Accounting for congestion and spillbacks in fixed-time traffic signal optimization: an analytical queueing model approach

Carolina Osorio [*] Michel Bierlaire [*]

October 26, 2008

Report TRANSP-OR 081026
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
`transp-or.epfl.ch`

[*] Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, {carolina.osoriopizano, michel.bierlaire }@epfl.ch

Abstract

We present and analyze a new aggregate model of urban traffic. The objective is to analytically capture the correlation between the different components of the network while maintaining a tractable model that can be used in an optimization framework.

Existing analytical queueing models for urban networks are formulated for a single intersection, and thus do no take into account the interactions among upstream and downstream roads. We formulate a model that considers a set of intersections and captures the correlation structure between consecutive roads based on finite capacity queueing theory. It therefore provides a detailed description of congestion. It identifies the sources of congestion (e.g. bottlenecks), describes how congestion propagates and dissipates; and quantifies the impact on the network performance.

We use the model in the context of fixed-time traffic signal optimization. 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. To the best of our knowledge, the existing signal control strategies based on analytical network models have not taken spillbacks into account. 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 analyze its performance. We compare its performance to that of several other methods. The results show the importance of taking the correlation between consecutive roads into account.

1 Introduction

Road traffic congestion is a costly phenomenon that is common to the vast majority of urban road networks. A recent European Commission report emphasizes that to alleviate congestion "in certain cases new infrastructure might be needed, but the first step should be to explore how to make better use of existing infrastructure" (CEC, 2007). Thus the importance of understanding the origins of congestion, of quantifying its effects and of controlling traffic in order to optimize the use of existing infrastructure. Within this context the contributions of this paper are two-fold.

Firstly, we present an analytical stochastic network model derived from the queueing model proposed by Osorio and Bierlaire (forthcoming). Existing analytical 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 no take into account the interaction between upstream and downstream roads. The framework that we present models a set of urban intersections. It captures the correlation structure between consecutive roads using *finite capacity queueing theory*. This correlation provides a detailed description of congestion: its sources, its propagation and its effects. In particular, it identifies both bottlenecks and spillbacks; and quantifies their impact upon the overall network performance.

The second contribution of this paper concerns the improvement of the use of existing infrastructure. 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 analytical network models have not taken spillbacks into account. More generally, most signal control strategies do not account for saturated or highly congested networks where spillbacks are likely to occur (Papageorgiou et al., 2003). We therefore believe that the considered queueing model is an appropriate tool to improve urban signal settings, namely during peak hours. Furthermore, the stochastic nature of this model allows it to take into account the variability of traffic flows, which is particularly important when designing fixed-time signal plans (Yin, 2008).

This paper is structured as follows. We present in Section 2 a literature review and the signal control optimization framework. In Section 3 we describe the network model and formulate the optimization problem. We then discuss the role of a microscopic traffic simulation tool used in this framework (Section 4.1). 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 4.2 analyzes the added value of the explicit modeling of correlation. In Section 4.3 it is compared with a pre-existing signal plan for the city of Lausanne, and to the plans derived by the methods proposed in Webster (1958) and in the Highway Capacity Manual (TRB, 1994; Tian, 2002).

2 Literature Review

2.1 Analytic queueing models

Queueing models have been used in transportation mainly to model highway traffic (Garber and Hoel, 2002). Several simulation models have been developed, but few studies have explored the potential of the queueing theory framework to develop analytical urban traffic models. Furthermore, existing urban queueing models have mainly focused on unsignalized intersections. Heidemann and Wegmann (1997) give an excellent literature review for exact analytical 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 (1962). Heidemann also contributed to the study of signalized intersections (1994), and presented a unifying approach to both signalized and unsignalized intersections (1996). 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 analytical 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 the traffic process for a set of urban intersections using an analytic queueing network framework. Nevertheless the methods proposed by Jain and Smith (1997) and Van Woensel and Vandaele (2007) which are both based on the Expansion Method (Kerbache and Smith, 2000) and formulated for highway traffic could be extended to consider an urban setting.

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 (Papageorgiou et al., 2003). 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 pre-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 (1958), and is still widely used. Other expressions include those of Newell (1965), Miller (1963), and McNeil (1968). Viti (2006) provides a review of delay models; Dion et al. (2004) compare the performance of different delay models, and Chow and Lo (2007) 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 (1997) 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 (Wong and Yang, 1997; Ziyou and Yifan, 2002).

The works of Allsop (1992) and of Shepherd (1994) review signal control methods. Allsop (1992) describes in detail the corresponding terminology as well as the different formulations for isolated methods. More recently the reviews of Papageorgiou et al. (2003) and Cascetta et al. (2006) cover different but complementary aspects of this research field. Papageorgiou et al. (2003) 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. (2006) review the more general problem of traffic control and demand assignment methods.

