Manuscript File Click here to view linked References

Link transmission model: a formulation with enhanced compute time for large-scale network optimization

Jing Lu[a] and Carolina Osorio[b]

[a]*Operations Research Center, Massachusetts Institute of Technology, USA*
Email: jl3724@alum.mit.edu
[b]*Department of Decision Sciences, HEC Montréal, Canada*
Email: carolina.osorio@hec.ca

July 22, 2023

Abstract

We formulate a traffic theoretic and probabilistic analytical link transmission model. The proposed model extends past work that is based on a stochastic formulation of the link transmission model, which itself is an operational formulation of Newell's simplified theory of kinematic waves. The proposed model yields a probabilistic description of the link's upstream and downstream boundary conditions. The model only tracks the transient probabilities of two of the link's boundary states. This leads to a model with a state space dimension that is constant, i.e., it does not depend on any link attributes, such as link length. In other words, the model has constant complexity, whereas past formulations have a complexity that scales linearly or cubically with link length. The gain in computational runtime is of at least one order of magnitude and it increases with link length. This makes the proposed model suitable for large-scale network optimization. The model is validated versus a simulation-based implementation of the stochastic link transmission model. Its performance is also benchmarked with other past analytical formulations. The proposed model yields estimates with comparable accuracy, while the computational efficiency is enhanced by at least one order of magnitude. It is also validated versus a microscopic traffic simulator, the results indicate that the proposed model accurately approximates the link's boundary conditions for realistic traffic situations, such as signalized links and platoon arrival patterns. The model is then used to address a city-wide traffic signal control problem. The performance of the proposed model is benchmarked versus various other models and traffic signal control approaches. It is shown to reduce optimization compute time by at least one order of magnitude, while also yielding solutions (i.e., signal plans) with improved performance.

1 Introduction

In the field of traffic flow modeling, there is recent and increased interest in the formulation of probabilistic models. This is facilitated and motivated by a number of factors, including increased availability of urban mobility data, advanced sensor technologies that enable increased data granularity

(i.e., resolution) such that more detailed models can be calibrated and validated, enhanced computing capabilities such that more elaborate models can be evaluated. Additionally, transportation agencies in the US and in Europe have recognized both the importance and the need to evaluate and to improve network robustness and reliability metrics (U.S. Department of Transportation; 2008; Transport for London; 2010). This calls for a probabilistic description of network performance.

Calvert et al. (2012) discuss the advantages and disadvantages of both deterministic and stochastic modeling approaches from both methodological and transportation practice perspectives. They identify the lack of computational efficiency as one of the main challenges current stochastic models face. Indeed, compared to their deterministic counterparts, stochastic models may suffer from the curse of dimensionality and are often computationally inefficient for the analysis, let alone the optimization, of large-scale networks. The goal of this paper is to propose an analytical stochastic traffic theoretic model that addresses these scalability and computational efficiency concerns.

Deterministic traffic model formulations and their solution methods have been extensively studied leading to seminal works such as Lighthill and Whitham (1955); Richards (1956); Daganzo (1994); Newell (2002).The formulation of their stochastic counterparts are in the early stages. Detailed reviews of stochastic traffic flow models, analytical or simulation-based, are provided by Sumalee et al. (2011); Jabari (2012); Calvert et al. (2012); Laval and Chilukuri (2014) and Chen et al. (2015). This paper focuses on analytical (i.e., not simulation-based) formulations. In this research area, recent work has proposed formulations based on the variational theory of Daganzo (2005) (Deng et al.; 2013; Laval and Chilukuri; 2014; Laval and Castrillón; 2015). The most popular approach to formulate a stochastic traffic model is to add stochasticity to a specific deterministic traffic flow model. For instance, Boel and Mihaylova (2006) formulate a stochastic cell-transmission model (CTM) (Daganzo; 1994) by adding Gaussian noise terms to the sending and receiving functions of the deterministic CTM. However, for such approaches, the expected traffic dynamics are not guaranteed to be consistent with their deterministic CTM counterparts. A detailed discussion of this, including the existence and implications of negative sample paths, are given in Jabari and Liu (2012). Formulations that add Gaussian noise terms to a specific deterministic model can become computationally inefficient due to the need to sample from high-dimensional Gaussian distributions. Rather than adding noise directly to the speed-density relationship, Jabari and Liu (2012) consider stochastic vehicle headways and Jabari et al. (2014) consider a stochastic formulation of Newell's simplified car-following model (Newell; 2002). Probabilistic assumptions are made at the microscopic level, and macroscopic probabilistic speed-density relationships are then derived. A probabilistic formulation of the deterministic model of Newell (1961) is proposed in Zheng et al. (2018). Like the work of Jabari et al. (2014), the formulations of Sumalee et al. (2011); Zheng et al. (2018) are based on assumptions of parametric uncertainty. Such formulations are often combined with uncertainty propagation methods as in Sayegh et al. (2017). Another common approach to stochastic modeling are gas-kinetic, Boltzmann-like, formulations (e.g., Paveri-Fontana (1975); Treiber et al. (1999); Hoogendoorn and Bovy (2001)). However, the challenges of embedding realistic stochastic dependency assumptions within kinetic models has, so far, precluded their use for large-scale networks (Helbing; 2001).

