A Computationally Efficient Simulation-Based Optimization Algorithm for Large-Scale Urban Transportation Problems

Carolina Osorio, Linsen Chong

TRANSPORTATION SCIENCE
Vol. 49, No. 3, August 2015, pp. 623–636
ISSN 0041-1655 (print) | ISSN 1526-5447 (online)
http://dx.doi.org/10.1287/trsc.2014.0550

A Computationally Efficient Simulation-Based Optimization Algorithm for Large-Scale Urban Transportation Problems

Carolina Osorio, Linsen Chong
Civil and Environmental Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 {osorioc@mit.edu, linsenc@mit.edu}

This paper proposes a computationally efficient simulation-based optimization (SO) algorithm suitable to address large-scale generally constrained urban transportation problems. The algorithm is based on a novel metamodel formulation. We embed the metamodel within a derivative-free trust region algorithm and evaluate the performance of this SO approach considering tight computational budgets. We address a network-wide traffic signal control problem using a calibrated microscopic simulation model of evening peak period traffic of the full city of Lausanne, Switzerland, which consists of more than 600 links and 200 intersections. We control 99 signal phases of 17 intersections distributed throughout the entire network. This SO problem is a high-dimensional nonlinear constrained problem. It is considered large-scale and complex in the fields of derivative-free optimization, traffic signal optimization, and simulation-based optimization. We compare the performance of the proposed metamodel method to that of a traditional metamodel method and that of a widely used commercial signal control software. The proposed method systematically and efficiently identifies signal plans with improved average city-wide travel times.

Keywords: simulation-based optimization; metamodel; large-scale urban transportation problems
History: Received: November 2012; revisions received: September 2013, January 2014; accepted: April 2014. Published online in *Articles in Advance* February 13, 2015.

1. Introduction

The massive amount and variety of mobility data that can now be collected through, for instance, ubiquitous mobile devices, is enhancing our fundamental understanding of individual mobility. For instance, it improves our understanding of the intricate behavior of travelers—e.g., how they make activity and thereby travel decisions and how these decisions are motivated by an underlying objective to enhance their well-being.

State-of-the-art microscopic traffic simulation models embed such disaggregate models of traveler behavior (e.g., departure time choice, multimodal route choice, and access and response to en route traffic information) and account for behavior heterogeneity. They represent individual vehicles and can therefore be coupled with vehicle-specific simulators (e.g., propulsion simulators) to yield detailed estimates of the performance of vehicles (e.g., energy consumption or emissions estimates) in networks with complex topologies and complex traffic dynamics. Additionally, microscopic simulators provide a detailed representation of the underlying supply (e.g., variable message signs and public transport priorities).

Microscopic traffic simulators describe in detail the interactions between (i) vehicle performance, (ii) traveler behavior, and (iii) the underlying transportation infrastructure, and they yield an elaborate description of traffic dynamics in urban networks. They are therefore suitable tools to address transportation problems that should account for a detailed representation of these three components.

Microscopic simulators are popular tools used in practice to evaluate the performance of a set of predetermined transportation strategies. Cities such as Toronto, New York, Boston, Stockholm, and Hong Kong have used these tools to inform their planning and operations decisions (Traffic Technology International 2012a, b; Papayannoulis et al. 2011; Toledo et al. 2003; Hasan 1999).

For a given strategy, these simulators can provide accurate and detailed performance estimates. Their use is mostly limited to what-if analysis (also called scenario-based analysis) or sensitivity analysis. That is, they are used to evaluate the performance of a set of predetermined transportation alternatives (e.g., traffic management or network design alternatives), such as in Bullock et al. (2004); Ben-Akiva et al. (2003); Hasan, Jha, and Ben-Akiva (2002); Stallard and Owen (1998); Gartner and Hou (1992) and Rathi and Lieberman (1989). See further references in Ben-Akiva et al. (2003).

The numerous models of disaggregate traveler behavior, vehicle-performance, and supply components lead to detailed performance estimates yet also to models that are expensive to develop and calibrate and computationally expensive to evaluate. Thus, an accurate estimation of performance is computationally costly to obtain. Additionally, these simulators derive stochastic nonlinear, and typically nonconvex, performance measures with no closed-form available. For these reasons, the use of these simulators to address optimization problems is a challenge.

Currently, the use of these simulation tools is mostly limited to what-if analysis. With the ubiquity of access to real-time traffic information, and the increasing number of prevailing and interacting traffic control strategies, traffic dynamics of congested networks are becoming more intricate. Thus, determining a priori a set of alternatives with good local and network-wide performance is no longer feasible. Thus, there is a need to embed these detailed simulators within optimization frameworks to systematically identify alternatives with improved local and network-wide performance. Additionally, given the high cost of developing large-scale simulation tools, transportation projects would benefit from computationally efficient methods that allow the use of simulators to go beyond a what-if analysis.

This paper proposes a simulation-based optimization (SO) method that allows large-scale urban transportation problems to be addressed with detailed microscopic traffic simulators. We focus on problems where the objective function is derived from the simulator and, thus, no closed-form analytical expression is available. The problems have general (e.g., nonconvex) constraints. Closed-form analytical and differentiable expressions are available for all constraints (i.e., no simulation-based constraints).

These urban transportation problems can be formulated as

$$\min_{x \in \Omega} f(x, z; p) \equiv E[F(x, z; p)], \tag{1}$$

where the purpose is to minimize the expected value of a given stochastic performance measure F, x denotes the deterministic continuous decision vector, z denotes other endogenous variables, and p denotes the deterministic exogenous parameters. For instance, in this paper we use the proposed SO approach to solve a traffic signal control problem, where F denotes trip travel time; x represents the green times of the signal phases; z accounts, for instance, for signalized link capacities and route choice decisions; and p accounts, for instance, for the network topology, the total traffic demand, and fixed lane attributes (e.g., length, grade, and maximum speed). The feasible space Ω consists of a set of general, typically nonconvex, deterministic, analytical, and differentiable constraints.

This paper proposes a technique that can efficiently address generally constrained large-scale simulation-based urban transportation problems. The performance of the technique is evaluated by considering a network-wide traffic signal control problem. This problem is considered large-scale and complex for derivative-free algorithms, signal control algorithms, and simulation-based optimization algorithms.

Additionally, the paper focuses on SO techniques with good short-term performance, i.e., computationally efficient methods that can identify alternatives with improved performance within a tight computational budget. The computational budget can be defined as a limited number of simulation runs or a limited simulation run time. Such techniques respond to the needs of transportation practitioners by allowing them to address problems in a practical manner.

We present a review of past work in this field in §2. In §3 of this paper, we present the methodology. We then present the traffic signal control problem, which is used to evaluate the scalability and short-term performance of this approach (§4). Empirical results are detailed in §5, followed by conclusions (§6).

2. Literature Review

Few SO methods that embed microscopic simulators have been developed (Li et al. (2010); Stevanovic et al. (2008); Branke, Goldate, and Prothmann (2007); Yun and Park (2006); Hale (2005); Joshi, Rathi, and Tew (1995)). The most common approach is the use of heuristic algorithms and, in particular, the use of genetic algorithms (see Yun and Park (2006) for a review). These methods embed microscopic simulators within general-purpose optimization algorithms. They treat the simulator as a black box, using no a priori structural information about the underlying transportation problem (e.g., network structure). They therefore require a large number of simulated observations to identify transportation strategies (i.e., trial points) with improved performance.

This paper proposes an SO technique with good short-term performance suitable for microscopic traffic simulators to be used to address complex high-dimensional problems. To derive computationally efficient methods that embed inefficient simulators, information from other more efficient (i.e., tractable) models that provide analytical structural information to the algorithm should be used throughout the optimization process.

In general, methods to address SO problems can be classified as direct-search methods, stochastic gradient methods, and metamodel methods. For reviews of SO methods, see Hachicha et al. (2010); Barton and Meckesheimer (2006); Fu, Glover, and April (2005). This paper focuses on metamodel methods. For a description of why metamodel techniques are a suitable approach to address complex simulation-based transportation problems, see Osorio and Bierlaire (2013).

Metamodel methods build an analytical approximation of the simulation-based components of the optimization problem (e.g., objective function, constraints). In this paper, the objective function is simulation-based. Thus, the metamodel provides an analytical approximation of the objective function. By resorting to a metamodel approach, the stochastic response of the simulation is replaced by an analytical response function (the metamodel) such that deterministic optimization techniques can be used. Metamodel techniques use an indirect-gradient approach—i.e., they compute the gradient of the metamodel, which is a deterministic function. Thus, traditional deterministic gradient-based optimization algorithms for generally constrained problems can be used.

