A surrogate model for traffic optimization of congested networks: an analytic queueing network approach *

C. Osorio † M. Bierlaire †

August 25th, 2009

Report TRANSP-OR 090825
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*This research is supported by the Swiss National Science Foundation grants 205321-107838 and 205320-117581
†TRANSP-OR, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, {carolina.osoriopizano, michel.bierlaire}@epfl.ch

Abstract

Congested networks involve complex traffic dynamics that can be accurately captured with detailed simulation models. However, when performing optimization of such networks the use of simulators is limited due to their stochastic nature and their relatively high evaluation cost. This has lead to the use of general-purpose analytical metamodels, that are cheaper to evaluate and easier to integrate within a classical optimization framework, but do not capture the specificities of the underlying congested conditions.

In this paper, we argue that to perform efficient optimization for congested networks it is important to develop analytical surrogates specifically tailored to the context at hand so that they capture the key components of congestion (e.g. its sources, its propagation, its impact) while achieving a good tradeoff between realism and tractability. To demonstrate this, we present a surrogate that provides a detailed description of congestion by capturing the main interactions between the different network components while preserving analytical tractable.

In particular, we consider the optimization of vehicle traffic in an urban road network. The proposed surrogate model is an approximate queueing network model that resorts to finite capacity queueing theory to account for congested conditions. Existing analytic queueing models for urban networks are formulated for a single intersection, and thus do not take into account the interactions between queues. The proposed model considers a set of intersections and analytically captures these interactions.

We show that this level of detail is sufficient for optimization in the context of signal control for peak hour traffic. Although there is a great variety of signal control methodologies in the literature, there is still a need for solutions that are appropriate and efficient under saturated conditions, where the performance of signal control strategies and the formation and propagation of queues are strongly related. We formulate a fixed-time signal control problem where the network model is included as a set of constraints. We apply this methodology to a subnetwork of the Lausanne city center and use a microscopic traffic simulator to validate its performance. We also compare it with several other methods. As congestion increases, the new method leads to improved average performance measures. The results highlight the importance of taking the interaction between consecutive roads into account when deriving signal plans for congested urban road networks.

1 Introduction

The complexity of the traffic dynamics in congested networks has lead to the development of simulation models capable of capturing in detail the interactions between the different network components. For instance, advanced microscopic urban traffic simulators, as well as pedestrian simulators, provide a precise representation of the interactions of individuals with both the network supply components and with other individuals. These simulation tools can provide accurate network performance estimates in the context of scenario-based analysis (i.e. what-if analysis) or sensitivity analysis. Unfortunately, their integration within an optimization framework remains an intricate task.

Indeed, a simulator can be seen as a stochastic, nonlinear, non-continuous function with both discrete and continuous variables, that is often expensive to evaluate. The optimization problems considered are often nonlinear mixed-integer constrained problems, which are already hard to solve in a deterministic setting. Fu [28] highlights the additional challenges that arise when the stochastic nature of the problem is accounted for.

Therefore, the field of simulation-based optimization has been investigated quite widely in the literature. Reviews of the existing methods include [23, 29, 28, 49]. Since performance measures

cannot be evaluated exactly, but only estimated, it is difficult to conclude about the optimality or even improvement of a specific design. Gradient estimations also become more involved, not to mention costly, although there have been significant advances and novel approaches for their estimation [29, 58, 22, 20].

Metamodel-based simulation optimization uses a surrogate model of the simulation model (i.e. a metamodel) to perform optimization. The surrogates are typically less accurate but cheaper to evaluate. They are often a linear combination of basis functions from a parametric family, such as polynomials or splines. Recent reviews of commonly used surrogates are given in [21, 7]. Nevertheless, by resorting to a general-purpose surrogate model, the simulator is used as a blackbox since no information with regards to the underlying structure of the problem is used.

Therefore, we claim that to optimize congested networks it is important to formulate metamodels that are context specific and that capture the key components of congestion (e.g. its sources, its propagation, its impact). The main challenge of such models is to achieve a good tradeoff between realism and tractability. In this paper we develop such a surrogate for congested urban road networks. Modeling urban congestion is a particularly intricate task, since it involves complex components such as driver behavior, endogenous road capacities, priority rules, as well as their interactions. Additionally, the increasing incidence and cost of urban congestion shows the importance of deriving surrogates for traffic management. To illustrate the use of this surrogate, we consider the problem of optimizing signal plans for peak hour traffic. Within this context the contributions of this paper are two-fold.

Firstly, the proposed surrogate model is an analytic stochastic network model, derived from the queueing model proposed in [50]. Existing analytic queueing network models have focused on the study of uninterrupted traffic flow. To the best of our knowledge, the few studies that consider interrupted traffic flow are formulated for a single intersection. They therefore do not take into account the interaction between flows on upstream and downstream roads. The framework that we present models a set of urban intersections. It captures the correlation structure between consecutive roads by combining approximation methods with *finite capacity queueing theory*. This correlation provides a detailed description of congestion. It identifies the sources of congestion (e.g. bottlenecks), describes how congestion propagates (e.g. spillbacks) and dissipates; and quantifies the impact on the network performance.

The second contribution of this paper is to provide evidence that the use of this model as a surrogate to perform traffic optimization indeed achieves a good compromise between realism and tractability. We formulate a fixed-time signal control problem where the network model is included as a set of constraints. To the best of our knowledge, the existing signal control strategies based on analytic network models have not taken spillbacks into account. More generally, most signal control strategies do not account for saturated or highly congested networks [52]. The considered model grasps spillbacks by accounting for between-queue correlation, it is therefore an appropriate surrogate for signal control under congested conditions.

This paper is structured as follows. We present in Section 2 a literature review of analytic queueing models for urban traffic and of signal control methodologies. We describe the surrogate network model (Section 3) and formulate the optimization problem (Section 4). We then present and discuss the role of a microscopic traffic simulator used in this framework (Section 5.1). The methodology is applied to a subnetwork of the Lausanne city center. The optimized signal plan is then compared with plans generated by several other methods. Section 5.2 compares it with the plan derived by the queueing model assuming independent queues. This section analyzes the added value of the explicit modeling of correlation. In Section 5.3 the signal plan is compared to an existing plan for the city of Lausanne, and to the plans derived by the methods proposed in [69, 60].

2 Literature Review

2.1 Analytic queueing models

Queueing models have been used in transportation mainly to model highway traffic [30]. Several simulation models have been developed, but few studies have explored the potential of the queueing theory framework to develop analytic urban traffic models. Furthermore, existing urban queueing models have mainly focused on unsignalized intersections. Heidemann and Wegmann [36] give an excellent literature review for exact analytic queueing models of unsignalized intersections. They model the minor stream as an M/G2/1 queue. They emphasize the importance of the pioneer work of Tanner [59]. Heidemann also contributed to the study of signalized intersections [34], and presented a unifying approach to both signalized and unsignalized intersections [35]. These models combine a queueing theory approach with a realistic description of traffic processes for a given lane at a given intersection. They yield detailed performance measures such as queue length distributions or sojourn time distributions. Nevertheless, as exact analytic methods, they are difficult to generalize to consider multiple lanes, not to mention multiple intersections.

To the best of our knowledge no method has been proposed to model urban traffic for a set of intersections using an analytic queueing network framework. Nevertheless the methods proposed in [39, 62] which are both based on the Expansion Method [40] and formulated for highway traffic could be extended to consider an urban setting.

As described in detail in [50], methods that allow the exact evaluation of the joint stationary distribution of the network of queues are difficult to obtain, let alone transient distributions [48]. The method proposed here is based on an approximation method that decomposes the network into single queues and estimates their marginal distributions. It describes congestion by using a novel formulation of the state space of the queues combined with what is known in queueing theory as the blocking mechanism [53].