An alternative approach has been the use of probabilistic queueing theory. Most work has considered a stationary analysis (Heidemann; 1991, 1994; Heidemann and Wegmann; 1997). Work that considers transient (i.e., non-stationary) analysis includes Olszewski (1994); Heidemann (2001); van Zuylen and Viti (2003); Viti and Van Zuylen (2010). Most work on the transient analysis of queueing systems

assumes infinity space capacity. Note that in queueing theory, the term capacity refers to space capacity. In this paper, unless stated otherwise the term capacity also refers to space capacity, rather than flow capacity. In reality, there is physical upper bound on the number of vehicles a road can hold, i.e., queueing capacity is finite. Finite capacity can lead to vehicular spillbacks, which are frequently observed in congested urban networks. Formulations based on both transient queueing theory and finite (space) capacity queueing network theory have also been proposed (Osorio et al.; 2011; Osorio and Flötteröd; 2015; Lu and Osorio; 2018). Transient queueing models can contribute to provide a probabilistic, analytical and time-dependent description of congestion build-up and dissipation. However, the formulation of a probabilistic transient model with sufficient computational efficiency to enable large-scale network analysis and optimization remains a challenge (Viti and Van Zuylen; 2010).

This paper tackles this challenge and extends this literature. The model of Osorio and Flötteröd (2015) is a stochastic formulation of the deterministic link transmission model of Yperman et al. (2007), which itself is an operational formulation of Newell's simplified theory of kinematic waves (Newell; 1993). The model considers a single link with space capacity ℓ and represents the link as a set of three queues with finite (space) capacity. It derives the joint transient probability distribution of the link's upstream and downstream boundary conditions. For a link with space capacity ℓ, the model complexity is in the order of $\mathcal{O}(\ell^3)$. The recent work of Lu and Osorio (2018) extends the model of Osorio and Flötteröd (2015) by making it more computationally efficient. Instead of deriving the joint distribution of the link's upstream and downstream boundary conditions, Lu and Osorio (2018) yield the marginal distribution of the link's upstream boundary conditions and the marginal distribution of the link's downstream boundary conditions. They provide a simplified description of the spatial and temporal dependencies between the upstream and the downstream boundary conditions. The model complexity is in the order of $\mathcal{O}(\ell)$. This reduction in model complexity enhances the computational efficiency of the model. In this paper, we formulate a model with further enhanced computational efficiency. The goal is to enable large-scale network optimization to be performed efficiently. We extend the model of Lu and Osorio (2018) and propose a formulation with constant complexity, i.e., the complexity no longer depends on the link's space capacity ℓ.

The paper is organized as follows. In Section 2, we motivate and formulate the proposed model. The proposed model is validated in Section 3. It is then used to address a city-wide signal control problem and is benchmarked versus other methods (Section 4). Section 5 summarizes the paper and discusses ongoing work. Appendices A and B provide detailed derivations of results presented in Section 2.2.1, and Appendices C, D and E provide detailed derivations of results presented in Section 2.2.2. Additional numerical results of the experiments of Section 3.1 are given in Appendix F. Appendix G discusses the congestion level of the Lausanne network in Section 4.

