Transportation Research Part B 46 (2012) 1393–1403

Contents lists available at SciVerse ScienceDirect

Transportation Research Part B

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

Exploiting the fundamental diagram of urban networks for feedback-based gating

Mehdi Keyvan-Ekbatani [a,*], Anastasios Kouvelas [b], Ioannis Papamichail [a], Markos Papageorgiou [a]

[a] Dynamic Systems and Simulation Laboratory, Department of Production Engineering and Management, Technical University of Crete, University Campus, 73100 Chania, Greece
[b] California Center for Innovative Transportation, University of California at Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 21 October 2011
Received in revised form 23 June 2012
Accepted 23 June 2012

Keywords:
Traffic signal control
Saturated traffic condition
Gating
Network fundamental diagram
Feedback control

ABSTRACT

Traffic signal control for urban road networks has been an area of intensive research efforts for several decades, and various algorithms and tools have been developed and implemented to increase the network traffic flow efficiency. Despite the continuous advances in the field of traffic control under saturated conditions, novel and promising developments of simple concepts in this area remains a significant objective, because some proposed approaches that are based on various meta-heuristic optimization algorithms can hardly be used in a real-time environment. To address this problem, the recently developed notion of network fundamental diagram for urban networks is exploited to improve mobility in saturated traffic conditions via application of gating measures, based on an appropriate simple feedback control structure. As a case study, the proposed methodology is applied to the urban network of Chania, Greece, using microscopic simulation. The results show that the total delay in the network decreases significantly and the mean speed increases accordingly.

1. Introduction

Traffic congestion in urban road networks is a persisting or even increasing problem of modern society. Congestion can be reduced either by increasing road capacity (supply), or by reducing traffic demand. On the supply side, the provision of new infrastructure is usually not a feasible solution, hence it is necessary to focus on a better utilization of the existing infrastructure (e.g. via traffic management), to mitigate congestion and improve urban mobility. The field of urban traffic control (UTC) has been studied and developed in a variety of ways during the past decades. In fact, the traffic flow conditions in large-scale urban networks depend critically on the applied signal control strategies. However, as the debate regarding urban mobility and the wish for a sustainable transport system indicate, the negative effects of congested transport networks, such as excessive delays, environmental impact and reduced safety, persist or even increase; hence, introducing improved traffic signal control methods and techniques continues to be a vital issue. In particular, the development of practicable and efficient real-time signal control strategies for urban road networks under saturated traffic conditions is a major challenge with significant scientific and practical relevance. The scientific relevance stems from the increased interest in the specific problem as well as recent, potentially valuable, models and insights that may contribute to improved signal control methods. The practical relevance stems from the congestion, degradation and gridlock problems encountered increasingly in modern urban road networks that could benefit highly from improved signal control under saturated traffic conditions.

* Corresponding author. Tel.: +30 28210 37421; fax: +30 28210 37584.
E-mail addresses: m_ekbatani@dssl.tuc.gr (M. Keyvan-Ekbatani), kouvelas@berkeley.edu (A. Kouvelas), ipapa@dssl.tuc.gr (I. Papamichail), markos@dssl.tuc.gr (M. Papageorgiou).

0191-2615/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2012.06.008

UTC systems constitute a scientific field with long-lasting and extensive research and development activities. Many methodologies have been proposed so far, but there is still space for new developments, particularly for saturated traffic conditions. In fact, widely used strategies like SCOOT (Hunt et al., 1982) and SCATS (Lowrie, 1982), although applicable to large-scale networks, are deemed less efficient under saturated traffic conditions. On the other hand, more advanced traffic-responsive strategies like OPAC (Gartner, 1983), PRODYN (Farges et al., 1983), and RHODES (Mirchandani and Head, 1998) use optimization algorithms with exponential complexity, which do not permit a straightforward central network-wide application of their optimization-based kernel. Thus, most available strategies face limitations when it comes to saturated traffic conditions that are frequently observed in modern metropolitan areas. A noteworthy and practicable attempt to address saturated traffic conditions was the more recently developed signal control strategy TUC (Diakaki et al., 2002, see also Aboudolas et al., 2010). Furthermore, a number of research approaches have been proposed, that employ various computationally expensive numerical solution algorithms, including genetic algorithms (Abu-Lebdeh and Benekohal, 1997; Lo et al., 2001), multi-extended linear complementary programming (De Schutter and De Moor, 1998), mixed-integer linear programming (Lo, 1999; Beard and Ziliaskopoulos, 2006) and ant colony optimisation (Putha et al., 2010); however, in view of the high computational requirements, the network-wide implementation of these optimization-based approaches might face some difficulties in terms of real-time feasibility. In fact, a recent FHWA report (as cited in Lieberman et al. (2010)) opined: "*No current generally available tool is adequate for optimizing [signal] timing in congested conditions*".

The notion of a fundamental diagram (e.g. in the form of a flow-density curve) for highways was recently found to apply (under certain conditions) to two-dimensional urban road networks as well; see Gartner and Wagner (2004) for simulation-based experiments; Geroliminis and Daganzo (2008) for real-data based investigations; Daganzo and Geroliminis (2008), Farhi (2008) and Helbing (2009) for analytical treatments. In fact, a fundamental-diagram-like shape of measurement points was first presented by Godfrey (1969), but also observed in a field evaluation study by Dinopoulou et al. (2005), see Fig. 6 therein. The concept is sometimes called MFD (macroscopic fundamental diagram), but since the ordinary fundamental diagram (for highways) is also macroscopic, we prefer to call it NFD (network fundamental diagram) for better distinction. Although the exact NFD curve may depend on the origin–destination demand (Ji et al., 2010), it may be quite stable from day to day, particularly if the traffic load is homogeneously distributed in network links (Geroliminis and Sun, 2011). In simulated environments, where different signal control strategies are tested, this homogeneity condition may call for activation of a dynamic traffic assignment device to reduce possible transient phenomena, such as a hysteresis between network filling and emptying data (Aboudolas et al., 2010; Geroliminis and Sun, 2011).

The NFD concept for urban road networks has been an issue of intensive investigations recently; indeed, the conditions under which it appears, the stability of its shape under different O–D patterns or at different peak periods or days-of-the-week, the impact of different signal control strategies, the possible hysteresis between the network filling and emptying phases, are still under the loop of ongoing analytical or empirical investigations and research. Nevertheless, based on what is known or observed in data, it is not too early for the NFD concept to be considered a basis for the derivation of traffic control strategies. Daganzo (2007) used the NFD concept to propose a control rule that maximizes the network outflow; however, as discussed later, that rule cannot be directly employed for practical use in urban networks. Other works (Haddad et al., 2012; Strating, 2010) pursued a model-predictive control (MPC) approach. However, MPC calls for sufficiently accurate model and external disturbance predictions, which may be a serious impediment for practicable control. In fact, Haddad et al. (2012) tested the MPC concept only on the basis of the same simple model used within the optimal control problem; while Strating (2010) used detailed microscopic simulation, but reported a failure to produce sensible control results.

