Efficient simulation-based toll optimization for large-scale networks

Carolina Osorio
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA02139, USA, osorioc@MIT.EDU

Bilge Atasoy
Department of Maritime and Transport Technology, Delft University of Technology, Delft, Netherlands, b.atasoy@tudelft.nl

This paper proposes a simulation-based optimization technique for high-dimensional toll optimization problems of large-scale road networks. We formulate a novel analytical network model. The latter is embedded within a metamodel simulation-based optimization (SO) algorithm. It provides analytical and differentiable structural information of the underlying problem to the SO algorithm. Hence, the algorithm no longer treats the simulator as a black-box.

The analytical model is formulated as a system of nonlinear equations that can be efficiently evaluated with standard solvers. The dimension of the system of equations scales linearly with network size. It scales independently of the dimension of the route choice set and of link attributes such as link length. Hence, it is a scalable formulation suitable for the optimization of large-scale networks. For instance, the model is used in the case study of the paper for toll optimization of a Singapore network with over 4050 OD (origin-destination) pairs and 18200 feasible routes. The corresponding analytical model is implemented as a system of 860 nonlinear equations.

The analytical network model is validated based on one-dimensional toy network problems. It captures the main trends of the simulation-based objective function, and more importantly, accurately locates the global optimum for all experiments. The proposed SO approach is then used to optimize a set of 16 tolls for the network of expressways and major arterials of Singapore. The proposed method is compared to a general-purpose algorithm. The proposed method identifies good quality solutions at the very first iteration. The benchmark method identifies solutions with similar performance after 2 days of computation or similarly after more than 30 points have been simulated. The case study indicates that the analytical structural information provided to the algorithm by the analytical network model enables it to: (i) identify good quality solutions fast, (ii) become robust to both the quality of the initial points and to the stochasticity of the simulator. The final solutions identified by the proposed algorithm outperform those of the benchmark method by an average of 18%.

Key words: toll optimization, simulation-based optimization
History:

1. Introduction

Transportation demand management (TDM), also referred to as travel demand management or traffic demand management, consists of strategies to reduce or redistribute travel demand, in time or space, such as to improve, for instance, the efficiency or the sustainability of the transportation network. The temporal and spatial dimensions of travel (e.g., departure time, mode, route) and even the decision of whether or not to travel can be shaped through TDM (Saleh and Sammer 2009). TDM strategies include pricing and incentives (negative pricing). Pricing strategies are deployed more widely than incentives. Congestion pricing has been extensively studied. A recent review of congestion pricing methods is given in de Palma and Lindsey (2011). Toll optimization is one of the most studied types of congestion pricing. Both offline and online toll optimization strategies have been proposed. Offline strategies can be static (i.e., they yield a single toll for the entire time horizon of interest at a given toll location) or dynamic (i.e., they yield a time-dependent toll, such as in time-of-day tolling). Online strategies can be reactive or proactive (also known as anticipatory or predictive). Reactive strategies use observed traffic conditions to determine tolls, while proactive strategies combine both observed and predictive traffic conditions to determine the tolls.

Table 1 summarizes some of the recent toll optimization literature. For each paper (i.e., each row), the table indicates whether it considers an offline static problem, an offline dynamic (i.e., time-dependent tolls) problem, an online problem (whether reactive or proactive). The table also indicates if the traffic model used for toll optimization is analytical or simulation-based, the dimension of the toll vector and a summary of the network of the largest case study in the paper.

The table indicates that the focus of recent work has mostly been on online (i.e., real-time) problems. Nonetheless, such approaches are mostly limited to simple applications that consider a single corridor with its neighboring arterials. Few approaches, including Gupta *et al.* (2016), Chen *et al.* (2016), and Vu *et al.* (2018), have considered more intricate and large-scale network topologies. Most literature has focused on low-dimensional problems. The recent works of Gupta *et al.* (2016) and Vu *et al.* (2018) consider a higher-dimensional problem that optimizes, respectively, 13 and 16 tolls distributed throughout Singapore.

	Offline		Online	Analytical.	Simulation-based	Toll	Network
Study	Static	Dynamic					
Lou *et al.* (2011)			✓	✓		1	One highway segment
Dong *et al.* (2011)			✓		✓	2	One part of the I-95 corridor and its surroundings with 3459 links
Hassan *et al.* (2013)			✓		✓	2	One corridor along with its adjacent arterials
Jang *et al.* (2014)			✓	✓		1	One 14-mile highway corridor in the San Francisco area
Toledo *et al.* (2015)			✓	✓		1	One 14-km highway corridor
Zheng *et al.* (2016)		✓		✓		1	Area-based toll for Sioux-Falls network
Chen *et al.* (2016)		✓			✓	5	Network of highways and arterials in Maryland (USA) with 2158 links
Gupta *et al.* (2016)			✓		✓	13	Network of Singapore's expressways and major arterials with 1150 links
Liu *et al.* (2017)	✓			✓		7	13 link network
Han *et al.* (2017)		✓		✓		10	10 link network
Vu *et al.* (2018)			✓		✓	16	Network of Singapore's expressways and major arterials with 1150 links
Chen *et al.* (2018)		✓			✓	5	Network of highways and arterials in Maryland (USA) with 2158 links
Zhang *et al.* (2018)			✓		✓	2	Highway corridor in Texas with 167 links
This paper	✓				✓	16	Network of Singapore's expressways and major arterials with 1150 links

Table 1 Recent toll optimization literature

The congestion pricing literature has extensively studied the importance of accounting for a detailed description of travel demand. For instance, accounting for the heterogeneity in the value of time across the population is important (Lou *et al.* 2011, Jang *et al.* 2014, Gupta *et al.* 2016). Furthermore, Vu *et al.* (2018) have enhanced the methods to account for elasticity of travel demand such that the travelers can change their mode, departure time and even cancel their trip in response to tolls. The work of Vu *et al.* (2018) considers distance-based tolling. The increased complexity of the demand modeling component has fostered the use of simulation-based models, which can embed detailed probabilistic travel demand models with random coefficients, such as value of time.

When the toll optimization problem is formulated as a simulation-based optimization problem, the most common approaches are the use of black-box (i.e., general-purpose) algorithms such as genetic algorithms (Gupta *et al.* 2016, Vu *et al.* 2018) in combination with a constrained local search technique (Zhang *et al.* 2018). Such approaches can be directly used to address a variety of formulations (e.g., changes in the objective function or the feasible region can be readily accounted for). Nonetheless, this limits their performance under tight computational budgets or small samples. Given the computational cost of running high-resolution or large-scale traffic simulators, algorithms that can yield good quality solutions within few simulations are essential to address intricate optimization problems such as toll optimization.

The recent approach of Chen *et al.* (2016) has considered a metamodel approach to address the simulation-based optimization (SO) problem. A general-purpose Kriging metamodel is used to address a 5-dimensional problem for a large-scale network with non-linear network topology. They also apply similar simulation-based optimization techniques for improving the travel time reliability of the network (Chen *et al.* 2018). The advantage of using a general-purpose metamodel is that the approach can be directly applied to a variety of problem formulations. Nonetheless, this generality also comes with the need to run a significant number of simulations prior to optimization. For instance, for the 5-dimensional case study of Chen *et al.* (2016), a set of 100 points are simulated prior to optimization. This limits the use of the approach for high-dimensional problems. Distance-based toll optimization problems have also been receiving recent attention, Liu *et al.* (2017) provide a review.

This paper focuses on the design of toll optimization problems with the following characteristics. First, we consider large-scale road networks, such that the large-scale (e.g., across a full city or metropolitan region) impact of the tolls is accounted for. In particular, the use of a large-scale network model allows to capture the impacts of the tolling on traffic assignment. This is particularly important when considering commuting patterns, where travelers with long-distance commutes will react to changes in tolls throughout the network. Second, we consider networks with intricate topologies. In particular, we aim to go beyond the analysis of linear-topology corridor studies. Third, we consider high-dimensional problems such that various tolls distributed throughout the network are simultaneously or jointly optimized. This is important such as to account for the global (or joint) impact of tolling on traffic assignment. This also allows to coordinate tolls such as to account for equity considerations. Fourth, we focus on the use of stochastic high-resolution simulation-based traffic models. The latter allows for a detailed (e.g., probabilistic, dynamic) description of travel demand, which is essential to forecast the impact of tolls on congestion patterns.

This paper contributes to this area by focusing on the design of computationally efficient algorithms. These are algorithms that are designed to yield a good quality solution within few simulation runs. The case study of this paper considers an offline static problem. We view the design of these computationally efficient offline algorithms as the building block for efficient real-time algorithms.

Computational efficiency can be improved through parallel computations such as in Gupta *et al.* (2016), Vu *et al.* (2018). Nonetheless, as our transportation systems and users become more real-time responsive and more connected, the intricacy of both the traffic simulation tools used and of the transportation optimization problems addressed increases. Hence, there is a need for computationally efficient algorithms.

Our approach to achieve computational efficiency is to allow the algorithms to exploit problem-specific structural information. More specifically, we propose to formulate an analytical network model that approximates the mapping between the tolls and the network-wide traffic conditions. We then embed this analytical structural information within the algorithm. In other words, this analytical information is combined with simulation-based information to identify suitable toll vectors.