2.2 Traffic signal control

Traffic signal setting strategies can be either fixed-time or traffic-responsive strategies. *Fixed-time* (also called *pre-timed*) strategies use historical traffic data, and yield one traffic signal setting for the considered time of day. The traffic signal optimization problem is solved offline. On the other hand *traffic-responsive* (also called *real-time*) methods use real-time data to define timings for immediate implementation that are used over a short time horizon. Furthermore, signal timings can be derived by considering either a single or a set of intersections. These methods are called *isolated* methods and *coordinated* methods, respectively [52]. Methods that handle individual intersections are based on models that capture the local dynamics of the network. They describe in detail the dynamics at an intersection, but at the expense of capturing less well the interactions among intersections.

A *phase* is defined as a set of 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. Methods where the stage structure (i.e. the sequence of stages) is given are known as *stage-based* approaches, whereas methods where the stage structure is endogenous are referred to as *phase-based* or *group-based* approaches.

Delay minimization and reserve capacity maximization are the most common objective functions used by existing methods. Delay may be directly measured, leading to a data-driven approach, or estimated (model-based approach). The first approximate expression for the delay at an intersection was given by Webster [69], and is still widely used. Other expressions include those given in [47, 45, 44]. Viti [63] provides a review of delay models; Dion et al. [26] compare the performance of different delay

models, and Chow and Lo [17] derive approximate delay derivatives that can be integrated within a simulation-based signal setting optimization context in order to reduce the computation time required to obtain numerical derivatives. The notion of the *reserve capacity* of an intersection is defined by Wong and Yang [72] as the greatest common multiplier of existing flows that can be accommodated subject to saturation and signal timing constraints. This notion has been extended to consider several intersections [72, 75].

The works of Allsop [6] and of Shepherd [57] review signal control methods. Allsop [6] describes in detail the corresponding terminology as well as the different formulations for isolated methods. More recently the reviews of Papageorgiou et al. [52] and Cascetta et al. [15] cover different but complementary aspects of this research field. Papageorgiou et al. [52] provide an excellent review of urban traffic control methods, while highlighting their applications (either via simulation or field implementations). They also consider freeways and route guidance methodologies. Cascetta et al. [15] review the more general problem of traffic control and demand assignment methods. A detailed review of signal control methodologies is provided in Appendix A of this paper.

The method proposed here belongs to the category of fixed-time coordinated methods. Traditionally, fixed-time strategies have been considered suitable only for undersaturated traffic conditions [3, 57, 17, 52]. Thus methods for saturated conditions have focused on real-time strategies. Nevertheless, we believe that the development of optimal fixed-time methods is of primary importance. First, they can be used as benchmark solutions to evaluate traffic-responsive strategies. Second, they represent robust control solutions [74]. Finally, they may be directly or indirectly used as building blocks to derive real-time methods.

Although there is a vast range of signal control methodologies in the literature, there is still a need for solutions that are appropriate and efficient under saturated conditions [25]. Under congested conditions the performance of signal control strategies and the formation and propagation of queues are strongly related. Models that ignore the spatial extension of queues fail to capture congestion effects such as spillbacks, and gridlocks. Adopting a vertical queueing model is therefore only reasonable when the degree of saturation is moderate. The effects of ignoring this spatial dimension are illustrated in [17, 3]. Therefore a signal control strategy suitable for congested conditions must take into account the correlation between queues. Nevertheless, most existing strategies do not account for this correlation and are thus unsuitable for highly congested networks [52, 2]. Furthermore Abu-Lebdeh and Benekohal [2] emphasize that accounting for the effects of queue propagation remains a secondary consideration within a signal timing framework.

The recently proposed TUC method [25] focuses on saturated traffic conditions. It overcomes the exponential complexity of the existing methods by avoiding the use of discrete variables, which "is of paramount importance because it opens the way to the application of a number of highly efficient optimization and control methods". Following these ideas the model proposed here also represents the outflow as a continuous variable. The consequences of this assumption are detailed in [25].

Additionally, the most recent real-time methods, such as TUC and RHODES, overcome the need for analytically grasping the interaction between queues by assuming that measurements are available either on every link or on every signalized link of the network. By capturing the between-queue interactions such assumptions could be relaxed, allowing these methods to be applicable on a wider range of networks. Therefore, the queueing model proposed in this paper is an appropriate tool both to improve urban signal settings during peak hours and to emphasize the importance of accounting for the between-queue interactions.

3 Surrogate network model

This section formulates the surrogate model. A detailed description is provided in order to assess the realism of this model. We consider an urban transportation network composed of a set of both signalized and unsignalized intersections. The traffic model consists of a set of queueing models organized in a network, or a *queueing network model*.

We study a fixed-time signal control problem where the offsets, the cycle times and the all-red durations are fixed. The stage structure is also given. In other words, the set of lanes associated with each stage as well as the sequence of stages are both known.

The objective is to minimize the average time T spent in the network by adjusting the green splits at each intersection (i.e. the proportion g_p of cycle time that is allocated to each phase p). The travel time is derived from a traffic model which combines both exogenous (fixed) parameters α, such as the total demand, the route choice decisions and the topological structure of the street network, with endogenous variables x, such as the capacities and the probability of spillbacks. The latter are directly linked with the decision variables g. We now present the traffic model that derives T from g, x and α. We then formulate the signal control problem.

3.1 The queueing model

The network model is an extension of the model described in a previous paper [50], where we have proposed an analytic approximate queueing network model that accurately describes the formation and the diffusion of congestion. This model was validated versus both existing methods and simulation results. We provide below a general description of the existing model, and then detail its adaptation for urban traffic networks.

In the original model, we assume both the total demand and the capacities to be given, and derive a set of performance measures such as stationary distributions and congestion indicators. Each queue is defined according to a set of exogenous structural parameters. The key feature is the description of the interactions among the different queues. Congestion and spillbacks are modeled by what is referred to 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 overall process is described by a set of nonlinear equations capturing these between-queue interactions. Given the exogenous parameters, the endogenous variables are evaluated by solving this system of nonlinear equations. We extend here the formulation by considering the capacities endogenous, as they are determined by the decision variables (i.e. the green splits).

3.2 Network topology

Each road in the network is divided into segments such that the number of lanes is constant on each segment. Segment boundaries are therefore either intersections, or locations where the number of lanes changes between intersections. They correspond to changes of capacity. In this paper the term *capacity* will be used according to its traffic theory definition [64], that is the maximum flow rate expected to cross a given roadway per unit of time. It is typically given in vehicles per hour.

A queue is associated with each lane of each segment in the network (similarly to the supply simulator in the DynaMIT system, see [8]). Each queue is connected to the downstream segments where a turning of the underlying lane is permitted. Note that connecting a queue to a segment means that it is connected to all of the queues in that segment. The interactions among the queues

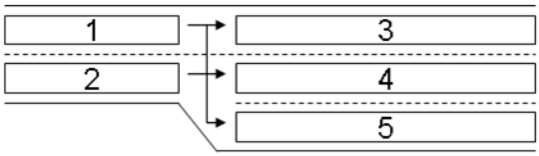

Figure 1: Example of how a road is mapped to a set of queues

are explicitly captured by linking the parameters of the queues (such as the capacity and the arrival flow) with the state of other queues.

Figure 1 illustrates how a road composed of 2 main lanes and 1 right turn lane is modeled as 2 upstream queues (indexed 1 and 2) followed by 3 downstream queues (indexed 3-5). In particular, if queue 5 spills back it will block the through movements of the upstream queues.

3.3 Bounded queues

In order to account for the limited physical space that a queue may occupy we resort to *finite capacity queueing theory*, where there is a finite upper bound on the length of each queue. The use of a finite bound allows us to capture the impact of queues on upstream segments (e.g. spillbacks), and to consider congested scenarios where traffic demand may exceed capacity. In queueing theory terms this corresponds to a traffic intensity that may exceed one. This is the key distinction between classical queueing theory and finite capacity queueing theory.

The upper bound of queue i is denoted k_i. It is known as the capacity of the queue in queueing theory. In this paper it will be referred to as the upper bound of the queue length.

