. We benchmark the approach versus two traditional algorithms: SPSA and a derivative-free pattern search algorithm. The proposed approach yields objective function estimates that are one order of magnitude lower than those of the benchmark methods. It improves fit to link counts by an average of 77%, compared to benchmark methods. This illustrates its scalability and its efficiency.

Recently, we formulated a metamodel idea for a time-independent calibration problem (i.e., a single OD matrix for a single time interval is calibrated for a dynamic stochastic traffic simulator) (Osorio, 2017). In this paper, we extend this static method and propose a method suitable for dynamic OD estimation, where multiple time-dependent OD matrices are calibrated for a dynamic stochastic traffic simulator. The main challenge is to formulate an analytical metamodel that accounts for the temporal dependencies in the dynamic problem, yet is sufficiently scalable and efficient so that the overall SO algorithm remains scalable and efficient. The distinctions between the proposed method and the static method are detailed in Section 2.4.

Section 2 presents the proposed methodology. The latter is validated for toy networks in Section 3 and applied to a Singapore network case study in Section 4. The main conclusions are summarized in Section 5.

2. Methodology

2.1. Formulation of the dynamic OD calibration problem

A traditional approach to modeling the spatial and temporal distribution of travel demand is through the use of origin-destination (OD) matrices. A city or region is segmented into zones. The OD matrix describes the expected demand between each pair of zones. To account for time-varying travel demand, one OD matrix is specified for each of T non-overlapping equal-length time intervals (e.g., 8–9 am, 9–10 am). A dynamic OD calibration problem aims to calibrate one OD matrix for each of the T time intervals, such that the simulated network performance (e.g., expected link counts or speeds) is similar to the field measurements (e.g., average link counts or speeds). We consider a problem that simultaneously (i.e., jointly) determines the OD matrices for all T time intervals. We consider an offline problem, which is solved offline based on historical traffic data.

We introduce the following notation.

d_{zt}	expected travel demand for OD pair z and departure time interval t (scalar);
d	decision vector composed of the scalars d_{zt};
f	simulation-based objective function;
F_{it}	flow on link i and time interval t as defined by the simulator;
y_{it}	average flow on link i and time interval t estimated from field data;
$\widetilde{d}_{zt}$	prior value for the expected demand for OD pair z and departure time interval t (scalar);
u_1	vector of endogenous simulation variables;
u_2	vector of exogenous simulation parameters;
δ	weight parameter for prior information (scalar);
$d^{\max}$	upper bound vector;
Z	number of OD pairs per time interval;
T	number of time intervals;
$\mathcal{I}$	set of indices of links with sensors;
$\mathcal{T}$	set of time interval indices $\mathcal{T} = \{1,2,..., \mathrm{T}\}$;
$\mathcal{Z}$	set of OD pair indices $\mathcal{Z} = \{1,2,..., \mathrm{Z}\}$.

The dynamic OD calibration problem is formulated as follows.

$$\min_{d} f(d) = \frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}}(y_{it} - E[F_{it}(d, u_1;u_2)])^2 + \delta\frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{Z}|}\sum_{t\in\mathcal{T}}\sum_{z\in\mathcal{Z}}(\widetilde{d}_{zt} - d_{zt})^2 \tag{1}$$

$$0 \leqslant d \leqslant d^{\max}, \tag{2}$$

where for a given set $\mathcal{I}$, $\mathcal{T}$ or $\mathcal{Z}$, the notation $|\mathcal{I}|$, $|\mathcal{T}|$ or $|\mathcal{Z}|$, denotes the cardinality of the set. The decision vector consists of OD values for each OD pair z and each departure time interval t. It is of dimension $T{\cdot}Z$. In the case study of Section 4, $T = 4$ and $Z = 4050$, leading to a high-dimensional problem of dimension 16200. Note that for $T = 1$, Problem (1)–(2) is equivalent to the time-independent problem addressed in Osorio (2017, Problem (1)–(2)).

The first summation of Eq. (1) represents the distance between the field measurements of link traffic conditions (terms y_{it}) and their corresponding simulation-based functions (terms $E[F_{it}(d, u_1;u_2)]$). This summation aims to estimate OD demand such that the simulator leads to similar traffic conditions than those observed in the field. The set of links with field measurements, $\mathcal{I}$, is usually a low dimensional set. Often, less than 10% of the links have field measurements.

The OD calibration problem is, for a realistic network, an underdetermined problem (i.e., there is an observability issue). Hence, there are an infinite set of OD vectors, d, that yield the same simulated traffic conditions. Among this set some solutions are more physically plausible than others. For example, some may be consistent with the underlying land-use of the region leading to plausible travel demand patterns for a specific time period. To address this underdetermination issue, the traditional approach is to include in the objective function a term which leads to OD solutions that are close to a plausible pre-defined OD matrix. This is the role of the second summation of Eq. (1). The pre-defined OD matrix, denoted $\widetilde{d}$, is known as a prior or seed matrix. It is most often estimated from other data sources such as census data or obtained through OD estimation of a simpler (e.g., static) traffic model. It is assumed to capture plausible travel demand patterns for the considered calibration time period. The scalar δ is a weight factor that defines the trade-off between the distance to field measurements (first summation) and the distance to the prior OD matrix (second summation).

The most common field measurements, y_{it}, used for calibration are link counts, followed by link speeds. With more granular mobility data becoming increasingly available, field measurements based on partial trajectory data are being used. Nonetheless, for the majority of urban areas, link counts remain the most commonly available and used type of calibration data. In this paper, we focus on the use of link counts, i.e., y_{it} represents the average flow on link i during time interval t. The simulation-based functions $E[F_{it}(d, u_1;u_2)]$ represent the expected link flow on link i during time interval t. The vector u_1 represents endogenous variables of the simulator (e.g., link travel times, densities, queue-lengths). The vector u_2 represents exogenous parameters of the simulator such as the network topology, supply parameters (e.g., traffic management strategies such as signal plans or congestion pricing schemes) and exogenous link or lane attributes (e.g., length, maximum speed).

The problem is formulated with bound constraints Eq. (2). In summary, the calibration problem consists of a high-dimensional SO problem with a nonlinear, most often non-convex, non-differentiable, simulation-based objective function with analytical bound constraints. This problem inherits the usual challenge of an SO problem: the objective function is not available analytically, instead it

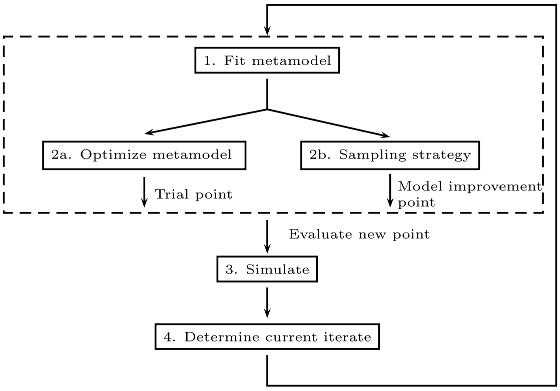


Fig. 1. General metamodel framework.

can only be estimated via simulation. For calibration problems, two additional challenges arise. First, the problem is high-dimensional. In the field of continuous SO, problems of dimension in the order of 200 are considered high-dimensional. The dimension of calibration problems are often 2–3 orders of magnitude higher. Second, traffic simulators are particularly costly to evaluate. In transportation, depending on the simulation model and on the type of performance metric to estimate, runtimes can range from hours to days. For instance, one evaluation of the Singapore network of Section 4 typically requires 4 h on a standard machine. This high computational cost combined with the high dimensionality of the problem highlights the need for algorithms that are computationally efficient and can derive good quality solutions within few simulation evaluations.

2.2. Metamodel framework

The general idea of a metamodel algorithm is to address the SO problem by solving a set of analytical optimization problems, for which traditional and efficient gradient-based algorithms can be used. To do so, the simulation-based objective function is approximated by an analytical parametric function, known as the metamodel. Fig. 1 summarizes the main steps of each iteration of an SO algorithm. Step 1 uses all simulation observations obtained so far to fit the parameters of the metamodel. Step 2a replaces the simulation-based objective function by the fitted metamodel and solves the analytical problem, which is known as the metamodel optimization problem or the subproblem. Step 3 simulates the solution to this subproblem. Step 4 determines the current iterate, which is the point (i.e., the OD value) that is considered to have best performance so far. Step 2b identifies points to simulate which may not be solutions to the subproblem. They may be derived from a general sampling strategy, or be defined such as to improve metamodel fit properties or geometric properties of the sampled space. As the SO iterations increase, so does the number of simulated points, which can then lead to an enhanced fit of the metamodel and ultimately to the identification of points with improved simulated performance.

The methodological challenge lies in the formulation of a suitable metamodel. For the calibration problem, the metamodel should: (i) be analytical and differentiable (such that traditional gradient-based algorithms can be used to solve the subproblem), (ii) be scalable (such that large-scale network problems can be solved), (iii) be efficient (because the high-dimensional subproblem needs to be solved at every iteration of the SO algorithm), and (iv) approximate well the simulation-based objective function in the entire feasible region. The main challenge is providing a formulation that balances both efficiency and accuracy (i.e., factors (iii) and (iv)).

For Problem (1)–(2), the formulation of a metamodel requires analytically approximating the functions $E[F_{it}(d, u_1;u_2)]$. These are intricate functions for the following reasons. Recall that $E[F_{it}(d, u_1;u_2)]$ represents the expected flow on link i during time interval t. This expectation depends on what route each individual traveler has chosen, which typically depends on the travel times across all plausible (or considered) routes. These travel times depend themselves on the route choices of all individuals both during the current and past time intervals. Hence, the function $E[F_{it}(d, u_1;u_2)]$ embeds both spatial and temporal dependency information of traffic conditions. Providing an analytical, relatively accurate yet also efficient approximation of this function is difficult.

Our past work for static OD calibration (Osorio, 2017) proposed a metamodel formulation that satisfied the above criteria. In particular, it achieved efficiency by embedding information from an analytical stationary network model, which approximated the mapping between travel demand (OD vector) and link performance (e.g., expected link flow). The use of a stationary network model led to a tractable and scalable formulation. The main challenge when going from a static (i.e., one time interval) OD calibration problem to a dynamic (i.e., multiple time intervals) OD calibration problem is the temporal correlation of network (e.g., link, path) performance metrics. The analytical network model of Osorio (2017) is a time-independent model. Hence, it does not capture the temporal evolution of congestion nor the temporal dependencies among the functions $E[F_{it}(d, u_1;u_2)]$.

A natural approach to overcome this limitation would be to formulate a time-dependent analytical network model, and then replace the stationary model of Osorio (2017) by the time-dependent model. As detailed above, this would require the formulation of a time-dependent model that remains differentiable, scalable and, most importantly, sufficiently efficient to address large-scale network optimization problems. As the vast literature on dynamic traffic assignment (DTA) shows, this remains an unresolved challenge. In particular, if the analytical network model is not efficient, one is better off allocating the computational resources to simulation evaluations rather than to, inefficiently, solving the approximate subproblem.

In this paper, we take a different approach. We propose to decouple the metamodel optimization problem into a set of T subproblems, one for each time interval. In other words, Step 1 of Fig. 1 fits a set of T metamodels and Step 2a solves T analytical optimization problems. For each time interval, we proceed as in Osorio (2017): we use a metamodel based on a stationary network model. We then capture temporal dependencies across time intervals through the parameters of the metamodel. The paper includes experiments with high temporal correlation across the time intervals. The results indicate that the proposed approach performs well under such conditions.

Hence, the main contribution of this paper is to show that if sufficient spatial correlation (or more generally spatial dependency) is captured analytically (in this case through the use of a time-independent analytical network model), then it can be sufficient to capture temporal correlation in a simple (i.e., more tractable) way, compared to the use of a time-dependent analytical network model. The proposed approach captures temporal correlation in very simple way: the parameters of the metamodel are fitted based on temporally correlated simulation observations. In particular, the metamodel does not rely on any time-dependent assumptions (such as the duration of the time intervals, or the variation of congestion across time). Hence, it can be used as is for any network and any specification of time-intervals. This makes it a broadly applicable approach.