This paper proposes an SO algorithm that enables high-dimensional toll optimization problems for large-scale networks to be addressed in a computationally efficient way. The essential component of the proposed methodology is the formulation of an analytical network model which provides an analytical and differentiable mapping between the toll vector and the network-wide performance metrics (such as revenue and traffic conditions). The proposed formulation is scalable, and hence suitable for large-scale networks. More specifically, for a network with n links, the analytical network model is formulated as a system of n nonlinear equations. Importantly, the analytical network model has endogenous traffic assignment, yet its complexity (i.e., the dimension of the corresponding system of equations) scales independently of the dimension of the route choice set, of link attributes (e.g., link length, number of lanes), and of origin-destination matrix dimensions. The proposed formulation is embedded within a metamodel SO algorithm and is used to address a high-dimensional offline toll optimization problem for a large-scale Singapore network. More specifically, 16 tolls that are distributed throughout the network of Singapore expressways and major arterials are optimized. The network is modeled as set of over 1150 links, 2300 lanes, 4050 OD pairs with over 18000 routes.

Section 2 presents the proposed methodology. Validation experiments are presented in Section 3, followed by a Singapore case study in Section 4. Conclusions are discussed in Section 5.

2. Methodology

2.1. Problem formulation

To formulate the toll optimization problem, we introduce the following notation:

$\boldsymbol{x}$	decision vector (i.e., toll vector) in a given currency unit;
$f(\boldsymbol{x})$	simulation-based objective function;
$F_i(\boldsymbol{x})$	hourly flow on link i;
$E[F_i(\boldsymbol{x})]$	expected hourly flow on link i;
$\boldsymbol{x}_L$	lower bound vector;
$\boldsymbol{x}_U$	upper bound vector;
$\mathcal{T}$	set of links with tolls.

The problem is formulated as follows:

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}) = \sum_{i \in \mathcal{T}} x_i E[F_i(\boldsymbol{x})] \tag{1}$$

$$\boldsymbol{x}_L \leq \boldsymbol{x} \leq \boldsymbol{x}_U. \tag{2}$$

The decision vector $\boldsymbol{x}$ is the vector of toll rates, with element i denoted by x_i. For the case study of Section 4, the dimension of $\boldsymbol{x}$ is 16. In other words, the tolls of 16 distinct locations are determined simultaneously. The objective function $f(\boldsymbol{x})$ (Eq. (1)) is an unknown simulation-based function, which in this paper is the expected hourly revenue. Term i in the summation represents the expected hourly revenue for toll i. The latter is defined as the product of the toll and of the expected number of vehicles per hour that travel on link i. This expected link flow is a function of the tolls because the tolls impact traffic assignment (i.e., when choosing their routes, the travelers account for the toll costs). This expectation is an unknown function, which is estimated via simulation. The toll optimization problem considers lower and upper bound constraints for the tolls (Eq. (2)). These bounds are available in analytical form, i.e., they are not simulation-based constraints. This problem formulation considers the main incentive from the perspective of the toll operator: revenue maximization. The proposed methodology is suited to address other problem formulations which account explicitly for how tolls impact network performance. This can be done by including in the objective function, additional network-wide performance metrics such as travel times and speeds. In summary, the problem is an SO problem with continuous decision variables, simulation-based objective function and analytical deterministic bound constraints. Other problem formulations that account, for instance, for the level of service or the social welfare can also be addressed with the proposed methodology.

In practice, dynamic tolls throughout a network are determined independently (i.e., the spatial dependencies between tolls is not explicitly accounted for in the optimization). Commonly used approaches in practice are rule-based approaches or lookup tables that determine tolls based on the prevailing traffic conditions. An example is the dynamic tolling at I-394 in the Minneapolis-St. Paul metropolitan area (Halvorson *et al.* 2011). Recent literature has addressed the spatial dependencies between tolls. However, many recent studies focus on the analysis of a single toll, as illustrated in Table 1. The proposed formulation simultaneously determines all tolls in the network. In the Singapore case study of this paper, a set of 16 tolls distributed throughout Singapore are jointly or simultaneously determined. This allows to account for the joint impact of tolls on the spatial and temporal propagation of congestion. For instance, a toll on an upstream link influences flow on downstream links, which influences downstream tolls. This illustrates the importance of accounting for these dependencies between the tolls for toll optimization. More generally, it illustrates the intricate mapping between the toll vector, the underlying traffic dynamics and the resulting objective function.

In practice, the toll bounds (Eq. (2)) are typically given by the context. The upper bound is usually set by regulations. It is not a decision variable. As an example of the discussion of such regulations we refer to Albalate *et al.* (2009) and an example from the US context is given by Zhang *et al.* (2018). The lower bound can be set based on tolling policies. For example, it can be set such as to avoid large toll fluctuations across time intervals.

2.2. Metamodel SO algorithm

We use in this paper the general metamodel SO approach of Osorio and Bierlaire (2013). To briefly describe its main ideas, we introduce the following notation. The index k refers to a given SO algorithm iteration.

m_k	metamodel function;
$\boldsymbol{\beta}_k$	parameter vector of metamodel m_k;
$\beta_{k,j}$	element j of the parameter vector $\boldsymbol{\beta_k}$;
$f_A(\boldsymbol{x})$	approximation of the SO objective function provided by an analytical traffic model;
$\phi(\boldsymbol{x};\boldsymbol{\beta}_k)$	polynomial component of the metamodel m_k;
$h(\boldsymbol{x};p)$	analytical traffic model formulated as a system of nonlinear equations.

To address Problem (1)-(2), each iteration k of the SO algorithm solves a metamodel optimization problem of the following form:

$$\max_{\boldsymbol{x}} m_k(\boldsymbol{x};\boldsymbol{\beta}_k) = \beta_{k,0} f_A(\boldsymbol{x}) + \phi(\boldsymbol{x};\boldsymbol{\beta}_k) \tag{3}$$

$$h(\boldsymbol{x};p) = 0 \tag{4}$$

$$\boldsymbol{x}_L \leq \boldsymbol{x} \leq \boldsymbol{x}_U. \tag{5}$$

Problem (1)-(2) differs from Problem (3)-(5) in two main ways. First, the SO objective function, $f(\boldsymbol{x})$, is replaced by an analytical, and differentiable, function known as the metamodel m_k. The latter is a parametric function that is defined as the sum of: (i) an analytical approximation of $f(\boldsymbol{x})$ provided by an analytical traffic model ($f_A(\boldsymbol{x})$ term), and (ii) a polynomial function (ϕ term). The metamodel can be seen as the approximation of f provided by an analytical traffic model and corrected for with both a scaling factor ($\beta_{k,0}$) and an additive error term (ϕ term). The polynomial, ϕ, is defined as a quadratic function with diagonal second-derivative matrix. More specifically:

$$\phi(\boldsymbol{x};\boldsymbol{\beta}_k) = \beta_{k,1} + \sum_{j=1}^{T} x_j \beta_{k,j+1} + \sum_{j=1}^{T} x_j^2 \beta_{k,j+T+1}, \tag{6}$$

where T is the dimension of the decision vector $\boldsymbol{x}$. Equation (6) defines the function ϕ of (3), which is the general-purpose (i.e., not problem-specific) component of the metamodel. Second, it has an additional set of constraints (Eq. (4)) which represent the analytical traffic model. The latter is formulated as a system of nonlinear equations.

Each iteration k of the SO algorithm carries out the following main steps: (i) use all simulation observations collected so far (i.e., all estimates of $f(\boldsymbol{x})$) to fit the parameter vector, $\boldsymbol{\beta}_k$, of the metamodel (the least squares problem that is solved to fit $\boldsymbol{\beta}_k$ is detailed in Appendix B, it aims to minimize a distance function between the metamodel predictions and the simulated estimates of $f(\boldsymbol{x})$); (ii) solve the metamodel optimization problem (3)-(5); (iii) simulate new points (for instance, simulate the optimal solution of the metamodel optimization problem).

The main component of the metamodel which can enable it to address high-dimensional and intricate SO problems (e.g., non-convex) in a computationally efficient way is the analytical traffic model approximation (i.e., f_A term). This term provides a problem-specific approximation of f, while the polynomial provides a general-purpose approximation. In other words, depending on the choice of f (e.g., revenue, consumer surplus, etc.) the functional form of f_A will vary, while that of ϕ will not. The analytical traffic model provides analytical structural information to the SO algorithm. More specifically, it provides an analytical and physically plausible (i.e., problem-specific) approximation of the mapping between the decision vector and the objective function. . Traditional SO algorithms treat the simulator as a black-box. The use of f_A enables problem- and network-specific information to be provided to the algorithm. Hence, the simulator is no longer treated as a black-box.

The main challenge in this metamodel SO approach is the formulation of an analytical traffic model that has the following properties. First, it should provide a good approximation of the unknown simulation-based objective function f (such that the optimal solutions to Problem (3)-(5) are good quality solutions to Problem (1)-(2)). Second, it should provide a good global approximation (i.e., a good approximation in the entire feasible region). This differs from local metamodels such as polynomials. Third, it should be computationally efficient to evaluate. Since Problem (3)-(5) is solved at *every* iteration of the SO algorithm it needs to be solved efficiently (otherwise, we are better off allocating the computing resources to running additional simulations). Fourth, it should be a scalable model such that it can be used for large-scale networks, i.e., the System of Equations (4) needs to be efficiently evaluated for large-scale networks. Fifth, it should be differentiable such that Problem (3)-(5) can be solved with a variety of traditional gradient-based algorithms. In Section 2.3, we formulate an analytical traffic model with all of the above properties for toll optimization problems. This general metamodel SO idea has been formulated and used to design efficient algorithms for various transportation problems, including various traffic signal control problem (Chong and Osorio 2017, Osorio *et al.* 2017, Osorio and Nanduri 2015, Osorio and Chong 2015), and more recently for model calibration problems (Zhang *et al.* 2017).