A practical tool, frequently employed against over-saturation of significant or sensitive links, arteries or urban network parts, is gating (Wood et al., 2002; Bretherton et al., 2003; Luk and Green, 2010). The idea is to hold traffic back (via prolonged red phases at traffic signals) upstream of the links to be protected from over-saturation, whereby the level or duration of gating may depend on real-time measurements from the protected links. The method is usually employed in an ad hoc way (based on engineering judgment and manual fine-tuning) regarding the specific gating policy and quantitative details, which may readily lead to insufficient or unnecessarily strong gating actions.

In this paper, the urban NFD concept is exploited to improve mobility in saturated traffic conditions via application of gating measures, based on an appropriate simple feedback control structure. More specifically, an operational NFD is defined, which may eventually allow for efficient gating control based on a limited amount of real-time measurements. The operational NFD is used to derive clear gating targets so as to maximize throughput in the protected network part. Moreover, an appropriate simple dynamic model is developed, that allows for the straightforward derivation of simple, efficient and practicable feedback regulators, suitable for smooth and efficient operations. The proposed methodology is applied and demonstrated for the urban road network of Chania, Greece, in a microscopic simulation environment under realistic traffic conditions.

2. Methodology

2.1. General gating task

The objective of the presented methodology is to mitigate urban traffic congestion via feedback gating, by exploiting the notion of the network fundamental diagram (NFD) for an urban network part that needs to be protected from the detrimen-

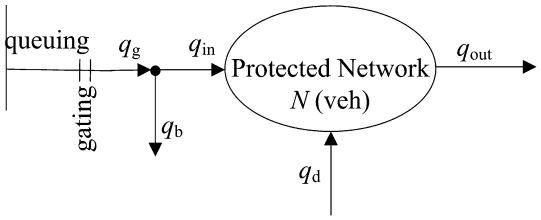

Fig. 1. General scheme of the protected network and gating.

tal effects of over-saturation. To gate the traffic flow (usually during the peak periods) in an urban network, the area to be protected from possible congestion and the locations where gating queues will be created, must be defined. The general scheme of gating, including the protected network (PN), is sketched in Fig. 1. To implement gating, the usual traffic lights settings must be modified at (one or more) upstream junctions, which may be located more or less close to the problematic area. In Fig. 1, the double line indicates the gating location, upstream of which vehicle queues may grow temporarily faster than without gating; q_g is the gated flow, a part of which (q_b) may not be bound for the protected network (PN); while q_{in} is the part of the gated flow that enters the protected network; q_d represents other (non-gated or internal) inflows to the PN (disturbances); finally q_{out} and N stand for the PN exit flow (both internal and external) and the number of vehicles included in the PN, respectively.

If N is allowed to grow beyond certain limits, the PN exit flow q_{out} decreases (according to the NFD) due to link queue spillovers and gridlock. To avoid this PN degradation, gating should reduce the PN inflow q_{in} appropriately, so as to maximize the PN throughput. This may incur some temporary vehicle delays in the queues of the gated junctions, which, however, may be eventually offset (at least for the q_{in} portion of the gated flow) thanks to the higher PN exit flow enabled by gating; on the other hand, the flow q_b will experience gating delays without any direct reward; these delays will be generally smaller if the gating junction is closer (or attached) to PN, due to accordingly smaller (or zero) flows q_b. Overall, gating will be beneficial if the saved delays in the PN are higher than the unnecessary delays incurred to the q_b portion of the gated flow. In some situations, e.g. when major problems in PN causes congestion to spread rapidly to adjoining areas, the use of gating could provide even higher benefits to the overall network.

2.2. Application network

In this study, the central business district (CBD) of the Chania urban road network, where the congestion usually starts during the peak period, is considered as the protected network. Eight gating links are specified exactly at the border of the protected network. A greater part of the Chania urban road network is modeled in the AIMSUN microscopic simulation environment (TSS, 2008), according to Fig. 2. The PN is separated from the rest of the network by the red[1] border in Fig. 2. The PN consists of 165 links, while the 8 traffic junctions selected for gating are indicated by arrows in Fig. 2. In the middle of every link inside the red border line, a loop detector has been installed, and the related measurements are collected at every cycle (in this case 90 s). The gating links have been chosen to provide sufficient space for vehicle queuing, so that further upstream junctions are not significantly obstructed. As indicated with small circled links in Fig. 2, multiple origins and destinations are introduced at the network boundaries, but also at internal network locations, including the PN area. These origins and destinations (O–D) account for various corresponding in- and outflows, including on-street and off-street parking arrivals and departures, that may partially affect the PN area. The introduced O–D flows are realistic (based on real measurements) but not exact (particularly with regard to the used O–D rates).

When running AIMSUN, the tool's embedded real-time dynamic traffic assignment option is activated, as this is deemed to lead to a more realistic distribution of the demand within the network. In particular, if gating measures create long queues and delays at the gated links, alternative routes (if available) may be selected by the drivers towards their respective destinations; clearly, this reflects the medium-term routing behavior of drivers to any introduced gating measures. Note also that this diversion may jeopardize to some extent the intended gating impact if drivers divert and enter the PN via non-gated links; therefore, the choice of gating links should also consider the availability and potential attractiveness of alternative routes that bypass the gating location.

2.3. Fundamental diagram of the PN

A network fundamental diagram may be an *ideal* NFD, if based on *exact* knowledge of the displayed quantities (this is practically only possible in analytic or simulation-based studies) for *all* links $z \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of all network links; or an *operational* NFD, if based on available (more or less accurate) measurements and estimates at a subset $\mathbb{M}$ of all links, i.e. $\mathbb{M} \subseteq \mathbb{Z}$; an operational NFD is called *complete*, if the measurements cover all network links, i.e. if $\mathbb{M} = \mathbb{Z}$.

[1] For interpretation of color in Figs. 2–7, the reader is referred to the web version of this article.

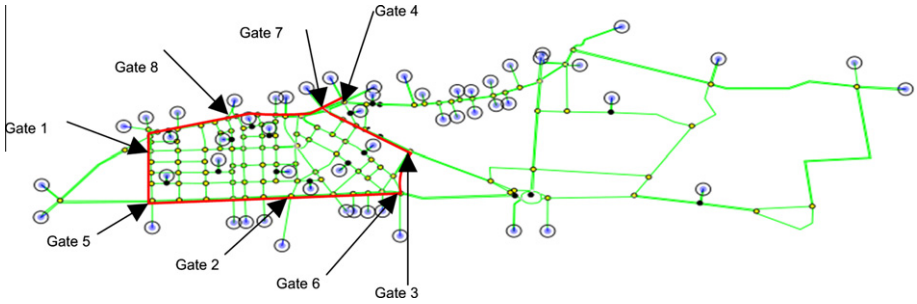

Fig. 2. Chania urban network modeled in AIMSUN.