Metamodel SO methods are iterative methods based on the two main steps depicted in Figure 1. (For more details see Osorio and Bierlaire (2013).) Step 1 fits the metamodel based on the current sample of simulated observations. Step 2 uses the fitted metamodel to perform optimization and derive a trial point (e.g., a suitable traffic management or network design alternative). The performance of the trial point is then evaluated by the simulator, which leads to new observations. As new observations become available, the metamodel is fitted again (step 1), leading to more accurate metamodels and ultimately to trial points with improved performance (step 2).

Reviews of metamodels are given by Conn, Scheinberg, and Vicente (2009b); Barton and Meckesheimer (2006); and Søndergaard (2003). Metamodels can be classified as either physical or functional (Søndergaard 2003). Physical metamodels are application- or problem-specific metamodels. Their functional form and parameters have a physical interpretation. Functional metamodels are general-purpose (i.e., generic) functions chosen based on their analytical tractability. The most common

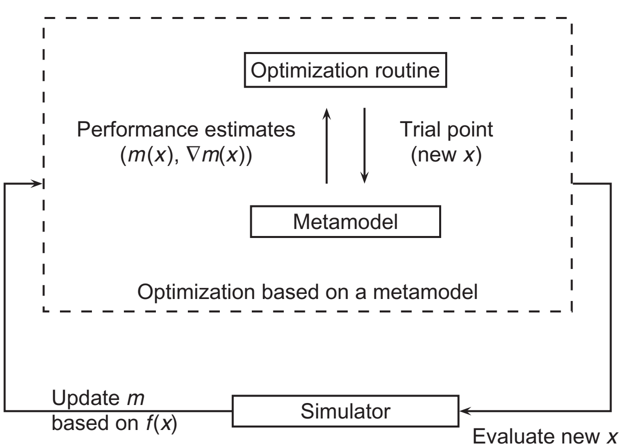

Figure 1 **Metamodel Simulation-Based Optimization Methods**
Source. Adapted from Alexandrov et al. (1999).

general-purpose metamodel is the use of low-order polynomials, particularly quadratic polynomials (Conn, Scheinberg, and Vicente (2009b); Kleijnen (2008); Marti (2008)). Other general-purpose metamodels include spline models, radial basis functions, and Kriging models (Kleijnen, van Beers, and van Nieuwenhuyse (2010); Wild, Regis, and Shoemaker (2008); Barton and Meckesheimer (2006)).

The existing metamodels consist of either physical or functional components. Recent work has proposed a metamodel that is a combination of a functional and physical metamodel (Osorio and Bierlaire 2013). The functional component ensures asymptotic metamodel properties necessary for convergence analysis (such as full linearity (Conn, Scheinberg, and Vicente 2009a)). The physical component is an analytical and differentiable macroscopic traffic model. It provides a problem-specific analytical approximation of the objective function, unlike the generic approximation provided by the functional component. The physical component therefore yields structural information about the problem at hand, which enables the identification of well-performing alternatives (i.e., trial points) with small samples (i.e., good short-term algorithmic performance). The physical component used here is an analytical differentiable queueing network model. This macroscopic traffic model is less detailed and accurate than the simulator, but it is computationally efficient to evaluate.

The combined use of functional and physical metamodels allows information from the detailed, yet inefficient microscopic simulator to be combined with analytical information from a more efficient macroscopic model. This leads to an algorithm with a good detail-tractability trade-off and good short-term performance.

This physical and functional metamodel approach has been used to efficiently address complex urban transportation problems, such as signal control problems that account for detailed (also called microscopic) vehicle-specific energy consumption patterns (Osorio and Nanduri 2015), emissions patterns (Osorio and Nanduri 2013), and reliable signal control problems that used detailed full distributional travel time estimates provided by the simulator to improve both average travel times and travel time reliability (Chen, Osorio, and Santos 2012).

This approach has been successfully used to control networks with approximately 50 roads, yet is unsuitable to address problems for much larger-scale networks. This paper builds on this existing metamodel SO technique (hereafter referred to as the *initial* method) and proposes a metamodel that can efficiently address high-dimensional simulation-based problems.

3. Methodology

3.1. Metamodel Functional Form

Recall the general form of the urban transportation problems that we address (Equation (1)). Since there is no closed-form available for the objective function, f, we use a metamodel to approximate it. The functional form of the metamodel used in this paper is that proposed by Osorio and Bierlaire (2013). It combines a physical and functional component. Its functional form is given by

$$m(x, y; \alpha, \beta, q) = \alpha T(x, y; q) + \phi(x; \beta), \tag{2}$$

where ϕ (the functional component) is a quadratic polynomial in x with diagonal second-derivative matrix, T (the physical component) represents the approximation of the objective function proposed by the analytical macroscopic traffic model, y are endogenous macroscopic model variables (e.g., queue length distributions), q are exogenous macroscopic parameters (e.g., total demand), and α and β are parameters of the metamodel. The metamodel m can be interpreted as a macroscopic approximation of the objective function provided by T, which is corrected parametrically by both a scaling factor α and a separable error term $\phi(x; \beta)$. For details regarding the choice of this functional form, we refer the reader to Osorio and Bierlaire (2013).

In this paper, we use the same functional component as in Osorio and Bierlaire (2013) (i.e., the quadratic polynomial ϕ). We propose a novel scalable physical component. In §3.2 we recall the formulation of the physical component of the initial metamodel and describe its limitations. We then present the new formulation of the physical component in §3.3.

3.2. Initial Queueing Network Model

The physical component of the initial metamodel is an urban traffic model based on queueing network theory. It combines ideas from existing traffic models, various national urban transportation norms, and queueing models. The detailed formulation of the model is given in Osorio and Bierlaire (2009b) (which is based on the more general queueing network model of Osorio and Bierlaire 2009a). We outline here the main ideas of its formulation.

Each lane of an urban road network is modeled as a queue (and in some cases as a set of queues). To account for the limited physical space that a queue of vehicles may occupy, we resort to *finite capacity queueing theory*, where there is a finite upper bound on the length of each queue. Each lane is modeled as a finite capacity $M/M/1/k$ queue. The network model analytically approximates the queue interactions among adjacent lanes. Congestion and spillbacks are modeled by what is known in queueing theory as *blocking*. This occurs when a queue is full and thus blocks arrivals from upstream queues at their current location. This blocking process is described by endogenous variables such as blocking probabilities and unblocking rates. The model consists of a set of nonlinear equations that capture these between-queue interactions.

In the following notation, the index i refers to a given queue:

- γ_i external arrival rate,
- λ_i total arrival rate,
- μ_i service rate,
- $\tilde{\mu}_i$ unblocking rate,
- μ_i^{eff} effective service rate (accounts for both service and eventual blocking),
- ρ_i traffic intensity,
- P_i^f probability of being blocked at queue i,
- k_i upper bound of the queue length,
- N_i total number of vehicles in queue i,
- $P(N_i = k_i)$ probability of queue i being full, also known as the blocking or spillback probability,
- p_{ij} transition probability from queue i to queue j, and
- $\mathcal{D}_i$ set of downstream queues of queue i.

The queueing network model is formulated as follows:

$$\begin{cases}
\lambda_i = \gamma_i + \dfrac{\sum_j p_{ji}\lambda_j(1 - P(N_j = k_j))}{1 - P(N_i = k_i)}, & \text{(3a)}\\[2ex]
\dfrac{1}{\tilde{\mu}_i} = \displaystyle\sum_{j\in\mathcal{D}_i} \dfrac{\lambda_j(1 - P(N_j = k_j))}{\lambda_i(1 - P(N_i = k_i))\mu_j^{\text{eff}}}, & \text{(3b)}\\[2ex]
\dfrac{1}{\mu_i^{\text{eff}}} = \dfrac{1}{\mu_i} + P_i^f \dfrac{1}{\tilde{\mu}_i}, & \text{(3c)}\\[2ex]
P(N_i = k_i) = \dfrac{1-\rho_i}{1-\rho_i^{k_i+1}}\rho_i^{k_i}, & \text{(3d)}\\[2ex]
P_i^f = \displaystyle\sum_j p_{ij} P(N_j = k_j), & \text{(3e)}\\[2ex]
\rho_i = \dfrac{\lambda_i}{\mu_i^{\text{eff}}}. & \text{(3f)}
\end{cases}$$

Equation (3a) is a flow conservation equation that relates flow transmission between upstream and downstream queues. The factor $(1 - P(N_i = k_i))$ represents the probability that queue i is not full (i.e., the queue can receive flow from its upstream queues). If the queue is full, it cannot receive flow from upstream queues, which may lead to spillbacks. Equation (3b) defines the rate at which spillbacks at queue i dissipate, $\tilde{\mu}_i$. Equation (3c) defines the rate at which queue i dissipates accounting for both spillback and nonspillback states, μ_i^{eff}. It is defined as a function of the service rate of the queue, μ_i. The latter is determined by combining ideas from national transportation norms and is a function, for instance, of the free flow capacity of the underlying lane. Equation (3d) defines the probability that a queue is full—i.e., the spillback probability of the underlying lane. This expression is derived from finite capacity queueing theory (Bocharov et al. 2004). Equation (3e) defines the probability of a vehicle being blocked while in queue i—i.e., the probability that a vehicle at the underlying lane is affected by spillback from a downstream lane. Equation (3f) defines the traffic intensity of a queue; it is also derived from traditional finite capacity queueing formulae.