2.3. Formulation of the metamodel optimization problems

At every iteration of the SO algorithm, we propose to solve a set of T subproblems. Let $\mathcal{P}_{kt}$ denote the analytical subproblem for SO iteration k and time interval t. Let m_{kt} denote the metamodel of problem $\mathcal{P}_{kt}$.

To provide an interpretation of the subproblems, we rewrite Problem (1)–(2) as:

$$\min_{d} f(d) = \sum_{t\in\mathcal{T}} \left\{\frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} (y_{it} - E[F_{it}(d,\, u_1;u_2)])^2 + \delta\frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} (\widetilde{d}_{zt} - d_{zt})^2\right\}. \tag{3}$$

We introduce the following notation.

k	iteration of the SO algorithm;
d_t	vector of expected demand for all OD pairs for departure time interval t, i.e., $d_t = (d_{1t}, d_{2t}, \ldots, d_{Zt})$;
m_{kt}	metamodel function for SO iteration k and time interval t;
β_{kt}	parameter vector of metamodel m_{kt}, $\beta_{kt} = (\beta_{kt0}, \beta_{kt1}, \ldots, \beta_{kt(Z+1)})$;
β_{ktj}	scalar element j of the parameter vector β_{kt};
f_t^A	analytical approximation, derived by the analytical traffic model, of the term within the curly brackets of Eq. (3);
$\phi(d;\beta)$	polynomial component of the metamodel with variables d and coefficients β;
h	analytical traffic model.

At iteration k of the SO algorithm, the subproblem for time interval t, $\mathcal{P}_{kt}$, is defined as:

$$\min_{d_t} m_{kt}\left(d_t;\beta_{kt}\right) = \frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{I}|}(\beta_{kt0}f_t^A(d_t) + \phi(d_t;\beta_{kt})) + \delta\frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} (\widetilde{d}_{zt} - d_{zt})^2 \tag{4}$$

$$h(d_t;r) = 0 \tag{5}$$

$$0 \leqslant d_t \leqslant d^{\max}. \tag{6}$$

The metamodel m_{kt} analytically approximates the term within the brackets of (3). This term differs from Eq. (4) only in that it replaces the first summation, which contains the simulation-based functions $E[F_{it}(d,\, u_1;u_2)]$, with an analytical function. The latter is defined as an approximation (denoted f_t^A) provided by an analytical network model that is parametrically corrected for: (i) by scaling by a factor of β_{kt0}, and (ii) by an additive error term $\phi(d_t;\beta_{kt})$. The latter is defined as a linear function in d_t with coefficients β_{kt}:

$$\phi\left(d_t;\beta_{kt}\right) = \beta_{kt1} + \sum_{z=2}^{Z+1} \beta_{ktz} d_{(z-1)t}. \tag{7}$$

The analytical network model is represented by h (Eq. (5)) and is specified as a system of nonlinear equations that depends on the decision vector d_t and on a vector of exogenous parameters denoted r. The exogenous parameters include information such as the network topology, the route choice set and the route choice model parameters. This system of equations (i.e., h function) is specified in Section 2.4.

In metamodel SO, it is common to use general-purpose analytical functions to approximate the simulation-based functions (i.e., to specify the metamodels). Common choices include low-order polynomials, radial-basis functions, Kriging functions (Jones et al., 1998; Barton and Meckesheimer, 2006; Wild et al., 2008; Kleijnen et al., 2010; Ankenman et al., 2010). A review of metamodel approaches appears in Osorio (2010, Chap. 5). The mathematical properties of such general-purpose functions are then used to derive asymptotic properties of the SO algorithms, such as convergence guarantees. Nonetheless, as general-purpose functions they do not embed any problem-specific structural information, and hence they typically fail to identify points with good performance under tight computational budgets (i.e., when few simulation observations are available).

Our past SO work has shown that the use of problem-specific models enables the design of computationally efficient algorithms for both continuous transportation problems (e.g., Osorio and Nanduri, 2015, Chong and and Osorio, 2018, Osorio and Atasoy, 2017) and for discrete transportation problems (Zhou et al., 2017). Moreover, this past body of work shows that the use of problem-specific analytical models enhances the robustness of the SO algorithms to the quality of the initial points and to the stochasticity of the underlying simulators. Thus, for the design of efficient algorithms, the essential building block is the formulation of a suitable function f_t^A. Since we need to solve Problem $\mathcal{P}_{kt}$ at *every* iteration of the SO algorithm and for each time interval t, we require a formulation for f_t^A that is analytical, differentiable, scalable and computationally efficient. Such a formulation is presented in Section 2.4.

Let us discuss the implications of decomposing the metamodel optimization subproblem into a set of subproblems. We no longer solve a single subproblem of dimension $T{\cdot}Z$, instead we solve T decoupled subproblems each of dimension Z. This contributes to the scalability and to the efficiency of the approach. Nonetheless, by solving decoupled problems, we no longer capture the temporal dependencies among the functions $E[F_{it}(d, u_1;u_2)]$. It is essential to capture this information, particularly when there is high temporal dependency among the traffic conditions. This occurs, for instance, when the time intervals are of short duration. To address this limitation, we propose to fit the parameters of m_{kt} using temporally correlated simulation observations. In other words, we use a time-independent metamodel and correct for the transients by fitting the metamodel parameters with simulation observations that embed temporal correlation.

2.4. Analytical network model

We introduce the following notation.

	Endogenous variables of the analytical traffic model:
λ_i	expected hourly demand for link i;
k_i	expected density per lane of link i;
v_i	expected (space-mean) speed per lane of link i;
t_r	expected travel time for route r;
$P(r)$	route choice probability for route r.
	Exogenous parameters of the analytical traffic model:
k_i^{jam}	jam density per lane of link i;
$v_i^{\max}$	maximum speed of link i;
q^{cap}	lane flow capacity;
n_i	number of lanes of link i;
ℓ_i	average lane length of link i;
z_r	toll cost for route r;
θ_1, θ_2	parameters of the route choice model;
α_{1i}, α_{2i}	parameters of the fundamental diagram of link i;
c	scaling parameter common to all links;
$O(r)$	OD pair of route r;
$\mathcal{R}_1(i)$	set of routes that include link i;
$\mathcal{R}_2(s)$	set of routes of OD pair s;
$\mathcal{L}(r)$	set of links of route r.

The expression for the analytical problem-specific approximation (f_t^A) of the simulation-based function is defined by the following system of nonlinear equations.

$$f_t^A(d_t) = \sum_{i\in\mathcal{I}} (y_{it} - \lambda_i)^2 - \sum_{i\in\mathcal{I}} y_{it}^2 \tag{8}$$

$$\lambda_i = \sum_{r\in\mathcal{R}_1(i)} P(r) d_{O(r)t} \tag{9}$$

$$P(r) = \frac{e^{\theta_1 t_r + \theta_2 z_r}}{\sum_{j\in\mathcal{R}_2(O(r))} e^{\theta_1 t_j + \theta_2 z_j}} \tag{10}$$

$$t_r = \sum_{i\in\mathcal{L}(r)} t_i \tag{11}$$

$$t_i = \frac{\ell_i}{v_i} \tag{12}$$

$$v_i = v_i^{\max}\left(1 - \left(\frac{k_i}{k_i^{\text{jam}}}\right)^{\alpha_{1i}}\right)^{\alpha_{2i}} \tag{13}$$

$$k_i = c\frac{k_i^{\text{jam}}}{q^{\text{cap}}}\frac{\lambda_i}{n_i} \tag{14}$$

Eq. (8) defines the analytical approximation f_t^A. Note that the first summation term directly approximates the first summation term of (3). More specifically, the expected (simulation-based) link flow ($E[F_{it}(d, u_1;u_2)]$) is approximated by the expected demand (λ_i) derived by the analytical network model. The second summation term of Eq. (8) was introduced to enhance the metamodel fit, its role is discussed in Appendix A. Note that this second summation does not depend on d. Hence, it will not have an impact on the optimal solution to problem $\mathcal{P}_{kt}$.

The expression for f_t^A is derived by evaluating an analytical traffic model which consists of the System of Eqs. (9)–(14). The latter is represented by the function h in Eq. (5). This analytical traffic model was formulated for a simulation-based toll optimization problem in Osorio and Atasoy (2017). It was used for a static OD calibration problem in Osorio (2017). Let us describe the interpretation of each equation.

Eq. (9) expresses the expected demand for link i as a function of the probability of choosing routes that travel through link i and of the expected OD demand for the OD pairs of those routes. Eq. (10) specifies a route choice model, which is defined as a simple multinomial logit model that depends on the route's expected travel time and toll cost. The use of toll costs is particularly relevant for the Singapore case study of Section 4, which has various tolled links. Eq. (11) expresses the expected route travel time as the sum of the expected link travel times. Eq. (12) defines the expected link travel time as the ratio of the link length and the expected link speed. The latter is defined by the fundamental diagram of Eq. (13). Eq. (14) relates the link's expected density (k_i) to its expected demand (λ_i). It assumes a linear relationship between the link density and the link demand. The proportionality constant depends on the link's jam density (k_i^{jam}), a lane flow capacity term q^{cap}, and a constant c. The latter is fitted based on toy network experiments. For the experiments of this paper, it is set to 1/6.

This analytical traffic model simplifies the demand and supply models of traditional simulation-based traffic models. In all numerical experiments of this paper, we use the mesoscopic dynamic simulator DynaMIT (DYnamic Network Assignment for the Management of Information to Travelers) (Ben-Akiva et al., 2010). The analytical model includes the following simplifications. Arguably, the most important distinction is that the analytical model is a stationary model, while the simulator is a dynamic traffic model. The analytical model assumes that all lanes of a link are homogenous, while the simulator allows for heterogeneous lanes (e.g., lanes may have different lengths, maximum speeds, etc.). Eq. (13) is a differentiable approximation of the non-differentiable fundamental diagram of the simulator. Unlike the simulator, the analytical model assumes that all lanes have common flow capacity q^{cap} and does not account for vehicular spillbacks. The route choice model is also simplified. For instance, the simulator accounts for the population's heterogeneity in value of time (VOT) by assuming a random VOT that is lognormal distributed, while the analytical model considers a unique value of time for the entire population.

Recall from the discussion of Section 2.1 that the functions $E[F_{it}(d, u_1;u_2)]$ are temporally correlated. Hence, when decoupling Problem (1)–(2) into the set of problems $\{\mathcal{P}_{kt}\}_t$, it is important to approximately capture this temporal dependency. We do so by fitting the metamodel parameters using simulation estimates that are spatially and temporally correlated. This is detailed in Appendix A.

Let us summarize the main ideas of our proposed approach. At every iteration k of the SO algorithm, we solve a set of T analytical and decoupled subproblems, $\mathcal{P}_{kt}$ (Eqs. (4)–(6)). Note that once the metamodel parameters (β_{kt} of (4)) are fitted, the T subproblems can be solved in parallel. In other words, Step 2b of Fig. 1 solves T independent problems. The metamodel is extended from that of Osorio (2017). In particular, it is based on the use of a stationary network model which is implemented as a set of n nonlinear equations, for a network with n links. Problem $\mathcal{P}_{kt}$ is therefore an analytical optimization problem with a decision vector of dimension Z, with n nonlinear equality constraints and bound constraints. The dimension of the constraints of $\mathcal{P}_{kt}$ scales linearly with the number of links in the network and, more importantly, it scales independently of the dimension of the route choice set. This makes it particularly scalable and suitable for large-scale networks.