2.3. Traffic model formulation

To formulate the analytical traffic model, we introduce the following additional notation:

	Endogenous variables of the analytical traffic model:
y_i	expected hourly demand per lane of 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;
z_r	toll cost for route r;
$P(r)$	route choice probability for route r.

	Exogenous parameters of the analytical traffic model:
d_s	expected hourly travel demand for OD pair s;
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;
θ_1, θ_2	coefficients of the route choice model;
$\alpha_{1,i}, \alpha_{2,i}$	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.
$\mathcal{T}$	set of links with tolls.

The analytical traffic model is formulated as follows:

$$\begin{cases} f_A(\boldsymbol{x}) = \sum_{i \in \mathcal{T}} x_i n_i y_i & \text{(7a)} \\ y_i = \dfrac{1}{n_i} \sum_{r \in \mathcal{R}_1(i)} P(r) d_{O(r)} & \text{(7b)} \\ P(r) = \dfrac{e^{\theta_1 t_r + \theta_2 z_r}}{\sum_{j \in \mathcal{R}_2(O(r))} e^{\theta_1 t_j + \theta_2 z_j}} & \text{(7c)} \\ t_r = \sum_{i \in \mathcal{L}(r)} t_i & \text{(7d)} \\ z_r = \sum_{i \in \mathcal{L}(r) \cap \mathcal{T}} x_i & \text{(7e)} \\ t_i = \dfrac{\ell_i}{v_i} & \text{(7f)} \\ v_i = v_i^{\max} \left(1 - \left(\dfrac{k_i}{k_i^{\text{jam}}}\right)^{\alpha_{1,i}}\right)^{\alpha_{2,i}} & \text{(7g)} \\ k_i = c \dfrac{k_i^{\text{jam}}}{q^{\text{cap}}} y_i. & \text{(7h)} \end{cases}$$

Eq. (7a) is the approximation of the expected hourly revenue provided by the analytical traffic model. It is the analytical counterpart of Eq. (1), where the simulation-based expected hourly flow of link i, $E[F_i(\boldsymbol{x})]$, is replaced by the analytical expected hourly demand of link i. The latter is defined as the product of the hourly demand per lane of link i, y_i, and the number of lanes on link i, n_i. The hourly demand per lane of link i is defined by Eq. (7b) as the ratio of the sum of the expected demand for all routes that travel through link i and of the number of lanes of link i. The expected route demand is the product of the route choice probability and the total demand for the underlying OD (Origin-Destination) pair. The route choice probability (Eq. (7c)) is approximated as a multinomial logit model with a utility that is a function of the route's expected travel time and the route's toll cost. The expected route travel time (Eq. (7d)) is the sum of the expected travel times of its links. The route toll cost is the sum of the tolls that are on the given route (Eq. (7e)). Equation (7f) defines the expected link travel time as the ratio of the link length and the expected (space-mean) link speed. The fundamental diagram of the link is given by Eq. (7g). It relates link densities to link speeds. Finally, Eq. (7h) relates the link demand to the link density. The underlying assumption is that the ratio of density to jam density is proportional to the ratio of hourly demand to flow capacity. The proportionality constant, c, is common for all links. In this model, a common flow capacity value is used for all links. This System of Equations (7) defines the system of equations denoted h in Eq. (4).

In summary, the above analytical traffic model provides a simplified (compared to the simulator) description of how tolls impact the spatial distribution of vehicular flow. The System of Equations (7) accounts for the dependencies between tolls (recall discussion in Section 2.1) through an analytical description of how tolls impact route choices, which in turn impact the spatial propagation of congestion and the corresponding toll revenues.

The above model (Eq. (7)) is implemented for a network with n links as a system of n nonlinear equations. In other words, the complexity of the model (i.e., the dimension of the corresponding system of equations) scales linearly with the number of links in the network, and is independent of other link attributes such as link length or lane attributes. What is particularly remarkable is that the analytical model has endogenous traffic assignment (i.e., endogenous route choice), yet its complexity does not depend on the size of the route choice set. For instance, for the Singapore case study of this paper, the analytical model considers a network with 860 links, 4050 OD pairs and over 18,200 routes; and is implemented as a system of 860 nonlinear equations. This is a scalable formulation suitable for large-scale network optimization.

A description of how this analytical model differs from the simulation model used for the case studies of this paper is given in Appendix A. The exogenous parameters of the analytical model are calibrated as follows. The simulator considers multi-lane links, with heterogeneous lanes, while the analytical model assumes all lanes of a link to be homogeneous. Hence, the link attributes ($k_i^{\text{jam}}, v_i^{\max}$ and ℓ_i) of the analytical model are estimated as the average, over all lanes of the link, of the individual lane attributes of the simulator. The parameters of its fundamental diagram of each link of the analytical model are defined based on those of the simulator, the specific equation that governs this is detailed in Appendix A. For all lanes, the analytical model uses a single lane flow capacity value (denoted q^{cap}), its value is set as the maximum flow capacity value, among all all of the simulators lanes. The simulator has a pre-determined route choice set, which defines the route choice set of the analytical model. Additional details on this are provided in Appendix A. The simulator allows for time-dependent OD demand matrices. The time-independent OD demand, d, of the analytical model is obtained as the average, over the time period of interest, OD demand. For the case studies of this paper, the proportionality coefficient c of Eq. (7h) is kept to the same value, which is set to 1/6. This value was obtained through insights from numerical experiments on a toy network. As is described in Appendix A, the simulator uses a route choice model with a probabilistic value of time, while the analytical model considers a deterministic value of time. The coefficients of the route choice model (denoted θ_1 and θ_2), include the value of time, are estimated as the expected value of the probabilistic coefficients of the simulator's route choice model.

3. Validation

We carry out experiments on a synthetic small toy network as illustrated in Figure 1. The network has 1 OD pair, and 3 multi-lane links in a diverge node topology. In other words, all trips have a common origin link. After traveling on link 1, the travelers can choose between links 2 and 3. The only difference between links 2 and 3 is that link 2 is tolled. This is a one-dimensional problem (i.e., a scalar toll value). We consider experiments with 2 different value of time (VOT) values (\$15/h and \$30/h). For each of the experiments, we consider four different demand scenarios. Three demand scenarios have a constant OD demand with values 3600, 4800 and 6000 vehicles per hour, respectively. The fourth demand scenario considers time-dependent OD matrices, where demand increases gradually (from 3600 to 6000), and the average demand is 4800 vehicles per hour. These experiments cover various levels of congestion ranging from free-flowing conditions to congested conditions.

Figure 2 displays eight plots. The top plots display the objective function (i.e., expected revenue function f of Eq. (1)) estimates obtained via simulation. The bottom plots display

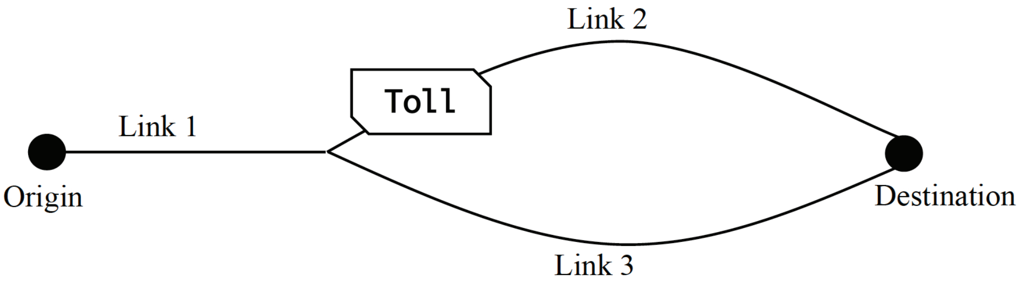


Figure 1 **Toy network topology**

the approximations derived by the analytical network model (i.e., function f_A of Eq. (3)). A given row of plots displays, from left to right, the demand scenarios with constant demand 3600, constant demand 4800, constant demand 6000 and increasing demand with average 4800, respectively. Each plot displays two curves that correspond to a given VOT value: a VOT of 15 corresponds to the solid curve, a VOT of 30 corresponds to the dashed curve. Note that the different plots have different y-axis limits. For a given demand scenario (i.e., a given plot), as the VOT increases, the maximum revenue increases and the toll for which maximum revenue is obtained also increases. This holds for all demand scenarios. This trend is well replicated by the analytical model. For a given demand scenario (i.e., when comparing a given column of plots), the analytical model has an accurate approximation of the value of the optimal toll. This holds for all demand scenarios, with both constant and time-dependent demand. Nonetheless, the magnitude of the revenue functions differ. The analytical model tends to overestimate the simulated revenue.