In this model, the exogenous parameters of a given queue are γ_i, μ_i, p_{ij}, and k_i. All other parameters are endogenous. When used to solve a signal control problem, the flow capacity of the signalized lanes become endogenous, which makes the corresponding service rates, μ_i, endogenous. In that case, the exogenous parameters are γ_i, p_{ij}, and k_i. This is a stationary model with exogenous traffic assignment (the turning probabilities p_{ij} are exogenous). As described in §6, analytical tractable formulations that describe both traffic dynamics and endogenous assignment are being developed as part of ongoing work.

As described in §2, this model has been used to solve signal control problems for medium-scale networks. However, it is not sufficiently tractable to address large-scale network problems. For instance, in the case of the Lausanne city network (with more than 600 links and 200 intersections), the time needed by a standard nonlinear optimization algorithm to solve the trust-region (TR) subproblem (detailed in §4.2) exceeds 20 minutes. Since this TR subproblem is solved at every iteration of the SO algorithm, it is critical to solve it efficiently.

In this paper, we propose a more tractable and scalable physical component of the metamodel. It is an approximation of this initial queueing network model. It consists of a simple system of one linear and two nonlinear equations. In particular, as is detailed in §5.2, the TR subproblem is now solved on average within less than two minutes. This significantly enhances the computational efficiency of the SO algorithm and allows us to efficiently address more complex high-dimensional constrained transportation problems.

3.3. Highly Tractable Queueing Network Model

We introduce the following two variables:

λ_i^{eff} effective arrival rate;

ρ_i^{eff} effective traffic intensity.

These two new variables are defined by

$$\lambda_i^{\text{eff}} = \lambda_i(1 - P(N_i = k_i)), \tag{4}$$

$$\rho_i^{\text{eff}} = \frac{\lambda_i^{\text{eff}}}{\mu_i^{\text{eff}}}. \tag{5}$$

The highly tractable queueing network model is given by

$$\begin{cases} \lambda_i^{\text{eff}} = \gamma_i(1 - P(N_i = k_i)) + \sum_j p_{ji}\lambda_j^{\text{eff}} & \text{(6a)} \\ \rho_i^{\text{eff}} = \dfrac{\lambda_i^{\text{eff}}}{\mu_i} + \left(\sum_{j\in\mathcal{D}_i} p_{ij}P(N_j = k_j)\right)\left(\sum_{j\in\mathcal{D}_i} \rho_j^{\text{eff}}\right) & \text{(6b)} \\ P(N_i = k_i) = \dfrac{1 - \rho_i^{\text{eff}}}{1 - (\rho_i^{\text{eff}})^{k_i+1}}(\rho_i^{\text{eff}})^{k_i}. & \text{(6c)} \end{cases}$$

Equation (6a) is obtained directly by inserting Equation (4) into Equation (3a). Equation (6b) is obtained as follows. Multiply Equation (3b) and (3c), respectively, by λ_i^{eff} to obtain

$$\frac{\lambda_i^{\text{eff}}}{\tilde{\mu}_i} = \sum_{j\in\mathcal{D}_i} \frac{\lambda_j^{\text{eff}}}{\mu_j^{\text{eff}}}, \tag{7}$$

$$\rho_i^{\text{eff}} = \frac{\lambda_i^{\text{eff}}}{\mu_i} + P_i^f \frac{\lambda_i^{\text{eff}}}{\tilde{\mu}_i}. \tag{8}$$

Insert Equation (7) into (8) to obtain

$$\rho_i^{\text{eff}} = \frac{\lambda_i^{\text{eff}}}{\mu_i} + P_i^f\left(\sum_{j\in\mathcal{D}_i} \rho_j^{\text{eff}}\right). \tag{9}$$

Insert the expression of P_i^f given by Equation (3e), and Equation (6b) results.

Equation (6c) is an approximation of Equation (3d) that is obtained by replacing the traffic intensity ρ with the effective traffic intensity ρ^{eff}. That is, we use the expression of the blocking probability of a finite capacity queue, yet approximate the traffic intensity with the effective traffic intensity.

Equation (5) defines ρ^{eff} and shows that it may underestimate ρ. For queues with light traffic, we have $\rho^{\text{eff}} \approx \rho$, and the two models will yield similar network performance estimates. For congested links, the scalable approximation may underestimate link congestion.

The proposed model consists of three endogenous variables per queue (λ_i^{eff}, ρ_i^{eff}, $P(N_i = k_i)$). When using this model to address signal control problems, μ_i also becomes endogenous. This model is defined by one linear and two nonlinear equations. This formulation results in increased computational efficiency, enabling us to address a full city-scale microscopic simulation-based optimization problem.

3.4. Example of Functional Form of T

As described in §2, one of the advantages of using a physical component in the metamodel is to have problem-specific approximations of the objective function. In this section, we give an example of the functional form of the analytical approximation of the objective function provided by the queueing model, $T(x, y; q)$. In §4, we address a signal control problem, where the objective is to minimize the expected trip travel time. The queueing approximation of this expectation is obtained by applying Little's law (Little (2011); Little (1961)) to the entire network. It is given by:

$$\frac{\sum_i E[N_i]}{\sum_i \gamma_i(1 - P(N_i = k_i))}, \tag{10}$$

where $E[N_i]$ represents the expected number of vehicles in lane i, γ_i is the rate of vehicles entering the network via lane i (i.e., the external arrival rate), and $P(N_i = k_i)$ is the probability that lane i is full (i.e., spillback or blocking probability). The numerator of Equation (10) represents the expected number of vehicles in the network, whereas the denominator represents the effective arrival rate to the network. Their ratio yields the expected time in the network.

The expected number of vehicles on lane i, $E[N_i]$, is given by

$$E[N_i] = \rho_i\left(\frac{1}{1-\rho_i} - \frac{(k_i+1)\rho_i^{k_i}}{1-\rho_i^{k_i+1}}\right). \tag{11}$$

This expression is derived in Appendix A. In the scalable model proposed in this paper, ρ_i is approximated by ρ_i^{eff} in Equation (11).

3.5. SO Algorithm

The SO algorithm used in this paper is that of Osorio and Bierlaire (2013). It is given in Appendix B and is based on the derivative-free trust-region (TR) algorithm proposed by Conn, Scheinberg, and Vicente (2009a). For an introduction to TR methods, we refer the reader to Conn, Gould, and Toint (2000). They summarize the main steps of a TR method in the *basic trust region algorithm*. The derivative-free method proposed by Conn, Scheinberg, and Vicente (2009a) builds on the basic TR algorithm by adding two additional steps: a model improvement step and a criticality step. This algorithm allows for arbitrary metamodels to be used, and unlike traditional TR algorithms, it makes no assumptions on how these metamodels are fitted (interpolation or regression). It is therefore particularly appealing for the simulation-based context where derivatives are costly to estimate and where metamodels fitted via regression are more suitable than their interpolated versions.

At a given iteration k of the SO algorithm, it solves a TR subproblem and approximates the objective function by the current metamodel m_k (defined in Equation (2)). The metamodel parameters (α_k and β_k) are fitted via regression based on the simulated observations collected so far. For a detailed description of the algorithm, see Osorio and Bierlaire (2013).

4. Traffic Signal Control Problem

This methodology is suitable to address a variety of simulation-based urban transportation optimization problems. In this section, we evaluate the performance of the methodology by considering a large-scale network-wide traffic signal control problem.

4.1. Problem Formulation

A detailed review of traffic signal control formulations is given in Appendix A of Osorio (2010). In this paper, we consider a fixed-time strategy. Fixed-time (also called time of day or pre-timed) strategies are predetermined based on historical traffic patterns. They yield one traffic signal setting for the considered time of day. The traffic signal optimization problem is solved offline.

In this paper, the signal plans of several intersections are determined jointly. For a given intersection and a given time interval (e.g., evening peak period), a fixed-time signal plan is a cyclic (i.e., periodic) plan that is repeated throughout the time interval. The duration of the cycle is the time required to complete one sequence of signals. The cycle times of the intersections controlled in the Lausanne network (used in the case study of this paper) are 80, 90, or 100 seconds.

A phase is defined as a set of traffic streams that are mutually compatible and that receive identical control. The cycle of a signal plan is divided into a sequence of periods called stages. Each stage consists of a set of mutually compatible phases that all have green times simultaneously. The stage sequence is defined so as to separate conflicting traffic movements at intersections. The cycle may also contain all-red periods, where all streams have red indications, as well as stages with fixed durations (e.g., for safety reasons). The sum of the all-red periods and the fixed periods is called the fixed cycle time.