Fixed-time isolated strategies

These strategies can be stage-based such as SIGSET (Allsop, 1971) and SIGCAP (Allsop, 1976). SIGSET minimizes delay using Webster's nonlinear formulation (Webster, 1958), whereas SIGCAP maximizes reserve capacity. Both methods consider a set of linear constraints. Improta and Cantarella (1984) consider a phase-based method formulated as a mixed-integer linear program. They give formulations for both delay minimization and reserve capacity maximization problems.

Fixed-time coordinated strategies

Optimizing a set of signals along an arterial is the focus of the arterial progression schemes MAXBAND (Little et al., 1981) and MULTIBAND (Gartner et al., 1991). 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 (Gartner and Stamatiadis, 2002). Heuristics have also been specifically developed to solve this problem (Pillai et al., 1998). 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 (Robertson and Bretherton, 1991).

Several phase-based strategies have been proposed (Wong et al., 2002; Wong, 1997; Wong, 1996). 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. (2002) 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 by Papageorgiou et al. (2003). 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 (Webster, 1958). These splits are then adjusted to improve progression. Boillot et al. (1992) highlight 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 (Abu-Lebdeh and Benekohal, 2003), Chow and Lo (2007) emphasize that the use of TRANSYT is appropriate only for low to moderate degrees of saturation.

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 (Boillot et al., 1992). 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 by Sen and Head (1997). 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 (Bretherton, 1989) is considered to be the traffic-responsive version of TRANSYT. A description of how TRANSYT evolved into SCOOT is given by Robertson and Bretherton (1991). 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 (Robertson and Bretherton, 1991)). The Australian method SCATS (Lowrie, 1982) 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 Aboudolas et al. (2007) and Dinopoulou et al. (2006) both describe, their performance deteriorates under congested conditions.

Dynamic programming methods are used in the French system PRODYN (Henry and Farges, 1989) as well as in the US systems OPAC and RHODES. RHODES (Mirchandani and Head, 2001) uses the COP algorithm (Sen and Head, 1997) 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 by Gartner et al. (2001) and Gartner et al. (1991). The Italian method UTOPIA is yet another method that has been evaluated and implemented (Mauro and Di Taranto, 1989). As Dinopoulou et al. (2006) describe, the exponential complexity of these methods does not allow for network-wide optimization. This is also emphasized by Boillot et al. (1992): "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 (Boillot et al., 2006; Boillot et al., 1992), and of the TUC method (Dinopoulou et al., 2006).

The method proposed in this paper belongs to the category of fixed-time coordinated methods. Traditionally, fixed-time strategies have been considered suitable only for undersaturated traffic conditions (Abu-Lebdeh and Benekohal, 1997; Shepherd, 1994; Chow and Lo, 2007; Papageorgiou et al., 2003). 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 (Yin, 2008). 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 (Dinopoulou et al., 2006). 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. Both Chow and Lo (2007) and Abu-Lebdeh and Benekohal (1997), illustrate the effects of ignoring this spatial dimension. 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 (Papageorgiou et al., 2003; Abu-Lebdeh and Benekohal, 2003). Furthermore Abu-Lebdeh and Benekohal (2003) emphasize that accounting for the effects of queue propagation remains a secondary consideration within a signal timing framework. We therefore believe that 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 correlation.

3 Methodological framework

We consider an urban transportation network, composed of a set of both signalized and unsignalized intersections. We capture the traffic dynamics with a set of queuing models organized in a network, or a *queueing network model.*

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.

A queue is then associated with each lane of each segment in the network. The interactions among the 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.

We consider 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 control problem consists in minimizing 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, such as the total demand, the route choice decisions and the topological structure of the street network, with endogenous variables, such as the capacities and the probability of spillbacks. The latter are directly linked with the decision variables. Consequently, we now formulate the model capturing the traffic dynamics that derives T from g, the exogenous parameters and the endogenous variables, as well as the constraints associated with the traffic signal settings.

3.1 The network model

In a previous paper (Osorio and Bierlaire, forthcoming), we have proposed a new analytic queueing network model that accurately describes the formation and the diffusion of congestion. 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 equations capturing the queue dynamics. Given the exogenous parameters, the values of the endogenous variables are evaluated by solving a system of nonlinear equations. We extend this formulation by considering the capacities endogenous, as they are determined by the decision variables (i.e. the green splits).

3.2 Queues

A queue is associated with each lane of each segment in the network. Each queue is connected to the downstream segments where a turning of the underlying lane is permitted. Note that connecting a queue to a segments means that it is connected to all of the queues in that segment.