Figure 3 considers the same set of experiments. It now displays in the same plot, the experiments with common VOT value yet different demand scenarios. As before, the top (resp. bottom) plots correspond to simulation-based estimates (resp. analytical approximations). The four demand scenarios are represented as follows: constant demand of 3600 (blue curve), constant demand of 4800 (black curve), constant demand of 6000 (green curve), increasing demand with an average of 4800 (red curve). The top plots indicate that for a given VOT value, as the average demand increases, so does the maximum revenue, yet the value of the optimal toll does not vary much. The analytical model captures these trends, as indicated by the bottom plots. Since the analytical model is time-invariant (in particular, it is a stationary model), the revenue values for both demand scenarios (constant demand and increasing demand) with an average demand of 4800 are identical. The simulator, which is a dynamic model, also yields similar expected revenue functions for these two demand scenarios.

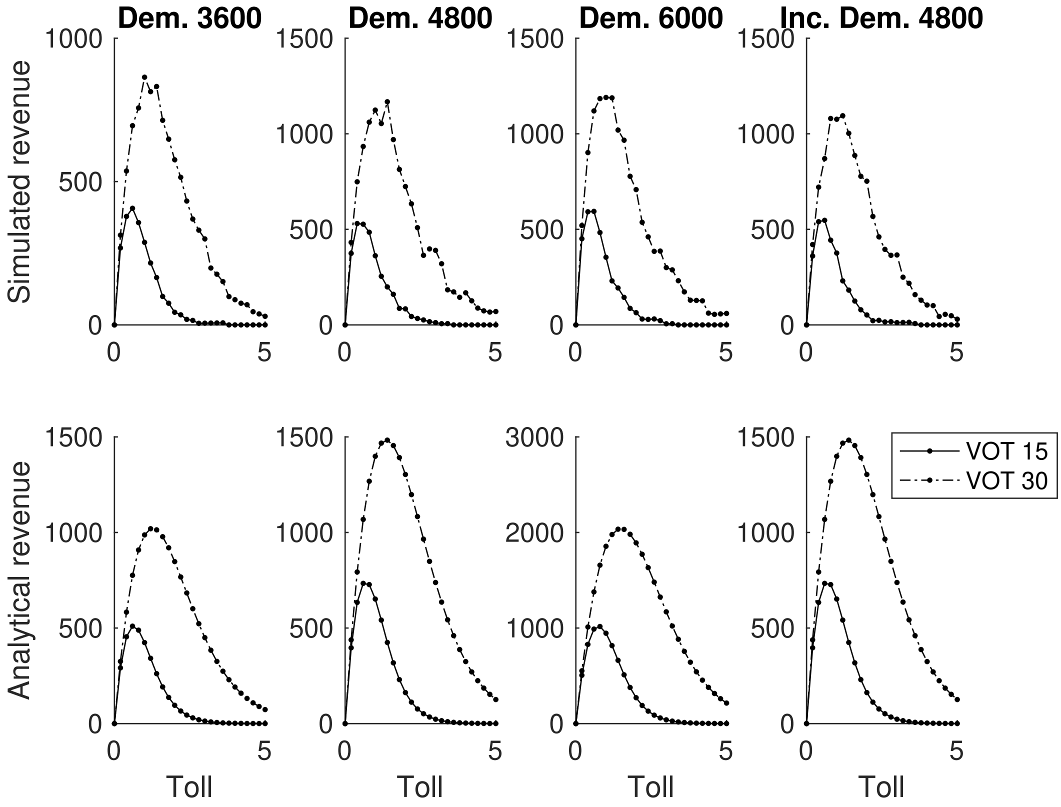


Figure 2 Comparison of the simulation-based objective function and the analytical objective function for various demand and value of time scenarios

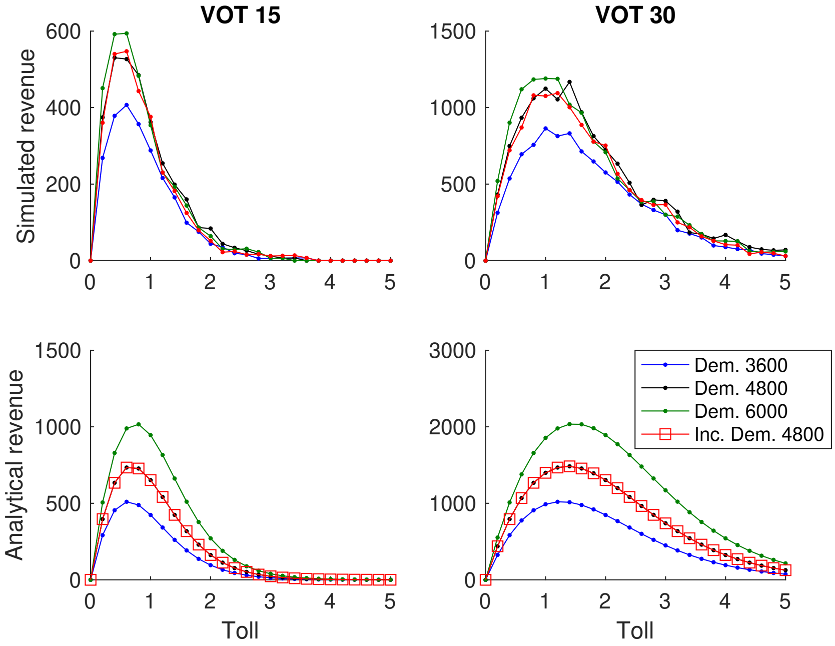


Figure 3 Comparison of the simulation-based objective function and the analytical objective function for various demand and value of time scenarios

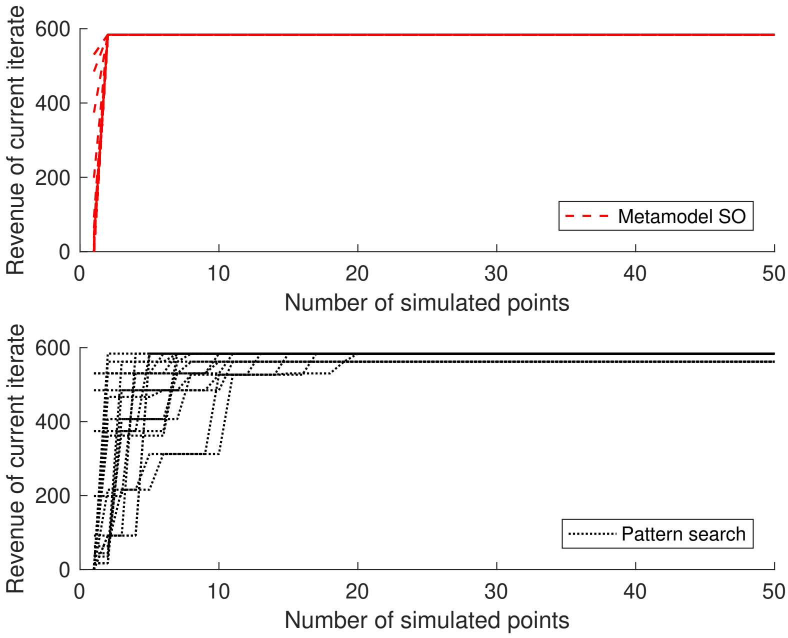


Figure 4 **Comparison of the performance of the algorithms for a toy network toll optimization problem**

We address the toll optimization problem for this toy network. We consider a VOT value of \$15/h and a constant demand of 4800 vehicles per hour. We compare our proposed approach with the derivative-free generalized pattern search algorithm (Mathworks, Inc. 2016). This algorithm was chosen as a benchmark for the following reasons. As a derivative-free algorithm, it does not rely on first-order derivative information, which can be computationally costly to estimate. It is suitable for non-continuous and non-differentiable objective functions, as is the case of our simulation-based objective function (the traffic simulator relies on non-continuous and non-differentiable functions). Moreover, recent work has highlighted the good performance of similar direct search techniques when used under tight computational budgets (Dong *et al.* 2017).

We consider 20 different initial points, which are uniformly drawn from the feasible region (Eq. (2)). For each initial point, we run each method (i.e., algorithm) allowing for a computational budget of 50 simulations (i.e., the algorithm is terminated once a total of 50 simulations are evaluated).

Figure 4 displays the performance of each method considering each of its 20 runs. Each plot displays the estimate of the objective function of the current iterate (i.e., the point considered to have best performance so far) as a function of the total number of simulated points (i.e., computational budget used so far). The top (resp. bottom) plot considers the proposed (resp. benchmark) method.