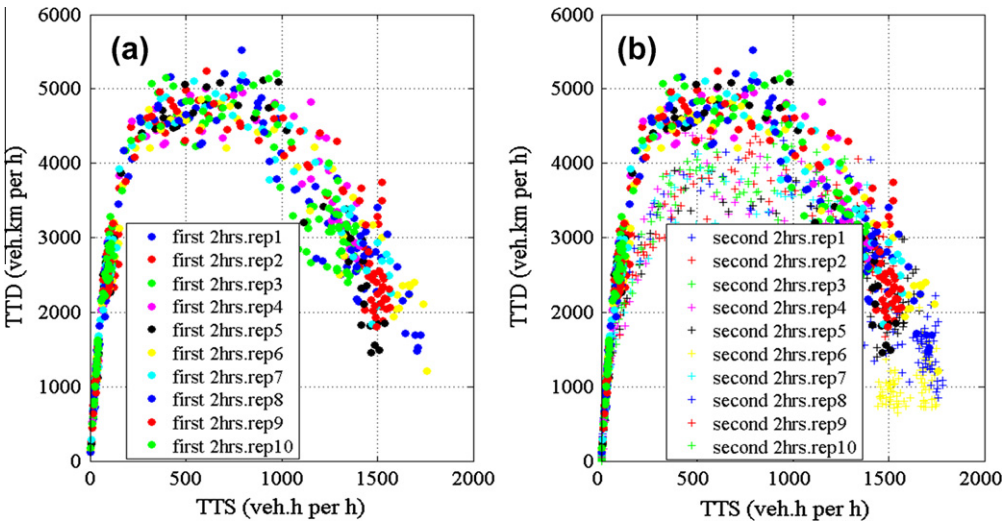

Fig. 3. (a) NFD of the PN for the first 2 h for 10 replications and (b) NFD of PN for the 4 h simulation for 10 replications.

The complete operational NFD of the Chania PN is obtained via a 4-h AIMSUN simulation scenario with realistic O–D demands and dynamic traffic assignment based routing, and is displayed in Fig. 3. The NFD's *y*-axis reflects the Total Traveled Distance (*TTD* in veh km per h), while the *x*-axis reflects the Total Time Spent (*TTS* in veh h per h) by all vehicles in the PN. The *TTD* and *TTS* are obtained from the emulated loop measurements via the following equations:

$$TTS(k) = \sum_{z\in\mathbb{M}} \frac{T\cdot \widehat{N}_z(k)}{T} = \sum_{z\in\mathbb{M}} \widehat{N}_z(k) = \widehat{N}_z(k) \tag{1}$$

$$TTD(k) = \sum_{z\in\mathbb{M}} \frac{T\cdot q_z(k)\cdot L_z}{T} = \sum_{z\in\mathbb{M}} q_z(k)\cdot L_z \tag{2}$$

where z is the link where a measurement is collected; $\mathbb{M}$ is the set of measurement links, here $\mathbb{M}=\mathbb{Z}$; $k = 0,1,2,\ldots$ is a discrete time index reflecting corresponding cycles; T is the cycle time; q_z is the measured flow in the link z during cycle k; L_z is the length of link z; and $\widehat{N}_z(k)$ is the estimated number of vehicles in link z during cycle k, which is derived from measured occupancy measurements via the following equation

$$\widehat{N}_z(k) = L_z\cdot\frac{\mu_z}{100\lambda}\cdot o_z(k-1) \tag{3}$$

where o_z is the measured time-occupancy (in%) in link z during cycle k; μ_z is the number of lanes of link z; and λ is the average vehicle length. Eq. (3) is reasonably accurate, particularly if the detector is located around the middle of the link (Papageorgiou and Vigos, 2008). According to the derivations in (1) and (2), *TTS* equals the number of vehicles in all PN links equipped with detectors; while *TTD* is a length-weighted sum of the corresponding PN link flows.

Fig. 3a displays the (complete operational) NFD for the Chania PN (assuming that all links are detector-equipped, i.e. $\mathbb{M}=\mathbb{Z}$) for the first 2 h of the employed scenario, i.e. the period during which the network is filled, and the congestion is created; 10 different replications (each with different seed in AIMSUN) were carried out. To build a comprehensive NFD that

includes free-flow conditions, the specified demand starts from very low values and increases gradually to levels that lead to heavy congestion in PN (as under typical real traffic conditions at the peak periods); eventually, the demand is gradually reduced until the network is virtually emptied at the end of the simulation (see Fig. 3b). Fig. 3a demonstrates that a fundamental diagram (asymmetric inverse-U) shape is indeed occurring during the 2-h network filling period, with quite moderate scatter even across different replications; Fig. 3b indicates that the inverse-U shape appears also during the decreasing demand period of 2 h, albeit with a visible hysteresis compared to the filling 2-h period. The hysteresis is limited and is due to different link load patterns that prevail in the emptying period compared to the filling period. Whatever the exact NFD (and despite some limited scatter), it can be seen in Fig. 3 that the maximum *TTD* values in the diagram occur in a *TTS* region of 600 to 800 veh h per h. If *TTS* (i.e. *N*) is allowed to increase beyond this limit, then *TTD* (and hence the PN throughput) decreases; this leads to an unstable escalation, as long as the PN inflows continue to be higher, that degrades increasingly the PN throughput and efficiency, leading them to accordingly low levels (or even to zero in the extreme total-gridlock case). To avoid this unstable degradation and, in fact, maximize the PN throughput and efficiency, the PN's *TTS* should be maintained in the mentioned optimal range, and this is exactly the goal pursued in this work.

2.4. System modeling for feedback control design

Gating may be enabled via very simple, but highly efficient and robust feedback regulators that are well-known in Control Engineering. The regulators are strictly based on real-time measurements, without any need for online model or demand predictions. On the other hand, for a proper choice of the feedback structure (among several offered in classical feedback theory), it is essential to know the basic dynamics of the process under control, and this task is indeed rendered quite simple and easy when using the notion of the NFD.