Note that for a given Problem $\mathcal{P}_{kt}$, the analytical network model (defined by (9)–(14)) approximates the expected demand for link

i (term λ_i) by using only the OD information of the same time interval t. In other words, the impact of demand from past time intervals is not accounted for. If the time intervals are short and if demand varies over time, then one would not expect this network model to capture the traffic conditions during a given time interval accurately. More specifically, the network model would not provide a good approximation of the simulation-based expected link flows $E[F_{it}(d, u_1;u_2)]$. Nonetheless, as is presented in Section 3, the metamodel performs well even under such conditions. This indicates that the simple parametric correction of the network model (recall from Eq. (4) that it is corrected for with a scaling factor and an additive linear error term) is sufficient to capture the temporal correlations of network-wide traffic conditions.

Our past work on OD calibration (Zhang and Osorio, 2017; Osorio, 2017) has proposed metamodel methods for problems with a single time interval ($T = 1$). This paper presents the first metamodel formulation suitable for a dynamic OD calibration problem (i.e., where $T > 1$). Additionally, the proposed formulation differs from that of past work in the following ways. The metamodel of Zhang and Osorio (2017) formulates one metamodel for each link with a sensor, while we propose a single metamodel regardless of the number of links with sensors. This leads to fewer metamodel parameters to fit, and overall increased computational efficiency of the method. The analytical traffic model of Zhang and Osorio (2017) consists of a system of linear (instead of nonlinear) equations, it is therefore more tractable. Nonetheless, it assumes exogenous assignment, while the traffic model used here accounts for endogenous assignment. For networks with intricate traffic dynamics, such as networks with dynamic tolls, like the Singapore network of Section 4, a formulation with endogenous assignment may be necessary.

In this paper, we extended the metamodel formulation of the static OD calibration method of Osorio (2017) to enable a set of time-dependent OD's to be calibrated. The proposed formulation solves one metamodel optimization problem for each time interval, while our past work considered a single metamodel optimization problem. We use the same functional form of the metamodel and the same stationary network model as in Osorio (2017). The second summation term of function f_t^A (see Eq. (8)) is not used in Osorio (2017). The introduction of this term led us to formulate and solve a different least squares problem to fit the metamodel parameters. The least squares problem we address is given in Appendix A.

3. Validation

We validate the proposed method on a toy network. We compare its performance to that of two benchmark methods: the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm (Spall, 1992, 2003) and a derivative-free generalized pattern search algorithm (GPS) (Mathworks, 2016). The parameters of each benchmark algorithm are set based on standard guidelines (Spall, 2003, Chap. 7). Additionally, for SPSA, we use values determined by MIT's ITS (Intelligent Transportation Systems) Lab members, which combine both: (i) general guidelines developed from using SPSA for various calibration problems (Balakrishna, 2006; Vaze et al., 2009) and (ii) specific SPSA values determined from past calibration efforts of the same Singapore network model (Lu et al., 2015) as that used in Section 4 of this paper.

We consider the synthetic toy network of Astarita et al. (2001) displayed in Fig. 2. It has 3 origin nodes (labeled o_1, o_2, o_3) and 3 destination nodes (labeled d_1, d_2, d_3). For each time interval, we consider an OD matrix with 9 entries ($Z = 9$), i.e., all origin-destination combinations are feasible. The network considers a multi-lane highway (top links) with on-ramps and off-ramps and a single-lane arterial (bottom horizontal links). There is a total of 28 one-way links and 43 lanes. We assume that all links have sensors. Note that even when all links have sensors, there are, for each time interval, multiple OD matrices that can lead to the same link counts (i.e., the problem is underdetermined). For all numerical experiments of this paper, we use the mesoscopic dynamic simulator DynaMIT (DYnamic Network Assignment for the Management of Information to Travelers) (Ben-Akiva et al., 2010).

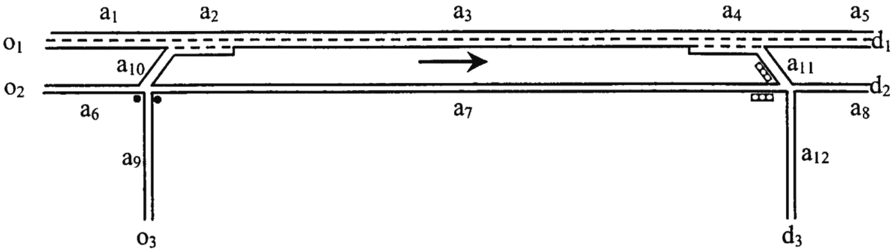


Fig. 2. Synthetic toy network proposed by Astarita et al. (2001).

We consider two sets of experiments. The first set evaluates how the performance of the methods varies with the dimension of the problem and with the temporal correlation of the link performance metrics across time intervals. The second set evaluates the impact on performance of transient versus stationary traffic conditions.

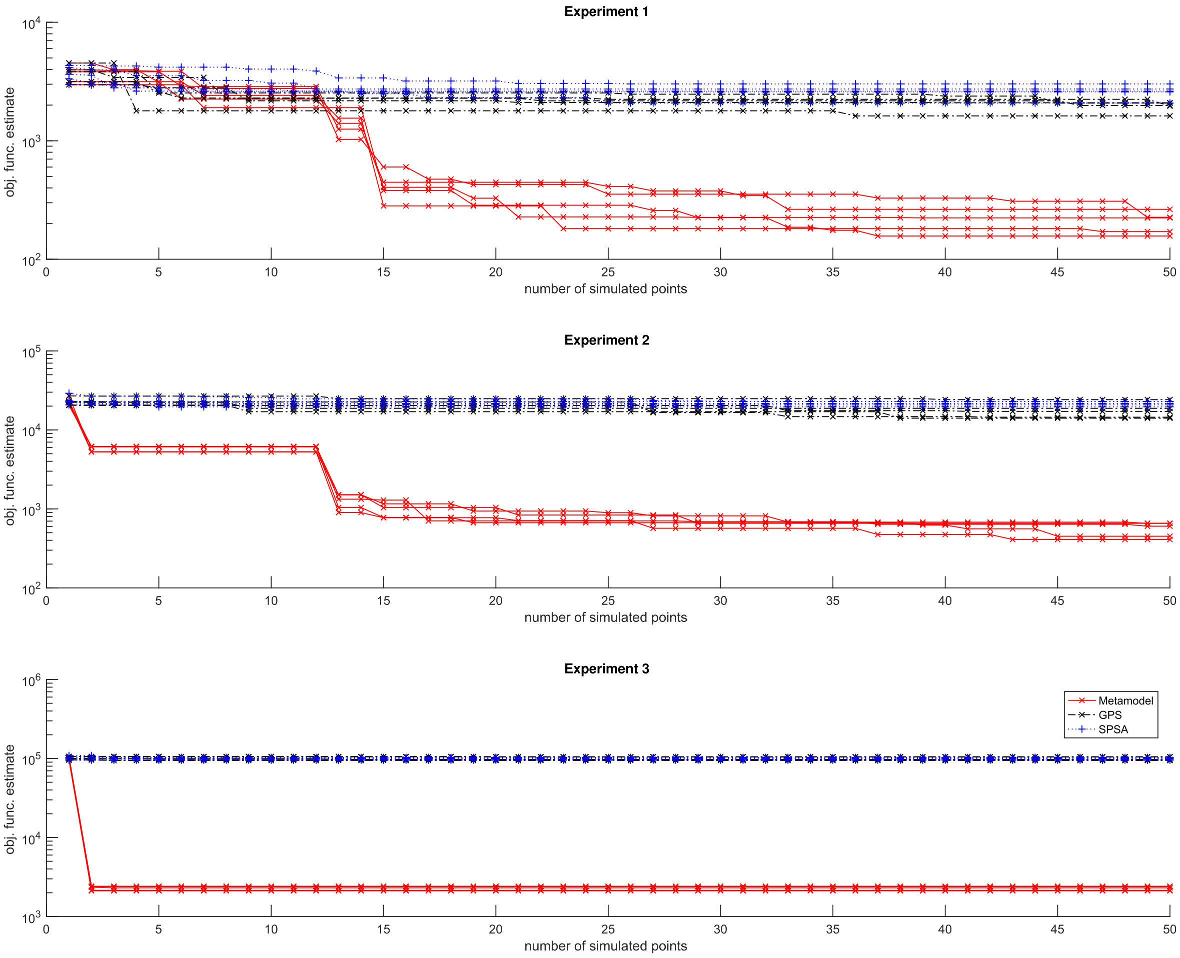


Fig. 3. Performance estimate of the current iterate of each method as a function of the number of simulated points. From top to bottom, the plots consider a higher-temporal resolution with the total number of time intervals, T, set to 5, 10 and 15, respectively.

In the first set of experiments, we consider a demand scenario that transitions from uncongested to congested and back to uncongested conditions. More specifically, let D_1 denote a given level of demand. We consider a 5 h time period, the expected true demand for each OD pair and for each of the 5 h is D_1, $1.5D_1$, $2D_1$, $1.5D_1$ and D_1, respectively. For this demand scenario, we consider three experiments. The first calibrates one OD matrix for every hour (i.e., it considers five 1-h time periods). The second (resp. third) calibrates an OD matrix for every 30 (resp. 15) minutes, i.e., it considers ten 30-min (resp. twenty 15-min) time periods. For experiments 1, 2 and 3, we have $T = 5$, $T = 10$ and $T = 20$, respectively. For a given problem with T calibration time-intervals and Z OD entries to calibrate per time interval, the dimension of the decision vector is $T{\cdot}Z$. Hence, experiments 1, 2 and 3 are of dimension 45, 90 and 180, respectively. As the duration of the time intervals decreases (i.e., as the number of time intervals increases), the temporal dependency between the simulation-based performance metrics (terms $E[F_{it}(d, u_1;u_2)]$ of Eq. (1) or of Eq. (3)) across different time intervals increases. Hence, for the proposed metamodel method, these experiments also evaluate the impact of decomposing the metamodel calibration problem into decoupled subproblems and of capturing limited temporal dependency across time intervals.

Based on the true demand, we estimate, via simulation, synthetic link counts. We consider these counts to be the "true" field counts (i.e., terms y_{it} of Eq. (1)). The prior OD matrix (term $\tilde{d}$ of Eq. (1)) is obtained as the sum of the corresponding true OD value and a randomly drawn normal distributed error term with an expectation of zero and a standard deviation of 20% of the true OD value. Hereafter, when performing optimization, we present results obtained by using one simulation replication to estimate the simulation-based functions. Experiments based on multiple replications led to the same trends across methods than those for a single replication. Nonetheless, the use of multiple replications may be necessary for other case studies.