All queues have one server, which represents the service due to the change of capacity at the boundary of a segment. The size of a queue i is denoted by k_i. It is composed of the server and the buffer. Note that k_i is known as the capacity of the queue in queueing theory. In this paper the term *capacity* will be used according to its traffic theory definition (VSS, 1998), and we therefore refer to k_i as the queue size. Heidemann (1996), as well as Van Woensel and Vandaele (2007), divide each road into segments of length $1/k_{\mathrm{jam}}$, where k_{jam} is the jam density, and thus $1/k_{\mathrm{jam}}$ represents the minimal length that each vehicle needs. We also follow this type of reasoning and define the queue size 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). This fraction is then rounded down to the nearest integer. In this model all queues have a finite size. This is referred to in queueing theory as *finite capacity queues*, and is necessary in order to account for congestion and spillback effects.

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 i 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 (Bierlaire and Frejinger, 2008).

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.

For unsignalized intersections (e.g. two-way stop controlled intersections, yield-controlled intersections) the standard VSS (1999a) 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 (2006). 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 (1999b), 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 9 of the Highway Capacity Manual (TRB, 1994).

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.

3.3 System of equations

In this paper all queues have a single server. Thus the equations presented by Osorio and Bierlaire (forthcoming) simplify. In the following notation all rates are average rates and the index i refers to a given queue.

$\pi(i)$	stationary distribution;
$\mathcal{S}(i)$	state space;
$Q(i)$	transition rate matrix;
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^f_i	probability of being blocked at queue i;
p_{ij}	transition probability from queue i to queue j;
k_i	queue size;
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 to a single value that is now denoted $\tilde{\mu}_i$. The system of equations is thefore given by:

$$\pi(i)Q(i) = 0, \tag{1}$$

$$\sum_{s\in\mathcal{S}(i)} \pi(i)_s = 1, \tag{2}$$

$$Q(i) = f(\lambda_i, \mu_i, P^f_i, \tilde{\mu}_i), \tag{3}$$

$$\lambda_i = \gamma_i + \frac{\sum_j p_{ji}\lambda_j(1 - P(N_j = K_j))}{(1 - P(N_i = k_i))}, \tag{4}$$

$$\frac{1}{\tilde{\mu}_i} = \sum_{j\in\mathcal{I}^+} \frac{\lambda_j(1 - P(N_j = K_j))}{\lambda_i(1 - P(N_i = k_i))\hat{\mu}_j}, \tag{5}$$

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P^f_i \frac{1}{\tilde{\mu}_i}, \tag{6}$$

$$P^f_i = \sum_j p_{ij} P(N_j = K_j).. \tag{7}$$

We briefly describe these equations, 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. Equations (1) and (2) are known as the *global* *balance equations*, they link the stationary distribution of a queue to its transition rate matrix, $Q(i)$. This matrix describes the rates at which a transition can take place between any pair of states. It is defined by Equation (3), where function f is detailed in Table 1 of Osorio and Bierlaire (forthcoming). We approximate the transition rates using structural parameters that capture the between-queue correlation (Equations (4)-(7)). 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 will dissipate. Each queue has 6 endogenous variables $(\lambda_i, \mu_i, \tilde{\mu}_i, \hat{\mu}_i, P(N_i = k_i), P_i^f)$. For each queue the dimension of its distribution is $2k_i + 1$. Thus the system of equations consists of $\sum_i (2k_i + 7)$ nonlinear equations.

3.4 Optimization problem

In order to formulate the signal control problem we intordouce 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:

$$\min_{g,x} \; T(g, x, \alpha) \tag{8}$$

subject to

$$\sum_{p \in \mathcal{P}_{\mathcal{I}}(i)} g_p = b_i, \ \forall i \in \mathcal{I} \tag{9}$$

$$\mu_\ell - \sum_{p \in \mathcal{P}_{\mathcal{L}}(\ell)} g_p s = 0, \ \forall \ell \in \mathcal{L} \tag{10}$$

$$h(x, \alpha) = 0 \tag{11}$$

$$g \geq g_L \tag{12}$$

$$x \geq 0. \tag{13}$$

The objective is to reduce the average time that vehicles spend in the network, which is represented by T (Equation (8)). T is a nonlinear function of the queueing model parameters. Resorting to the notation of the previous section, T is given by

$$\sum_i \frac{E[N_i]}{\lambda_i}.$$

The linear constraints (9) link the green times with the available cycle time for each intersection. Equation (10) links the green times of the signalized lanes to their capacities. The bounds (12) correspond to minimal green time values for each phase. These have been set to 4 seconds according to the Swiss standard VSS (1992). Equation (11) represents the network model, presented in Section 3.3.

The optimization problem is solved with the Matlab routine for constrained nonlinear problems, *fmincon*, which resorts to a sequential quadratic programming method (Coleman and Li, 1996, 1994). A feasible initial point is obtained by fixing a control plan and solving the network model (Equation (11)). We refer the reader to Osorio and Bierlaire (forthcoming) for more details on the solution procedure of this system of equations as well as for its own initialization settings.