Heidemann [35], as well as Van Woensel and Vandaele [62], divide each road into segments of length $1/k_{\text{jam}}$, where k_{jam} is the jam density, and thus $1/k_{\text{jam}}$ represents the minimal length that each vehicle needs. We also follow this type of reasoning and define k_i as:

$$k_i = \lfloor (\ell_i + d_2)/(d_1 + d_2) \rfloor,$$

where ℓ_i denotes the length of lane i, d_1 is the average vehicle length (e.g. 4 meters), and d_2 is the minimal inter-vehicle distance (e.g. 1 meter). The fraction is then rounded down to the nearest integer.

The physical space occupied by a queue is represented by a server followed by a buffer. All queues have one server, which represents the service due to the change of capacity at the boundary of a segment.

3.4 Arrival and service rates

The exogenous parameters used to describe the distribution of the demand throughout the network are the external arrival rates and the transition probabilities. The external arrival rate of a queue corresponds to vehicles reaching the queue coming from outside of the network, and not from another queue. This typically applies to the boundaries of the network, or parking lots inside the network. The transition probability between queue i and queue j, is the proportion of flow from queue i that goes to queue j, which may be obtained from a route choice model [9].

The service rates of the queues are defined as the capacities of the underlying lanes. For segments that lead to intersections the service rate of its queues is defined as the capacity of the intersection for that approach or lane. We derive formulations for the capacities of the different types of intersections based on the Swiss national transportation standards. A detailed description is given in Appendix B.

When a segment does not lead to an intersection (e.g. segments where all of the vehicles leave the network, or segments that lead directly to another segment) the service rate of its queues is set to the saturation flow of the corresponding lane.

3.5 System of equations

We adapt the equations proposed in [50] to this context. In the following notation all rates are average rates and the index i refers to a given queue.

γ_i	external arrival rate;
λ_i	total arrival rate;
μ_i	service rate of a server;
$\tilde{\mu}_i$	unblocking rate;
$\hat{\mu}_i$	effective service rate (accounts for both service and eventual blocking);
P_i^f	probability of being blocked at queue i;
p_{ij}	transition probability from queue i to queue j;
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 probability;
$\mathcal{I}^+$	set of downstream queues of queue i.

Since we consider a single server network, the vector denoted by $\tilde{\mu}(i, b)$ in the initial model reduces here to a single value that is now denoted $\tilde{\mu}_i$. We calculate the blocking probability based on the closed form expression available for *finite capacity queues* [10], instead of resorting to the *global balance equations* as in [50]. The system of equations is therefore given by:

$$\lambda_i = \gamma_i + \frac{\sum_j p_{ji}\lambda_j(1 - P(N_j = k_j))}{(1 - P(N_i = k_i))}, \tag{1}$$

$$\frac{1}{\tilde{\mu}_i} = \sum_{j \in \mathcal{I}^+} \frac{\lambda_j(1 - P(N_j = k_j))}{\lambda_i(1 - P(N_i = k_i))\hat{\mu}_j}, \tag{2}$$

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i}, \tag{3}$$

$$P_i^f = \sum_j p_{ij} P(N_j = k_j), \tag{4}$$

$$P(N_i = k_i) = \frac{1 - \rho_i}{1 - \rho_i^{k_i+1}} \rho^{k_i}, \tag{5}$$

$$\rho_i = \frac{\lambda_i}{\hat{\mu}_i}. \tag{6}$$

We briefly describe these equations below. For more details the reader is referred to the initial paper. The exogenous parameters are γ_i, p_{ij} and k_i. All other variables are endogenous. We approximate the transition rates between states using structural parameters that capture the between-queue interactions (Equations (1)-(4)). These equations link the endogenous parameters of a given queue (e.g. arrival rate, service rate) with the parameters of its upstream and downstream queues. In particular, P_i^f gives the probability with which a spillback can occur, while $\tilde{\mu}_i$ describes the rate at which such a spillback dissipates. These endogenous parameters also allow to identify the sources of congestion (by using conditional transition probabilities) and its impact (by using the expected number of

blocked vehicles). Equations (5) and (6) are the closed form expressions for the blocking probability. Each queue has 6 endogenous variables $(\lambda_i, \mu_i, \tilde{\mu}_i, \hat{\mu}_i, P(N_i = k_i), P_i^f)$. Thus the system consists of $6n$ equations, where n is the total number of queues.

4 Optimization problem

In order to formulate the signal control problem we introduce the following notation:

y_i	available cycle time of intersection i (cycle time minus the all-red times of intersection i) [seconds];
b_i	available cycle ratio of intersection i (ratio of y_i and the cycle time of intersection i);
g_p	green split of phase p (green time of phase p divided by the cycle time of its corresponding intersection);
g_L	vector of minimal green splits for each phase (minimal green time allowed for each phase divided by the cycle time of its corresponding intersection);
s	saturation flow rate [veh/h];
x	endogenous queueing model variables;
α	exogenous queueing model parameters;
$\mathcal{I}$	set of intersection indices;
$\mathcal{L}$	set of indices of the signalized lanes;
$\mathcal{P}_\mathcal{I}(i)$	set of phase indices of intersection i;
$\mathcal{P}_\mathcal{L}(\ell)$	set of phase indices of lane ℓ.

The problem is formulated as follows:

$$\min_{g,x} \; T(g, x, \alpha) \tag{7}$$

subject to

$$\sum_{p \in \mathcal{P}_\mathcal{I}(i)} g_p = b_i, \; \forall i \in \mathcal{I} \tag{8}$$

$$\mu_\ell - \sum_{p \in \mathcal{P}_\mathcal{L}(\ell)} g_p s = 0, \; \forall \ell \in \mathcal{L} \tag{9}$$

$$h(x, \alpha) = 0 \tag{10}$$

$$g \geq g_L \tag{11}$$

$$x \geq 0, \tag{12}$$

where h represents the traffic model presented in Section 3.5.

The objective is to reduce the average time that vehicles spend in the network, which is represented by T (Equation (7)). The expression for T is derived based on the use of Little's law applied to the network and taking into account the *finite capacity queueing framework*. This leads to the following nonlinear objective function:

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

The linear constraints (8) link the green times with the available cycle time for each intersection. Equation (9) links the green times of the signalized lanes to their capacities. Equation (10) consists of

the system of nonlinear equations given in Section 3.5. The bounds (11) correspond to minimal green time values for each phase. These have been set to 4 seconds according to the Swiss standard VSS [68].

The optimization problem is solved with the Matlab routine for constrained nonlinear problems, *fmincon*, which resorts to a sequential quadratic programming method [19, 18]. A feasible initial point is obtained by fixing a control plan and solving the network model (Equation (10)). We refer the reader to [50] for more details on the solution procedure of this system of equations as well as for its own initialization settings. At the solution both the maximum constraint violation and the relative gradient of the Lagrangian are smaller than the threshold, 10^{-6}. The use of relative gradient information to test for optimality is detailed in [24], and further described in the context of constrained optimization in [20]. The choice of the threshold is based on the criteria given in [24].

5 Empirical analysis

5.1 Microscopic traffic simulation model of the city of Lausanne

To perform the empirical analysis, we use a calibrated microscopic traffic simulation model of the Lausanne city center. This model [27] is implemented with the AIMSUN simulator [61]. It contains 652 roads and 231 intersections, 49 of which are signalized. We use this model for two purposes.

Firstly, we use it to extract the network data (e.g. road characteristics, demand distribution) needed to estimate the exogenous parameters of the queueing model. The intersection characteristics include an existing fixed-time signal control plan of the city of Lausanne. For more information concerning this control plan we refer the reader to [27]. Based on this control plan we give initial values to the capacities of the signalized lanes. More details with regards to the calibration of the queueing model are given in Appendix C.

Secondly, we use this simulation model to evaluate and compare the performance of different signal plans. Once a new plan is determined, it is integrated in the simulation model, its performance is evaluated and then compared with that of other plans.