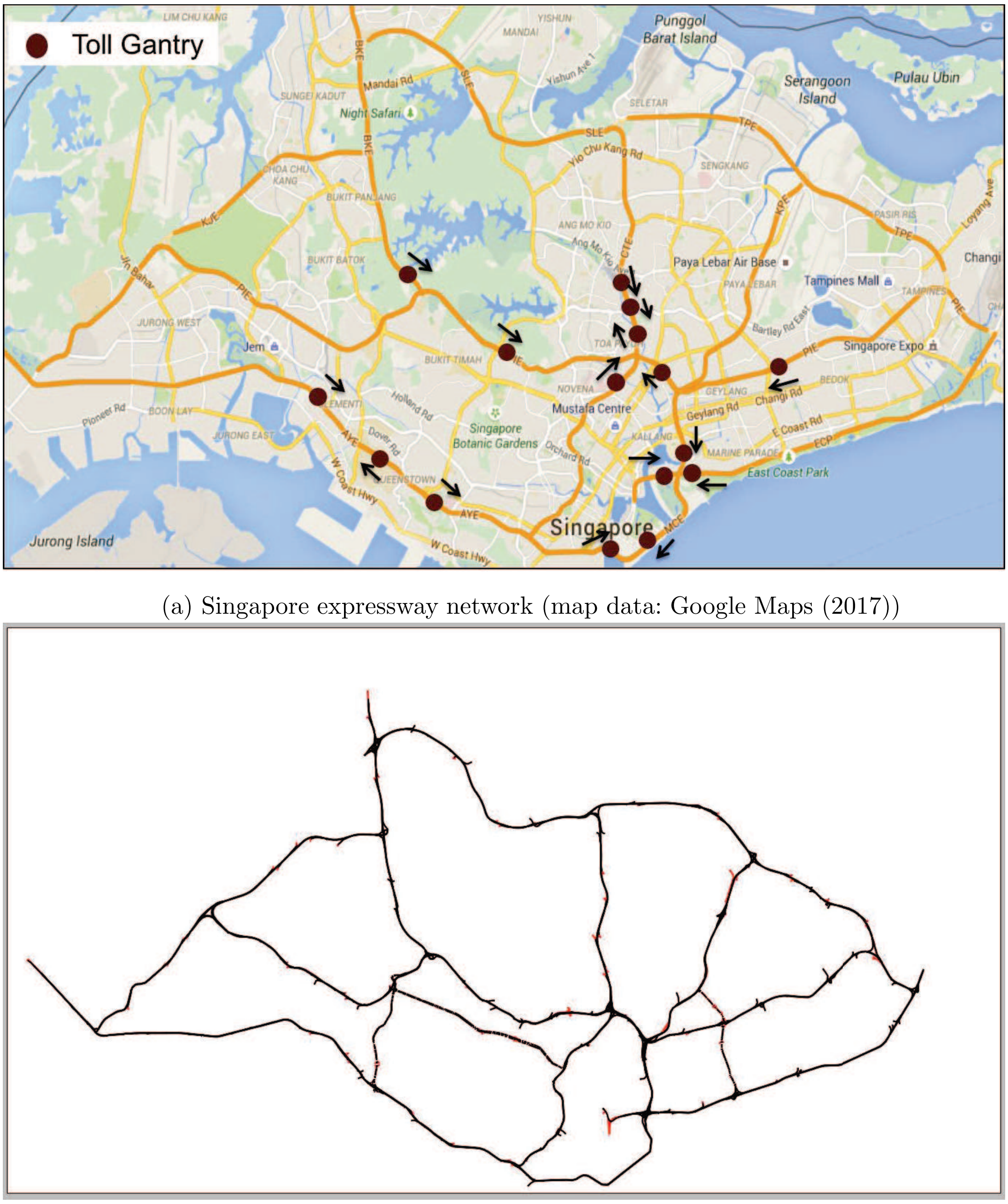
The top plot indicates that all 20 runs of the proposed method identify at iteration 1 a solution with excellent performance. This shows the ability of the proposed method to perform well regardless of the quality of the initial solution. In other words, it is robust to the quality of the initial point. The benchmark method gradually finds points that have improved performance compared to the initial point. Its performance for low computational budgets (e.g., 10 simulation runs) is sensitive to the quality of the initial point. For larger budgets, it has a performance similar to that of the proposed method. All 20 runs of the proposed method converge to the same final solution, which is the global optimal solution. For the benchmark method, 16 of the 20 runs converge to this value, the other 4 runs converge to a local optimum with an estimated objective function that is 4% lower than the global optimal solution.

4. Singapore case study

We consider a case study with a network model of the major arterials and expressways of Singapore as seen in Figure 5a. We consider a set of 16 tolls to optimize. The 16 toll gantries are distributed throughout the network and labeled in the figure with red circles. The expressway links are represented with orange lines, the major arterials are represented with yellow lines. The simulation model of the network is displayed in Figure 5b. It accounts for all expressway links and a few arterial links. Currently, Electronic Road Pricing (ERP) is deployed in Singapore such that travelers are charged at gantries when they enter a zone. The toll to be charged varies across the day (time-of-day tolling) and location. The pricing is pre-determined: it is reviewed quarterly and published so that the travelers know exactly the toll they will pay. We refer to Seik (2000) for the history of the ERP system in Singapore.

The network model consists of 1150 links with over 2300 lanes, 4050 OD pairs with over 18000 routes. Note that of the 1150 links of the simulator only 860 belong to routes of the pre-determined route choice set. Details on how this choice set is defined are given in Appendix A.

Hence, the analytical model only accounts for those 860 links. Demand is defined, in the simulator, as calibrated time-dependent OD matrices for every 5 minute interval. We simulate a weekday 8am-9:30am period, during this period expected travel demand is of the order of 226247 trips. We determine the tolls such as to optimize revenue for the 9am-9:30am period. We set the lower toll bound to its smallest value of S$0 and upper toll bound to a relatively high value of S$5 (Singapore dollars). As a reference, toll values in Singapore for the period of May-Aug 2018 are provided by the Land Transport Authority



(a) Singapore expressway network (map data: Google Maps (2017))

(b) Simulation network model

Figure 5 **Singapore network**

(2018) and toll values tend to be significantly lower than S$5. We consider 5 initial points. Each initial point is uniformly drawn from the feasible region (Eq. (2)). For a given point, we run each algorithm once and allow for a computational budget of 80 simulations.

We compare the performance of the algorithms as a function of the number of simulated points (i.e., the amount of computational budget depleted). This allows for a comparison that is hardware (e.g., computational resources) and software (e.g., code implementation) agnostic. We also compare their performance as a function of the total computation time.

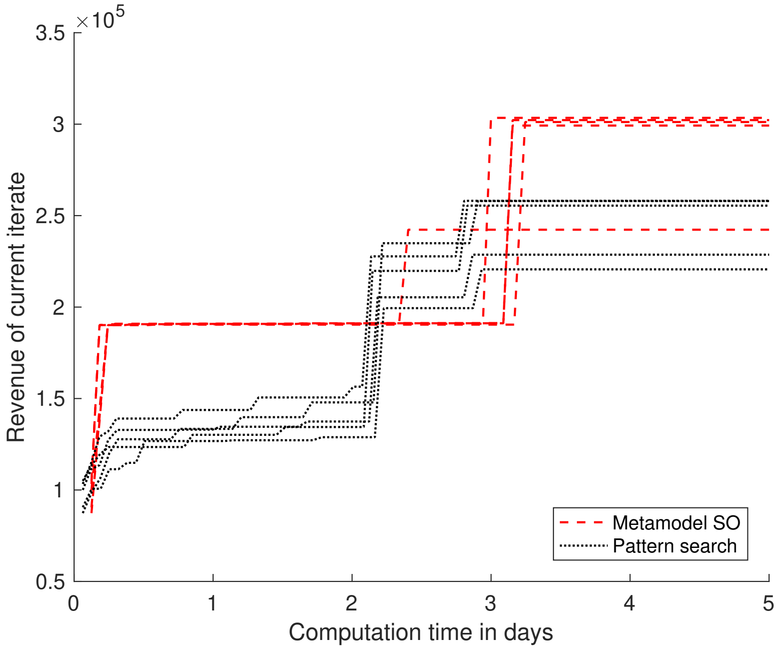


Figure 6 **Revenue of the current iterate as a function of the cumulative computation time**

The runs are carried out on various servers, which do not guarantee equal CPU allocation to each job. Hence, the computation times are indicative of efficiency yet the differences may not be solely attributable to the algorithm.

Figure 6 displays the performance of each algorithmic run as a function of computation time. The red dashed lines represent the proposed metamodel SO method. The black dotted lines represent the benchmark method. For each algorithm, there are 5 lines, which correspond to 5 different initial points. The x-axis displays the total computation time in days. The y-axis displays the simulation-based estimate of the objective function of the current iterate. Note that on average one simulation run takes approximately 2 hours to compute. Figure 6 indicates that all 5 runs of the proposed method significantly outperform the benchmark runs for the first two days of computation. Then, there is a phase where all runs of the benchmark outperform 4 runs of the proposed method. After day 3.5, four of the proposed method runs outperform all benchmark runs. The fifth run outperforms two of the five benchmark runs.

This figure illustrates that the proposed method often (in four out of five instances) identifies final solution that significantly outperform the solutions proposed by the benchmark method. When comparing the objective function value of the final solution and averaging over the 5 runs, the proposed method yields an average improvement of 19% (average revenues of 289678 versus 244158). Similarly, when comparing the revenue of the best solution (out of the 5 solutions) proposed by each method, the proposed method yields an improvement of 18% (revenue of 303450 versus 258090). The *best* solution is defined as that with the highest objective function estimate. Four of the five proposed solutions (i.e., solutions derived by the proposed method) have very similar objective function values. This indicates robustness of the method to both the quality of the initial point and the stochasticity of the simulator. There is higher variability among solutions derived by the benchmark method.