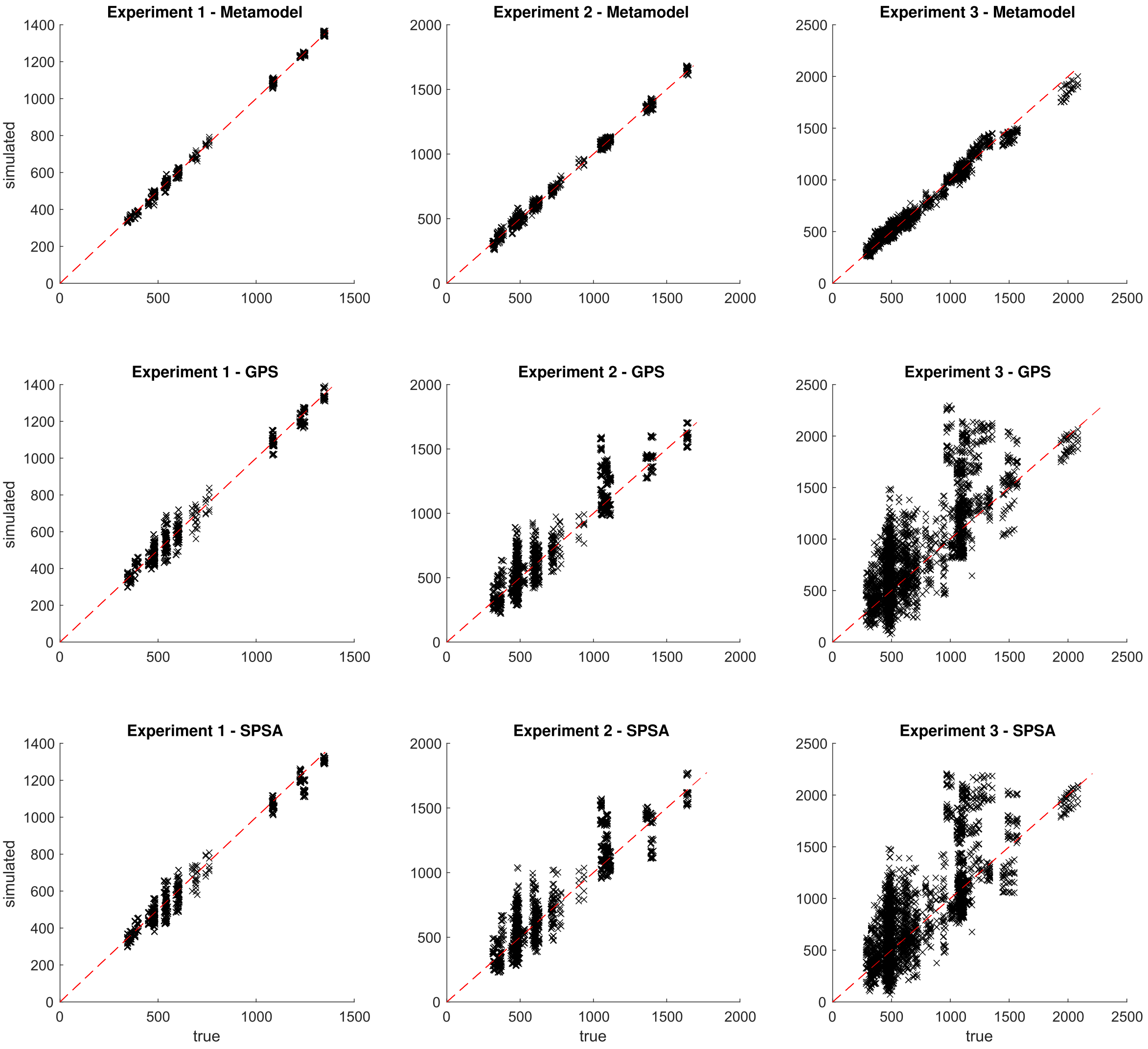


Fig. 4. Fit to link counts of the solutions proposed by each method. Experiments 1, 2 and 3 correspond, respectively, to the left-most, middle and right-most column of plots. The top, middle and bottom row of plots correspond, respectively, to the plots of the metamodel, GPS and SPSA.

For each experiment, we consider 5 initial points that are uniformly and randomly drawn from the feasible region (Eq. (2)). For each experiment, we initialize each method with each of the 5 initial points. This leads to 5 runs of each method for each experiment. We terminate an algorithmic run once a total of 50 points have been simulated. That is, the simulation, or computational, budget is set to 50. The top, middle and bottom plots of Fig. 3 consider, respectively, the first, second and third experiment. The 5 runs of the proposed metamodel method are displayed as a red solid red line. Those of the GPS (resp. SPSA) method are displayed as a dash-dotted black line (resp. dotted blue line). For each plot, the x-axis displays the number of points simulated and the y-axis displays the objective function estimate of the current iterate (i.e., of the best point identified so far by the method). Note that the y-axis has a logarithmic scale.

The top plot indicates that the performance of all three methods is similar for the first 10–15 simulation points. Thereafter, GPS and SPSA have similar performance, while the metamodel identifies points with one order of magnitude improvement in the objective function. For the middle plot, GPS and SPSA have similar performance. They are both outperformed by the metamodel. The latter outperforms by 1 order of magnitude as of the second simulation and by 2 orders of magnitude as of the 13th simulation. For the bottom plot, GPS and SPSA have similar performance and are both outperformed by the metamodel by one order of magnitude as of the second simulation.

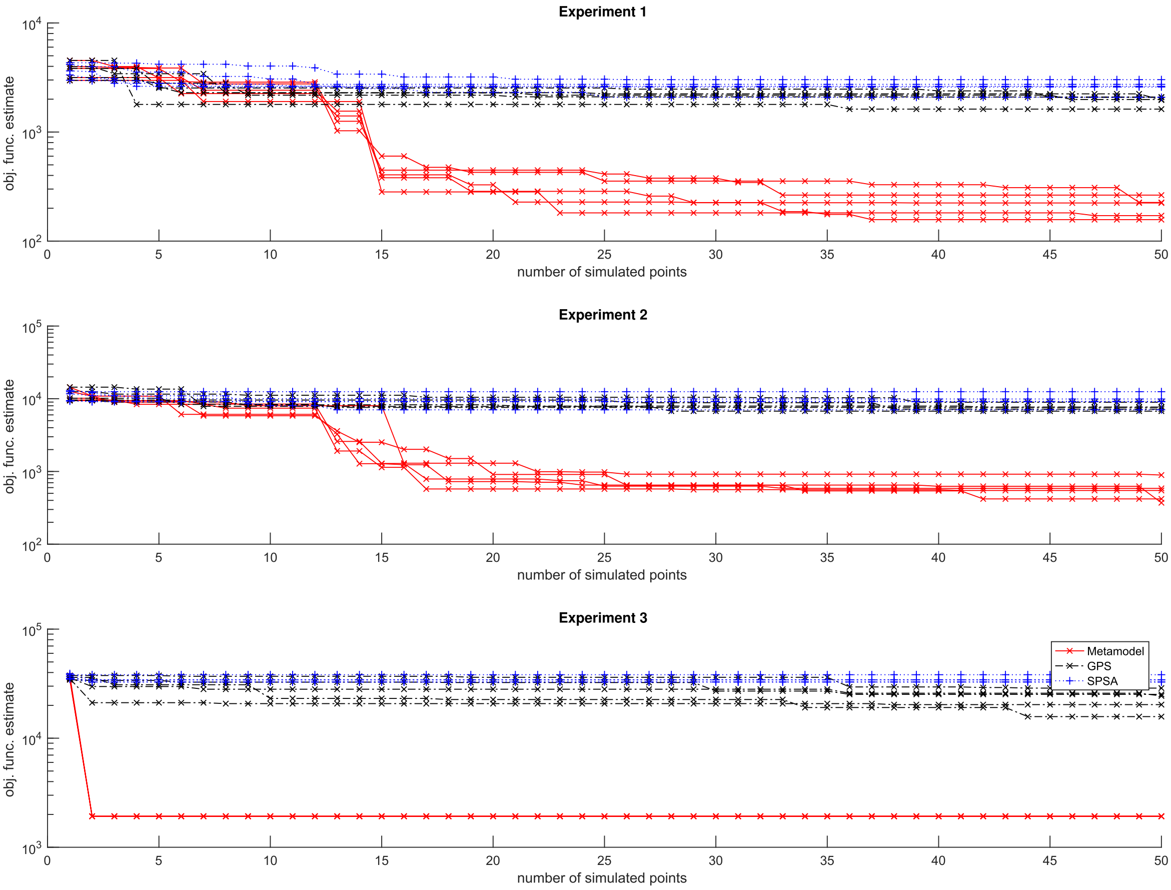


Fig. 5. Performance estimate of the current iterate of each method as a function of the number of simulated points. From top to bottom, the plots consider time-intervals of duration 1 h, 30 min and 15 min, respectively.

Experiments 1 through 3 have increasing dimension and increasing temporal correlation (of the link performance metrics across time intervals). These plots indicates that as these two factors increase, so does the magnitude by which the metamodel outperforms the benchmark methods. When considering the objective function estimates, the metamodel tends to outperform the benchmark methods by 1 to 2 orders of magnitude. Both benchmark methods have similar performance.

Fig. 4 considers, for each experiment, the 5 solutions proposed by each method (there is one for each initial point). For a given method and a given initial point, the proposed solution is the current iterate (i.e., solution with best performance) upon depletion of the simulation budget. Fig. 4 evaluates the performance of the proposed solutions in terms of how they fit the true counts. The left-most (resp. middle and right-most) column of plots considers experiment 1 (resp. 2 and 3). The top (resp. middle and bottom) row of plots considers the metamodel (resp. GPS and SPSA) method. Each plot displays along the x-axis the true counts and along the y-axis the simulated counts of the proposed solutions. For each method, the best fit to the counts is observed for experiment 1, followed by experiments 2 and 3. For a given experiment, GPS and SPSA have similar performance, while the metamodel outperforms both benchmark methods. Just as for the previous figure, as the number of calibration periods and the temporal correlation increase, so does the ability of the metamodel method to outperform the benchmark methods. Recall that experiment 1 is that with the longest time intervals. In other words, it has the least temporal dependence between the simulation-based performance metrics across time intervals (terms $E[F_{it}(d, u_1;u_2)]$ of Eq. (1)). Experiment 2 has higher dependency, and experiment 3 has the highest. These results therefore show that there is a negative correlation between the duration of the calibration intervals and the performance of the general-purpose methods. The metamodel method identifies points with good performance for all 3 levels of temporal dependency.

In the second set of experiments, we carry out three experiments that consider 5 calibration periods ($T = 5$). All experiments have a common decision vector dimension of 45 ($T{\cdot}Z = 5{\cdot}9$). The true demand levels vary from uncongested conditions to congested conditions and back to uncongested conditions. Let D_1 denote a given level of demand. Then, for each of the 5 calibration time