The developed model and feedback controller structures are summarized in Fig. 4. The model input is the gated flow q_g (see Fig. 1); the model output is the PN's *TTS*; while the main external disturbance is the uncontrolled PN inflow q_d. The model is first developed in a continuous-time environment for convenience. To start with, we have in the general case

$$q_{in}(t) = \beta \cdot q_g \cdot (t - \tau) \tag{4}$$

where β is the portion of gated flow (q_g) that enters the PN; t is the time argument; τ is the travel time needed for gated vehicles to approach the PN (when the gating link is not directly at the PN boundary). The conservation equation for vehicles in the PN (see Fig. 1) reads:

$$\dot{N}(t) = q_{in}(t) + q_d(t) - q_{out}(t) \tag{5}$$

As in the discrete-time case, we have also for the ideal values $TTS_{id}(t) = N(t)$ (where N is the real number of vehicles within PN), but *TTS* in Fig. 4 denotes the operational value, which differs from the ideal value in two respects: firstly, detectors may not be available in each and every PN link, hence the operational *TTS* will be smaller by some factor $A \leqslant 1$; secondly, the occupancy measurement and, most importantly, the estimation (3) may not be exact, hence we introduce a measurement/estimation error ε_1; which finally yields

$$TTS(t) = A \cdot N(t) + \varepsilon_1(t) \tag{6}$$

From this operational $TTS(t)$, we may derive, using the operational NFD, the corresponding (operational) *TTD*, i.e.

$$TTD(t) = F[TTS(t)] + \varepsilon_2(t) \tag{7}$$

where $F(\cdot)$ is a nonlinear best-fit function of the operational NFD's measurement points, and ε_2 denotes the corresponding fitting error (due to NFD scatter). Since *TTD* in (7) and Fig. 4 is the operational quantity, the ideal TTD_{id} (considering all PN links, not just the ones equipped with detectors) will be bigger, i.e.

$$TTD_{id}(t) \cdot B = TTD(t) \tag{8}$$

where $B \leqslant 1$ is the flow-analogous factor of A earlier.

To proceed, we will now introduce the modeling assumption that the PN outflow q_{out} is proportional to TTD_{id}, i.e.

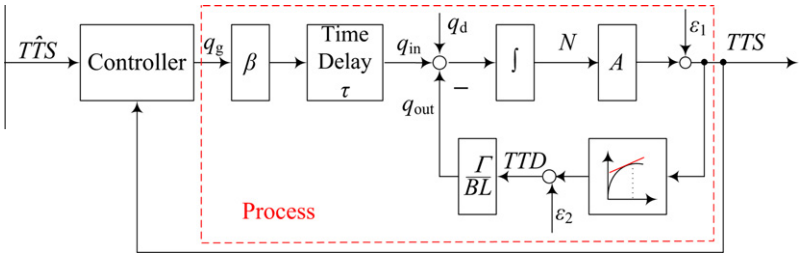

Fig. 4. Block diagram of the system and the feedback controller.

$$q_{out}(t) = \frac{\Gamma}{L} TTD_{id}(t) \tag{9}$$

where Γ is a sort of network exit rate, $0 \leqslant \Gamma \leqslant 1$, and L is the average PN link length. Replacing (8) in (9), we complete the process model derivation according to Fig. 4. The overall model (from q_g to TTS) is obtained by replacing (4), (6)–(9) into (5) and turns out to be a time-delayed nonlinear first-order system. Its portion from q_{in} to TTS (i.e. without the time delay) reads

$$\frac{d}{dt}(TTS(t)) = \left(q_{in}(t) + q_d(t) - \frac{\Gamma}{BL} F[TTS(t)]\right) \cdot A + \varepsilon(t)$$

where ε may be derived from the previous errors ε_1 and ε_2.

This model may be linearized around an optimal steady state that is within the aforementioned maximum *TTD region* of the NFD. The introduction of a desired steady state is quite usual in Control Engineering to enable the derivation of a linearized model and subsequent linear feedback control design. In fact, it is the goal of the feedback regulator, to be eventually derived, to maintain the system state around this steady state; thus, if successful, the steady-state and the linearized system dynamics assumptions are actually verified via the intended control action. Denoting steady-state variables with bars, we have

$$\bar{q}_{in} + \bar{q}_d = \bar{q}_{out} \tag{10}$$

$$\bar{q}_{out} = \frac{\Gamma}{BL} \overline{TTD} \tag{11}$$

while $\bar{\varepsilon}_1$ and $\bar{\varepsilon}_2$ are set equal to zero. With the notation $\Delta x = x - \bar{x}$ used analogously for all variables, the linearization yields

$$\frac{d}{dt}(\Delta TTS) = \left(\Delta q_{in} + \Delta q_d - \frac{\Gamma \bar{F}'}{BL} \Delta TTS\right) \cdot A + \varepsilon \tag{12}$$

where $\bar{F}'$ is the slope of the NFD at the optimal set-point $\widehat{TTS}$, i.e. $\overline{TTS} = \widehat{TTS}$. This set-point should be selected within the optimal *TTS*-range of the NFD, e.g. within [600, 800] for the Chania PN, for maximum efficiency. Note that $\bar{F}'$ may be virtually equal to zero if the set-point is optimal; nevertheless, we will assume $\bar{F}' > 0$ here, in order to enable proper linearized modeling. This assumption has no impact on the employed regulator (14), whose operation is only governed by the regulation error $\widehat{TTS} - TTS$.

The continuous-time state Eq. (12) of the protected network (using the conservation equation and the NFD) may be directly translated in discrete time by use of standard formulas (Seborg et al., 1989) as follows

$$\Delta TTS(k+1) = \mu \cdot \Delta TTS(k) + \zeta \cdot [\Delta q_{in}(k) + \Delta q_d(k)] + \varepsilon(k) \tag{13}$$

where $\mu = \exp(-\Gamma \bar{F}' TA/BL)$ and $\zeta = (1-\mu)BL/\Gamma \bar{F}'$. It is trivial to include in these models the time delay, by replacing q_{in} from (4).

The derived simple model includes a number of parameters that have clear physical meaning; nevertheless, the precise value of some of these parameters may be difficult to obtain in practice, particularly if the PN is a sizeable network (as in the Chania example). However, the main reason for developing the gating model is to deduce the basic structure of the underlying dynamics, which is essential for a proper choice of the regulator structure.

2.5. Controller design

To avoid congestion-caused degradation (i.e. a *TTD* decrease), the critical value (i.e. the value of *TTS* at which the maximum *TTD* is attained) in the NFD is considered as the set value for the controller. The control goal is to keep the traffic state of the PN around the set value, so that *TTD* is maximized and the network does not enter the over-saturation area in the NFD according to Fig. 3. To this end, given the derived model structure in the previous section, the following standard proportional–integral-type (PI) feedback controller is well suitable

$$q_{in}(k) = q_{in}(k-1) - K_P[TTS(k) - TTS(k-1)] + K_I[\widehat{TTS} - TTS(k)] \tag{14}$$

where K_P and K_I are the (non-negative) proportional and integral gains, respectively. Good regulator gain values may be found with appropriate Control Engineering methods or manual fine-tuning; model parameter estimation (e.g. of μ and ζ in (13)), by use of real q_{in} versus *TTS* measurements, may be useful in this endeavor; in any case, feedback regulators are quite robust to moderate parameter value changes.