2 Link model formulation

In this section, we present the proposed model formulation. We first briefly discuss the past link model formulation which the proposed model is built upon (Section 2.1). We then present the proposed model formulation in Section 2.2. The proposed formulation contains two key parts: downstream boundary conditions discussed in Section 2.2.1 and upstream boundary conditions discussed in Section 2.2.2. Section 2.2.3 summarizes the proposed algorithm.

2.1 Past link model formulations

Let us first outline the main ideas of the models of Lu and Osorio (2018) (hereafter referred to as the mixture model) and of Osorio and Flötteröd (2015) (hereafter referred to as the multivariate model). These link models are stochastic counterparts of the deterministic link transmission model of Yperman et al. (2007) that assumes a triangular fundamental diagram. Hence, there are only two possible characteristic kinematic wave speeds (a forward one and a backward one). A detailed description of this is given in Section 4.5 of Yperman (2007), summary descriptions appear in Section 2.1 of Himpe et al. (2016) or in Section 2 of Osorio and Flötteröd (2015).

Consider an isolated link with a triangular fundamental diagram, free flow speed v, backward wave speed w, flow capacity $\hat{q}$, link length L and jam density $\hat{\rho}$. The models are continuous-space discrete-time models, where k denotes the discrete time interval index. They consider the stochasticity in the link's arrival and departure processes as follows. The arrival process at the upstream boundary of the link is an inhomogeneous Poisson process with rate $\lambda(k)$. The service times at the downstream end of the link are independent and identically distributed exponential random variables with rate $\mu(k)$. The process that vehicular flow experiences through the link is illustrated in Figure 1. During time interval k, the link has an expected inflow (resp. outflow) of $q^{in}(k)$ (resp. $q^{out}(k)$). The proposed model, just like the models it builds upon, is a probabilistic macroscopic model. Hence, it tracks the dynamics of vehicular flows, and does not track individual vehicles. These models use queueing-theoretic tools to describe the dynamics of the link's boundary conditions through stochastic (queueing) processes. Hence, the term queue, unless stated otherwise, does not refer to a vehicular queue, instead it refers to a probabilistic queueing system. Upon entrance to the link, the vehicular inflow experiences a delay of $L/(v\delta)$ time intervals before it becomes ready to leave the link at the downstream end of the link. This delay incurred upon entrance is represented by the lagged inflow queueing system, referred to as the *lagged inflow queue* LI. This delay process can be viewed as if the flow traveled sequentially through a set of $k^{fwd} = \lceil \frac{L}{v\delta} \rceil$ cells. In other words, LI is represented by a set of k^{fwd} discrete cells. We denote the most downstream cell of LI as LLI in Figure 1. After this delay, the vehicular flow is ready to depart the link. This flow is represented by the downstream queueing system, referred to as the *downstream queue* DQ. Upon departure from the link, the newly available space is delayed for $L/(|w|\delta)$ time intervals before it becomes available at the upstream end of the link. This space that is not yet available at the upstream end of the link is represented by the lagged outflow queueing system, referred to as the *lagged outflow queue* LO. In other words, vehicular departures from the link lead to two events: (i) vehicular flow departs DQ and (ii) "newly available space" flow enters LO. Similar to LI, LO can be viewed as a set of $k^{bwd} = \lceil \frac{L}{|w|\delta} \rceil$ discrete cells that are traversed sequentially. In Figure 1, the most downstream cell of LO is denoted LLO. The upstream boundary conditions of the link are described through the upstream queueing system, referred to as the *upstream queue* UQ, which is defined as:

$$UQ(k) = LI(k) + DQ(k) + LO(k), \tag{1}$$

where $UQ(k), LI(k), DQ(k)$ and $LO(k)$ are random variables representing the queue length of the corresponding queueing system at time interval k. In other words, $UQ(k)$ accounts for all vehicular flow in the link (i.e., $LI(k) + DQ(k)$) as well as all "newly available spaces" that are not yet available at the upstream end of the link (i.e., $LO(k)$). Equivalently, for a link with space capacity ℓ (which