4 Empirical analysis

4.1 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 (Dumont and

Bert, 2006) is implemented with the AIMSUN simulator (TSS, 2008). 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 Dumont and Bert (2006). Based on this control plan we give initial values to the capacities of the signalized lanes.

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

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.

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

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.

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. 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).

We now compare the performance of several methodologies, by considering a subnetwork of the Lausanne city center. For each methodology we derive the optimal signal plan for the subnetwork, and then use the simu-

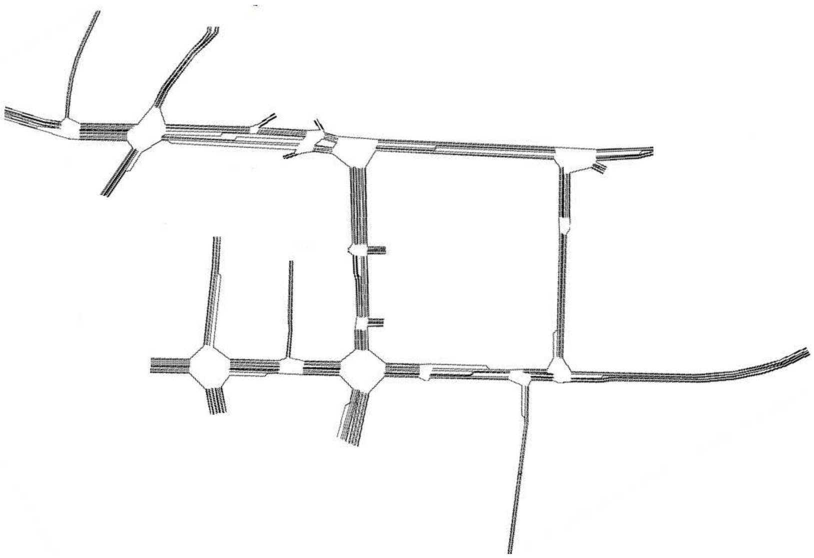

Figure 1: Subnetwork of the Lausanne city center

lation model to evaluate its effect upon the entire Lausanne network. The subnetwork (Figure 1) 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. The queueing model of this network consists of 102 queues. The optimization problem consists of 2288 endogenous variables with their corresponding lower bound constraints, 1829 nonlinear equality constraints, and 417 linear equality constraints.

4.2 Between-queue correlation

The queueing model proposed in this paper describes congestion by taking into account the correlation between upstream and downstream roads. In this section we illustrate the added value of accounting for the correlation. 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 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 2 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 3 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 2 and 3 are evidence of the added value of accounting for the dependence of queues in signal optimization.

We have also looked at the densities in the network. We did not expect any noticeable difference, as the network is highly congested, so that only the global throughput could be affected. Nevertheless, we found 3 locations with a significant difference in densities. Each plot of Figure 4 displays the density of one of these 3 roads as a function of time. For all three cases, there is a significantly smaller density under the correlated plan. Figure 5 displays errorbars such that the distance from the average to the upper (respectively, the lower) limit of the bar is equal to the standard deviation. The plots in the left column correspond to the independent plan, those in the right column correspond to the correlated plan. Each row of plots considers one of the 3 previously mentioned roads. These plots illustrate how with increasing congestion there is less variability in the density across replications under the correlated plan.

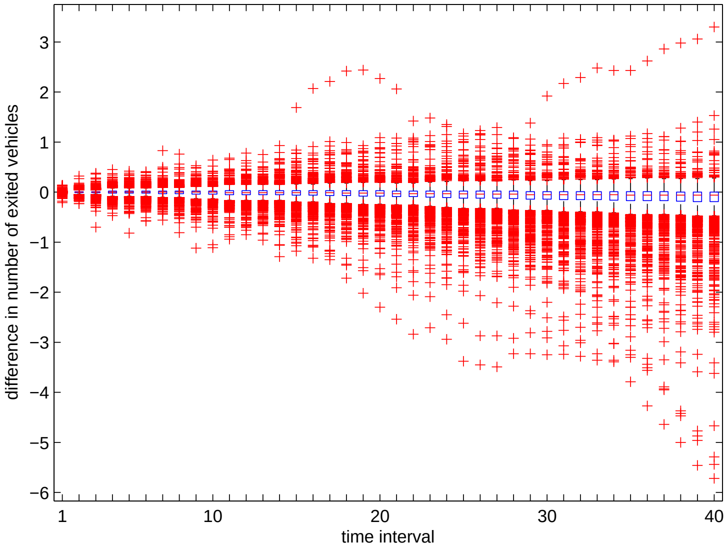

Figure 2: Difference in the average number of vehicles that have exited each OD pair versus time