If gating is applied at multiple links, the flow calculated by the (unique) regulator (14) must be split among the gated links according to some pre-specified policy. As long as the feedback-ordered total inflow is roughly followed via the gating traffic signal actions, the performance of the control (in terms of delay reduction) is not expected to depend significantly on the inflow splitting policy, except perhaps for special cases of network topology or demand patterns. What the splitting significantly affects, is the resulting queuing and delays at the individual gated links. For example, one may envisage the application of delay-balancing or queue-balancing splitting policies as in Papamichail and Papageorgiou (2011). This paper is mainly concerned with the overall gating control design and impact; while splitting and queue management issues,

which may include a variety of policies or wishes by the responsible traffic authorities and may incur corresponding requirements on detector equipment and overall cost, are left for more detailed future investigations. Thus, the splitting of the total ordered inflow in the present study is simply conducted in proportion to the respective saturation flows of the gated links.

The flow calculated by the regulator (14) must be constrained by pre-specified minimum and maximum values to account for physical or operational constraints. For the lower bound q_{min}, one may choose the flow corresponding to the minimum-green settings of the gated links (as in the Chania application here), or higher, e.g. if some gated links need to be protected from over-spilling. The upper bound has two components, a constant and a variable one, similarly to ALINEA ramp metering (Wang et al., 2010), and it is decided in real time which of the two is to be applied at each control step; the constant upper bound may be specified according to the maximum-green settings of the gated links (as in the Chania application here), or lower, e.g. if some downstream links need to be protected from over-spilling; the variable upper bound aims at activating the regulator more promptly under certain circumstances, see Wang et al. (2010) for further details on the reasoning and method. It should be noted that, in the Chania application of this paper, upper and lower flow bounds are actually specified also for every individual gated link. If the regulator flow distribution is found to violate some of these individual bounds, then the surplus flows are re-distributed among the rest of the gated links.

The necessary and sufficient conditions on the (non-negative) regulator parameters K_P and K_I for closed-loop stability of the linearized system (13) under regulator (14) can be easily established by use of the Jury-Blanchard criterion (Seborg et al., 1989) as follows. If $K_I > 0$, i.e. if an I-term is actually employed in the regulator, then the stability condition reads $2K_P + K_I < 2(\mu + 1)/\zeta$; this implies that the system can be stabilized even if the P-term in (14) is dropped, i.e. for $K_P = 0$, which would lead to an I-type regulator (as in ALINEA ramp metering, see Papageorgiou et al., 1991) and could render the fine-tuning task easier. If gating is applied directly at the PN border, i.e., if the involved process time-delay is zero, an I-type regulator would perform reasonably well; but with increasing time-delay, i.e. when gating is applied further upstream, the inclusion of a P-term becomes increasingly important for efficient regulation, as with ALINEA ramp metering with a far downstream bottleneck (Wang et al., 2010). On the other hand, if $K_I = 0$, i.e. if only a P-regulator is employed, then the stability condition reads $K_P < (1 - \mu)/\zeta$; one well-known advantage offered by the inclusion of an integral term in the regulator is that the stationary regulator error becomes automatically zero, i.e. $TTS = \widehat{TTS}$ under stationary conditions, as can be readily deduced from (14).

The rigorous proof of stability for the nonlinear system is more cumbersome and is left for a more control-oriented publication; roughly speaking, if TTS is higher than the set-point $\widehat{TTS}$, then both last terms on the right of the regulator Eq. (14) will continuously reduce the inflow q_{in}, such that TTS approaches its set-point; however, this action reaches its limit when q_{in} reaches its lower bound q_{min} mentioned earlier; then, if the sum $q_{min} + q_d$ of controlled and uncontrolled inflows happens to be higher than the TTS-dependent outflow q_{out}, the network cannot be fully protected from over-saturation as Eq. (5) indicates. More specifically, if the imposed q_{min} constraint is "too high" and/or the uncontrolled inflow q_d is "too high", the controller (in fact, any controller) will not be able to maintain TTS close to $\widehat{TTS}$, and hence to maximize the PN throughput. Having said that, the controller may still enable improvements if it can maintain TTS at lower values than without gating. This circumstance sets according limits to the level of the inflow q_d, that is left uncontrolled, in conjunction with the lower admissible bound q_{min} for the controlled inflow.

It is interesting at this point to consider the optimal rule of Daganzo (2007) for saturated network control. Translated in the present notation and context, that rule suggests: At each k, set $q_{in}(k) = 0$ if $TTS(k) > \widehat{TTS}$; else set $q_{in}(k)$ as high as possible, subject to the constraint $TTS(k+1) \leqslant \widehat{TTS}$. The first part of this rule may be readily implemented in practice, using of course a positive lower bound q_{min} for $q_{in}(k)$ instead of zero for obvious reasons; but the literal implementation of the second part of the rule would call for an exact model and uncontrolled inflow q_d information to guarantee that $TTS(k+1)$ will not exceed $\widehat{TTS}$, which is not practicable. One way to render the rule practicable, is to apply an upper bound to $q_{in}(k)$ when $TTS(k) \leqslant \widehat{TTS}$, i.e., overall,

$$q_{in}(k) = \begin{cases} q_{min} & \text{if } TTS(k) > \widehat{TTS} \\ q_{max} & \text{else} \end{cases} \tag{15}$$

This is a bang-bang regulator (like the one deployed in electric irons) which is equivalent to the regulator (14) in terms of set-point and real-time data requirements. In fact, the bang-bang controller can be interpreted as a (constrained) P regulator with very high K_P value. Such a bang-bang regulator would incur a stationary oscillation of $TTS(k)$ around $\widehat{TTS}$, but, given the relatively wide range of throughput-maximizing TTS values (within [600, 800] here), the oscillation may not really affect the resulting efficiency. However, the implied frequent switching of the gated link green phase between a minimum and a maximum value may not be desirable with the drivers and the road authorities. In contrast to (15), the regulator (14) offers a smooth control behavior and $TTS(k) = \widehat{TTS}$ under stationary conditions.

Gating could be activated only within specific time windows (e.g. at the peak periods) or if some real-time measurement-based conditions are satisfied. After distributing the regulator-ordered flow to the gated links, the individual sub-flows must be converted to appropriate green times by modifying the usual traffic signal settings in the corresponding junctions. This was done in the present study simply by modifying the duration of the signal stages where gated inflows are involved; while more elaborated procedures involving changes of the stage structure, e.g. so as to reduce delays for PN exiting flows, are considered in ongoing research.