Cycle times, green splits, and offsets are the three main signal timing control variables. The green split corresponds to the ratio of green times (i.e., total duration of a phase) to cycle time. Offsets are defined as the difference in time between the start of cycles for a pair of intersections. Offset settings are especially important in coordinating the signals of adjacent intersections (e.g., to create green waves along arterials or corridors).

In this paper cycle times, offsets, and all-red durations are kept constant. The stage structure is also given—i.e., the set of lanes associated with each stage as well as the sequence of stages are both known. This is known as a stage-based approach. The decision variables consist of the endogenous green splits of the different intersections.

To formulate this problem, we introduce the following notation:

- c_i cycle time of intersection i,
- d_i fixed cycle time of intersection i,
- e_l ratio of fixed green time to cycle time of signalized lane l,
- s saturation flow rate [veh/h],

$x(j)$ green split of phase j,
x_L vector of minimal green splits,
$\mathcal{I}$ set of intersection indices,
$\mathcal{L}$ set of indices of the signalized lanes,
$\mathcal{P}_I(i)$ set of endogenous phase indices of intersection i, and
$\mathcal{P}_L(l)$ set of endogenous phase indices of lane l.

The problem is traditionally formulated as follows:

$$\min_x \quad f(x;p) \equiv E[F(x;p)] \tag{12}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{P}_I(i)} x(j) = \frac{c_i - d_i}{c_i}, \quad \forall i \in \mathcal{I}, \tag{13}$$

$$x \geq x_L. \tag{14}$$

The decision vector x consists of the green splits for each phase. The objective is to minimize the expected trip travel time (Equation (12)). The linear constraints (13) link the green times of the phases with the available (i.e., nonfixed) cycle time for each intersection. Equation (14) ensures lower bounds for the green splits. These bounds are determined based on the prevailing transportation norms.

4.2. Trust-Region Subproblem

This section presents the trust-region (TR) subproblem that is solved at each iteration of the SO algorithm. It is a variation of the signal control problem defined in §4.1. At a given iteration k, the SO algorithm considers a metamodel $m_k(x, y; \alpha_k, \beta_k, q)$, an iterate x_k (point considered to have best performance so far), and a TR radius Δ_k. The TR subproblem is formulated as follows:

$$\min_{x,y} \quad m_k = \alpha_k T(x, y; q) + \phi(x; \beta_k) \tag{15}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{P}_I(i)} x(j) = \frac{c_i - d_i}{c_i} \quad \forall i \in \mathcal{I}, \tag{16}$$

$$h(x, y; q) = 0, \tag{17}$$

$$\mu_l - \sum_{j\in\mathcal{P}_L(l)} x(j)s = e_l s, \quad \forall l \in \mathcal{L}, \tag{18}$$

$$\|x - x_k\|_2 \leq \Delta_k, \tag{19}$$

$$y \geq 0, \tag{20}$$

$$x \geq x_L. \tag{21}$$

The TR subproblem approximates the objective function by the metamodel at iteration k, m_k. It contains the constraints of the signal control problem and includes three additional constraints. Equations (16) and (21) are the signal control constraints; they correspond to Equations (13) and (14). The function h of Equation (17) represents the queueing network model (Equations (6a)–(6c)). Equation (18) relates the green splits of a phase to the flow capacity of the underlying lanes (i.e., the service rate of the queues). Constraint (19) is the trust region constraint. The endogenous variables of the queueing model are subject to positivity constraints (Equation (20)). Thus, the TR subproblem consists of a nonlinear objective function subject to nonlinear and linear equalities, a nonlinear inequality, and bound constraints.

Implementation notes. This problem is solved with the Matlab routine for constrained nonlinear problems, fmincon, and its sequential quadratic programming method (Coleman and Li 1994, 1996). We set the tolerance for relative change in the objective function to 10^{-3} and the tolerance for the maximum constraint violation to 10^{-2}. For further details on the TR subproblem formulation and its implementation, see Osorio and Bierlaire (2013).

We implement the lower bound constraints of Equation (21) as nonlinear equations by introducing a new variable g and implementing Equation (21) as

$$x = x_L + g^2. \tag{22}$$

We do not enforce the positivity of all endogenous variables (Equation (20)) but check a posteriori that all endogenous variables are positive. In our numerous experiments, we have not encountered a case with a negative value. We insert Equation (18) into Equation (6b) and implement the two constraints as a single constraint.

For a problem with n endogenous phases, l lanes, and b signalized intersections, where each lane is modeled by a single queue (i.e., we have l queues), there are $3l + n$ endogenous variables, which consist of 3 endogenous queueing variables per lane, and the green splits for each phase. There are l linear equations, $2l + b$ nonlinear equations, and one nonlinear inequality (TR constraint).

5. Empirical Analysis

5.1. Lausanne City Network

We evaluate the scalability and short-term algorithmic performance of this framework by solving a large-scale signal control problem. We solve a problem for the entire Swiss city of Lausanne. The map is displayed in Figure 2; the considered area is delimited in white.

We use a microscopic traffic simulation model of the Lausanne city center developed by Dumont and Bert (2006). It is implemented with the Aimsun simulator (TSS 2008) and is calibrated for evening peak period demand. Details regarding this Lausanne network are given in Osorio and Bierlaire (2009b). In this paper, the considered demand scenario consists of the first hour of peak period traffic, 17 h–18 h.

The road network consists of 603 links and 231 intersections. The signals of 17 intersections are controlled in

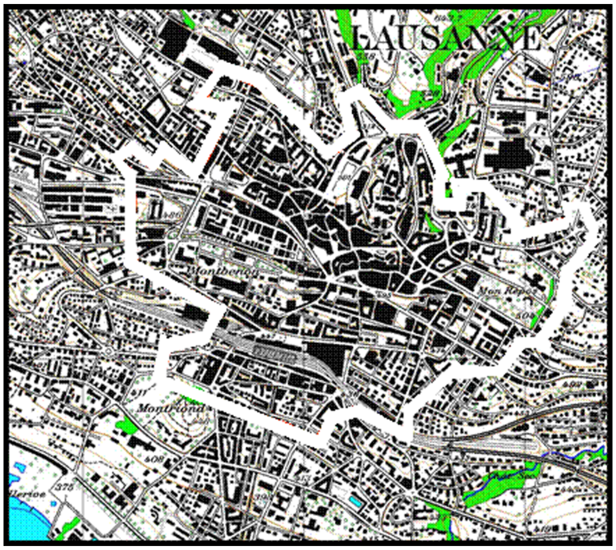

Figure 2 **(Color online) Lausanne City Road Network**
Source. Adapted from Dumont and Bert (2006).

this problem. The modeled road network is displayed in Figure 3, where the 17 intersections are depicted as filled squares. This leads to a total of 99 endogenous phase variables (i.e., the dimension of decision vector is 99).

The queueing model consists of 902 queues. The TR subproblem consists of 2,805 endogenous variables with 1,821 nonlinear equality constraints and 902 linear equality constraints. The lower bounds of the green splits (Equation (14)) are set to 4 seconds according to the Swiss transportation norm (VSS 1992).

Performing network-wide signal control of networks with around 70 links and 16 intersections is currently considered large-scale in the field of signal control, as illustrated by recent studies (Aboudolas et al. (2010); Aboudolas, Papageorgiou, and Kosmatopoulos (2007)). Thus, the simulation-based signal control problem of this paper is a challenging large-scale network-wide signal control problem that considers a congested network with a complex topology.

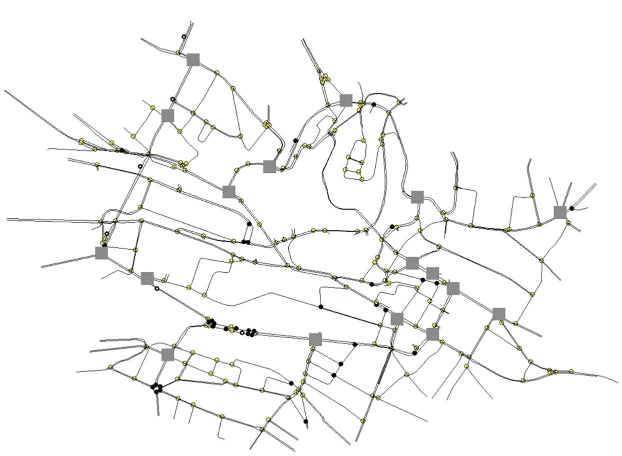

Figure 3 **Lausanne Network Model**

This is considered a large-scale problem for existing unconstrained derivative-free algorithms, where the most recent methods are limited to problems with around 200 variables (Conn, Scheinberg, and Vicente 2009b), not to mention the added complexity of nonlinear constraints and stochasticity. Given the complexity of the underlying simulator, this problem is also considered complex for simulation-based optimization algorithms.