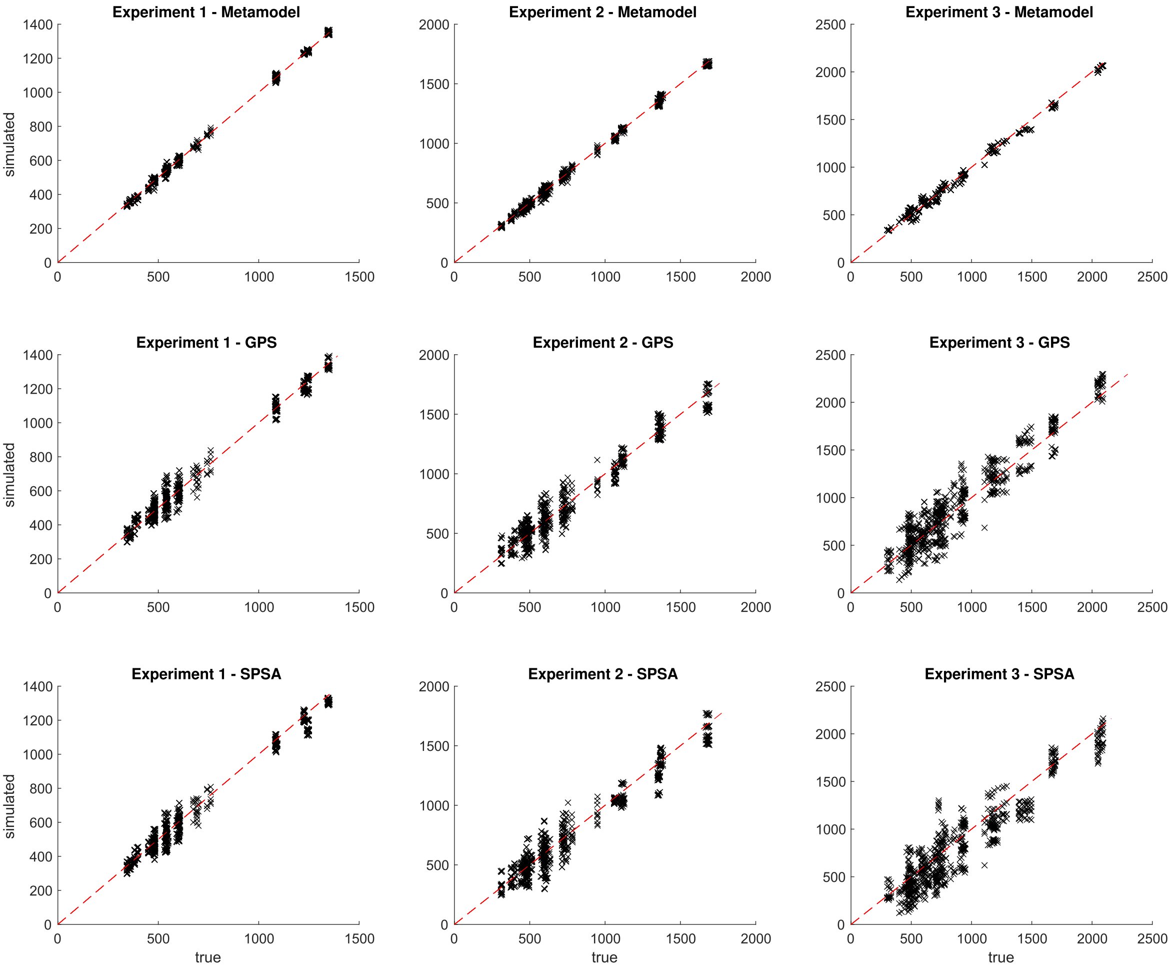
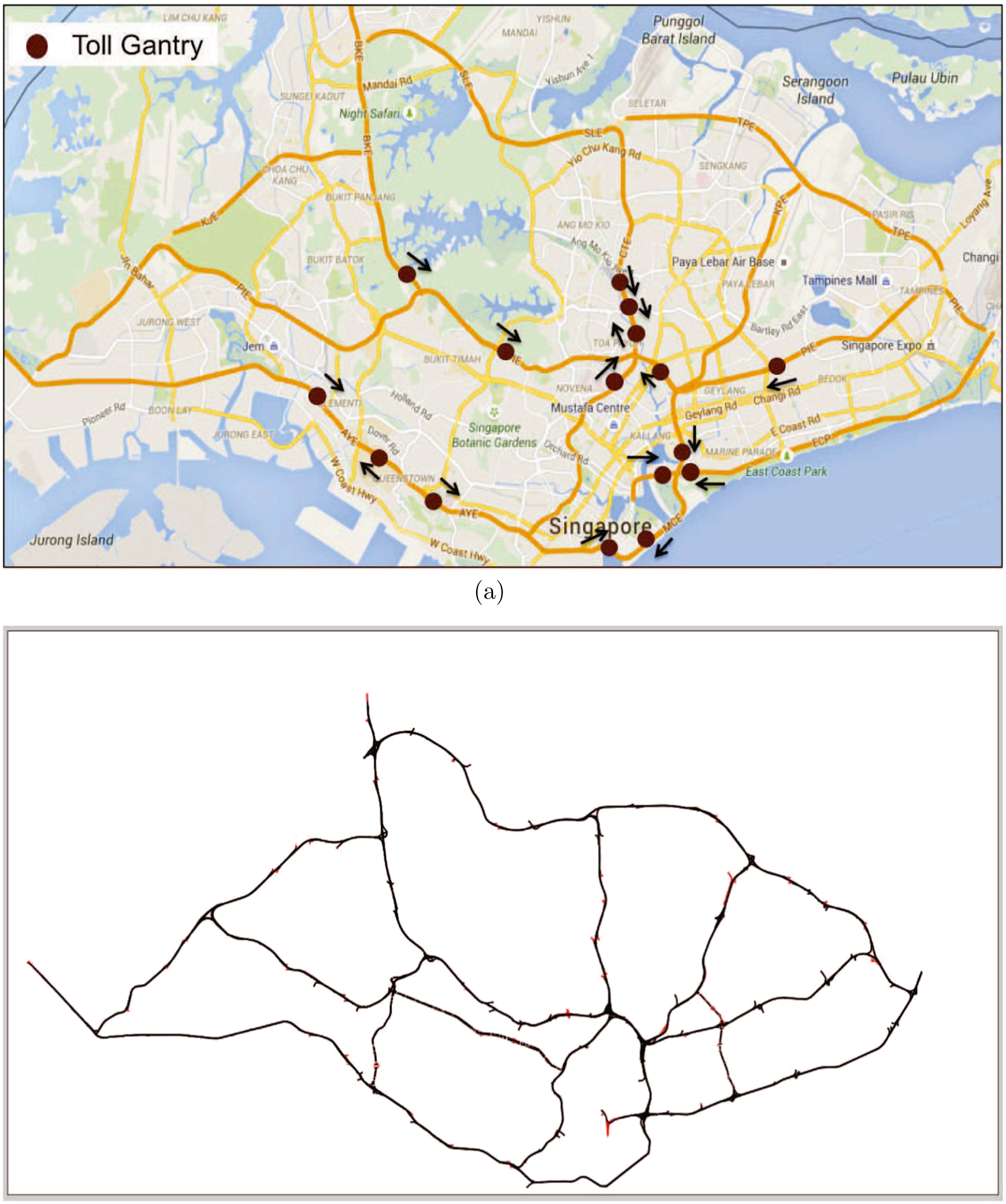


Fig. 6. Fit to link counts of the solutions proposed by each method. Experiments 1, 2 and 3 correspond, respectively, to the left-most, middle and right-most column of plots. The top, middle and bottom row of plots correspond, respectively, to the plots of the metamodel, GPS and SPSA.

periods, the OD demand is D_1, $1.5D_1$, $2D_1$, $1.5D_1$ and D_1, respectively. The 3 experiments differ in the time duration of a given true OD demand. More specifically, for experiment 1 (resp. 2 and 3) each OD demand lasts for 1 h (resp. 30 min and 15 min). Since the underlying true demand varies over time, the experiments with shorter time duration of the true OD demand serve to evaluate the ability of the methods to perform well under transient, rather than stationary, traffic conditions. Just as for the previous set of experiments, the shorter the duration of the time intervals, the higher the temporal dependency between the simulation-based performance metrics (terms $E[F_{it}(d, u_1;u_2)]$) across different time intervals. Hence, these experiments also evaluate the impact of decomposing the metamodel calibration problem into decoupled subproblems.

As before, for each experiment, we initialize each method with each of the 5 initial points and the simulation budget is set to 50 simulation runs. Fig. 5 has a similar layout than Fig. 3. The top, middle and bottom plots of Fig. 5 consider, respectively, experiments 1, 2 and 3. The line styles and colors are the same as those of Fig. 3. Again, the y-axis has a logarithmic scale. The following trends hold. For all experiments, GPS and SPSA have similar performance. The proposed method outperforms the benchmark methods by 1 to 2 orders of magnitude. As the effect of the transients increases (i.e., going from the top to the bottom plots), so does the outperformance of the metamodel compared to the benchmark methods. This figure indicates, again, that as the temporal dependency of the simulation-based performance metrics increases, the performance of the general-purpose methods decreases. For all 3 levels of temporal dependency, the proposed metamodel method identifies points with good performance.

Fig. 6 has a similar layout than Fig. 4. Each plot displays, for a given experiment and a given method, the performance of the 5 solutions. The top (resp. middle and bottom) row corresponds to the metamodel (resp. GPS and SPSA) method. The plots of column 1 (resp. 2 and 3) correspond to experiment 1 (resp. 2 and 3). For each plot, the true counts are displayed along the x-axis and the



(a)

(b)

Fig. 7. Singapore network. (a) Singapore expressway network (map data: Google Maps (2017)). (b) Simulation network model.

simulated counts are along the *y*-axis. For a given experiment, GPS and SPSA have similar performance. As the effect of the transients increases (i.e., going from the left to the right plots), the performance of GPS and SPSA deteriorates significantly. That of the metamodel deteriorates slightly, yet it remains good for all 3 experiments. Thus, as the effect of the transients increases, so does the outperformance of the metamodel compared to the benchmark methods. Both benchmark methods are outperformed by the metamodel approach. This holds for all 3 experiments. These plots indicate that the metamodel method identifies points with good performance even under transient traffic conditions with high temporal correlation across time intervals.

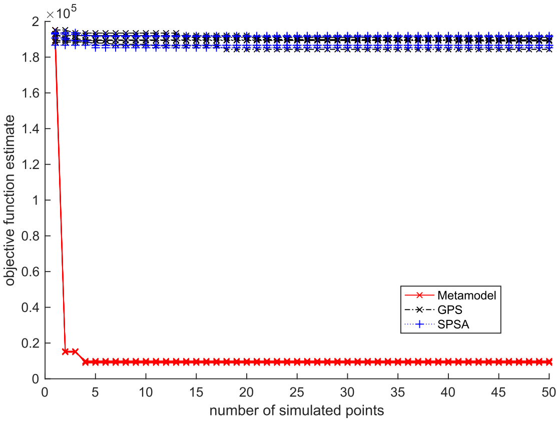


Fig. 8. Performance estimate of the current iterate of each method as a function of the number of simulated points.

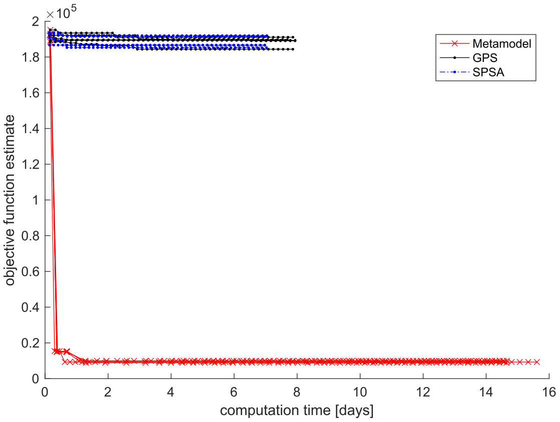


Fig. 9. Performance estimate of current iterate of each method as a function of computation runtime.

4. Singapore case study

We consider an OD calibration problem for the Singapore network of major arterials and expressways, a map of which is displayed in Fig. 7a. The network topology of the model is displayed in Fig. 7b. Fig. 7a represents the expressway links as dark orange links and the major arterials as yellow links. The red circles indicate the location of tolls in the network. The network model consists of 1150 links, over 2300 lanes and over 18,000 routes. Of the 1150 links of the simulator, only 860 are considered in the analytical network model. These 860 links are those that appear in the pre-determined route choice set.

For each time interval, each OD matrix is defined as a set of 4050 OD pairs (i.e., $Z = 4050$). We estimate 4 OD matrices (i.e., $T = 4$). This leads to a total of 16200 unknown ODs (i.e., the decision vector of Problem (1)–(2) is of dimension 16200). We focus on the morning peak period. Each OD matrix represents demand for one of the following 30 min time intervals: 7:30–8 am, 8–8:30 am, 8:30–9 am and 9–9:30 am. When simulating we use 6–7:30 am as a warm-up period, the OD matrix for this warm-up period is fixed. There are 172 links with sensors (i.e., $|\mathcal{I}| = 172$). We do not have access to field data, hence we use an existing time-dependent OD matrix as the "true" OD matrix. Based on this matrix, we generate via simulation synthetic link counts (i.e., terms y_{it} of Eq. (1)). Each entry of the prior OD matrix (terms $\widetilde{d}_{zt}$ of Eq. (1)) is obtained as the sum of the corresponding entry of the true OD matrix plus a randomly drawn error term from a normal distribution with expectation 0 and standard deviation equal to 20% of the true value. The upper bound ($d^{\max}$ of Eq. (2)) is set to 2000 veh/hr. The weight parameter of the prior OD matrix (δ of Eq. (1)) is set to 0.01.