During the first 2 days of computation the benchmark method slowly finds solutions that gradually improve performance. As of day 2, it finds solutions with performance comparable to those of the proposed method and that significantly outperform the initial solutions. On the other hand, the proposed method immediately (i.e., at iteration 1) identifies solutions with significantly improved performance compared to the initial solution (at iteration 1, the revenue improves on average by 95%; i.e., average revenues of 190150 versus 97229). This initial improvement is entirely due to the analytical network model. More specifically, when the algorithm starts, there are no simulation observations available, hence the first current iterate is defined as the solution to the analytical network model problem (i.e., a problem that maximizes $f_A(\boldsymbol{x})$ of Eq. (3) subject to constraints (4)-(5)). This shows the added value of the structural information provided by the analytical network model. Also note that the solutions to this analytical network model problem are the same for all 5 initial points. This shows the added value of using an analytical network model that provides a good global approximation of the objective function (rather than using local models, such as polynomials). It is this analytical network model that leads to an SO algorithm that is robust to the quality of the initial points.

Figure 7 differs from the previous figure in that it considers the performance as a function of the total number of simulated points (instead of the total computation time) (i.e., the two figures differ in their x-axis). This figure serves to compare the performance of the algorithm independently of hardware and software considerations. The lines are almost identical to those of the previous figure. All conclusions from the previous figure also hold here.

Figure 8 considers for each algorithm, the 5 solutions it yields upon termination and displays the coordinates of the solutions (i.e., toll values). Each plot contains 16 boxplots, one for each toll (i.e., each i value). For a given i, the boxplot illustrates the variability (across the 5 solutions) of the toll values, x_i. The top, middle and lower plot display, respectively, the 5 initial points, the 5 final solutions of the proposed method, and the 5 final solutions of the benchmark method. The top plot confirms that the 5 initial points

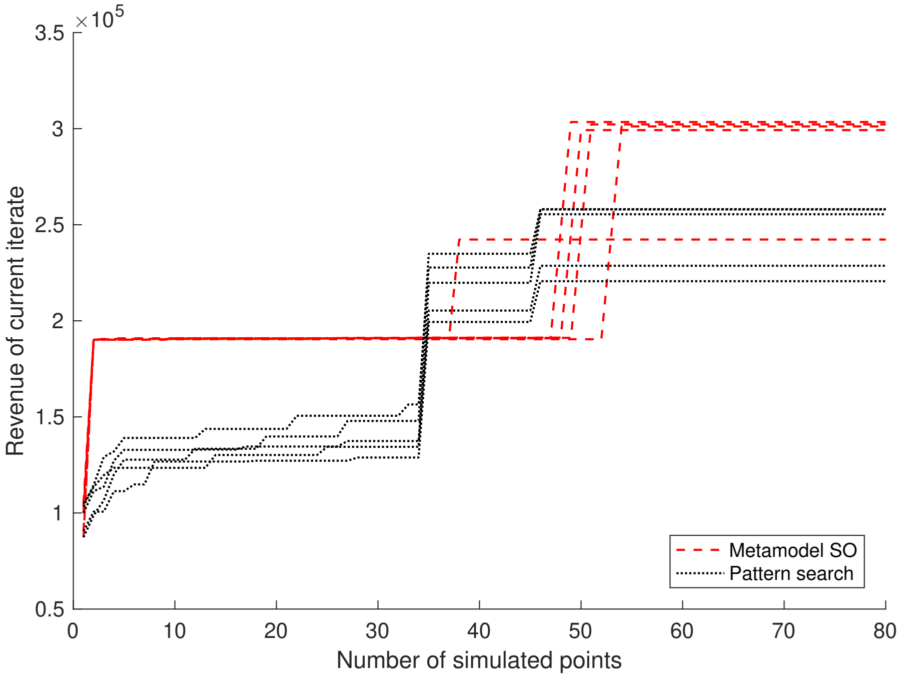


Figure 7 **Revenue of the current iterate as a function of the number of simulated points**

are very different (i.e., high variability). The middle plot shows that for most tolls there is little variability across toll values. More specifically, 13 of the 16 tolls have low variability (the tolls with high variability are indexed 5, 8 and 16). The lower plot shows that the benchmark method yields solutions with lower variability than the initial point, but with higher variability than the proposed method. More specifically, 10 of the 16 tolls have lower variability under the proposed method than under the benchmark method. For the proposed method, 8 of the tolls have a variance smaller than 0.1, while for the benchmark method, only 1 of the tolls is below this threshold.

Figure 9 considers all the points simulated across all 10 runs (5 for the benchmark method and 5 for the proposed method). For each of these points, it evaluates the analytical approximation of the objective function provided by the analytical network model. In other words, it considers a given $\boldsymbol{x}$ and solves the system of nonlinear equations (7) to obtain $f_A(\boldsymbol{x})$, which is the analytical network model approximation of $f(\boldsymbol{x})$ of Eq. (1). For each simulated point, $\boldsymbol{x}$, Figure 9 displays $f_A(\boldsymbol{x})$ along the x-axis and the simulated estimated of the expected revenue (i.e., an estimate of $f(\boldsymbol{x})$) along the y-axis. Each point (i.e., each $\boldsymbol{x}$ value) is displayed as a cross. The figure also displays the diagonal line (dashed blue line). This figure shows that most points lie along a line with positive slop. This indicates that the analytical network model provides approximations that have high positive

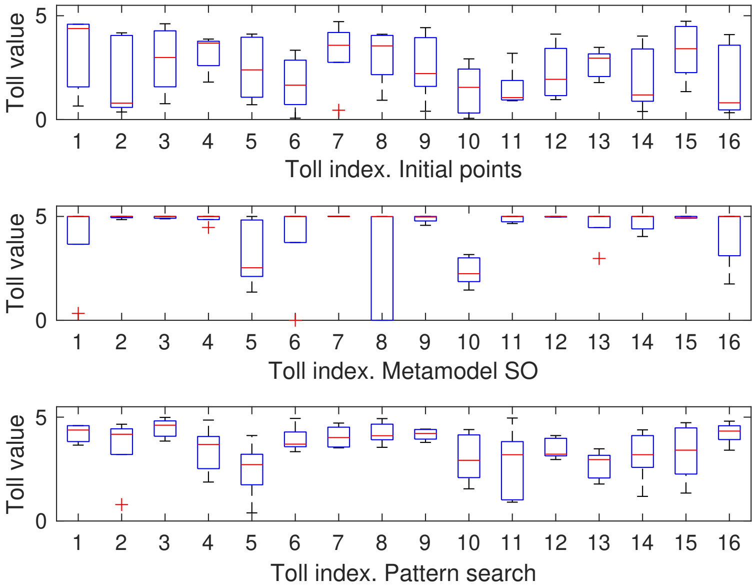


Figure 8 **Variability of the toll vectors for the initial points and the solutions proposed by each method.**

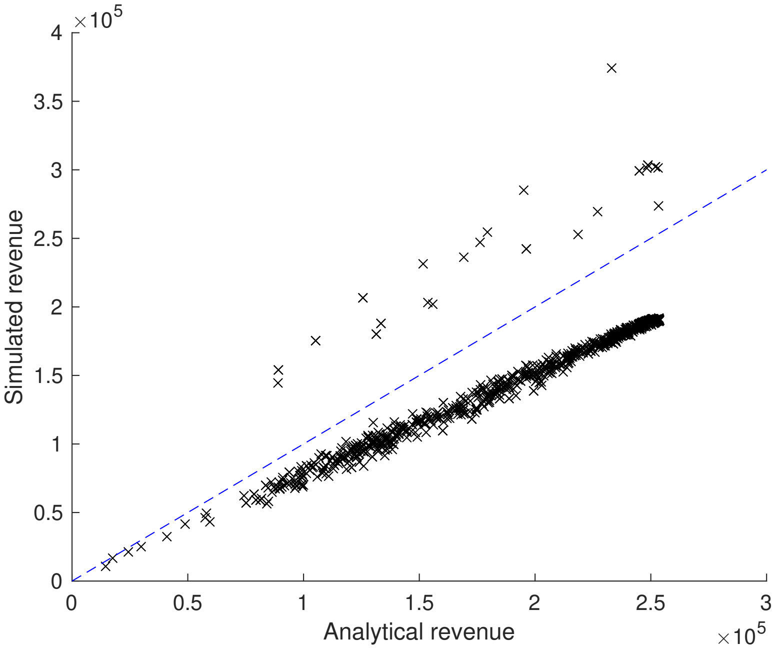


Figure 9 **Comparison of the simulated revenue and the revenue approximated by the analytical network model**

correlation with the simulation-based estimate. It is remarkable that such a simple analytical formulation captures the main trends of such an intricate simulation-based objective function.

5. Conclusions

This paper proposed a simulation-based optimization technique for high-dimensional toll optimization problems of large-scale road networks. The main novelty is the formulation of an analytical network model. The latter is embedded within a metamodel simulation-based optimization (SO) algorithm. It provides analytical and differentiable structural information of the underlying problem to the SO algorithm. Hence, the algorithm no longer treats the simulator as a black-box.

The analytical model is formulated as a system of nonlinear equations that can be efficiently evaluated with standard solvers. The dimension of the system of equations scales linearly with network size and independently of the dimension of the route choice set, and of link attributes, such as link length. Hence, it is a scalable formulation suitable for the optimization of large-scale networks. For instance, the model is used in Section 4 for toll optimization of a Singapore network with over 4050 OD pairs and 18200 feasible routes. The corresponding analytical model is implemented as a system of 860 nonlinear equations.