We now compare the performance of several methodologies, by considering a subnetwork of the Lausanne city center. The subnetwork (Figure 2) contains 48 roads and 15 intersections. Nine intersections are signalized and control the flow of 30 roads. There are a total of 51 phases that are considered variable. The intersections have a cycle time of either 90 or 100 seconds. For each methodology we derive the optimal signal plan for the subnetwork, and then use the simulation model to evaluate its effect upon the entire Lausanne network.

The simulation setup consists of 100 replications of the evening peak period (17h-19h), preceded by a 15 minute warm-up time. Within this time period congestion gradually increases. The average flow of the roads in the subnetwork steadily decreases from 339 to 25 (veh/h); and the average density increases from 10 to 57 (veh/km).

The queueing model of this subnetwork consists of 102 queues. The optimization problem consists of 621 endogenous variables with their corresponding lower bound constraints, 408 nonlinear equality constraints, and 171 linear equality constraints.

5.2 Between-queue interactions

The queueing model proposed in this paper describes congestion by taking into account the interactions between upstream and downstream roads. In this section we illustrate the added value of accounting for these interactions. We compare this model with the same model where independence of the queues is assumed. The optimization problem is solved for both queueing models (correlated queues versus

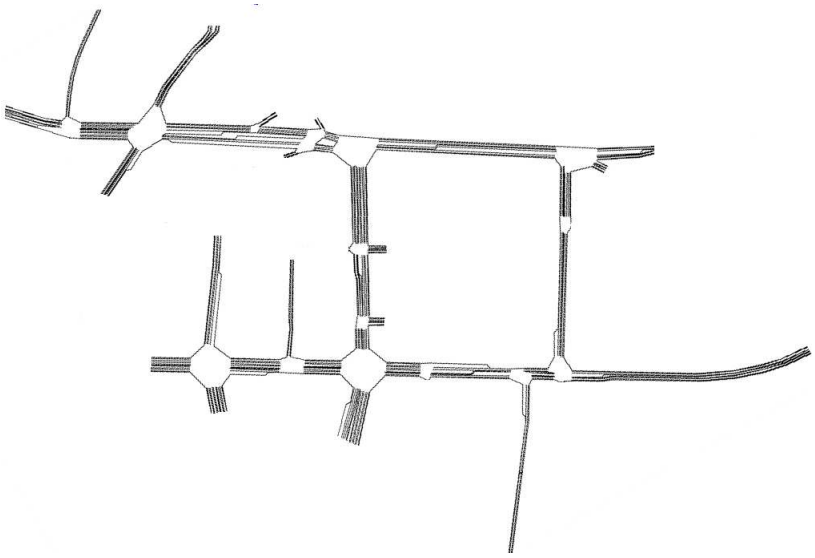

Figure 2: Subnetwork of the Lausanne city center

independent queues), and the performance of the corresponding signal plans are compared. We will denote these as the *correlated* and the *independent* plans, respectively.

Assuming independent queues leads to the following simplifications:

- the arrival rates are now exogenous;
- the effective service rates, are no longer linked to the potential spillbacks of downstream roads, i.e. the total time spent on a road is entirely determined by its capacity.

We consider the average number of vehicles that have exited each origin-destination (OD) pair at a given time. The simulation time is segmented into 40 3-minute intervals. Figure 3 displays for each time interval a boxplot of the difference between the average number of vehicles for the independent and the correlated plans. Each point within a boxplot represents this difference for a given OD pair. This figure illustrates how the number of OD pairs that have a higher flow under the correlated plan than under the independent one increases as congestion increases.
This figure also shows that there is no difference for the majority of the OD pairs. It makes sense, since only 51% of the 2096 OD pairs have more than 2 trips assigned per hour, 14% have more than 10 trips, and 6.6% have more than 20 trips. Thus for the majority of the OD pairs we would not expect a difference larger than a couple of vehicles.
Figure 4 displays the empirical cumulative distribution function of these differences for the intervals 10, 20, 30 and 40. It also shows that as congestion increases there is a higher proportion of OD pairs that perform better when the correlation is taken into account. The asymmetry of Figures 3 and 4 are evidence of the added value of accounting for the dependence of queues in signal optimization.
Figure 5 displays the density averaged across replications and across the roads of the network and of the subnetwork, respectively. These averages are plotted versus time. The crosses denote the independent plan, the circles represent the correlated plan. We did not expect any noticeable network-wide differences, as the network is highly congested, so that only the global throughput could be affected. Nevertheless, at the subnetwork level the average density is decreased. These results illustrate well the added value of the method, not only on the global throughput but also locally.

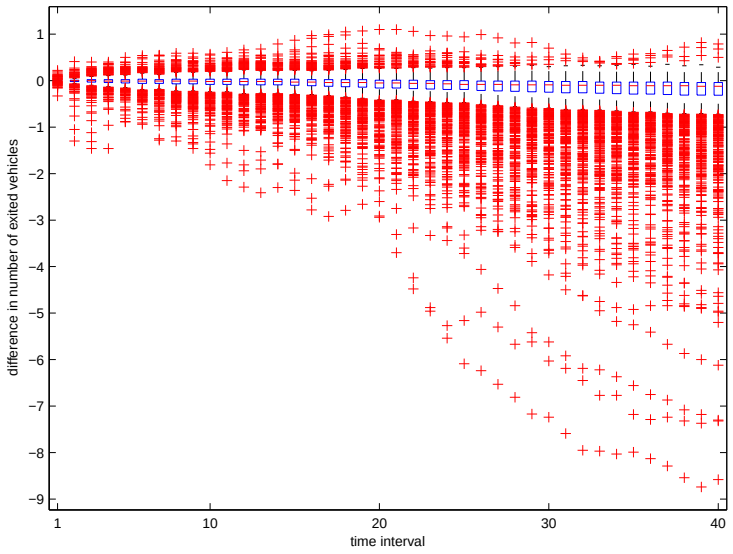

Figure 3: Difference in the average number of vehicles that have exited each OD pair versus time

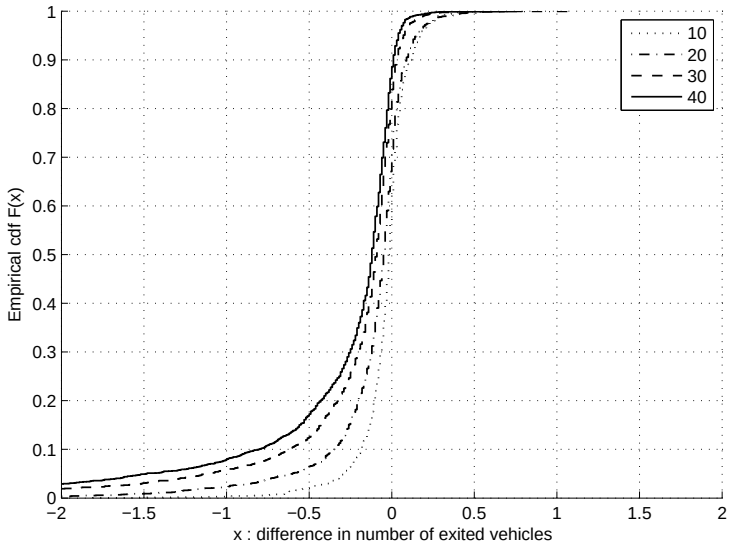

Figure 4: Empirical cumulative distribution function of the difference in the average number of vehicles that have exited the OD pairs for time intervals 10, 20, 30 and 40