Just as for Section 3 we compare the performance of the metamodel method to that of GPS and of SPSA. The algorithmic

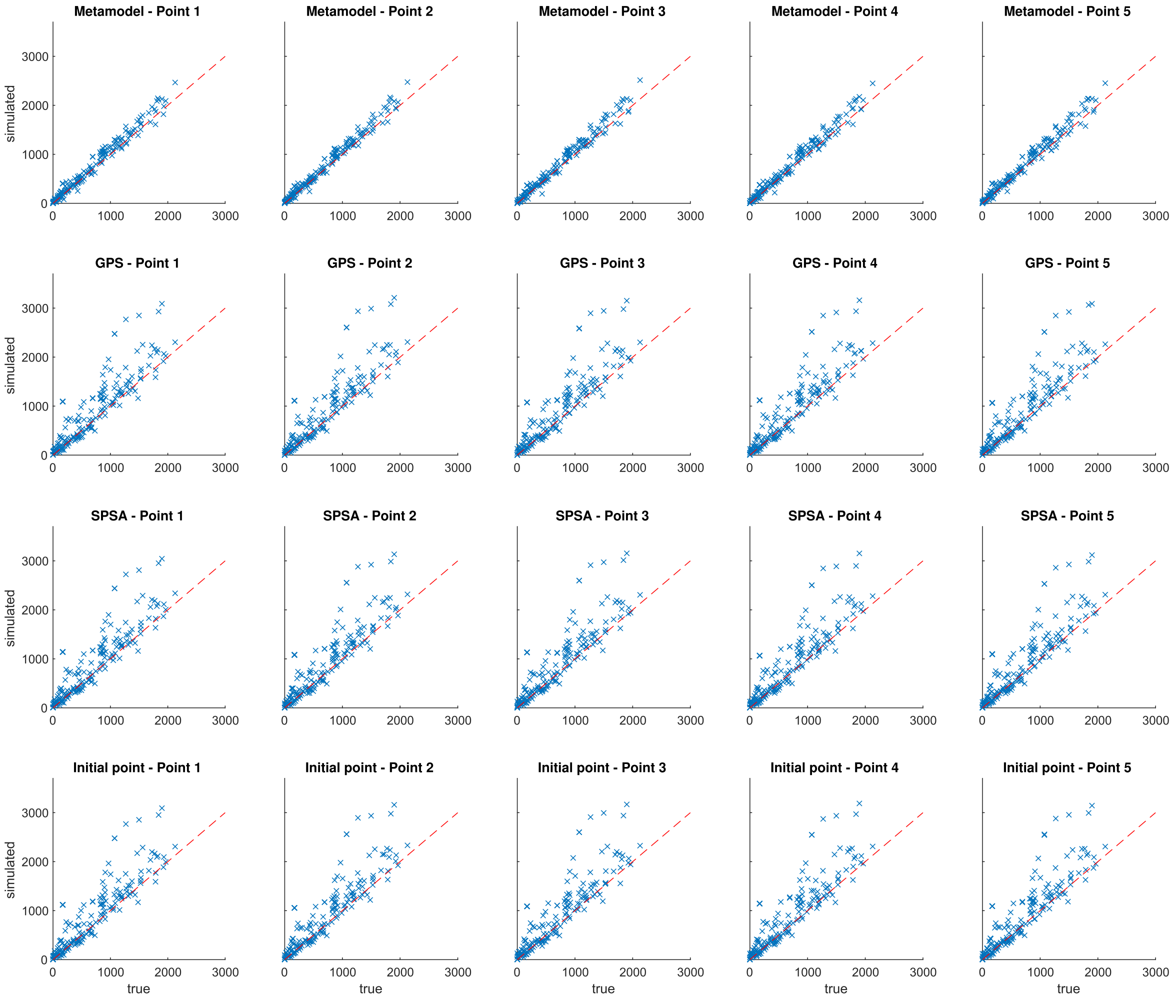


Fig. 10. Fit to link counts for time interval 1. The top (resp. 2nd, 3rd and bottom) row corresponds to the metamodel method (resp. GPS method, SPSA method and initial point).

parameters of GPS and SPSA are set following both general guidelines and calibration guidelines, as discussed in Section 3. In order to evaluate the performance of the methods under tight computational budgets, we proceed as follows. We consider 5 initial points (i.e., initial OD matrices) that are drawn randomly from the feasible region, which is defined by Eq. (2). For each initial point, we run each method and terminate it once 50 points have been simulated. Fig. 8 displays the number of simulated points along the x-axis and the estimate of the objective function of the current iterate along the y-axis. The red (resp. black solid and blue dotted) lines correspond to the metamodel (resp. GPS and SPSA) method. For each line style, there are 5 lines that correspond to running the corresponding method for each of the 5 initial points. This figure illustrates that from the first simulations the metamodel method identifies points with objective function estimates that are improved by one order of magnitude compared to the traditional methods. The traditional methods gradually, yet slowly, identify points with improved performance. Again, keep in mind that these traditional methods are designed based on asymptotic performance properties, they are not designed to perform well under tight computational budgets, let alone for such high-dimensional problems.

From a computational perspective, the metamodel method requires solving, at every iteration, a set of optimization problems (problems $\mathcal{P}_{kt}$ defined by Eqs. (4)–(6)). Hence, Fig. 9 evaluates the performance of the methods in terms of the total computational runtime. Recall, that one simulation evaluation typically takes 4 h to compute. Fig. 9 differs from Fig. 8 in that it considers total computation runtime along the x-axis. It illustrates that the metamodel method indeed takes longer to deplete the computational budget of 50 simulation runs. The average computation time needed to deplete the budget is 14.8 days for the metamodel method, versus 7.9 days for GPS and 7.0 days for SPSA. Nonetheless, for all computation times, the metamodel

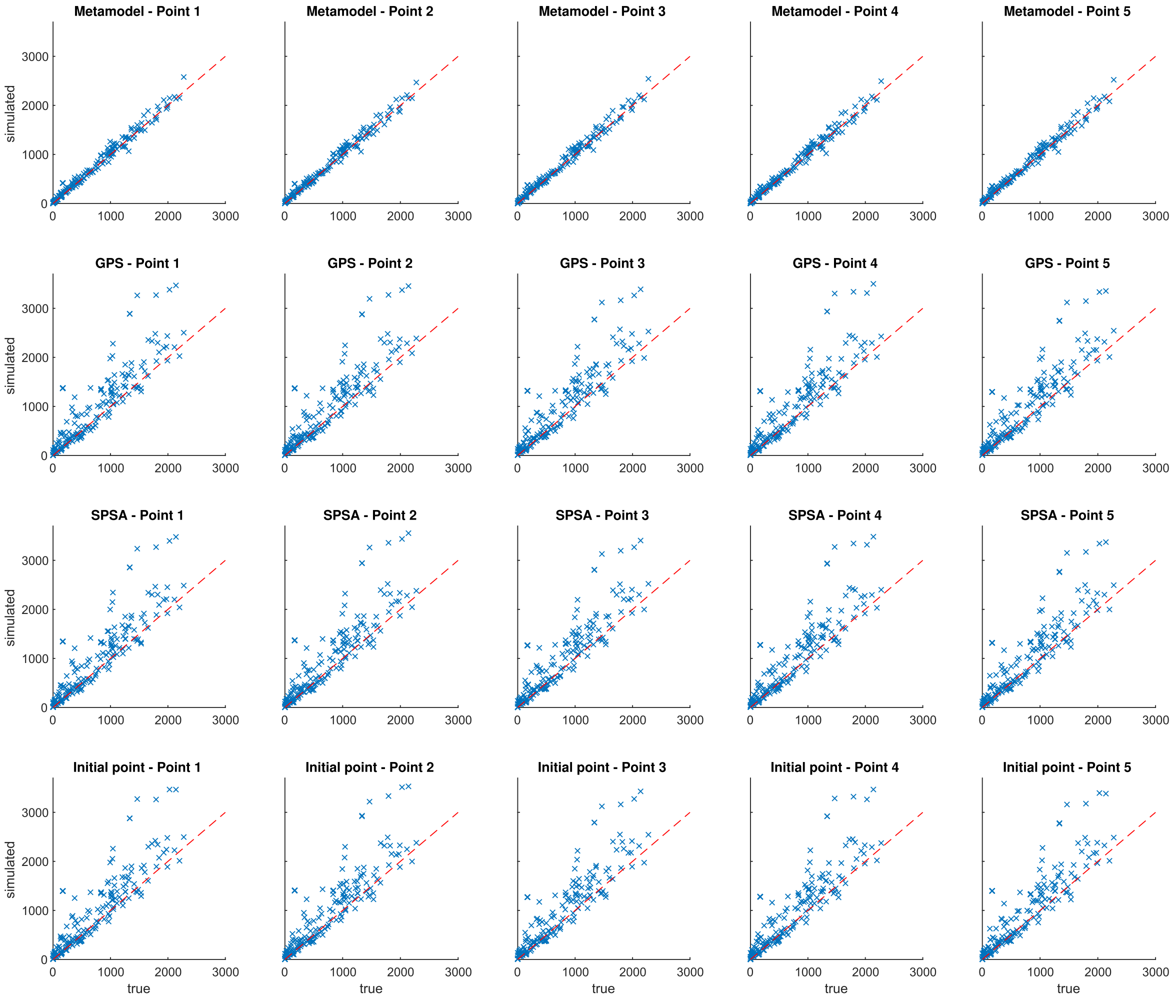


Fig. 11. Fit to link counts for time interval 2. The top (resp. 2nd, 3rd and bottom) row corresponds to the metamodel method (resp. GPS method, SPSA method and initial point).

method identifies points with performance that is improved by an order of magnitude compared to the traditional benchmark methods.

Recall that for each initial point, a method is terminated once the computational budget of 50 simulation runs is depleted. Upon depletion of the budget, the current iterate is referred to as the “best” solution. Figs. 10–13 evaluate the performance of the best solution, for each method and each initial point. The figures compare the link counts of the best solution to the “true” (synthetic) link counts. Fig. 10 (resp. 11, 12 and 13) considers time interval 1 (resp. 2, 3 and 4). For each figure, the plots in the first (i.e., top) row correspond to the metamodel method, those in the 2nd, 3rd and bottom row correspond, respectively, to GPS, SPSA and the initial point. Each figure contains 5 columns, one for each initial point. The trends for all figures are similar: (i) the metamodel method identifies ODs with good fit to the true counts, (ii) GPS and SPSA yield solutions that have performance similar to that of the initial point. Recall that we are considering tight computational budgets. More specifically, we allow for a total of 50 points to be simulated, while the problem is of dimension 16,200. In other words, the computational budget is in the order of 0.3% of the problem dimension. It is therefore remarkable that the metamodel method can identify points with such good fit to the counts under such tight computational budgets. Moreover, the method yields good performance for all 5 initial points. This illustrates the robustness of the method to both the quality of the initial points and to the stochasticity of the simulator.