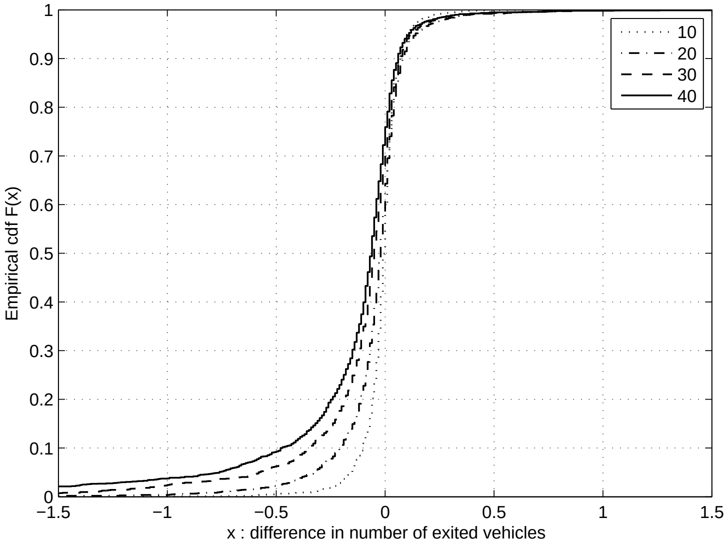

Figure 3: Empirical cumulative distribution function of the difference in the average number of vehicles that have exited the OD pairs for time intervals 10, 20, 30 and 40

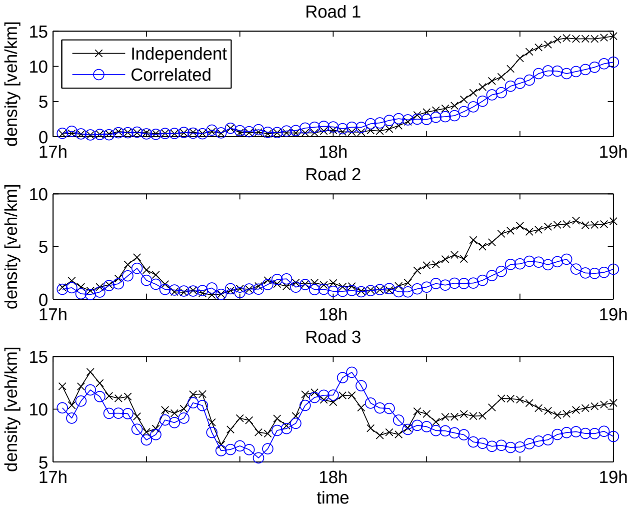

Figure 4: Average density versus time for 3 roads of the subnetwork

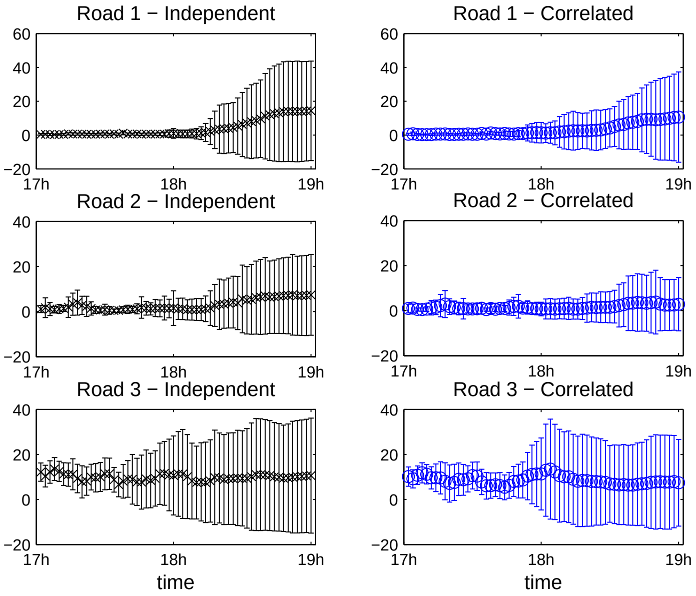

Figure 5: Errorbars for the average density [veh/km], plotted versus time for 3 roads of the subnetwork