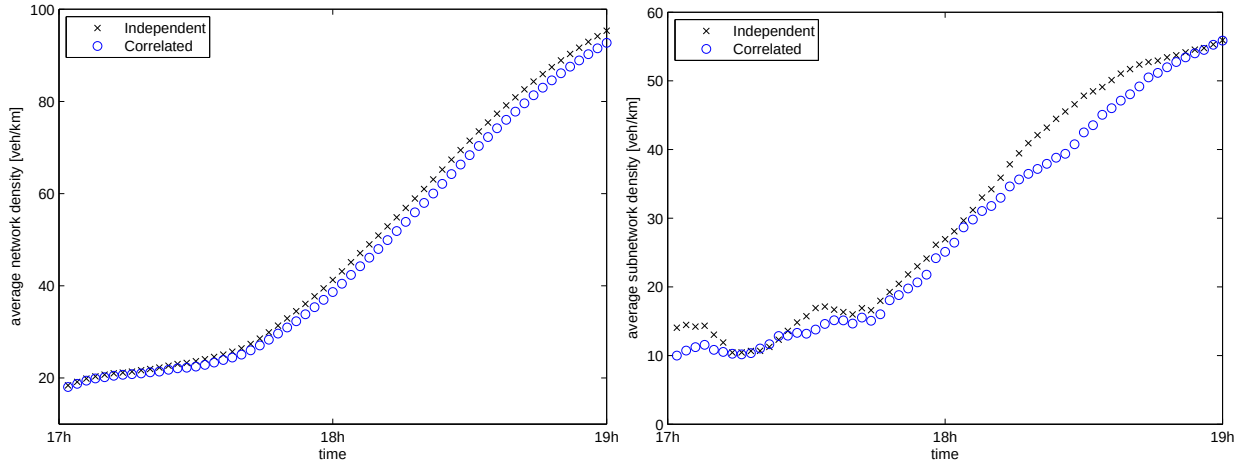

Figure 5: Average network and subnetwork density plotted versus time

5.3 Comparison with existing methods

We now compare the signal settings derived by the method proposed in this paper with an existing fixed-time signal settings for the city of Lausanne, the method derived by Webster [69] and with the method suggested in the Highway Capacity Manual [60].

Base plan The calibrated simulation model of the Lausanne city center is based on an existing fixed-time signal control plan. For more information concerning this control plan see [27]. This signal plan will be referred to as the *base* plan.

HCM/Webster Webster's method is described in detail in Appendix D. By allocating the green times such that the flow to capacity ratios for the critical movements of each phase are equal, the method suggested in the Highway Capacity Manual [60] leads to the same green split equations as Webster's method [69]. This equivalence is detailed in [51].

We consider the subnetwork and simulation setup described in Section 5.1. We compare the methods in terms of the average number of vehicles that have exited each OD pair across time. The description of how these comparisons are carried out has been described in Section 5.2. Figure 6 displays the empirical cumulative distribution functions when comparing the new plan to the base plan (left plot) and to the HCM/Webster plan (right plot). This figure shows that there is a high proportion of OD pairs for which the new plan yields an increase in outflow. Furthermore, this proportion increases with congestion. The asymmetry of this figure illustrates the superiority of the proposed method.

Figure 7 displays the average density across the network and the subnetwork. The densities are plotted versus time. The crosses, squares and circles denote the base plan, the HCM/Webster plan and the new plan, respectively. As in the previous experiment, we observe a slight improvement at the network level and a significant one at the subnetwork level. Figure 8 considers the flow and the travel time averaged across the roads of the subnetwork and across replications. These plots illustrate how the new plan leads to improved average travel times, whereas for the flow there is no trend.

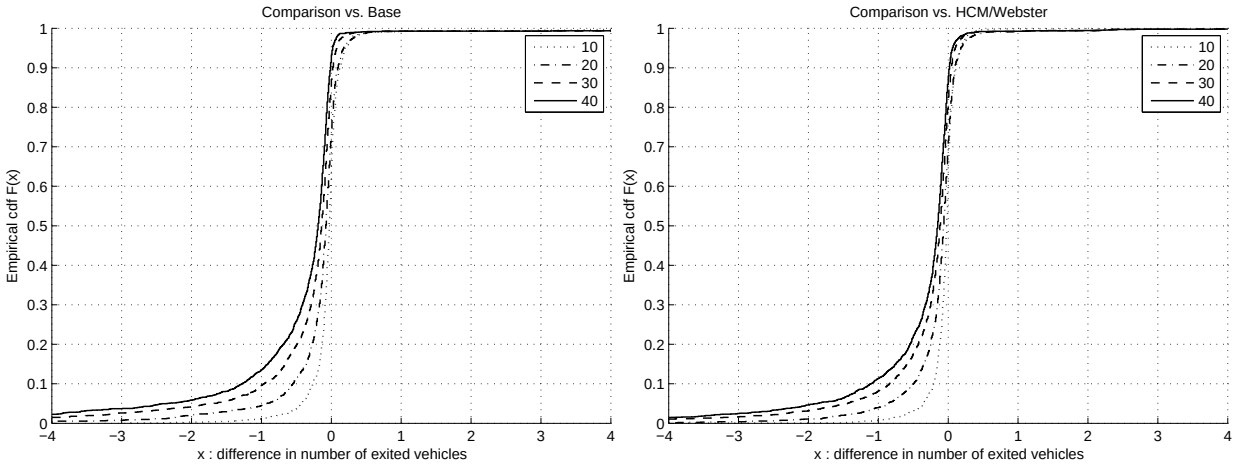

Figure 6: Empirical cumulative distribution function of the difference in the average number of vehicles that have exited the OD pairs for time intervals 10, 20, 30 and 40