5.2. Numerical Results

We compare the performance of the proposed metamodel method with a traditional metamodel method that consists only of a functional component, which is a quadratic polynomial with diagonal second derivative matrix (i.e., the metamodel consists of ϕ, defined in Equation (2)). To compare the two methods, we consider a tight computational budget, which is defined as a maximum of 150 simulation runs that can be carried out.

We consider three different initial points (i.e., signal plans). These points are uniformly drawn from the feasible space defined by Equations (13) and (14). For each initial point, we run the SO algorithm five times, each time allowing for 150 simulation runs. Thus, for each method and each initial point, we derive five "optimal" (or proposed) signal plans. We then use the simulator to evaluate in detail the performance of the proposed signal plans. For each proposed signal plan, we run 50 replications. We compare the empirical cumulative distribution function (cdf) of the average travel times obtained from these 50 replications.

Each plot of Figure 4 considers a different randomly drawn initial point. Each curve of each plot displays the empirical cdf's of a given signal plan. The solid thick curve corresponds to the empirical cdf of the initial signal plan (denoted x_0), the dashed curves (resp. solid thin curves) are the empirical cdf's of signal plans proposed by the traditional metamodel, i.e., the polynomial ϕ (resp. the proposed metamodel m).

Figure 4(a) indicates that all five plans derived by both the proposed metamodel and the traditional metamodel yield improved performance when compared with the initial signal plan. All five plans derived by the proposed metamodel also have better performance compared with those proposed by the traditional metamodel.

Figure 4(b) indicates that all five signal plans derived by the proposed metamodel yield improved performance when compared with the initial plan. Four of them outperform all five plans derived by the traditional metamodel. Two of the signal plans derived by the traditional metamodel outperform the initial plan,

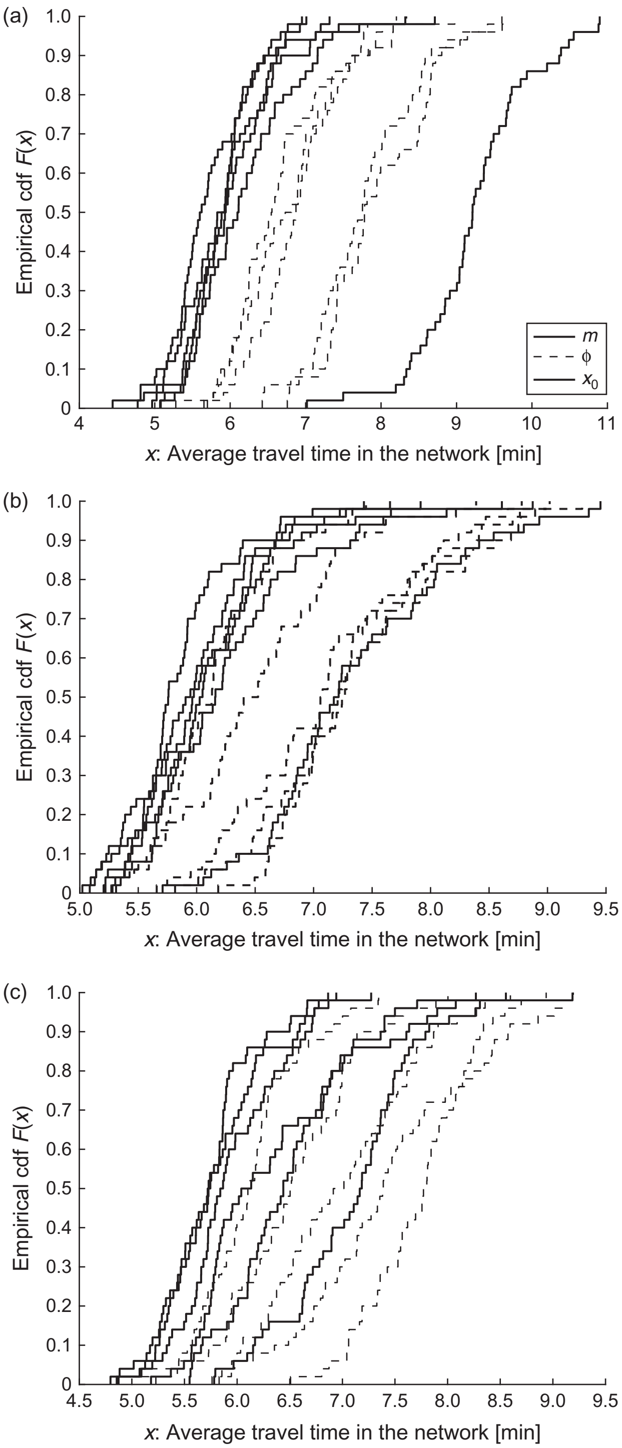

Figure 4 **Empirical cdf's of the Average Travel Times Considering Initial Random Signal Plans and Allowing for 150 Simulation Runs**

and the other three have performance similar to the initial plan.

In Figure 4(c), all five plans derived by the proposed metamodel yield improvement compared with the initial plan, and three of them outperform all five signal plans proposed by the traditional metamodel. Two of the signal plans proposed by the traditional metamodel have worse performance than the initial signal plan, one has similar performance, and two have improved performance.

For all three initial points, the proposed method systematically derives signal plans with improved performance when compared with the initial plan and, most often, when compared with the plans obtained from the traditional metamodel. Additionally, the plans derived by the proposed method have good and very similar performance across all SO runs and all initial points, whereas the performance of the plans proposed by the traditional metamodel varies depending on both the initial point and the SO run. This illustrates the robustness of the proposed method to both the initial points and the stochastics of the simulator.

We evaluate the performance of the proposed approach for larger sample sizes. We run the SO algorithm once and allow for a total of 1,500 simulation runs. We choose two random initial signal plans. We evaluate the performance of the signal plans proposed at sample sizes 50, 150, 200, 400, 600, 800, 1,000, and 1,500. We evaluate their performance just as before—i.e., for a given proposed plan we run 50 replications of the simulator and plot the empirical cdf (over these 50 replications) of the average travel times.

Figure 5(a) displays the corresponding cdf's of the initial signal plan used in Figure 4(a). The proposed approach identifies a signal plan with excellent performance already at sample size 50 (cdf labeled m 50). The signal plan identified as of sample size 150 remains the best up to sample size 1,500. It has slightly improved performance, and in particular reduced variability, compared with that of sample size 50.

The performance of the signal plans proposed by the traditional metamodel (dashed curves) improves as the sample size increases. The traditional metamodel requires a much larger sample size to identify signal plans with good performance.

We carry out a paired t-test to evaluate whether the difference in performance of the signal plans proposed by each method at sample size 1,500 is statistically significant. We assume that the observed average travel times arise from a normal distribution with common but unknown variance. The null hypothesis assumes that the expected travel time is the same for both methods, whereas the alternative hypothesis assumes that they differ. The confidence level is 0.05, and there are 49 degrees of freedom. The sample average and sample standard deviation of our proposed signal plan (resp. that proposed by the polynomial metamodel) are 5.73 minutes and 0.51 minutes (resp. 5.95 minutes and 0.47 minutes). The critical value of the test is 1.96. The difference is statistically significant (t-statistic of -2.38, p-value of 0.02).

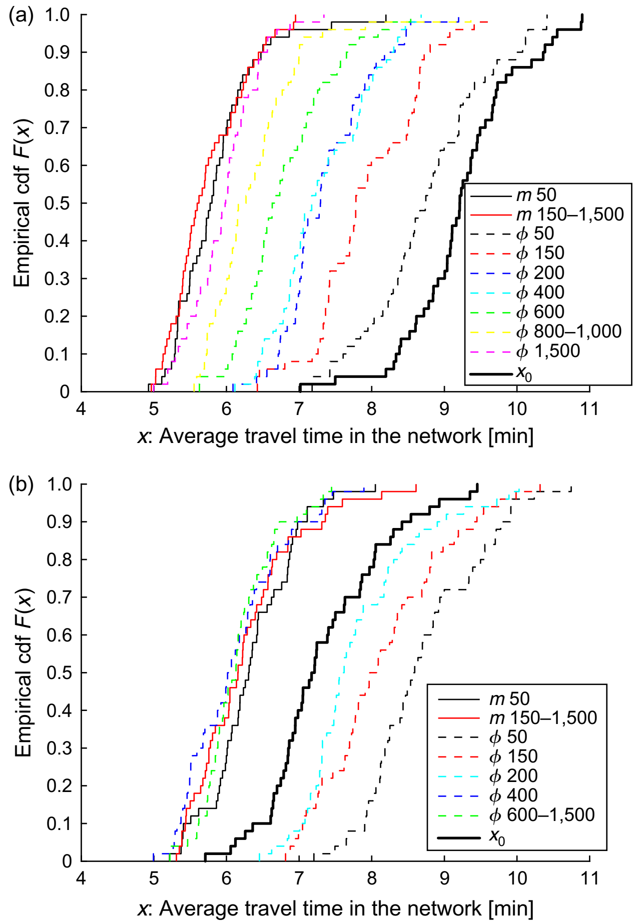

Figure 5 (Color online) Empirical cdf's of the Average Travel Times Considering Initial Random Signal Plans and Allowing for 1,500 Simulation Runs

Thus, at sample size 1,500 the proposed method still outperforms its traditional counterpart. That is, the signal plan identified by the proposed method as of sample size 150 outperforms that identified by the traditional method at sample size 1,500.