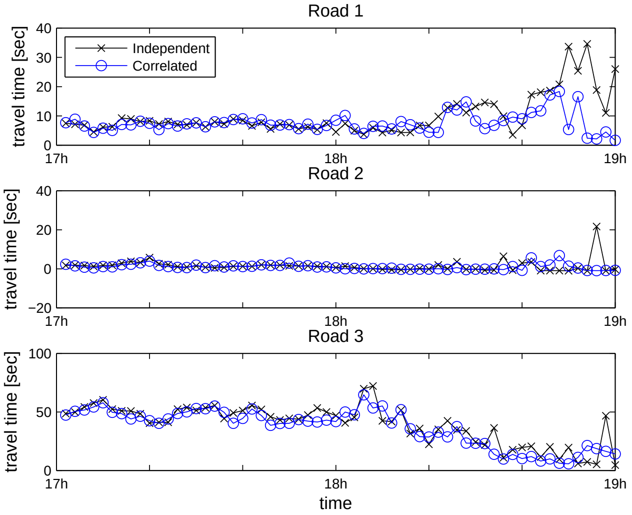

Figure 6: Average travel time per vehicle for 3 roads of the subnetwork

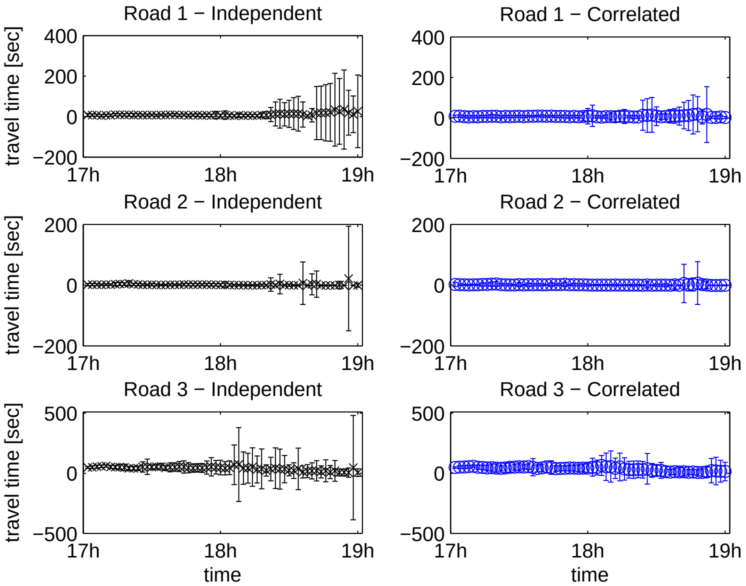

Figure 7: Errorbars for the average travel time, plotted versus time for 3 roads of the subnetwork

We have also performed an analysis of the impact on the average travel time per vehicle on these roads. In this case the average travel times do not exhibit a significant difference (Figure 6), except for the end of the simulation period on road 1. The added value of the method with correlated queues clearly appears in the analysis of the standard deviations, as illustrated in Figure 7. These results illustrate well the added value of the method, not only on the global throughput but also locally.

4.3 Comparison with pre-existing methods

We now compare the signal settings derived by the method proposed in this paper with a pre-existing fixed-time signal settings for the city of Lausanne, the method derived by Webster (1958) and with the method suggested in the Highway Capacity Manual (TRB, 1994; Tian, 2002).

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 we refer the reader to Dumont and Bert (2006). This signal plan will be referred to as the *base* plan.

HCM/Webster 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 (TRB, 1994; Tian, 2002) leads to the same green split equations as Webster's method (1958). This equivalence is detailed in Osorio and Bierlaire (2008).

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 softwares such as SYNCHRO and PASSER V (Chaudhary et al., 2002); and the delay estimate is one of the best known (Cascetta, 2001). 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

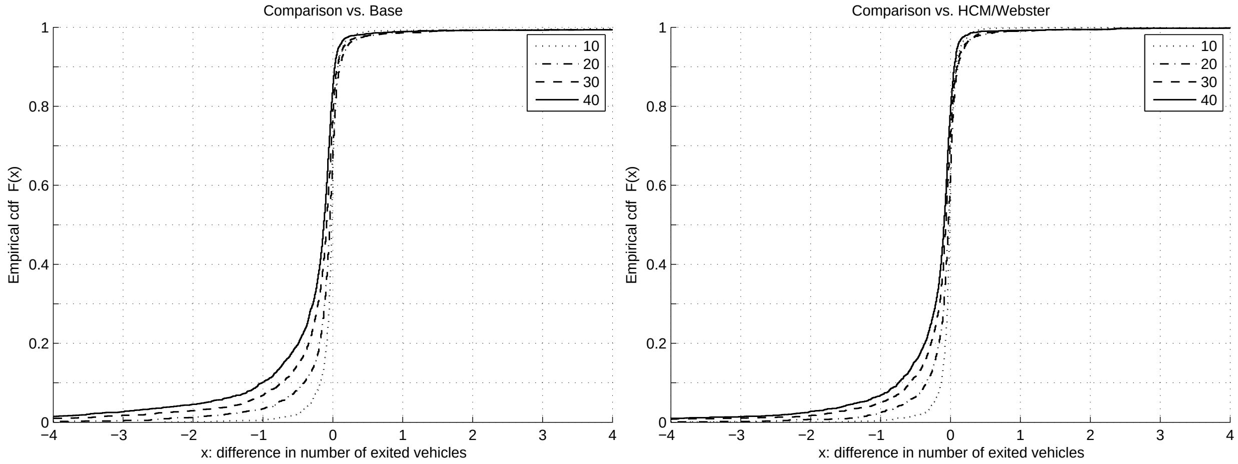

Figure 8: Empirical cumulative distribution function of the difference in the average number of vehicles that have exited the OD pairs for time intervals 10, 20, 30 and 40

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:

$$g_p = \frac{Y_p}{\sum_{j \in \mathcal{P}_{\mathcal{I}}(i)} Y_j} b_i \quad \forall p \in \mathcal{P}_{\mathcal{I}}(i). \tag{14}$$

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 (1999b). The approach flows are set using the observed flows derived by the simulation model.