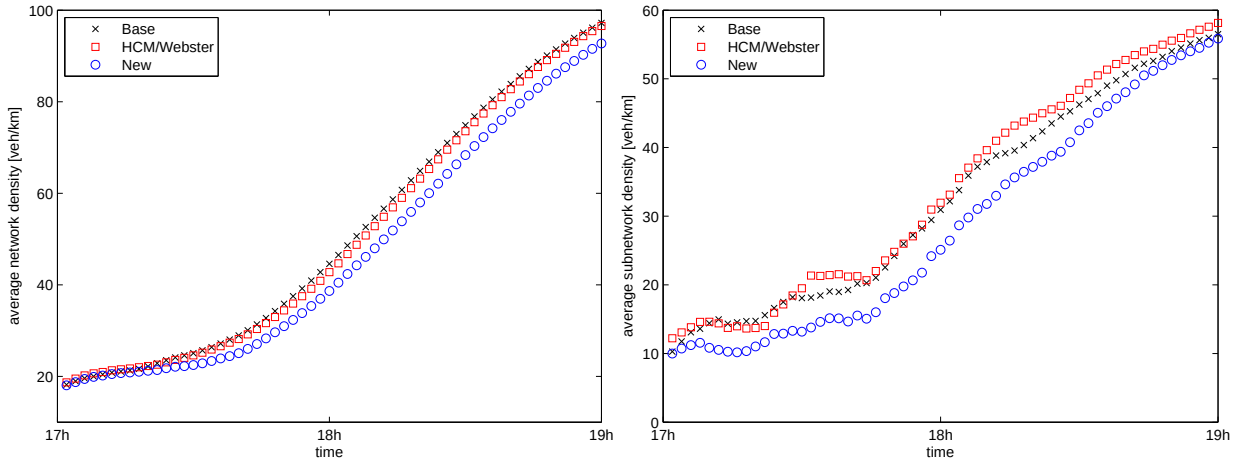

Figure 7: Average network and subnetwork density plotted versus time

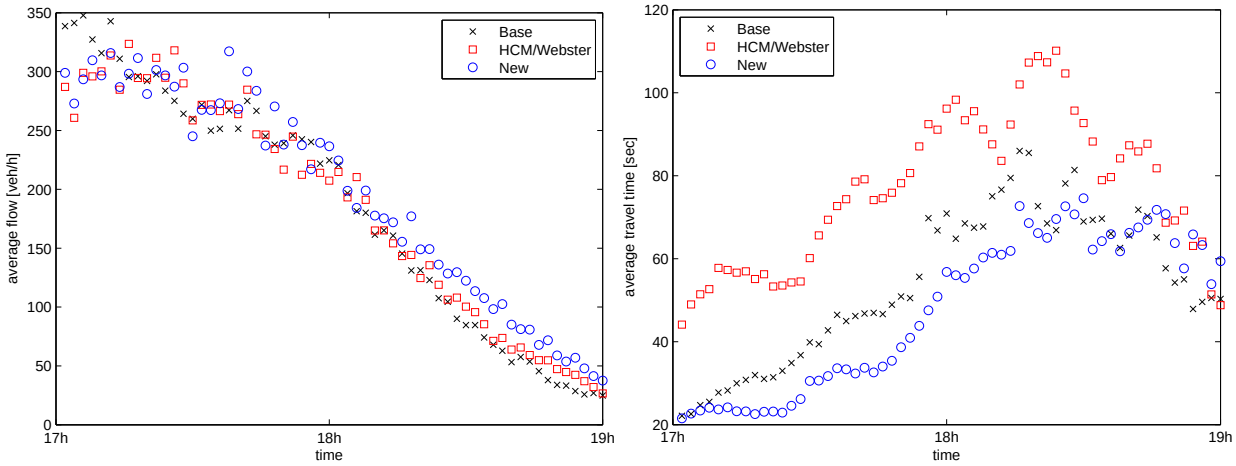

Figure 8: Average flow and travel time of the roads of the subnetwork, plotted versus time

6 Conclusion

We have proposed a model based on a queueing network framework, which is to be used as a surrogate of traffic simulators to perform optimization for congested urban networks. As a specific example of such an approach, we have formulated a fixed-time traffic signal optimization problem, and have used the surrogate as the network model. We have solved the signal control problem for a subnetwork of the city of Lausanne. The new signal plan has been evaluated with a microscopic traffic simulation tool. Its performance has been compared with the same model assuming independent queues, with a fixed-time plan that exists for the city of Lausanne, with Webster's method and with the method proposed by the Highway Capacity Manual. As congestion increases, the new method leads to performance measures that improve on average.

The formulation of an urban road network using finite capacity queueing theory and accounting for multiple intersections is novel. Additionally, by using a set of structural parameters that capture the between-queue interactions, this queueing model approximates how congestion arises and how it spreads, sufficiently well to be used as an appropriate surrogate. It characterizes congestion in terms of its sources, its frequency, its propagation and its impact. This approach, based on a fine decomposition of the phenomenon of congestion, is of general interest for traffic control, and particularly appropriate for the study and management of congested urban networks. There is a need for operational signal control methodologies that are suitable for congested conditions and that capture the complex features of congested traffic flows such as spillbacks. This model contributes to the development of these methodologies thanks to an analytical framework, that makes an attractive trade-off between capturing the complexity of congested traffic flows and analytical tractability.

Clearly, the proposed model is not a fully realistic representation of spillbacks in signalized arterials. Therefore, there is a potential to investigate if more sophisitication would preserve the tractability of the model, while enhancing the optimization. The realism of the model can be increased by allowing for endogenous upper bounds for the queues or by accounting for the dynamic nature of traffic. Future work will also focus on further combining the use of this analytic traffic model with the traffic simulator for performing simulation-based optimization.

Acknowledgments

This research was supported by the Swiss National Science Foundation grants 205321-107838 and 205320-117581.

Appendices

A Review of traffic signal control methodologies

A.1 Fixed-time isolated strategies

These strategies can be stage-based such as SIGSET [4] and SIGCAP [5]. SIGSET minimizes delay using Webster's nonlinear formulation [69], whereas SIGCAP maximizes reserve capacity. Both methods consider a set of linear constraints. A phase-based method formulated as a mixed-integer linear program is considered in [38], where formulations for both delay minimization and reserve capacity maximization problems are given.

A.2 Fixed-time coordinated strategies

Optimizing a set of signals along an arterial is the focus of the arterial progression schemes MAXBAND [41] and MULTIBAND [33]. These methods aim at maximizing the bandwidth of through traffic along an arterial. MULTIBAND is an extension of MAXBAND allowing, among others, for different bandwidths for each link of the arterial. These problems are formulated as mixed-integer linear programs. They have been extended to consider a set of intersecting arterials [31]. Heuristics have also been specifically developed to solve this problem [54]. Nevertheless under congested scenarios where there is a strong interaction among the different queues, the calculated bands fail to grasp this complexity. Furthermore in dense urban networks with complex traffic movements bandwidth has little meaning [55].

Several phase-based strategies have been proposed [73, 71, 70]. The phase-based approach, although more general, is limited due to the exponential number of integer variables needed to describe the precedence constraints of incompatible phases.

Chaudhary et al. [16] compares the performance of 3 fixed-time coordinated stage-based methods: TRANSYT, PASSER and SYNCHRO. TRANSYT is the most widely used signal timing optimization package. It is a macroscopic model that aims at minimizing both delay and stops. A descriptive figure of its underlying methodology is given in [52]. SYNCHRO and TRANSYT have similar traffic models. SYNCHRO seeks to minimize stops and queues, by using an exhaustive search technique to determine the optimal signal timings. PASSER determines the green splits (also known as the green ratios), stage structure, cycle length, and offsets that maximize arterial progression (i.e. bandwidth-based method) for signalized arterials. PASSER performs an exhaustive search over the range of cycle lengths provided by the user, and sets the green splits using Webster's method [69]. These splits are then adjusted to improve progression. It is highlighted in [11] that in congested conditions, TRANSYT and PASSER do not grasp the queue length appropriately. Traditionally TRANSYT's traffic model considered vertical queueing (i.e. the spatial extension of the queue is ignored), thus not capturing spillbacks, making this software suitable only for undersaturated scenarios. Although, more recent versions now take into account the effects of queue formation using horizontal queueing models [2], Chow and Lo [17] emphasize that the use of TRANSYT is appropriate only for low to moderate degrees of saturation.

A.3 Traffic-responsive methods

Traffic-responsive methods use real-time measurements to drive the underlying optimization algorithm. The signal plans of these methods are derived either by making small adjustments to a predefined plan, by choosing between a set of pre-specified plans or by deciding when to switch to the next stages over a future time horizon [11]. The trend of real-time methods is the latter, where the optimization parameters are no longer cycle time, splits or offsets, but rather the switching times. These methods are referred to as non-parametric methods [56]. Nevertheless these methods are limited by the exponential size of the search space, due to the introduction of the integer variables that describe the switching times.

The British software SCOOT [13] is considered to be the traffic-responsive version of TRANSYT. A description of how TRANSYT evolved into SCOOT is given in [55]. SCOOT seeks to minimize the total delay by carrying out incremental changes to the off-line timings derived by TRANSYT. It therefore makes a large number of small optimization decisions (typically over 10000 per hour in a network of 100 junctions [55]). The Australian method SCATS [42] modifies signal timings on a cycle-by-cycle basis by minimizing stops and delay while constraining the formation of queues. Both SCOOT and SCATS are widely used strategies suitable for undersaturated conditions, but as is described in

[1, 25], their performance deteriorates under congested conditions.

Dynamic programming methods are used in the French system PRODYN [37] as well as in the US systems OPAC and RHODES. RHODES [46] uses the COP algorithm [56] to determine the switching times at a given intersection. This method does not react to traffic conditions just observed but rather proactively sets phase durations for predicted traffic conditions. A description of the OPAC model and algorithm, as well as its implementation are given in [32, 33]. The Italian method UTOPIA is yet another method that has been evaluated and implemented [43]. Nevertheless, the exponential complexity of these methods does not allow for network-wide optimization [25]. This is also emphasized in [11]: "the existing systems are not capable of controlling a zone of several junctions in a complete and coordinated manner. The chosen compromise is to control only one junction as OPAC or to use a decentralized optimization method as UTOPIA, PRODYN or to make little changes of the fixed-time signal plan as SCOOT and SCATS." Acknowledging the importance and lack of efficient control strategies under saturated conditions has lead to the development of the French system CRONOS [12, 11], and of the TUC method [25].