It should be noted that, under any signal implementation policy and conditions, the total implemented PN inflow may be different than the flow ordered by the regulator for a number of reasons, including limited accuracy of signal specification, low demand, over-spilling downstream link or flow constraints; however, the regulator is largely robust to these potential occurrences thanks to its feedback structure, as it will be demonstrated in the next section.

3. Results

The microscopic simulator AIMSUN is stochastic, thus different simulation runs (replications) with different random seeds may lead to different results. For this reason, it is common to use a number (10 in this work) of replications for each investigated scenario and then calculate the average value of the 10 runs for each evaluation criterion in order to compare different control cases with non-gated cases. Three performance indexes are used here (as provided by AIMSUN): the average vehicle delay per km and the mean speed, both for the entire Chania urban network (not only the PN); and the total number of vehicles that exit the overall network during the whole scenario.

3.1. Non-gating case

Signal control for the non-gating case corresponds to the usual fixed-time settings used in the real Chania network. Table 1 displays the aforementioned indexes for every replication; as well as the average, maximum and minimum values of each index. Since AIMSUN is a stochastic tool, link over-spilling and partial gridlocks may be more or less pronounced in individual replications. In some replications (e.g. Rep. 2 and Rep. 6 in Table 1), the created congestion in the peak period leads to more serious gridlocks in the PN, consequently the delay is higher and the mean speed is lower than average, while the (lower) number of exited vehicles in these replications indicates that the network is not yet empty at the end of the simulation.

To enable an illustrative comparison of the non-gating versus gating cases, the respective detailed results of Replication 1 are displayed in Fig. 5, namely the PN's *TTS* (Fig. 5a and d), the PN inflow (from 8 gated links) q_{in} (Fig. 5b and e), and the PN's *TTD* (Fig. 5c and f). Concentrating on the left column of Fig. 5 (parts (a)–(c)), the 1st hour is characterized by a gradual increase of all three displayed quantities, as it is typical for increasing demand in under-saturated traffic conditions. At time $t = 1$ h, the abrupt increase of q_{in} leads to according increases of *TTS* and *TTD*, the latter reaching soon capacity values according to Fig. 3a, while the former is traversing the aforementioned critical region of [600, 800]. However, as q_{in} continues to be high, *TTS* continues to increase to very high values (i.e. the PN becomes increasingly crowded with vehicles); as a conse-

Table 1
Non-gating performance indexes for each replication.

	Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Rep. 6	Rep. 7	Rep. 8	Rep. 9	Rep. 10	Ave.	Max	Min
Delay (s)	513	722	362	360	431	674	467	372	350	424	467	722	350
Speed (km/h)	6.62	5.05	8.91	8.88	7.81	5.55	7.25	8.84	9.06	7.83	7.58	9.06	5.05
Vehicles out	14,545	12,997	14,575	14,593	14,771	11,092	14,684	14,719	14,338	14,844	14,115	14,844	11,092

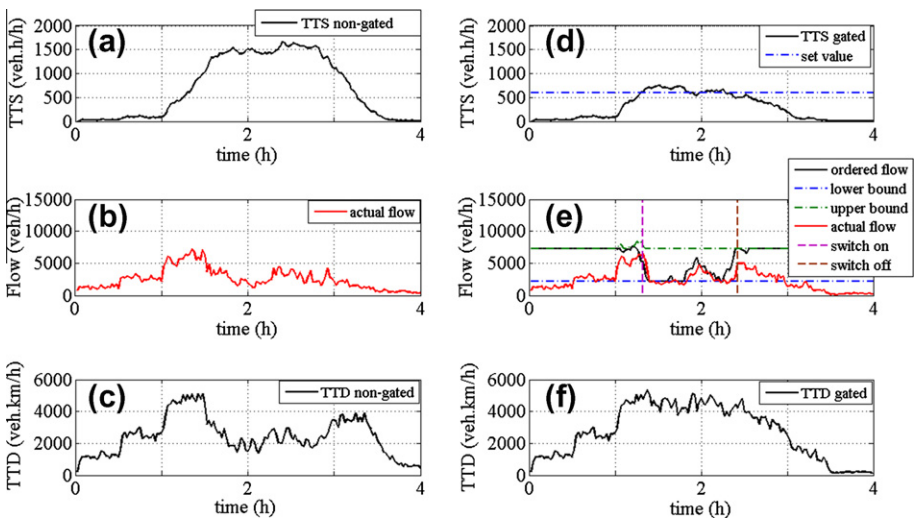

Fig. 5. (a) PN's *TTS* versus time in non-gating case; (b) actual PN inflow versus time for the non-gating case; (c) PN's *TTD* versus time for the non-gating case; (d) PN's *TTS* versus time for the gating case; (e) ordered and actual PN inflow versus time for the gating case; (f) PN's *TTD* versus time for the gated case.

quence, link over-spilling and gridlock lead to a sensible *TTD* reduction to low levels, that are persisting until $t = 3$ h, when the network starts de-congesting due to low demand. Remarkably, during the congested period $t \in [1.5\text{ h}, 3\text{ h}]$, the inflow q_{in} is also reduced due to over-spilling links of PN, i.e. as a result of the congestion that extends beyond the PN; but this reduction of the inflow comes too late, too little to reverse the already advanced PN degradation.

3.2. Gating case

In the Chania gating example, the time delay τ is zero, hence the K_P and K_I values may be specified for the controlled system (13), (14) to exhibit a time-optimal dead-beat regulator behavior, i.e. so as to reach the set-point within one single step, see also Papageorgiou et al. (1991) for a similar feedback control design by use of the system's z-transforms. A dead-beat behavior is established for set-point step-changes with $K_P = \mu/\zeta$ and $K_I = (1-\mu)/\zeta$; and for disturbance q_d step-changes with the same K_P but $K_I = 1/\zeta$. Thus, a least-squares parameter estimation was first conducted for μ and ζ in (13), using time-series of (q_{in}, *TTS*)-measurements within and around the critical *TTS*-range of [600, 800]. Once the parameters μ and ζ have been specified ($\mu = 0.807$ and $\zeta = 0.038$), the regulator parameters for (set-point) dead-beat behavior were calculated to be $K_P = 20\ \text{h}^{-1}$ and $K_I = 5\ \text{h}^{-1}$. Note anyhow that the feedback controller is quite robust to parameter variations, as it can be verified from the stability and optimality conditions provided.

The regulator's maximum and minimum bounds are visible in Fig. 5e. The regulator runs continuously in the background, but gating is actually activated only when *TTS* exceeds a threshold, that is selected slightly lower than (in this case 85% of) the set point; and is de-activated when *TTS* falls below a 2nd, slightly lower threshold. At all other times, fixed-time signal control is applied, as in the non-gating case. A set point of $\widehat{TTS} = 600$ veh h per h is selected for the gating operation.