We consider the network and simulation setup described in Section 4.1. 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 4.2. The empirical cumulative distribution functions of Figure 8 show that there is a high proportion of OD pairs for which the new plan yields an increase in outflow. This figure also shows that this proportion increases with congestion. The asymmetry of this figure illustrates the significant superiority of the proposed method.

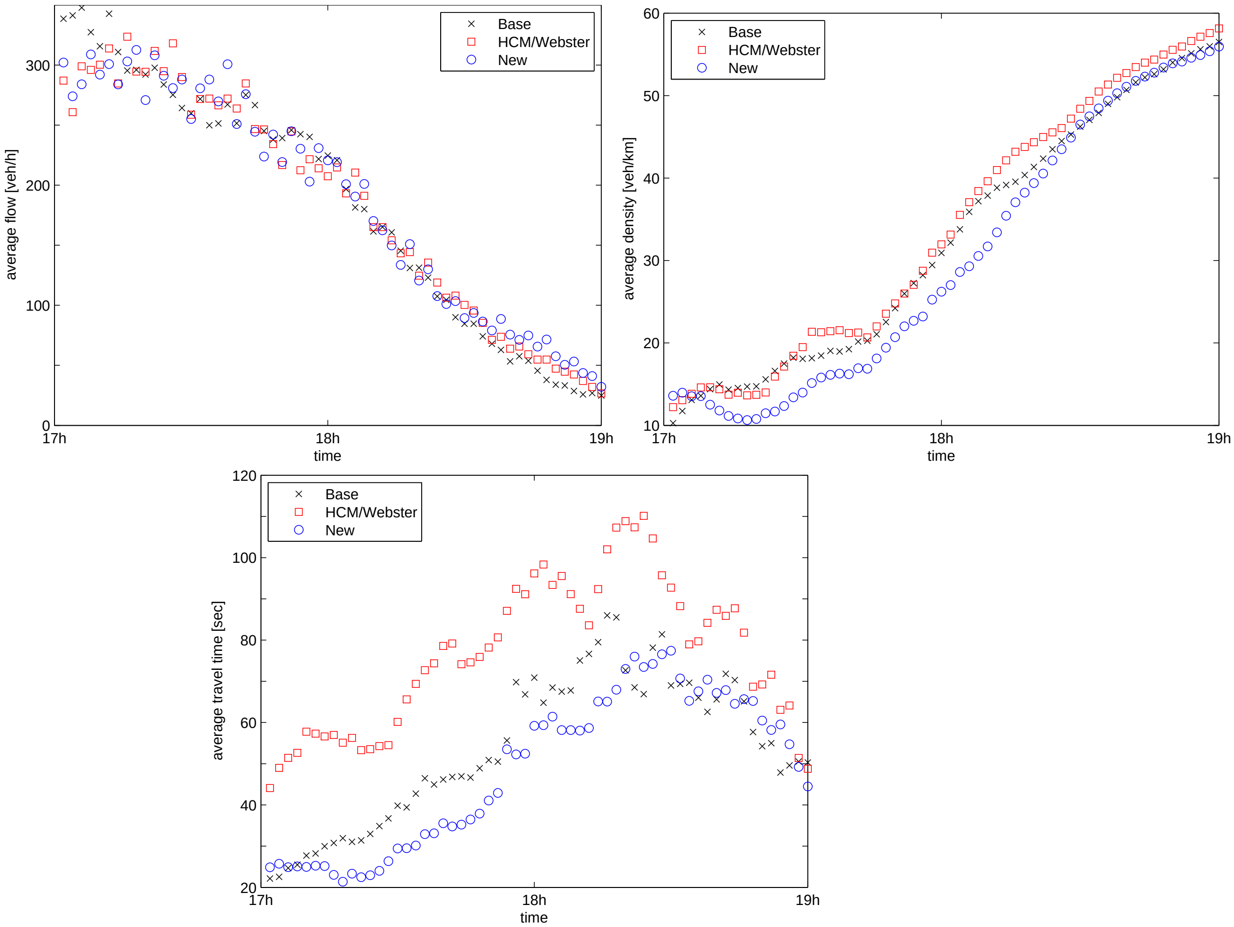

Figure 9: Average flow, density and travel time of the roads of the subnetwork, plotted versus time

Contrarily to the previous experiment, we observe here a significant improvement in terms of densities and travel times. The 3 plots of Figure 9 consider the flow, the density and the travel time of the roads of the subnetwork, plotted across time. The crosses, squares and circles denote the base plan, the HCM/Webster plan and the new plan, respectively. These plots illustrate how the new plan leads to improved subnetwork density and travel times, whereas for the flow there is no trend.