B Service rates

The service rates of the queues are defined as the capacities of the underlying lanes. We describe how we derive formulations for the capacities based on the Swiss national transportation standards.

For unsignalized intersections (e.g. two-way stop controlled intersections, yield-controlled intersections) the standard VSS [65] is used. The turning movements are ranked. For each movement the conflicting flow is calculated based on a set of equations that depend on the type of movement and its rank. Then their potential capacity and their movement capacity is calculated. Finally the capacity of the lanes with multiple turnings are adjusted to take into account the lack of side lanes.

The capacity of the lanes leading to, on, or exiting roundabouts are derived based on the standard VSS [67]. They take into account the same parameters as for unsignalized intersections but are based on a different set of equations. This standard accounts for roundabouts with either one lane or one large lane. For networks that contain roundabouts with two lanes, the capacity of these lanes is calculated based on the equations for roundabouts with one large lane.

For signalized intersections we use the standard VSS [66], which defines the capacity of a lane as the product of the saturation flow and the proportion of green time allocated to that lane per cycle. This approach is also proposed in Chapter 16 of the Highway Capacity Manual [60].

C Calibration details

The demand distribution in the traffic simulator is described in terms of roads, whereas the calibration of the surrogate model requires lane specific distributions. We describe how we convert the road-specific distribution to the lane-specific distribution.
For each road we have three types of flow data: external inflow (flow that arises from outside of the network), road-to-road turning flow, external outflow (flow that leaves the network). In order to obtain lane specific distributions we disaggregate the flow data as follows.

External inflow We assume that this flow is distributed with equal probability across all the lanes of the road. If the road is modeled with several segments the inflow is associated with the first segment. In other words arrivals only occur at the beginning of the road.

Turning flow We consider that this flow is distributed with equal probability across all the lanes involved in the turning.

External outflow We assume that this flow is distributed with equal probability across all the lanes of the road. If the road is modeled with several segments the outflow is associated with the last (most downstream) segment. In other words departures only occur at the end of the road.

D Webster's method

Webster's method is based on an estimate of the average delay per vehicle at a signalized intersection. It determines cycle times and green-splits of pre-timed signals that minimize delay. These green splits are used in signal setting software packages such as SYNCHRO and PASSER V [16]; and the delay estimate is one of the best known [14]. The analysis is based on isolated intersections under the assumption of the number of arrivals following a Poisson distribution, and undersaturated conditions (traffic intensity $\rho < 1$).
In this approach each phase is represented by one approach only: the one with the highest degree of saturation (ratio of flow to saturation flow). This maximum ratio for phase p is denoted Y_p. More specifically, assuming no yellow times and no lost times per phase, Webster's method leads to:

$$g_p = \frac{Y_p}{\sum_{j \in \mathcal{P}_{\mathcal{I}}(i)} Y_j} b_i \quad \forall p \in \mathcal{P}_{\mathcal{I}}(i). \tag{13}$$

This method requires as input the flows and saturation flows for each approach. These have been derived as follows. For a signalized intersection the saturation flow is set to a common value for all approaches, this value is based on the standards VSS [66]. The approach flows are set using the observed flows derived by the simulation model.

References

[1] K. Aboudolas, M. Papageorgiou, and E. Kosmatopoulos. Control and optimization methods for traffic signal control in large-scale congested urban road networks. In *American Control Conference*, pages 3132–3138, 2007.

[2] G. Abu-Lebdeh and R. Benekohal. Design and evaluation of dynamic traffic management strategies for congested conditions. *Transportation Research A: Policy and Practice*, 37(2):109–127, 2003.

[3] G. Abu-Lebdeh and R. Benekohal. Development of traffic control and queue management procedures for oversaturated arterials. *Transportation Research Record*, 1603:119–127, 1997.

[4] R. Allsop. SIGSET: A computer program for calculating traffic signal settings. *Traffic Engineering and Control*, 13(2), 1971.

[5] R. Allsop. SIGCAP: A computer program for assessing the traffic capacity of signal-controlled road junctions. *Traffic Engineering & Control*, 17:338–341, 1976.

[6] R. Allsop. Evolving application of mathematical optimisation in design and operation of individual signal-controlled road junctions. In J. D. Griffiths, editor, *Mathematics in Transport Planning and Control*. Institute of Mathematics and its Applications, University of Wales College of Cardiff, Oxford Clarendon, 1992.

[7] R. R. Barton and M. Meckesheimer. Metamodel-based simulation optimization. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in operations research and management science: Simulation*, volume 13, chapter 18, pages 535–574. Elsevier, Amsterdam, 2006.

[8] M. Ben-Akiva, M. Bierlaire, M. Burton, H. Koutsopoulos, and R. Mishalani. Network state estimation and prediction for real-time transportation management applications. *Networks and Spatial Economics*, 1(3-4):293–318, 2001.

[9] M. Bierlaire and E. Frejinger. Route choice modeling with network-free data. *Transportation Research C: Emerging Technologies*, 16(2):187–198, 2008.

[10] P. P. Bocharov, C. D'Apice, A. V. Pechinkin, and S. Salerno. *Queueing theory*, chapter 3, pages 96–98. Modern Probability and Statistics. Brill Academic Publishers, Zeist, The Netherlands, 2004.

[11] F. Boillot, J. Blosseville, J. Lesort, V. Motyka, M. Papageorgiou, and S. Sellam. Optimal signal control of urban traffic networks. *Road Traffic Monitoring (IEE Conf. Pub. 355)*, 1992.

[12] F. Boillot, S. Midenet, and J. Pierrelée. The real-time urban traffic control system CRONOS: Algorithm and experiments. *Transportation Research C: Emerging Technologies*, 14(1):18–38, 2006.

[13] R. D. Bretherton. SCOOT - urban traffic control system - philosophy and evaluation. In *IFAC Symposium of Control Communications in Transportation*, pages 237–239. Pergamon Press, Oxford, September 1989.

[14] E. Cascetta. *Transportation Systems Engineering: theory and methods*, volume 49 of *Applied Optimization*, chapter 2, pages 50–51. Kluwer academic publishers, Dordrecht, 2001. ISBN 0792367928.

[15] E. Cascetta, M. Gallo, and B. Montella. Models and algorithms for the optimization of signal settings on urban networks with stochastic assignment models. *Annals of Operations Research*, 144(1):301–328, 2006.

[16] N. A. Chaudhary, V. G. Kovvali, and S. M. Alam. Guidelines for selecting signal timing software. Technical Report 0-4020-P2, Texas Transportation Institute, U.S. Department of Transportation, Federal Highway Administration, September 2002.

[17] A. H. F. Chow and H. K. Lo. Sensitivity analysis of signal control with physical queuing: Delay derivatives and an application. *Transportation Research B: Methodological*, 41(4):462–477, May 2007.

[18] T. F. Coleman and Y. Li. On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, 67(2):189–224, 1994.

[19] T. F. Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6:418–445, 1996.

[20] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-region methods*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadelphia, PA, USA, 2000.

[21] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to derivative-free optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadelphia, PA, USA, 2009.

[22] G. F. Corliss, C. Faure, A. Griewank, and L. Hascoet. *Automatic Differentiation of Algorithms: From Simulation to Optimization*. Springer, 2002.

[23] G. Deng. *Simulation-based optimization*. PhD thesis, University of Wisconsin, 2007.

[24] J. E. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 1996.

[25] V. Dinopoulou, C. Diakaki, and M. Papageorgiou. Applications of the urban traffic control strategy TUC. *European Journal of Operational Research*, 175(3):1652–1665, 2006.

[26] F. Dion, H. Rakha, and Y. Kang. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transportation Research B: Methodological*, 38(2): 99–122, 2004.

[27] A. G. Dumont and E. Bert. Simulation de l'agglomération Lausannoise SIMLO. Technical report, Laboratoire des voies de circulation, ENAC, Ecole Polytechnique Fédérale de Lausanne, Mai 2006.