By running AIMSUN with the control law (14) for the gated traffic signals, the results displayed in Table 2 are obtained for each of the 10 replications. The improvements, compared to Table 1, are significant, and, in fact, even the worse gating replication is superior to the best non-gating replication. An average speed increase of 40% results for the whole network (not just the protected part thereof).

Fig. 5d–f displays the detailed results of Replication 1 of the gating case and illustrate its way of functioning and impact. Traffic conditions are *identical* as in the non-gating case up to around $t = 1.2$ h, when gating is switched on (Fig. 5e), as *TTS* approaches its set value; the gating regulator orders low inflow values to maintain *TTS* around its set point, and, as a consequence, *TTD* is maintained at high levels, in clear contrast to the non-gating case. As mentioned earlier, the gating action creates temporary queues (and corresponding temporary vehicle delays) at the gated links; however, this proves highly beneficial for the PN throughput, and, as a consequence, the overall network delay (including the gating queue regions) is strongly reduced, as Table 2 indicates; in fact, even gated vehicles may have a net gain, as their temporary gating delays may be more than offset by enabled higher speeds once they enter the PN.

It is visible in Fig. 5e that the q_{in} values ordered by the regulator, differ from the implemented ones for various reasons mentioned earlier, but this has a minor impact on the regulator's efficiency, as expected. At $t = 2.3$ h, *TTS* is moving to lower values, gating is switched off, and traffic flow returns to under-saturated conditions; in contrast to the non-gating case where over-saturated conditions are seen in Fig. 5 to persist for 1 h longer.

Fig. 6 displays the simulation results obtained for Replication 1 using the bang-bang controller (15). The resulting delay is 341 s, which is higher than for any replication with the PI controller (14) (Table 2), but lower than for any replication without gating (Table 1). Fig. 6a indicates that the bang-bang controller maintains the corresponding *TTS* near the set-point; while Fig. 6b displays the expected bursty behavior of the bang-bang controller with regard to the controlled inflow.

Fig. 7 displays the simulation results obtained for Replication 1 using the PI controller (14), but now with an increased lower bound of $q_{min} = 4000$ veh/h (in place of 2180 veh/h used before). The regulator is seen to saturate at the new lower bound (i.e. it applies the maximum admissible gating action) for most of the active gating period. Fig. 7 indicates that the actual inflow deviates slightly, but increasingly with time, from the ordered flow q_{min}, which is due to long downstream link queues at a couple of gating locations. Overall, the constrained gating action is now not enough to maintain *TTS* close to its set-point because q_{min} is too high. Thus, except for two tiny periods of time at the start and the end, respectively, of the active gating period, *TTS* is essentially uncontrolled. Nevertheless, this constrained gating action is sufficient to enable lower *TTS*-values than without gating; in fact, *TTS* reaches a maximum of slightly above 1000 veh, as opposed to 1600 veh in the corresponding no-control case. The resulting delay is 330 s, which is higher than for any replication with lower q_{min} (Table 2), but lower than for any replication without gating (Table 1).

Table 2
Performance indexes using the proposed gating control strategy.

	Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Rep. 6	Rep. 7	Rep. 8	Rep. 9	Rep. 10	Ave.	Max	Min
Delay (s)	293	287	310	298	317	310	328	314	299	292	304	328	287
Upgrade (%)	42.8	60.2	14.4	17.2	26.5	54.0	29.8	15.6	14.6	31.1	35.0	54.0	14.4
Speed (km/h)	10.3	10.6	9.9	10.4	10.0	10.0	9.4	9.9	10.3	10.4	10.1	10.6	9.4
Upgrade (%)	55.6	110	11.1	17.1	28.0	80.2	29.7	12.0	13.7	32.8	39.2	110	11.1
Vehicles out	14,721	14,521	14,623	14,516	14,508	14,569	14,783	14,632	14,441	14,515	14,582	14,783	14,441

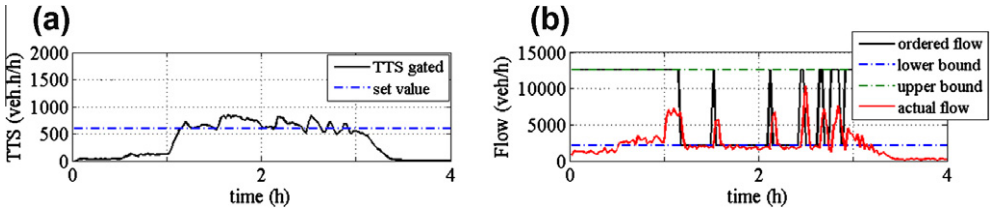

Fig. 6. Bang-bang gating results: (a) *TTS* versus time and (b) ordered and actual PN inflow versus time.

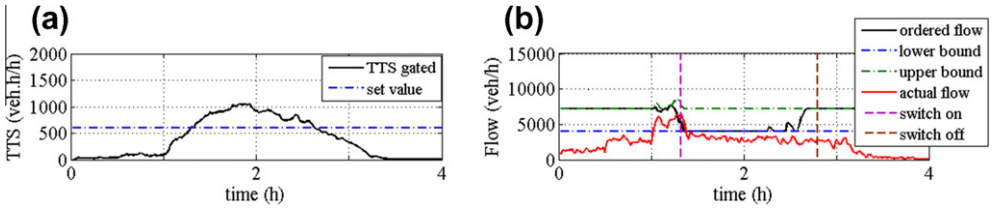

Fig. 7. Higher lower bound results: (a) *TTS* versus time and (b) ordered and actual PN inflow versus time.

4. Conclusion

Gating aims at protecting urban road networks from over-saturation, or, more specifically, at maximizing the network throughput. Based on the previously developed concept of a network fundamental diagram (NFD), an *operational* urban NFD has been defined to enable simple, practicable and efficient gating control, potentially even by use of a very limited amount of real-time measurements. A simple (nonlinear and linearized) control design model, incorporating the operational NFD, has been developed, which allows for the gating problem to be cast in a proper feedback control design setting. This allows for application and comparison of a variety of linear or nonlinear, feedback or predictive (e.g. Smith predictor, internal model control and other) control design methods from the Control Engineering arsenal; among them, a simple but efficient PI feedback regulator was developed and successfully tested in a fairly realistic microscopic simulation environment. More specifically, the Chania urban road network was modeled in the microscopic simulator AIMSUN as a test-bed for this research, to protect its most sensitive part from spillovers, gridlock, and the resulting strong degradation. Application of the developed gating strategy is demonstrated to lead to significant improvements (in the order of 35% of delay reduction) compared to the non-gating control case.

The protected network in this investigation was selected in an ad hoc way, based on related experience with the real traffic conditions. Further investigations and, hopefully, field implementations, with different network types and sizes, as well as different demand and congestion patterns and different gating locations may shed more light on the most beneficial practices to be applied.