5 Conclusion

In this paper we have formulated a fixed-time traffic signal optimization problem, where the underlying traffic model is based on a queueing network model. By using a set of structural parameters that capture the between-queue correlation, this queueing model provides a detailed description of how congestion arises and how it spreads.

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 and are less variable.

This model makes an attractive trade-off between a detailed description of congestion and analytical tractability. It is therefore particularly appropriate for the study and management of congested urban networks.

Future research will follow two main tracks. On the one hand, the framework will be extended to consider coordinated and traffic-responsive signal settings. On the other hand, it will be adjusted for large-scale scenarios by considering decomposition techniques.

Acknowledgments

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

References

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

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

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

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

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

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

Bierlaire, M. and Frejinger, E., 2008, Route choice modeling with network-free data, *Transportation Research Part C* **16**(2), 187–198.

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

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

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

Cascetta, E., 2001, *Transportation Systems Engineering: theory and methods*, Vol. 49 of *Applied Optimization*, Kluwer academic publishers, Dordrecht, chapter 2, pp. 50–51.

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

CEC, 2007, *Green Paper. Towards a new culture for urban mobility.* COM (2007) 551. Office for Official Publications of the European Communities, Luxembourg.

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

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

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

Coleman, T. F. and Li, Y., 1996, An interior, trust region approach for nonlinear minimization subject to bounds, *SIAM Journal on Optimization* **6**(2), 418–445.

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

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

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

Garber, N. J. and Hoel, L. A., 2002, *Traffic and Highway Engineering*, 3rd edn, Books Cole, Thomson Learning, chapter 6, pp. 204–210.

Gartner, N. H., Assman, S. F., Lasaga, F. and Hou, D. L., 1991, A multiband approach to arterial traffic signal optimization, *Transportation Research Part B* **25**(1), 55–74.

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

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

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

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

Heidemann, D. and Wegmann, H., 1997, Queueing at unsignalized intersections, *Transportation Research Part B* **31**(3), 239–263.

Henry, J. J. and Farges, J. L., 1989, PRODYN, *IFAC Symposium of Control Communications in Transportation*, Pergamon Press, Oxford, pp. 253–255.

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

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

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

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

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

Mauro, V. and Di Taranto, C., 1989, UTOPIA, *IFAC Symposium of Control Communications in Transportation*, Pergamon Press, Oxford, pp. 245–252.

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

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

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

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

Osorio, C. and Bierlaire, M., 2008, Network performance optimization using a queueing model, *Proceedings of the European Transport Conference (ETC)*, Noordwijkerhout, The Netherlands.

Osorio, C. and Bierlaire, M., forthcoming, An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, *European Journal Of Operational Research* . Accepted for publication.

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

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

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

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

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

Tanner, J. C., 1962, A theoretical analysis of delays at an uncontrolled intersection, *Biometrika* **49**(1–2), 163–170.

Tian, Z., 2002, *Capacity analysis of traffic-actuated intersections*, Master's thesis, Massachusetts Institute of Technology, Cambridge, USA.

TRB, 1994, *Highway capacity manual. Special report 209*, 3rd edn. Chapters 1,2,9,11.

TSS, 2008, *AIMSUN NG and AIMSUN Micro Version 5.1.*

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

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

VSS, 1992, *Norme Suisse SN 640837 Installations de feux de circulation; temps transitoires et temps minimaux.*

VSS, 1998, *Norme Suisse SN 640017a Capacité, niveau de service, charges compatibles; norme de base.*

VSS, 1999a, *Norme Suisse SN 640022 Capacité, niveau de service, charges compatibles; carrefours sans feux de circulation.*

VSS, 1999b, *Norme Suisse SN 640023 Capacité, niveau de service, charges compatibles; carrefours avec feux de circulation.*

VSS, 2006, *Norme Suisse SN 640024a Capacité, niveau de service, charges compatibles; carrefours giratoires.*

Webster, F. V., 1958, Traffic signal settings, *Technical Report 39*, Road Research Laboratory.

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

Wong, S., 1997, Group-based optimisation of signal timings using parallel computing, *Transportation Research Part C* **5**(2), 123–139.

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

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

Yagar, S., 1975, CORQ - a model for predicting flows and queues in a road corridor, *Transportation Research Record* **533**, 77–87.

Yin, Y., 2008, Robust optimal traffic signal timing, *Transportation Research Part B* **42**(10), 911–924.

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