Figure 5(b) displays the results considering the initial plan used in Figure 4(b). Similarly, the proposed approach identifies a signal plan with an excellent performance even at sample size 50. The signal plan with the best performance derived by the proposed metamodel is obtained at sample size 150 and remains the same until sample size 1,500. It has a similar performance to that of sample size 50.

For sample sizes smaller than 400, the traditional metamodel yields signal plans with worse performance than the initial plan. Their performance significantly improves with increasing sample size until size 400. The performance of the derived signal plans with samples larger than 400 are similar. The signal plans proposed by the traditional metamodel method for sample sizes 600 to 1,500 are the same.

We carry out the same paired t-test as before to evaluate whether the difference in performance of the signal plans proposed by each method at sample size 1,500 is statistically significant. The sample average and sample standard deviation of our proposed signal plan (resp. that proposed by the polynomial metamodel) are 6.25 minutes and 0.73 minutes (resp. 6.16 minutes and 0.50 minutes). The difference is not statistically significant (t-statistic of 0.72, p-value of 0.48).

Figure 6 displays two instances of the Lausanne city map. The links are colored based on average link travel times (averaged over the 50 replications). The left (resp. right) map considers the average link travel times for the initial (resp. proposed) signal plan. Here the proposed plan is that obtained with the initial plan and sample size of 150 of Figure 5(a). Green links have average travel times below 40 seconds, yellow links have travel times between 40 and 80 seconds, while red links have travel times greater than 80 seconds. This figure shows how the proposed plan yields city-wide travel time improvements.

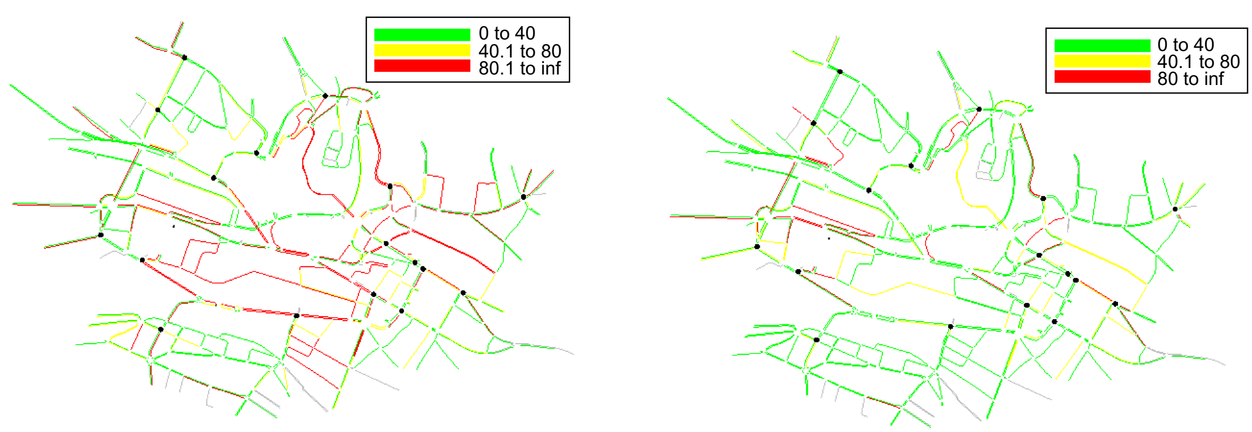

Figure 6 (Color online) Average Link Travel Times Using the Initial Signal Plan (Left Map) and the Signal Plan Proposed by the SO Approach (Right Map)
Note. The averages (in seconds) are taken over 50 simulation replications.

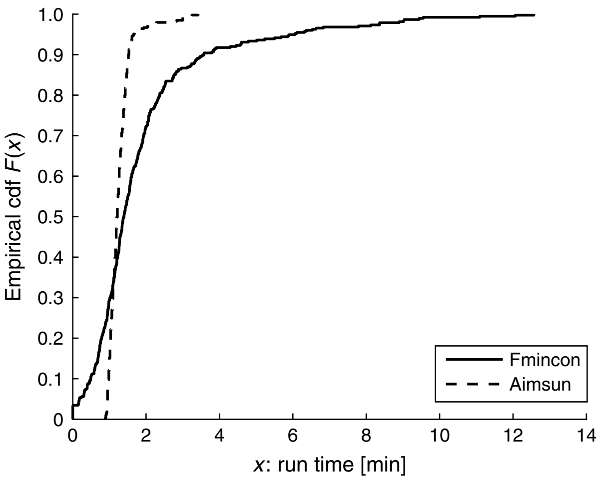

Figure 7 Simulation and Trust Region Subproblem Run Times

At each iteration of the SO algorithm, the two most computationally expensive tasks are the evaluation of the simulator as well as the solution of the trust-region (TR) subproblem (i.e., call of the fmincon routine). We consider the first initial plan (used in Figures 4(a) and 5(a)) and account for all five runs. Figure 7 displays the cdf of the simulation runs and the TR subproblem runs. On average one simulation run takes 1.3 minutes, and it takes 1.9 minutes to solve the TR subproblem. The experiments were run on a standard laptop (with a 2.7 GHz processor and 4 GB of RAM). Thus, the metamodel can be used to efficiently solve the TR subproblem at each iteration of the SO algorithm. Additionally, the structural information that it provides through the queueing network model allows the SO algorithm to identify signal plans with excellent performance under tight computational budgets.

5.3. Synchro Comparison

In this section we compare the performance of the signal plans derived by our approach with those derived by the mainstream, commercial, and widely used traffic signal control software Synchro (Trafficware 2011, Synchro 8). Synchro is a traffic signal control optimization software based on a macroscopic, deterministic, and local traffic model. It is widely used across the United States (NYCDOT (2012); Riniker, Eisenach, and Hannan (2009); Abdel-Rahim and Dixon (2007); ATAC (2003)). For details on the split optimization technique within Synchro, we refer the reader to Chapter 14 of Trafficware (2011).

The Synchro version used does not allow for any fixed (i.e., exogenous) phase durations. Hence, we solve a signal control problem without fixed phases. For each intersection we take as cycle time its available (i.e., nonfixed) cycle time, $c_i - d_i$. The problem formulation is given by Equations (12)–(14) and by replacing the right-hand side of Equation (13) by $(c_i - d_i)/(c_i - d_i)$, which equals 1. Synchro and our proposed SO method address the same problem. The corresponding TR subproblem is given by Equations (15)–(21) and by replacing the right-hand side of (16) by 1 and the right-hand side of (18) with zero.

The Lausanne network is coded in Synchro. All signal plan information needed for Synchro (e.g., phase structure) is obtained from the existing Lausanne signal plan. The minimum splits are set to 4 seconds as in §5.1. Lane saturation flows (denoted s in §4.1) are set to 1,800 vehicles per hour, following Swiss transportation norms. Synchro also needs, as inputs, estimates of prevailing movement flows. This was also needed when calibrating the analytical queueing model (e.g., to obtain turning probabilities). Hence, we use the same estimates as those provided to the queueing model. These are obtained from the simulator using the existing Lausanne signal plan.

To initialize the proposed SO approach, we consider the same three random initial signal plans as used in Figure 4. For each initial plan, we run the SO algorithm once, each time allowing for 150 simulation runs. To evaluate the performance of a plan, we use the simulator and proceed as described in §5.2.

Figure 8 presents the corresponding cdf curves. The three solid thin curves correspond to the plans derived by our proposed metamodel approach (denoted m). The dashed curves correspond to the three random initial signal plans (denoted x_0). The solid thick curve corresponds to the Synchro plan. All three plans derived by the purposed metamodel approach yield improved performance when compared with all three initial plans. All three plans derived by the SO approach also outperform the plan proposed by Synchro. The Synchro plan has performance similar to two of the three randomly drawn signal plans.