Table 1 quanitifies the quality of the fit to link counts for each method. It considers the commonly used root mean square normalized error (RMSN) function, which is defined for a given method and a given time interval t as:

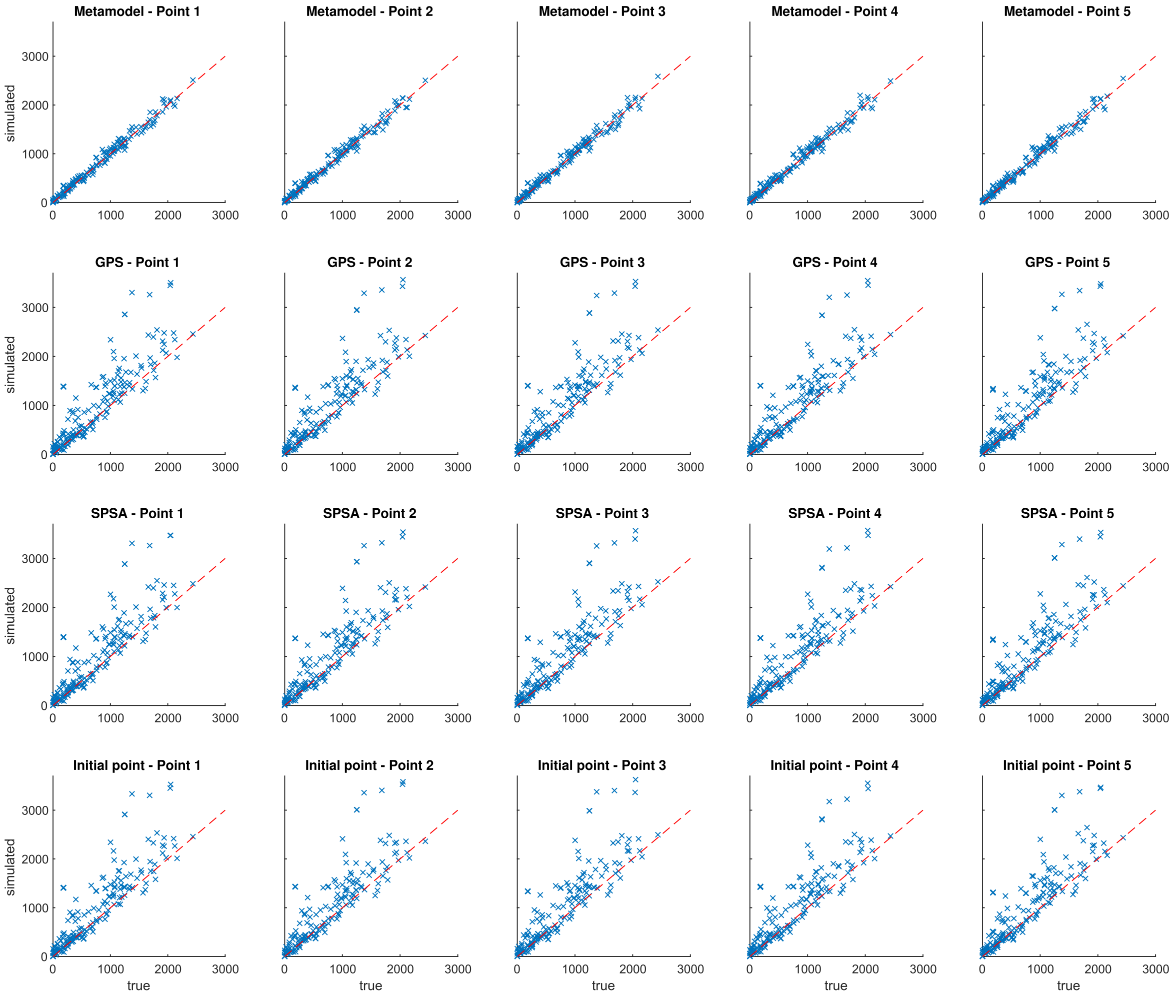


Fig. 12. Fit to link counts for time interval 3. The top (resp. 2nd, 3rd and bottom) row corresponds to the metamodel method (resp. GPS method, SPSA method and initial point).

$$100\frac{\sqrt{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(y_{it}-\widehat{E}\,[F_{it}])^2}}{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}y_{it}},$$

where $\widehat{E}\,[F_{it}]$ denotes the simulation-based estimate of the expected flow on link i during time interval t obtained for the best solution of a given method. The smaller the RMSN the better the fit to the link counts (terms y_{it}). Table 1 computes for each method and each time interval, the average RMSN (averaged over all 5 initial points). Columns 1–4 display the average RMSN for time intervals 1–4, respectively. Rows 1–4 display the average RMSN for the metamodel method, GPS, SPSA and the initial points, respectively. This table indicates that the solutions of the benchmark methods have similar performance than the initial points. For time intervals 1–4, the proposed method improves the fit to the counts compared to the initial points and to the benchmark methods by 69%, 81%, 83% and 77%, respectively. This leads to an average (across time intervals) improvement of 77%. Although we allow for only 50 simulation evaluations, for a problem of dimension 16,200, the proposed method achieves within such a tight computational budget an average improvement of 77% compared to both the initial points and to the benchmark methods. This highlights its efficiency.

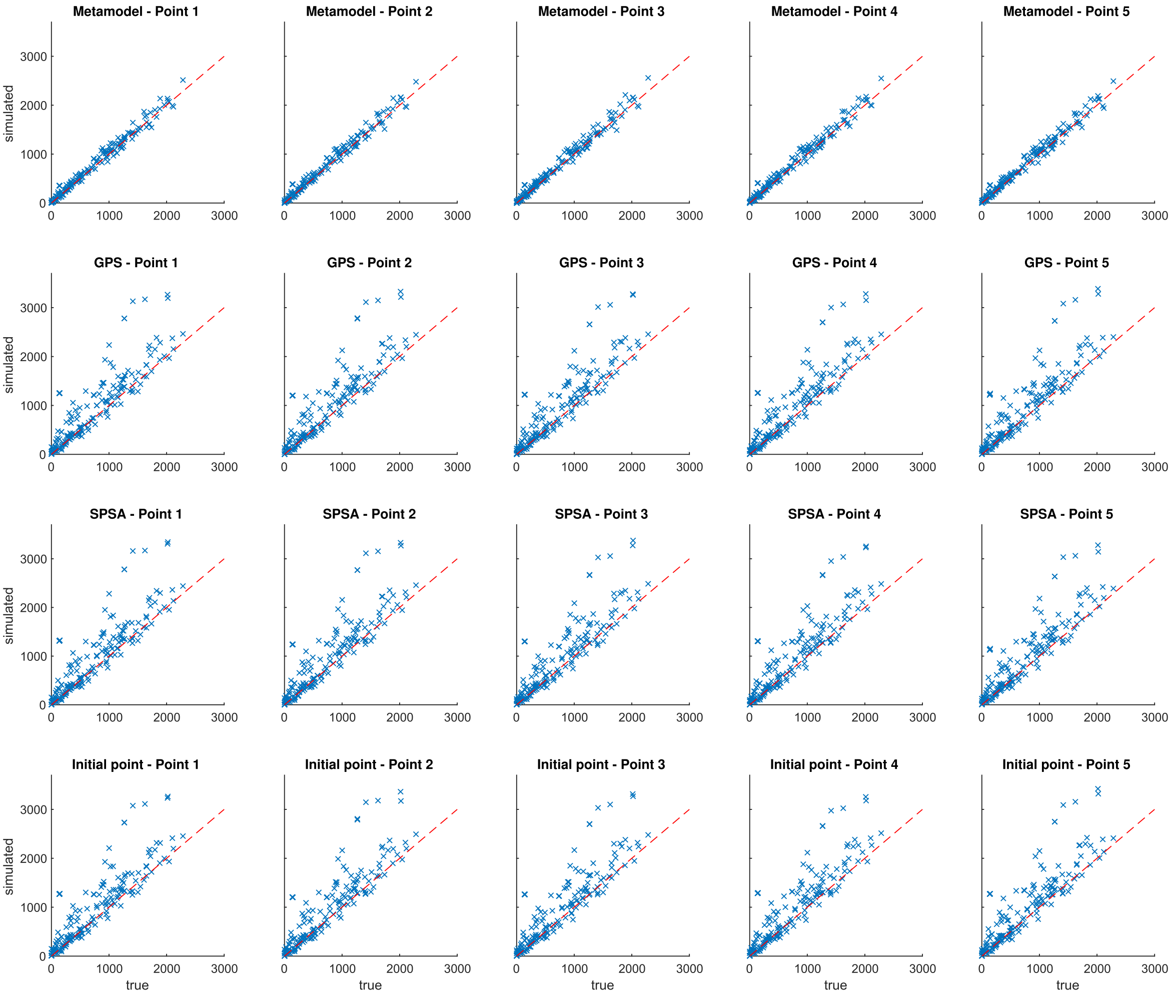


Fig. 13. Fit to link counts for time interval 4. The top (resp. 2nd, 3rd and bottom) row corresponds to the metamodel method (resp. GPS method, SPSA method and initial point).

Table 1
Average (across initial points) root mean square normalized error (RMSN) percentage

	Time interval 1	Time interval 2	Time interval 3	Time interval 4
Metamodel	17	11.4	10.8	12.9
GPS	55.6	58.9	62.1	56.6
SPSA	54.9	59	62.2	56.9
Initial point	55.5	59.5	62.9	57.1

5. Conclusions

This paper proposes a scalable and efficient algorithm for dynamic OD calibration problems. It is a metamodel SO algorithm. For a problem with T time intervals and Z OD pairs per time interval, the calibration problem is a $T \cdot Z$ dimensional SO problem. The approach solves, at each iteration of the SO algorithm, a set of T independent analytical and differentiable optimization problems, each of dimension Z. This decomposition across time contributes to the scalability of the SO approach. Each analytical problem consists of a nonlinear objective function constrained by lower and upper bounds, and n nonlinear equality constraints, where n is the number of links in the network. The temporal correlation, of the link performance metrics across time intervals, is approximately captured through the parameters of the analytical objective function (i.e., the metamodel). The metamodel relies on the use of stationary network model. The latter is defined as a system of nonlinear equations with a dimension that scales linearly with n and independently of other network attributes such as the dimension of the route choice set. Hence the metamodel is scalable and can be evaluated efficiently with standard solvers for systems of equations.

We benchmark the approach versus two general-purpose algorithms SPSA and a derivative-free pattern search algorithm. We validate the approach with experiments on a toy synthetic network. The experiments evaluate the impact of the problem dimension, the level of temporal correlation across time intervals, and the presence of transient versus stationary traffic conditions within each time interval. When comparing the objective function estimates, the proposed method outperforms the benchmark methods by 1 to 2 orders of magnitude. As the problem dimension and the temporal correlation increase, so does the magnitude by which the proposed method outperforms the benchmark methods. The proposed method performs well even when there is high temporal correlation across time intervals.

The 3 methods are applied to a high-dimensional large-scale Singapore case study. The proposed method identifies solutions with estimated objective functions that are 1 order of magnitude lower than the benchmark methods. Compared to the benchmark methods, the proposed method yields solutions that improve the fit to link counts by an average of 77%. The performance of the method is also shown to be robust to the quality of the initial points.

As discussed in Section 2.2, the main challenge when going from static to dynamic OD calibration is the temporal dependency of the network (e.g., link, path) performance metrics. More generally, this is true when going from a general static optimization problem to a dynamic one. The results of this paper indicate that analytically describing spatial dependency for dynamic optimization problems can help to limit, or even overcome, the need to analytically describe temporal dependency. More specifically, the proposed approach captures temporal correlation in a simple parametric way, without the formulation of a time-dependent analytical network model. This is encouraging, given the challenge of formulating analytical network models that are time-dependent, while also remaining tractable and scalable. The proposed approach lays the foundation for new research directions to address dynamic, and even real-time, optimization problems. Ongoing work extends the proposed approach to: (i) enable other types of mobility data, such as turning fraction data or partial trajectory data, to be used for calibration, (ii) enable real-time problems to be efficiently tackled.

Dynamic OD calibration problems are high-dimensional SO problems. As discussed in Section 1, the literature has proposed the use of various dimensionality reduction techniques. Nonetheless, their application often comes with a high-computational cost (due to the high cost of performing simulation). It would be of interest to explore the use of dimensionality reduction techniques based on analytical network models, which would avoid the use of simulation; or more generally the iterative use (within an SO algorithm) of dimensionality reduction techniques based on metamodels, which would allow us to learn from past simulation observations when performing dimensionality reduction.

Recent extensions to SPSA can lead to an increase in the total computational cost. For example, for c-SPSA the number of simulation evaluations increases linearly with the number of clusters (Tympakianaki et al., 2015), similarly the numerical estimation of the weight matrix of w-SPSA can be computationally demanding (Lu et al., 2015). It would be of interest to explore the use of the analytical network model to mitigate the computational cost of these new and promising extensions of SPSA.

In this paper, the analytical network model is used, iteratively (i.e., at every iteration of the SO algorithm), to solve an analytical (and approximate) optimization problem. Another promising research direction, is to use it, again as part of an iterative SO algorithm, to define a sampling distribution. For example, points with low analytical objective function values would have high sampling probabilities. Ongoing work shows the potential of this line of thinking for high-dimensional continuous SO problems. In particular, this could lead to metamodel techniques with further enhanced computational tractability because it can overcome the need to solve a series of (analytical) optimization problems. The study of other general purpose components of the metamodel (function ϕ of Eq. (4)) are also of interest. Examples include general-purpose functions that capture temporal dependencies or, if abundant high-resolution data is available, general-purpose machine learning techniques, such as in Wu et al. (2018).