[28] M. C. Fu. Optimization for simulation: Theory vs. practice (feature article). *INFORMS Journal on Computing*, 14(3):192–215, 2002.

[29] M. C. Fu, F. W. Glover, and J. April. Simulation optimization: a review, new developments, and applications. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 83–95, Piscataway, New Jersey, USA, 2005.

[30] N. J. Garber and L. A. Hoel. *Traffic and Highway Engineering*, chapter 6, pages 204–210. Books Cole, Thomson Learning, 3rd edition, 2002.

[31] N. Gartner and C. Stamatiadis. Arterial-based control of traffic flow in urban grid networks. *Mathematical and Computer Modelling*, 35(5):657–671, 2002.

[32] N. Gartner, F. Pooran, and C. Andrews. Implementation of the OPAC adaptive control strategy in a trafficsignal network. In *Intelligent Transportation Systems, IEEE*, pages 195–200, 2001.

[33] N. H. Gartner, S. F. Assman, F. Lasaga, and D. L. Hou. A multi-band approach to arterial traffic signal optimization. *Transportation Research B: Methodological*, 25(1):55–74, 1991.

[34] D. Heidemann. Queue length and delay distributions at traffic signals. *Transportation Research B: Methodological*, 28(5):377–389, 1994.

[35] D. Heidemann. A queueing theory approach to speed-flow-density relationships. In *Proceedings of the* 13^{th} *International Symposium on Transportation and Traffic Theory*, pages 103–118, Lyon, France, July 1996.

[36] D. Heidemann and H. Wegmann. Queueing at unsignalized intersections. *Transportation Research B: Methodological*, 31(3):239–263, 1997.

[37] J. J. Henry and J. L. Farges. PRODYN. In *IFAC Symposium of Control Communications in Transportation*, pages 253–255. Pergamon Press, Oxford, September 1989.

[38] G. Improta and G. E. Cantarella. Control system design for an individual signalized junction. *Transportation Research B: Methodological*, 18(2):147–167, 1984.

[39] R. Jain and J. M. Smith. Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation science*, 31(4):324–336, 1997.

[40] L. Kerbache and J. M. Smith. Multi-objective routing within large scale facilities using open finite queueing networks. *European Journal of Operational Research*, 121(1):105–123, 2000.

[41] J. Little, M. Kelson, and N. Gartner. MAXBAND: a program for setting signals on arteries and triangular networks. *Transportation Research Record*, 795:40–46, 1981.

[42] P. Lowrie. SCATS: The sydney co-ordinated adaptive traffic system. In *IEE International conference on road traffic signaling*, pages 67–70, 1982.

[43] V. Mauro and C. Di Taranto. UTOPIA. In *IFAC Symposium of Control Communications in Transportation*, pages 245–252. Pergamon Press, Oxford, September 1989.

[44] D. R. McNeil. A solution to the fixed-cycle traffic light problem for compound poisson arrivals. *Journal of Applied Probability*, 5:624–635, 1968.

[45] A. J. Miller. Settings for fixed-cycle traffic signals. *Operational Research Quarterly*, 14(4):373–386, 1963.

[46] P. Mirchandani and L. Head. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transportation Research C: Emerging Technologies*, 9(6):415–432, 2001.

[47] G. Newell. Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Review*, 7(2):223–240, 1965.

[48] G. F. Newell. *Approximate behavior of tandem queues*, volume 171 of *Lecture notes in economics and mathematical systems*. Springer-Verlag, Berlin, 1979.

[49] S. Ólafsson and J. Kim. Simulation optimization. In E. Y. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Proceedings of the 2002 Winter Simulation Conference*, pages 79–84, San Diego, California, USA, 2002.

[50] C. Osorio and M. Bierlaire. An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal Of Operational Research*, 196(3):996–1007, 2009.

[51] C. Osorio and M. Bierlaire. Network performance optimization using a queueing model. In *Proceedings of the European Transport Conference (ETC)*, Noordwijkerhout, The Netherlands, October 6-8 2008.

[52] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 91(12):2043–2067, 2003.

[53] H. Perros. Open queueing networks with blocking - a personal log. In G. Kotsis, editor, *Performance Evaluation - Stories and Perspectives*, pages 105–115, Vienna, Austria, December 2003. Austrian Computer Society. ISBN 3-85403-175-0.

[54] R. Pillai, A. Rathi, and S. L. Cohen. A restricted branch-and-bound approach for generating maximum bandwidth signal timing plans for traffic networks. *Transportation Research B: Methodological*, 32(8):517–529, 1998.

[55] D. Robertson and R. Bretherton. Optimizing networks of traffic signals in real time - the SCOOT method. *Vehicular Technology, IEEE Transactions on*, 40(1):11–15, 1991.

[56] S. Sen and K. Head. Controlled optimization of phases at an intersection. *Transportation science*, 31(1):5–17, 1997.

[57] S. Shepherd. Traffic control in over-saturated conditions. *Transport Reviews*, 14(1):13–43, 1994.

[58] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. Wiley-Interscience series in discrete mathematics and optimization. John Wiley & Sons, New Jersey, USA, 2003.

[59] J. C. Tanner. A theoretical analysis of delays at an uncontrolled intersection. *Biometrika*, 49: 163–170, 1962.

[60] TRB. *Highway capacity manual*. Transportation Research Board, National Research Council, Washington, D.C., USA, 2000. Chapter 16.

[61] TSS. *AIMSUN NG and AIMSUN Micro Version 5.1*. Transport Simulation Systems, May 2008.

[62] T. Van Woensel and N. Vandaele. Modelling traffic flows with queueing models: A review. *Asia-Pacific Journal of Operational Research*, 24(4):1–27, 2007.

[63] F. Viti. *The Dynamics and the Uncertainty of Delays at Signals*. PhD thesis, Delft University of Technology, November 2006. TRAIL Thesis Series, T2006/7.

[64] VSS. *Norme Suisse SN 640017a Capacité, niveau de service, charges compatibles; norme de base*. Union des professionnels suisses de la route, VSS, Zurich, Décembre 1998.

[65] VSS. *Norme Suisse SN 640022 Capacité, niveau de service, charges compatibles; carrefours sans feux de circulation*. Union des professionnels suisses de la route, VSS, Zurich, Mai 1999.

[66] VSS. *Norme Suisse SN 640023 Capacité, niveau de service, charges compatibles; carrefours avec feux de circulation*. Union des professionnels suisses de la route, VSS, Zurich, Août 1999.

[67] VSS. *Norme Suisse SN 640024a Capacité, niveau de service, charges compatibles; carrefours giratoires*. Union des professionnels suisses de la route, VSS, Zurich, Août 2006.

[68] VSS. *Norme Suisse SN 640837 Installations de feux de circulation; temps transitoires et temps minimaux*. Union des professionnels suisses de la route, VSS, Zurich, Mai 1992.

[69] F. V. Webster. Traffic signal settings. Technical Report 39, Road Research Laboratory, 1958.

[70] S. Wong. Group-based optimisation of signal timings using the TRANSYT traffic model. *Transportation Research B: Methodological*, 30(3):217–244, 1996.

[71] S. Wong. Group-based optimisation of signal timings using parallel computing. *Transportation Research C: Emerging Technologies*, 5(2):123–139, 1997.

[72] S. Wong and H. Yang. Reserve capacity of a signal-controlled road network. *Transportation Research B: Methodological*, 31(5):397–402, 1997.

[73] S. Wong, W. Wong, C. Leung, and C. Tong. Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control. *Transportation Research B: Methodological*, 36 (4):291–312, 2002.

[74] Y. Yin. Robust optimal traffic signal timing. *Transportation Research B: Methodological*, 42(10): 911–924, December 2008.

[75] G. Ziyou and S. Yifan. A reserve capacity model of optimal signal control with user-equilibrium route choice. *Transportation Research B: Methodological*, 36(4):313–323, 2002.