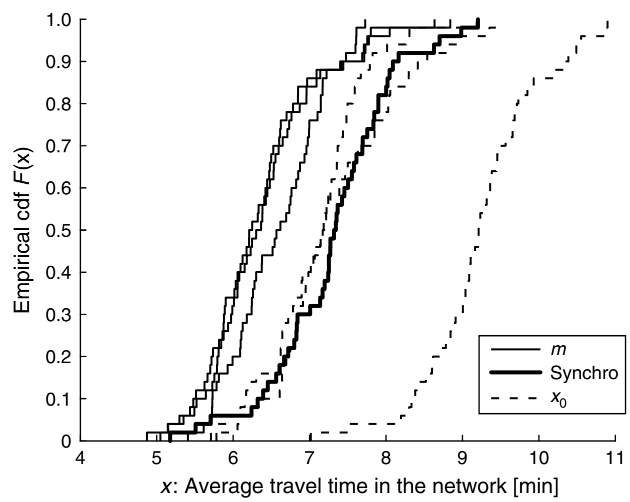

Figure 8 Emprical cdf's of the Average Travel Times of the Signal Plans Proposed by the SO Approach and by Synchro

6. Conclusions

This paper proposes a metamodel for large-scale simulation-based urban transportation optimization problems. It is a computationally efficient technique that identifies trial points (e.g., signal plans) with improved performance under tight computational budgets. This metamodel SO technique is based on the use of a highly tractable metamodel that combines a general-purpose component (a quadratic polynomial) with a physical component (a highly tractable analytical queueing network model).

We evaluate the performance of this approach by addressing a large-scale network-wide signal control problem for the Swiss city of Lausanne. This problem considers a congested network (evening peak period demand) with an intricate topology. We compare the performance of the proposed metamodel with that of a traditional metamodel. The proposed method identifies signal plans that improve the distribution of average travel times compared with both the initial signal plans and, most often, the signal plans derived by the traditional method. This network-wide signal control problem is considered high-dimensional for SO algorithms, derivative-free algorithms, and signal control algorithms. We also compare the performance of the proposed approach with that of a widely used signal control software, Synchro. All proposed signal plans outperform the plan derived by Synchro.

In this paper, random uniformly drawn signal plans are used as initial points for the SO algorithm. The results illustrate the robustness of the proposed metamodel method to initial points. This allows practitioners to use the method to address a variety of signal control problems without requiring any field knowledge to initialize the method.

As part of ongoing research, we are investigating the use of the proposed method to address a variety of generally constrained simulation-based transportation problems, including microscopic model calibration, multimodal traffic management, and multimodal network design problems. We are also developing SO algorithms with improved short-term performance by using information from analytical probabilistic traffic models, such as the queueing network model used in this paper, to inform both sampling strategies and statistical tests.

We are also investigating novel analytical traffic model formulations with increased accuracy. The model used in this manuscript is a stationary model. We are currently working on a time-dependent formulation based on the use of transient finite capacity queueing theory. Ongoing work is also developing a formulation with endogenous analytical traffic assignment. The main challenge in this analytical work is to derive a differentiable and highly tractable formulation suitable for large-scale simulation-based optimization.

Acknowledgments

The authors thank Emmanuel Bert and André-Gilles Dumont (Traffic Facilities Laboratory (LAVOC), Ecole Polytechnique Fédérale de Lausanne (EPFL)) for providing the Lausanne simulation model. This research was partially supported by the Center for Complex Engineering Systems at the King Abdulaziz City for Science and Technology (KACST) and the Massachusetts Institute of Technology (MIT).

Appendix A. Derivation of $E[N]$

In this section we omit the index i that refers to a given queue. $E[N]$ is defined as

$$E[N] = \sum_{n=0}^{k} nP(N=n). \tag{A1}$$

The stationary probabilities for each queue, $P(N=n)$, are given in Bocharov et al. (2004) by

$$P(N=n) = \frac{1-\rho}{1-\rho^{k+1}}\rho^n. \tag{A2}$$

Inserting Equation (A2) into (A1) and then rearranging the terms yields

$$E[N] = \sum_{n=0}^{k} n\frac{1-\rho}{1-\rho^{k+1}}\rho^n, \tag{A3}$$

$$= \sum_{n=1}^{k} n\frac{1-\rho}{1-\rho^{k+1}}\rho^n, \tag{A4}$$

$$= \frac{1-\rho}{1-\rho^{k+1}}\sum_{n=1}^{k} n\rho^n, \tag{A5}$$

$$= \frac{1-\rho}{1-\rho^{k+1}}\rho\sum_{n=1}^{k} n\rho^{n-1}. \tag{A6}$$

We then derive an expression for the last summation as follows. For a geometric series, such that $\rho \neq 1$, we have

$$\sum_{n=0}^{k} \rho^n = \frac{\rho^{k+1}-1}{\rho-1}. \tag{A7}$$

We differentiate this formula with respect to ρ and obtain

$$\sum_{n=1}^{k} n\rho^{n-1} = \frac{1-\rho^{k+1}}{(1-\rho)^2} - \frac{(k+1)\rho^k}{1-\rho}. \tag{A8}$$

Inserting the expression of Equation (A8) into Equation (A6), and rearranging the terms gives

$$E[N] = \frac{1-\rho}{1-\rho^{k+1}}\rho\left(\frac{1-\rho^{k+1}}{(1-\rho)^2} - \frac{(k+1)\rho^k}{1-\rho}\right) \tag{A9}$$

$$= \rho\left(\frac{1}{1-\rho} - \frac{(k+1)\rho^k}{1-\rho^{k+1}}\right). \tag{A10}$$

Appendix B. SO Algorithm

This SO algorithm is formulated in detail in Osorio and Bierlaire (2013) and is based on the derivative-free trust-region (TR) algorithm of Conn, Scheinberg, and Vicente (2009a). The parameters of the algorithm are set according to the values in Osorio and Bierlaire (2013).

0. *Initialization*. Define for a given iteration k: $m_k(x, y; \alpha_k, \beta_k, q)$ as the metamodel (denoted hereafter as $m_k(x)$), x_k

as the iterate, Δ_k as the TR radius, $\nu_k = (\alpha_k, \beta_k)$ as the vector of parameters of m_k, n_k as the total number of simulation runs carried out up until and including iteration k, u_k as the number of successive trial points rejected, and ε_k as the measure of stationarity (norm of the derivative of the Lagrangian function of the TR subproblem with regard to the endogenous variables) evaluated at x_k.

The constants η_1, γ, γ_{inc}, ε_c, $\bar{\tau}$, $\bar{d}$, $\bar{u}$, and $\Delta_{\max}$ are given such that $0 < \eta_1 < 1$, $0 < \gamma < 1 < \gamma_{\mathrm{inc}}$, $\varepsilon_c > 0$, $0 < \bar{\tau} < 1$, $0 < \bar{d} < \Delta_{\max}$, and $\bar{u} \in \mathbb{N}^*$. Set the total number of simulation runs permitted (across all points) $n_{\max}$; this determines the computational budget. Set the number of simulation replications per point $\tilde{r}$ (here we use $\tilde{r} = 1$).

Set $k = 0$, $n_0 = 1$, and $u_0 = 0$. Determine x_0 and Δ_0 ($\Delta_0 \in (0, \Delta_{\max}]$).

Given the initial point x_0, compute $f_A(x_0)$ (analytical approximation of Equation (12)) and $\hat{f}(x_0)$ (simulated estimate of Equation (12)), fit an initial model m_0 (i.e., compute ν_0).

1. *Criticality step*. If $\varepsilon_k \le \varepsilon_c$, then switch to *conservative mode*.

2. *Step calculation*. Compute a step s_k that reduces the model m_k and such that $x_k + s_k$ (the trial point) is in the TR (i.e., approximately solve the TR subproblem).

3. *Acceptance of the trial point*. Compute $\hat{f}(x_k + s_k)$ and

$$\rho_k = \frac{\hat{f}(x_k) - \hat{f}(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

—If $\rho_k \ge \eta_1$, then accept the trial point: $x_{k+1} = x_k + s_k$, $u_k = 0$.

—Otherwise, reject the trial point: $x_{k+1} = x_k$, $u_k = u_k + 1$.

Include the new observation in the set of sampled points ($n_k = n_k + \tilde{r}$), and fit the new model m_{k+1}.

4. *Model improvement*. Compute $\tau_{k+1} = \|\nu_{k+1} - \nu_k\| / \|\nu_k\|$. If $\tau_{k+1} < \bar{\tau}$, then improve the model by simulating the performance of a new point x, which is uniformly drawn from the feasible space. Evaluate f_A and $\hat{f}$ at x. Include this new observation in the set of sampled points ($n_k = n_k + \tilde{r}$). Update m_{k+1}.

5. *TR radius update*.

$$\Delta_{k+1} = \begin{cases} \min\{\gamma_{\mathrm{inc}}\Delta_k, \Delta_{\max}\} & \text{if } \rho_k > \eta_1 \\ \max\{\gamma\Delta_k, \bar{d}\} & \text{if } \rho_k \le \eta_1 \text{ and } u_k \ge \bar{u} \\ \Delta_k & \text{otherwise.} \end{cases}$$

If $\rho_k \le \eta_1$ and $u_k \ge \bar{u}$, then set $u_k = 0$.
If $\Delta_{k+1} \le \bar{d}$, then switch to *conservative mode*.