Acknowledgments

The work of C. Osorio is partially supported by the U.S. National Science Foundation under Grant No. 1351512. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Fitting of the metamodel parameters

This appendix formulates the least-squares problem that is solved, at every iteration k of the SO algorithm, to fit the parameters, β_{kt}, of the metamodel m_{kt}. For a given point d, define $\widehat{E}\,[F_{it}(d,\,u_1;u_2)]$ as the simulation-based estimate of the expected flow on link i during time interval t for point d and define $\widehat{g}_t(d)$ as:

$$\widehat{g}_t(d) = \sum_{i\in\mathcal{I}} (y_{it} - \widehat{E}\,[F_{it}(d,\,u_1;u_2)])^2 - \sum_{i\in\mathcal{I}} y_{it}^2. \tag{15}$$

Let f_t^1 denote the term within the curly brackets of Eq. (3). The first summation in Eq. (15) is an estimate of f_t^1. The second summation is a constant term, i.e., it does not depend on the decision vector d or on the simulation outputs, it only depends on the field measurements.

At iteration k of the SO algorithm, let $\mathcal{S}_k$ denote the set of points (i.e., OD matrices) simulated up until iteration k. The least-squares problem is formulated as:

$$\min_{\beta_{kt}} \sum_{d\in\mathcal{S}_k} \{w_k(d)(\hat{g}_t(d) - m_{kt}(d_t;\beta_{kt}))\}^2 + w_0^2\left((\beta_{kt0}-1)^2 + \beta_{kt1}^2 + \sum_{z=2}^{Z+1} \beta_{ktz}^2\right), \tag{16}$$

where w_0 is an exogenous (fixed) scalar weight coefficient (set to 10^{-4}) and $w_k(d)$ is a scalar weight for point d defined as in Osorio and Bierlaire (2013) by the following equation:

$$w_k(d) = \frac{1}{1 + \widetilde{c}\ \|d - d^k\|_2}, \tag{17}$$

where d^k denotes the current iterate and $\widetilde{c}$ denotes an exogenous scaling coefficient (set to 10^{-4}).

Problem (16) fits the parameters of m_{kt} by solving a weighted least squares problem. The first term of Problem (16) represents the weighted distance between the estimates of the simulation-based function and its corresponding metamodel approximation. For a given point d, its weight is proportional to its distance from the current iterate d^k. This aims to improve the local (i.e., in the vicinity of the current iterate) fit of the metamodel. The second term of Problem (16) accounts for the distance between the parameter β_{kt} and initial (or prior) values. This second term ensures that the least square matrix is of full rank. The initial values used correspond to an initial metamodel that is solely based on the analytical network model (i.e., $\beta_{kt0} = 1$ and $\forall\, j \geqslant 1\ \ \beta_{ktj} = 0$).

The metamodel can be viewed as a linear regression model that aims to predict the simulation-based function $(f_t^1(d) - \sum_{i\in\mathcal{I}} y_{it}^2)$. We have included the summation term because it led to a better model fit. Since this summation term does not depend on d (more generally, it does not depend on the simulation outputs), it will not impact the optimal solution of the metamodel optimization problem (i.e., it will simply shift the objective function by a constant term).

References

Ankenman, B., Nelson, B.L., Staum, J., 2010. Stochastic Kriging for simulation metamodeling. Oper. Res. 58 (2), 371–382.

Antoniou, C., Lima Azevedo, C., Lu, L., Pereira, F., Ben-Akiva, M., 2015. W-SPSA in practice: approximation of weight matrices and calibration of traffic simulation models. Transp. Res. Part C 59, 129–146.

Ashok, K., Ben-Akiva, M.E., 2002. Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transport. Sci. 36 (2), 184–198.

Astarita, V., Er-Rafia, K., Florian, M., Mahut, M., Velan, S., 2001. Comparison of three methods for dynamic network loading. Transp. Res. Rec. 1771, 179–190.

Balakrishna, R., 2006. Off-line calibration of dynamic traffic assignment models (Ph.d. thesis). Massachusetts Institute of Technology.

Balakrishna, R., Ben-Akiva, M.E., Koutsopoulos, H.N., 2007. Offline calibration of dynamic traffic assignment: simultaneous demand-and-supply estimation. Transp. Res. Rec. 2003, 50–58.

Barceló, J., Montero, L., 2015. A robust framework for the estimation of dynamic OD trip matrices for reliable traffic management. In Procedia Social and Behavioral Sciences. Papers selected for the 18th meeting of the EURO Working Group on Transportation.

Barton, R.R., Meckesheimer, M., 2006. Metamodel-based simulation optimization. In: In: Henderson, S.G., Nelson, B.L. (Eds.), Handbooks in Operations Research and Management Science: Simulation, vol. 13. Elsevier, Amsterdam, pp. 535–574 (chapter 18).

Ben-Akiva, M., Koutsopoulos, H.N., Antoniou, C., Balakrishna, R., 2010. Traffic Simulation with DynaMIT. Springer, New York, NY pp. 363–398.

Ben-Akiva, M., Gao, S., Wei, Z., Wen, Y., 2012. A dynamic traffic assignment model for highly congested urban networks. Transp. Res. Part C 24, 62–82.

Bierlaire, M., Crittin, F., 2004. An efficient algorithm for real-time estimation and prediction of dynamic OD tables. Oper. Res. 52 (1), 116–127.

Chong, L., Osorio, C., 2018. A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. Transportation Science 52 (3), 637–656.

Cipriani, E., Florian, M., Mahut, M., Nigro, M., 2011. A gradient approximation approach for adjusting temporal origin-destination matrices. Transp. Res. Part C 19 (2), 270–282.

Djukic, T., 2014. Dynamic OD Demand Estimation and Prediction for Dynamic Traffic Management (Ph.D. thesis). Delft University of Technology, Delft, The Netherlands.

Djukic, T., van Lindt, J.W.C., Hoogendoorn, S.P., 2012. Application of principal component analysis to predict dynamic origin-destination matrices. Transp. Res. Rec. 2283, 81–89.

Dong, N.A., Eckman, D.J., Poloczek, M., Zhao, X., Henderson, S.G., 2017. Comparing the finite-time performance of simulation-optimization algorithms. In: Chan, W.K.V., D'Ambrogio, A., Zacharewicz, G., Mustafee, N., Wainer, G., Page, E. (Eds.), Proceedings of the Winter Simulation Conference. IEEE.

Frederix, R., Viti, F., Himpe, W.W., Tampère, C.M., 2014. Dynamic origindestination matrix estimation on large-scale congested networks using a hierarchical decomposition scheme. J. Intell. Transport. Syst. 18 (1), 51–66.

Google Maps, 2017. Singapore expressway network.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. J. Global Optim. 13 (4), 455–492.

Kattan, L., Abdulhai, B., 2006. Noniterative approach to dynamic traffic origin-destination estimation with parallel evolutionary algorithms. Transp. Res. Rec. 1964, 201–210.

Kim, H., Baek, S., Lim, Y., 2001. Origin-destination matrices estimated with a genetic algorithm from link traffic counts. Transp. Res. Rec. 1771, 156–163.

Kleijnen, J.P.C., van Beers, W., van Nieuwenhuyse, I., 2010. Constrained optimization in expensive simulation: novel approach. Eur. J. Oper. Res. 202 (1), 164–174.

Lee, J.B., Ozbay, K., 2009. New calibration methodology for microscopic traffic simulation using enhanced simultaneous perturbation stochastic approximation approach. Transp. Res. Rec. 2124, 233–240.

Lu, L., Xu, Y., Antoniou, C., Ben-Akiva, M., 2015. An enhanced SPSA algorithm for the calibration of dynamic traffic assignment models. Transp. Res. Part C 51, 149–166.

Mathworks, Inc., 2016. Global Optimization Toolbox User's Guide Matlab (R2016b), Natick, MA, USA.

Osorio, C., 2010. Mitigating Network Congestion: Analytical Models, Optimization Methods and Their Applications (Ph.D. thesis). Ecole Polytechnique Fédérale de Lausanne.

Osorio, C., 2017. High-dimensional Offline od Calibration for Stochastic Traffic Simulators of Large-scale Urban Networks (Technical report) Massachusetts Institute of Technology. (Under review). Available at: <http://web.mit.edu/osorioc/www/papers/osoODCalib.pdf>.

Osorio, C., Atasoy, B., 2017. Efficient Simulation-based Toll Optimization for Large-scale Networks. Technical report Massachusetts Institute of Technology (Under review). Available at: <http://web.mit.edu/osorioc/www/papers/osoAtaTollSO.pdf>.

Osorio, C., Bierlaire, M., 2013. A simulation-based optimization framework for urban transportation problems. Oper. Res. 61 (6), 1333–1345.

Osorio, C., Nanduri, K., 2015. Urban transportation emissions mitigation: coupling high-resolution vehicular emissions and traffic models for traffic signal optimization. Transp. Res. Part B 81, 520–538.

Prakash, A.A., Seshadri, R., Antoniou, C., Pereira, F.C., Ben-Akiva, M.E., 2017. Reducing the dimension of online calibration in dynamic traffic assignment systems. Transp. Res. Rec. 2667, 96–107.

Prakash, A.A., Seshadri, R., Antoniou, C., Pereira, F.C., Ben-Akiva, M.E., 2018. Improving scalability of generic online calibration for real-time dynamic traffic assignment systems. Transport. Res. Rec (forthcoming).

Spall, J., 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transa. Automatic Control 37 (3), 332–341.

Spall, J.C., 2003. Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, New Jersey, USA.

Stathopoulos, A., Tsekeris, T., 2004. Hybrid meta-heuristic algorithm for the simultaneous optimization of the O-D trip matrix estimation. Comput.-Aided Civil Infrastruct. Eng. 19 (6), 421–435.

Tympakianaki, A., 2018. Demand Estimation and Bottleneck Management Using Heterogeneous Traffic Data (Ph.D. thesis). KTH Royal Institute of Technology.

Tympakianaki, A., Koutsopoulos, H., Jenelius, E., 2015. c-SPSA: cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin-destination matrix estimation. Transp. Res. Part C 55, 231–245.

Vaze, V., Antoniou, C., Wen, Y., Ben-Akiva, M., 2009. Calibration of dynamic traffic assignment models with point-to-point traffic surveillance. Transp. Res. Rec. 2090, 1–9.

Wang, W., Wan, H., Chang, K.H., 2016. Randomized block coordinate descendant STRONG for large-scale stochastic optimization. In: 2016 Winter Simulation Conference (WSC), pp. 614–625.

Wild, S.M., Regis, R.G., Shoemaker, C.A., 2008. ORBIT: optimization by radial basis function interpolation in trust-regions. SIAM J. Sci. Comput. 30 (6), 3197–3219.

Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: a forward and backward propagation algorithmic framework on a layered computational graph. Transport. Res. Part C (forthcoming).

Zhang, C., Osorio, C., 2017. Efficient Offline Calibration of Origin-destination (demand) for Large-scale Stochastic Traffic Models. Technical report Massachusetts Institute of Technology (Under review). Available at: <http://web.mit.edu/osorioc/www/papers/zhaOsoODcalib.pdf>.

Zhou, T., Osorio, C., Fields, E., 2017. A Data-driven Discrete Simulation-based Optimization Algorithm for Large-scale Two-way Car-sharing Network Design (Technical report) Massachusetts Institute of Technology (Under review). Available at: <http://web.mit.edu/osorioc/www/papers/zhoOsoFieCarSharing.pdf>.

Zhou, X., Mahmassani, H.S., 2007. A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. Transp. Res. Part B 41, 823–840.