The analytical network model is validated based on one-dimensional problems of a toy network. It captures the main trends of the simulation-based objective function, and more importantly, accurately locates the global optimum for all experiments. The proposed SO approach is then used to optimize a set of 16 tolls for the network of expressways and major arterials of Singapore. The proposed approach identifies good quality solutions at the very first iteration of the algorithm. This is entirely due to the analytical structural information provided by the analytical network model. More specifically, at iteration 1, the proposed method improves the objective function value by 95%. The benchmark method is slow to identify solutions with good performance: this is achieved after 2 days of computation or similarly after 30 points are simulated. The proposed algorithm yields final solutions that outperform those of the benchmark method by an average of 18%. The experiments also indicate that the analytical structural information enables the SO algorithm to become robust to both the quality of the initial points and the simulators stochasticity.

This general idea of formulating and using structural analytical problem-specific information is promising for the design of real-time algorithms. Cities worldwide are increasingly collecting real-time data and adapting their operations to respond to the real-time variations of demand and of supply. Hence, there is a pressing need to develop algorithms that both: are real-time feasible and capture the network performance at a large-scale (e.g., an entire city or metropolitan region). Coupling general-purpose black-box techniques (e.g., machine learning, simulation-based optimization) with such analytical structural information is a promising approach to enable them to achieve real-time feasibility for both high-dimensional problems and large-scale networks. Such structural information can also improve the robustness of these methods (e.g., to measurement errors, to sparse data).

There are various ways in which the structural information of the analytical traffic model can be used to enhance the performance of real-time algorithms. One promising approach is to use the analytical approximation of the simulation-based objective function to design a sampling distribution that balances the traditional exploration-exploitation trade-off. Ideas along these lines have been developed, for simulation-based signal control problems, in Tay and Osorio (2017). This allows to simulate both: (i) points with good performance (i.e., good objective function estimates) and (ii) points that further explore the feasible region.

The majority of the computational effort lies in the evaluation of the simulation model. For online optimization, it is essential to find more efficient ways of performing simulation evaluations. The traditional approach is to focus on parallelization. Promising, and recently proposed ideas, include marginal simulation (Corthout *et al.* 2014) and multi-model optimization (Osorio and Selvam 2017).

This paper focused on the design of algorithms that can identify points with good performance within a tight computational budget (i.e., within few simulation runs). Hence, the algorithm is terminated once this computational budget is reached (i.e., once a maximum number of simulations are performed). This does not guarantee a locally optimal solution. If larger computational budgets are considered (e.g., at least 2 orders of magnitude higher), then the algorithm can incorporate local optimality stopping criteria. Performing such local optimality tests is computationally demanding. Therefore, it is essential to devise strategies that trade-off the frequency of performing these tests with the associated computational costs. A discussion of these trade-offs in the context of discrete SO is discussed in Xu *et al.* (2013).

In practice, changes in tolls can impact not only route choices, but also departure time and mode choices. The simulator used in this paper can account for changes of departure time, mode and trip cancellations. This can also be accounted for in the analytical model, just as we did in this paper for route choice. This entails using a simplified mode and/or departure time choice model. This is particularly important when considering a dynamic problem (where tolls vary over time) such as to properly account for how tolls impact the temporal distribution of demand.

The simulator accounts for time-varying network demand, link demand and link states. The analytical model used to build the metamodel is a stationary model, which does not account for temporal variations in demand and link states. The extension of these ideas to address a dynamic problem (where tolls vary across time intervals) will require the use of an analytical model that captures these temporal variations. A straightforward approach would be to formulate a time-dependent analytical traffic model. The main challenge would then be to derive a sufficiently tractable analytical formulation. A different approach, which has recently been proposed for a class of high-dimensional SO problems (OD calibration problems), is to use one stationary model per time interval and to account for the temporal variations through the metamodel parameters (Osorio 2018). The extension of these ideas for dynamic toll optimization is part of ongoing work.

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. We acknowledge the support of the Intelligent Transportation Systems (ITS) Lab both at MIT and SMART (Singapore-MIT Alliance for Research and Technology). Special thanks to Dr. Ravi Seshadri for his help in using the Singapore network model, which was developed and calibrated by his DynaMIT team at SMART.

Appendix A: Main differences between the analytical model and the simulator used for the case studies of this paper

For the case studies of this paper, we use the mesoscopic simulator DynaMIT (DYnamic Network Assignment for the Management of Information to Travelers) which is a simulation-based DTA model that estimates and predicts traffic conditions, generates traffic management strategies and provides consistent guidance information to travelers (Ben-Akiva *et al.* 2010). DynaMIT uses time-dependent origin-destination matrices to specify demand. The traffic conditions are predicted for a specified prediction horizon, e.g., 30 min, and the predicted travel times are used for the route choice models. Traffic dynamics are determined based on the use of speed-density relationships and queueing models. The simulator considers both aggregate and disaggregate representation of demand. The aggregate demand is defined based on the OD matrix, which is then disaggregated into a population of drivers with heterogeneous behavior. Hence, individual vehicles are simulated with their corresponding and unique behavioral characteristics.

The analytical traffic model (Eq. (7)) is a stationary model which does not describe temporal variations in traffic patterns, while the simulator is a dynamic model. The analytical model assumes homogeneous supply conditions (e.g., common lane attributes such as length and maximum speed) and homogeneous traffic conditions for all lanes of a link (e.g., common speeds, densities, etc.). Hence, for the analytical model, the term lane and link can be used interchangeably. On the other hand, the simulator allows for heterogeneous lanes. For instance, possible turnings to downstream-lanes are defined at the lane level. Note that a more detailed analytical model which accounts for heterogeneous lanes is straightforward to formulate, yet will become less tractable. The case study of Section 4 considers a simulation network model with links that have heterogeneous lanes. The results indicate that the above analytical model provides a good approximation of the total expected hourly revenue.

The simulator provides a detailed description of vehicular travel time and accounts for the occurrence and impact of downstream congestion (e.g., spillbacks) on link travel time. More specifically, it considers each link to be comprised of two components: (i) a *moving part* that represents homogeneous traffic conditions governed by the link's supply (i.e., the links fundamental diagram); and (ii) a *queueing part* that accounts for the impact of downstream traffic conditions (e.g., build-up and dissipation of vehicular queues due to spillbacks). The analytical model defines the expected link travel time as the ratio of the link length and the expected (space-mean) link speed (Eq. (7f)). Unlike the simulator, the expected link travel time of the analytical model (Eq. (7f)) does not account for delays due to vehicular queueing or spillbacks. Instead each lane is assumed to have homogeneous traffic conditions along the lane.

The fundamental diagram (Eq. (7g)) is a differentiable simplification of the more intricate, non-differentiable, fundamental diagram of the simulator (Ben-Akiva *et al.* 2010). More specifically, the *moving part* of the simulator's link (or lane) supply model is governed by the following speed-density relationship:

$$v_i = \begin{cases} v_i^{\max} & \text{if } k < k_i^0 \\ v_i^{\max}\left[1 - \left(\dfrac{k - k_i^0}{k_i^{jam}}\right)^{\beta_i}\right]^{\alpha_i} & \text{otherwise.} \end{cases}$$

The parameters of the model are the critical density k_i^0, the maximum speed $v_i^{\max}$, the jam density k_i^{jam}, and two scalar coefficients α_i and β_i. The index i refers to a given link (or lane). This non-differentiable function is approximated by the differentiable function of Eq. (7g). The scalar coefficients of Eq. (7g), $\alpha_{1,i}$ and $\alpha_{2,i}$, are defined as functions of the parameters of the simulator's supply model as follows.

$$\alpha_{1,i} = \alpha_i + \frac{k_i^0}{2k_i^{jam}} \tag{8}$$

$$\alpha_{2,i} = \beta_i + 2.5\frac{k_i^0}{k_i^{jam}}. \tag{9}$$

The DynaMIT simulator generates a predetermined route choice set for each OD pair as the universal choice set. This process is carried out offline and it involves three steps as stated by Balakrishna (2002): "The shortest path computation step generates the shortest path connecting each link in the network to all destination nodes. This set represents the most probable paths chosen by drivers under uncongested conditions. A link elimination step augments the paths from the shortest path set with alternative paths. This step involves the elimination of each link in the network and the subsequent re-computation of the shortest path, and ensures that an incident on any link will still leave alternative paths open for every O-D pair. A further random perturbation step is performed in order to obtain a richer path set. The impedance of the links are perturbed randomly to simulate varying travel times. Another set of shortest paths are now computed, and appended to the existing set. The number of random perturbations performed could be controlled by the user. The algorithm also screens the final path set for uniqueness, and eliminates unreasonably long paths."