Set $n_{k+1} = n_k$, $u_{k+1} = u_k$, and $k = k + 1$.

If $n_k < n_{\max}$, then go to Step 1. Otherwise, stop.

References

Abdel-Rahim A, Dixon M (2007) Guidelines for designing and implementing traffic control systems for small- and medium-sized cities in Idaho. Technical Report N06-18, Idaho Department of Transportation, Boise, ID.

Aboudolas K, Papageorgiou M, Kosmatopoulos E (2007) Control and optimization methods for traffic signal control in large-scale congested urban road networks. *Amer. Control Conf.* 3132–3138.

Aboudolas K, Papageorgiou M, Kouvelas A, Kosmatopoulos E (2010) A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks. *Transportation Res. Part C* 18(5):680–694.

Alexandrov NM, Lewis RM, Gumbert CR, Green LL, Newman PA (1999) Optimization with variable-fidelity models applied to wing design. Technical Report CR-1999-209826, NASA Langley Research Center, Hampton, VA.

ATAC (2003) Signal coordination strategies final report. Technical report, Advanced Traffic Analysis Center, Upper Great Plains Transportation Institute, North Dakota State University, Fargo.

Barton RR, Meckesheimer M (2006) Metamodel-based simulation optimization. Henderson SG, Nelson BL, eds. *Handbooks in Operations Research and Management Science: Simulation*, Vol. 13 (Elsevier, Amsterdam), 535–574.

Ben-Akiva M, Cuneo D, Hasan M, Jha M, Yang Q (2003) Evaluation of freeway control using a microscopic simulation laboratory. *Transportation Res. Part C* 11(1):29–50.

Bocharov PP, D'Apice C, Pechinkin AV, Salerno S (2004) *Queueing Theory*, Modern Probability and Statistics (Brill Academic Publishers, Zeist, Netherlands), 96–98.

Branke J, Goldate P, Prothmann H (2007) Actuated traffic signal optimization using evolutionary algorithms. *Proc. 6th Eur. Congress and Exhibition on Intelligent Transport Systems and Services, Aalborg, Denmark*, June 18–20.

Bullock D, Johnson B, Wells RB, Kyte M, Li Z (2004) Hardware-in-the-loop simulation. *Transportation Res. Part C* 12(1):73–89.

Chen X, Osorio C, Santos BF (2012) A simulation-based approach to reliable signal control. *Proc. Internat. Sympos. Transportation Network Reliability (INSTR)*.

Coleman TF, Li Y (1994) On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. *Math. Programming* 67(2):189–224.

Coleman TF, Li Y (1996) An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* 6(2):418–445.

Conn AR, Gould NIM, Toint PL (2000) *Trust-Region Methods*, MPS/SIAM Series on Optimization (SIAM and MPS, Philadelphia).

Conn AR, Scheinberg K, Vicente LN (2009a) Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points. *SIAM J. Optim.* 20(1):387–415.

Conn AR, Scheinberg K, Vicente LN (2009b) *Introduction to Derivative-Free Optimization*, MPS/SIAM Series on Optimization (SIAM and MPS, Philadelphia).

Dumont AG, Bert E (2006) Simulation de l'agglomération Lausannoise SIMLO. Technical report, Laboratoire des voies de circulation, ENAC, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Fu MC, Glover FW, April J (2005) Simulation optimization: A review, new developments, and applications. Kuhl ME, Steiger NM, Armstrong FB, Joines JA, eds. *Proc. 2005 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 83–95.

Gartner NH, Hou DL (1992) Comparative evaluation of alternative traffic control strategies. *Transportation Res. Record* 1360:66–73.

Hachicha W, Ammeri A, Masmoudi F, Chachoub H (2010) A comprehensive literature classification of simulation optimisation methods. *Proc. Internat. Conf. Multiple Objective Programming and Goal Programming MOPGP10, Sousse, Tunisia*, May 24–25.

Hale D (2005) Traffic network study tool TRANSYT-7F. Technical report, McTrans Center, University of Florida, Gainesville.

Hasan M (1999) Evaluation of ramp control algorithms using a microscopic traffic simulation laboratory, MITSIM. Master's thesis, Massachusetts Institute of Technology, Cambridge.

Hasan M, Jha M, Ben-Akiva M (2002) Evaluation of ramp control algorithms using microscopic traffic simulation. *Transportation Res. Part C* 10(3):229–256.

Joshi S, Rathi A, Tew J (1995) An improved response surface methodology algorithm with an application to traffic signal optimization for urban networks. Alexopoulos C, Kang K, Lilegdon WR, Goldsman D, eds. *Proc. 1995 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1104–1109.

Kleijnen JPC (2008) Response surface methodology for constrained simulation optimization: An overview. *Simulation Modelling Practice and Theory* 16(1):50–64.

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

Li P, Abbas M, Pasupathy R, Head L (2010) Simulation-based optimization of maximum green setting under retrospective approximation framework. *Transportation Res. Record* 2192:1–10.

Little JDC (1961) A proof for the queueing formula: $L = \lambda W$. *Oper. Res.* 9(3):383–387.

Little JDC (2011) Little's law as viewed on its 50th anniversary. *Oper. Res.* 59(3):536–549.

Marti K (2008) *Stochastic Optimization Methods* (Springer, Berlin).

NYCDOT (2012) Downtown Flushing, mobility and safety improvement project. Technical report, New York City Department of Transportation, New York.

Osorio C (2010) Mitigating network congestion: Analytical models, optimization methods and their applications. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Osorio C, Bierlaire M (2009a) An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *Eur. J. Oper. Res.* 196(3):996–1007.

Osorio C, Bierlaire M (2009b) A surrogate model for traffic optimization of congested networks: An analytic queueing network approach. Technical Report 090825, Transport and Mobility Laboratory, ENAC, Ecole Polytechnique Fédérale de Lausanne, Lausaane, Switzerland. http://web.mit.edu/osorioc/www/papers/osorBier09TechRepQgTraf.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 (2013) Emissions mitigation: Coupling microscopic emissions and urban traffic models for signal control. Technical Report 19082013, Massachusetts Institute of Technology, Cambridge. http://web.mit.edu/osorioc/www/papers/osoNanEmissionsSO.pdf.

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

Papayannoulis V, Marsico M, Maguire T, Strasser J, Scalici S (2011) An integrated travel demand, mesoscopic and microscopic modeling platform to assess traffic operations for Manhattan, New York. *Proc. Transportation Res. Board (TRB) Conf.* (TRB, Washington, DC).

Rathi AK, Lieberman EB (1989) Effectiveness of traffic restraint for a congested urban network: A simulation study. *Transportation Res. Record* 1232:95–102.

Riniker K, Eisenach P, Hannan T (2009) City of Winchester, VA traffic signal upgrade project. Technical report, Institute of Transportation Engineers, Washington, DC.

Søndergaard J (2003) Optimization using surrogate models—By the space mapping technique. Ph.D. thesis, Technical University of Denmark, Lyngby, Denmark.

Stallard C, Owen L (1998) Evaluating adaptive signal control using CORSIM. Medeiros D, Watson E, Carson J, Manivannan M, eds. *Proc. 1998 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1147–1153.

Stevanovic J, Stevanovic A, Martin PT, Bauer T (2008) Stochastic optimization of traffic control and transit priority settings in VISSIM. *Transportation Res. Part C* 16(3):332–349.

Toledo T, Koutsopoulos HN, Davol A, Ben-Akiva ME, Wilco B, Andreasson I, Johansson T, Lundin C (2003) Calibration and validation of microscopic traffic simulation tools: Stockholm case study. *Transportation Res. Record* 1831:65–75.

Traffic Technology International (2012a) Admirable solution. *Traffic Technology International* (February/March). http://viewer.zmags.com/publication/e8a8bd2e#/e8a8bd2e/1.

Traffic Technology International (2012b) In the frame. *Traffic Technology International (April/May)*. http://viewer.zmags.com/publication/e1a5fab8#/e1a5fab8/1.

Trafficware (2011) Synchro Studio 8 User Guide. Trafficware, Sugar Land, TX.

TSS (2008) AIMSUN NG and AIMSUN Micro Version 5.1. Transport Simulation Systems, Barcelona, Spain.

VSS (1992) Norme Suisse SN 640837 Installations de feux de circulation; temps transitoires et temps minimaux. Union des professionnels suisses de la route, VSS, Zurich.

Wild SM, Regis RG, Shoemaker CA (2008) ORBIT: Optimization by radial basis function interpolation in trust-regions. *SIAM J. Scientific Comput.* 30:3197–3219.

Yun I, Park B (2006) Application of stochastic optimization method for an urban corridor. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1493–1499.