It must be stressed that gating can only be successful in reducing the overall network delay if a couple of conditions are actually met in the network to be protected from oversaturation. Firstly, a congestion-caused degradation, i.e. a reduction of throughput (e.g. due to link spillover and gridlock), must actually occur without gating, else there would be no potential for improvement. Secondly, the occurring degradation must be (at least partly) reversible if the number of vehicles in the network is maintained at a certain optimal level; in other words, the targeted high efficiency and throughput must be sustainable, not merely transient phenomena.

In this research, the measurements of all links within the protected network are fed to the regulator; on-going research reveals that this is not necessary, and that gating may be applied similarly efficiently with far less real-time measurements. However, this may open the question on the sort of measurements that are most beneficial for gating and hence contradict to some extent the usage of the NFD, which, by definition, reflects the traffic conditions on a whole network, not only on selected parts thereof.

Further research directions in this area include controller design in presence of time-delay (due to gated links that are remote from the protected network), queue balancing and management at the gating positions, comparison with more comprehensive traffic-responsive signal control strategies, and, most importantly, field implementation and evaluation of the proposed gating strategy.

Acknowledgment

The research leading to these results has been funded by the European Commission FP7 program NEARCTIS (Network of Excellence for Advanced Road Cooperative Traffic management in the Information Society).

References

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

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

Beard, C., Ziliaskopoulos, A., 2006. A system optimal signal optimization formulation. In: 85th TRB Annual Meeting, Washington, DC, USA.

Bretherton, D., Bowen, G., Wood, K., 2003. Effective urban traffic management and control: recent developments in SCOOT. In: 82nd TRB Annual Meeting, Washington, DC, USA.

Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transportation Research Part B 41 (1), 49–62.

Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for macroscopic fundamental diagram of urban traffic. Transportation Research Part B 42 (9), 771–781.

De Schutter, B., De Moor, B., 1998. Optimal traffic light control for a single intersection. European Journal of Control 4 (3), 260–276.

Diakaki, C., Papageorgiou, M., Aboudolas, K., 2002. A multivariable regulator approach to traffic-responsive network-wide signal control. Control Engineering Practice 10 (2), 183–195.

Dinopoulou, V., Diakaki, C., Papageorgiou, M., 2005. Application and evaluation of the signal traffic control strategy TUC in Chania. Journal of Intelligent Transportation Systems 9 (3), 133–143.

Farges, J.L., Henrry, J.J., Tufal, J., 1983. The PRODYN real-time traffic algorithm. In: 4th IFAC Symposium on Transportation Systems. Baden-Baden, Germany, pp. 307–312.

Farhi, N., 2008. Modélisation minplus et commande du trafic de villes régulières, Ph.D. Thesis. Université de Paris I-Panthéon-Sorbonne, Paris, France.

Gartner, N.H., 1983. OPAC: a demand-responsive strategy for traffic signal control. Transportation Research Record 906, 75–84.

Gartner, N.H., Wagner, P., 2004. Analysis of traffic flow characteristics on signalized arterials. Transportation Research Record 1883, 94–100.

Geroliminis, N., Daganzo, C.F., 2008. Existance of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B 42 (9), 759–770.

Geroliminis, N., Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transportation Research Part B 45 (3), 605–617.

Godfrey, J.W., 1969. The mechanism of a road network. Traffic Engineering and Control 11 (7), 323–327.

Haddad, J., Ramezani, M., Geroliminis, N., 2012. Model predictive perimeter control for two-region urban cities. In: 91th TRB Annual Meeting, Washington, DC, USA.

Helbing, D., 2009. Derivation of a fundamental diagram for urban traffic flow. The European Physical Journal B 70 (2), 229–241.

Hunt, P.B., Robertson, D.I., Bretherton, R.D., Royle, M.C., 1982. The SCOOT on-line traffic signal optimization technique. Traffic Engineering and Control 23 (4), 190–192.

Ji, Y., Daamen, W., Hoogendoorn, S., Hoogendoorn-Laser, S., Qian, X., 2010. Investigating the shape of the macroscopic fundamental diagram using simulation data. Transportation Research Record 2161, 40–48.

Lieberman, E., Chang, J., Bertoli, B., Xin, W., 2010. New signal control optimization policy for oversaturated arterial systems. In: 89th TRB Annual Meeting, Washington, DC, USA.

Lo, H.K., 1999. A novel traffic signal control formulation. Transportation Research Part A 33 (6), 433–448.

Lo, H.K., Chang, E., Chan, Y.C., 2001. Dynamic network traffic control. Transportation Research Part A 35 (8), 721–744.

Lowrie, P.R., 1982. SCATS: the Sydney co-ordinated adaptive traffic system-principles, methodology, algorithms. In: IEEE International Conference on Road Traffic Signalling, London, England, pp. 67–70.

Luk, J., Green, D., 2010. Balancing Traffic Density in a Signalized Network. Austroads Research Report AP-R369/10, Sydney, Australia.

Mirchandani, P., Head, L., 1998. RHODES – a real-time traffic signal control system: architecture, algorithm. In: TRISTAN III (Triennial Symposium on Transportation Analysis), vol. 2, San Juan, Puerto Rico.

Papageorgiou, M., Hadj-Salem, H., Blosseville, J.-M., 1991. ALINEA: a local feedback control law for on-ramp metering. Transportation Research Record 1320, 58–64.

Papageorgiou, M., Vigos, G., 2008. Relating time-occupancy measurements to space-occupancy and link vehicle-count. Transportation Research Part C 16 (1), 1–17.

Papamichail, I., Papageorgiou, M., 2011. Balancing of queues or waiting times on metered dual-branch on-ramps. IEEE Transactions on Intelligent Transportation Systems 12 (2), 438–452.

Putha, R., Quadrifoglio, L., Zechman, E., 2010. Using ant optimization for solving traffic signal coordination in oversaturated networks. In: 89th TRB Annual Meeting, Washington, DC, USA.

Strating, M., 2010. Coordinated Signal Control for Urban Networks by Using MFD, M.Sc. Thesis. Delft University of Technology, Delft, The Netherlands.

Seborg, D., Edgar, T.F., Mellichamp, D.A., 1989. Process Dynamics and Control. Wiley, New York.

TSS Transport Simulation Systems, 2008. AIMSUN User Manual Version 6, Barcelona, Spain.

Wang, Y., Papageorgiou, M., Gaffney, J., Papamichail, I., Rose, G., Young, W., 2010. Local ramp metering in random-location bottlenecks downstream of a metered on-ramp. Transportation Research Record 2178, 90–100.

Wood, K., Bretherton, D., Maxwell, A., Smith, K., Bowen, G., 2002. Improved Traffic Management and Bus Priority with SCOOT, TRL Staff Paper PA 3860/02. Transport Research Laboratory, London, UK.