Given the predetermined route choice set, three levels of route choice behavior are modeled in DynaMIT: (i) habitual route choice, which represents the behavior based on historical travel times, (ii) pre-trip route choice, which is the response of the habitual behavior to the available pre-trip information, and (iii) en-route choice, which represents the reaction to real-time information after the traveler starts traveling. The behavioral models for each of these levels use a similar set of attributes. Their distinction is the information provided to travelers which determines which travel times and costs are used in decision making. As an example, we specify here the pre-trip route choice model. It is a logit model where the deterministic part of the utility function for route r is given by:

$$V_r = \theta_1 t_r + \theta_2 z_r \tag{10}$$

where t and z are the travel time and cost for route r. When we normalize the utility by the travel time parameter we get θ_1/θ_2 which is proportional to value of time (VOT). For the ease of explanation we refer to this ratio, θ_1/θ_2 , as VOT. Therefore we obtain:

$$V_r = \theta_1 \left[t_r + \frac{1}{VOT} z_r \right]. \tag{11}$$

The simulator assumes that the value of time is distributed randomly across the population following a lognormal distribution. For each traveler, the value of time is sampled from a lognormal distribution. This accounts for the heterogeneity, across the population, in willingness to pay the toll. This specification of the route choice model of the simulator (Eq. (10) or equivalently (11)) considers the same attributes as that of the analytical model (Eq. (7c)). Nonetheless, the analytical model considers a deterministic value of time.

Appendix B: Metamodel parameter fit

In this Section, we detail how the metamodel parameters are fitted. For a given SO iteration, k, the metamodel parameter vector ($\boldsymbol{\beta_k}$) is fitted by solving the following least squares problem.

$$\min_{\boldsymbol{\beta}_k} \sum_{\boldsymbol{x} \in \mathcal{S}_k} \left\{ w_k(\boldsymbol{x}) \left(\hat{f}(\boldsymbol{x}) - m_k(\boldsymbol{x};\boldsymbol{\beta}_k) \right) \right\}^2 + w_0^2 \left((\beta_{k,0} - 1)^2 + \beta_{k,1}^2 + \sum_{j=1}^{T} \beta_{k,j+1}^2 + \sum_{j=1}^{T} \beta_{k,j+T+1}^2 \right), \tag{12}$$

where S_k denotes the set of points (i.e., tolls) simulated up until iteration k, $\hat{f}(\boldsymbol{x})$ denotes the estimate of the SO objective function (f of Eq. (1)) for point $\boldsymbol{x}$, $m_k(\boldsymbol{x};\boldsymbol{\beta}_k)$ is given by Eq. (3), w_0 is an exogenous scalar weight parameter (set to 10^{-2}), and $w_k(\boldsymbol{x})$ is a scalar point-specific weight defined as in Osorio and Bierlaire (2013) by the following equation:

$$w_k(\boldsymbol{x}) = \frac{1}{1 + \|\boldsymbol{x} - \boldsymbol{x}^k\|_2}, \tag{13}$$

where $\boldsymbol{x}^k$ denotes the current iterate (i.e., the point with best performance so far).

The first term of Problem (12) is a traditional least squares expression of a weighted distance between the estimates of the simulation-based function and its corresponding metamodel approximation. For a given point $\boldsymbol{x}$, the weight $(w_k(\boldsymbol{x}))$ is proportional to the distance between the considered point $\boldsymbol{x}$ and the current iterate $\boldsymbol{x}^k$. The use of this weight serves to improve the local (i.e., in the vicinity of the current iterate) fit of the metamodel. The second term of Problem (12) is used such as to ensure that the least squares matrix is of full rank. It can be interpreted as the distance between the parameter $\boldsymbol{\beta_k}$ and prior values. The prior values can be interpreted as giving more weight to the analytical network model than to the polynomial function (i.e., $\beta_{k,0} = 1$ and $\forall j \geq 1 \;\; \beta_{k,j} = 0$).

References

Albalate, D., Bel, G., and Fageda, X. (2009). Privatization and regulatory reform of toll motorways in europe. *An International Journal of Policy, Administration, and Institutions*, **22**(2), 295–318.

Balakrishna, R. (2002). *Calibration of the Demand Simulator in a Dynamic Traffic Assignment System*. Master's thesis, Massachusetts Institute of Technology.

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

Chen, X., Zhang, L., He, X., Xiong, C., and Zhu, Z. (2018). Simulation-based pricing optimization for improving network-wide travel time reliability. *Transportmetrica A: Transportation Science*, **14**(1-2), 155–176.

Chen, X. M., Xiong, C., He, X., Zhu, Z., and Zhang, L. (2016). Time-of-day vehicle mileage fees for congestion mitigation and revenue generation: A simulation-based optimization method and its real-world application. *Transportation Research Part C*, **63**, 71–95.

Chong, L. and Osorio, C. (2017). A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Science*. Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoChoDynSOsubmitted.pdf .

Corthout, R., Himpe, W., Viti, F., Frederix, R., and Tampere, C. M. (2014). Improving the efficiency of repeated dynamic network loading through marginal simulation. *Transportation Research Part C*, **41**, 90–109.

de Palma, A. and Lindsey, L. (2011). Traffic congestion pricing methodologies and technologies. *Transportation Research Part C*, **19**(6), 1377–1399.

Dong, J., Mahmassani, H. S., Erdogan, S., and Lu, C. C. (2011). State-dependent pricing for real-time freeway management: Anticipatory versus reactive strategies. *Transportation Research Part C*, **19**(4), 644–657.

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

Google Maps (2017). Singapore Expressway Network.

Gupta, S., Seshadri, R., Atasoy, B., Pereira, F. C., Wang, S., Vu, V., Tan, G., Dong, W., Lu, Y., Antoniou, C., and Ben-Akiva, M. (2016). Real time optimization of network control strategies in DynaMIT2.0. In *Transportation Research Board Annual Meeting*, Washington DC, USA.

Halvorson, R., Nookala, M., and Buckeye, K. R. (2011). High occupancy toll lane innovations: I-394 mnpass. In *Transportation Research Board Annual Meeting*, Washington DC, USA.

Han, L., Wang, D. Z. W., Lo, H. K., Zhu, C., and Cai, X. (2017). Discrete-time day-to-day dynamic congestion pricing scheme considering multiple equilibria. *Transportation Research Part B*, **104**, 1–16.

Hassan, A., Abdelghany, K., and Semple, J. (2013). Dynamic road pricing for revenue maximization: Modeling framework and solution methodology. *Transportation Research Record*, **2345**, 100–108.

Jang, K., Chung, K., and Yeo, H. (2014). A dynamic pricing strategy for high occupancy toll lanes. *Transportation Research Part A*, **67**, 69–80.

Land Transport Authority (2018). Erp Rates.

Liu, Z., Wang, S., Zhou, B., and Cheng, Q. (2017). Robust optimization of distance-based tolls in a network considering stochastic day to day dynamics. *Transportation Research Part C*, **79**, 58–72.

Lou, Y., Yin, Y., and Laval, J. A. (2011). Optimal dynamic pricing strategies for high-occupancy/toll lanes. *Transportation Research Part C*, **19**, 64–74.

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

Osorio, C. (2018). Dynamic OD calibration for large-scale network simulatos. Technical report, Massachusetts Institute of Technology. Under review. Available at: http://web.mit.edu/osorioc/www/papers/osoDynamicOD.pdf .

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

Osorio, C. and Chong, L. (2015). A computationally efficient simulation-based optimization algorithm for large-scale urban transportation. *Transportation Science*, **49**(3), 623–636.

Osorio, C. and Nanduri, K. (2015). Energy-efficient urban traffic management: a microscopic simulation-based approach. *Transportation Science*, **49**(3), 637–651.

Osorio, C. and Selvam, K. (2017). Simulation-based optimization: achieving computational efficiency through the use of multiple simulators. *Transportation Science*, **51**(2), 395–411.

Osorio, C., Chen, X., and Santos, B. F. (2017). Simulation-based travel time reliable signal control. *Transportation Science*. Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoCheSanReliableSO.pdf .

Saleh, W. and Sammer, G. (2009). *Travel Demand Management and Road User Pricing: Success, Failure and Feasibility*. Routledge, New York, USA.

Seik, F. T. (2000). An advanced demand management instrument in urban transport: Electronic road pricing in Singapore. *Cities*, **17**, 33–45.

Tay, T. and Osorio, C. (2017). An efficient sampling method for stochastic simulation-based transportation optimization. In *First INFORMS Transportation Science and Logistics (TSL) Conference*, Chicago IL, USA.

Toledo, T., Mansour, O., and Haddad, J. (2015). Simulation-based optimization of hot lane tolls. *Transportation Research Procedia*, **6**, 189–197.

Vu, V., Hashemi, H., Seshadri, R., Gupta, S., Tan, G., Prakash, A. A., and Ben-Akiva, M. (2018). Predictive distance-based toll optimization under elastic demand for real-time traffic management. In *Transportation Research Board Annual Meeting*, Washington DC, USA.

Xu, J., Nelson, B. L., and Hong, J. L. (2013). An adaptive hyperbox algorithm for high-dimensional discrete optimization via simulation problems. *INFORMS Journal on Computing*, **25**(1), 133–146.

Zhang, C., Osorio, C., and Flötteröd, G. (2017). Efficient calibration techniques for large-scale traffic simulators. *Transportation Research Part B*, **97**, 214–239.

Zhang, Y., Atasoy, B., and Ben-Akiva, M. (2018). Calibration and optimization for adaptive toll pricing. In *Transportation Research Board Annual Meeting*, Washington DC, USA.

Zheng, N., Rérat, G., and Geroliminis, N. (2016). Time-dependent area-based pricing for multimodal systems with heterogeneous users in an agent-based environment. *Transportation Research Part B*, **62**, 133–148.