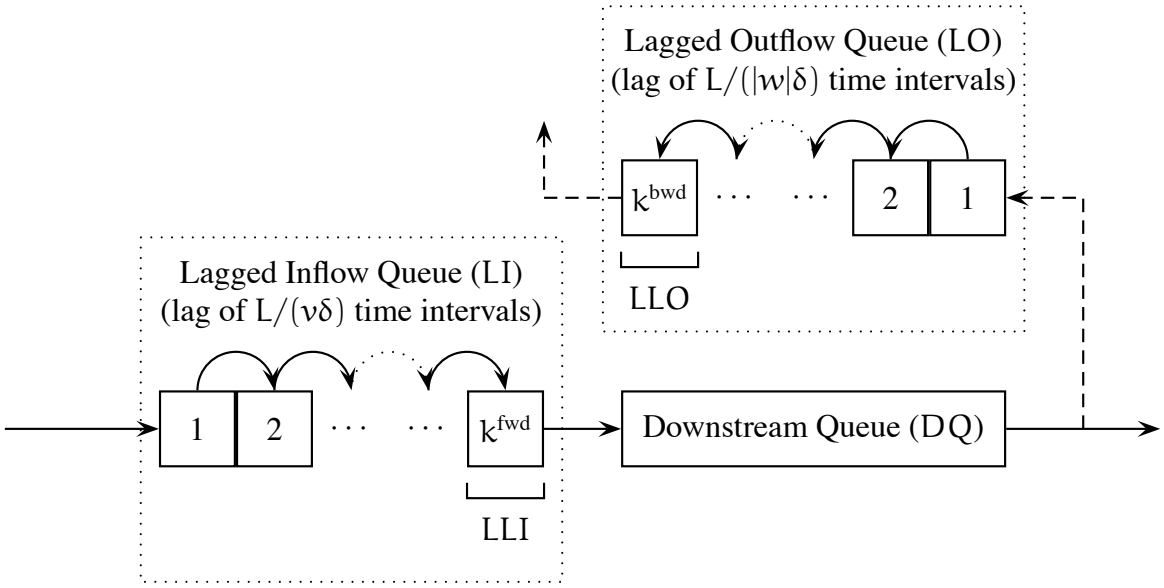


Figure 1: Link dynamics from Lu and Osorio (2018)

is defined as the rounded version of $\hat{\rho}L$), $\ell - UQ(k)$ represents the vehicular space available at the upstream end of the link.

All queues (LI, DQ, LO and UQ) are time-dependent stochastic processes. Thus, the flows on the link are stochastic and so are the cumulative flows both upstream and downstream of the link. Thus, the link models provide a stochastic description of the link's upstream and downstream boundary conditions.

In summary, the overall within-link dynamics are described by four queues (LI, DQ, LO, UQ), which are linearly constrained through Eq. (1). Thus, we need only to track three of these four queues. The model of Osorio and Flötteröd (2015) tracks three queues and yields the joint distribution of (LI, DQ, LO). For a given link with space capacity ℓ, each queue has a state space defined as $\{0, 1, ..., \ell\}$. Thus, the dimension of the state space of this joint distribution (LI, DQ, LO) is in the order of $\mathcal{O}(\ell^3)$.

Lu and Osorio (2018) note that the link's upstream (resp. downstream) boundary conditions are described by UQ (resp. DQ). The model is formulated as a mixture of two independent univariate models: a univariate model of UQ and a univariate model of DQ. The model tracks the full marginal distributions of UQ and of DQ, over time. The dimension of the state space for the model is $2(\ell+1)$, i.e., the model complexity is in the order of $\mathcal{O}(\ell)$. In other words, Lu and Osorio (2018) enhance the scalability of the multivariate model of Osorio and Flötteröd (2015) by formulating a model with linear, rather than cubic, complexity in the link's space capacity ℓ.

In this paper, we propose a formulation with a state space of dimension two. In other words, the model complexity is now independent of the link's space capacity ℓ. This leads to enhanced scalability and improves the ability of these models to be used efficiently for large-scale network optimization. This proposed formulation is simpler than past formulations, yet as illustrated in Section 3, it still captures sufficient dependency between the link's upstream and downstream boundary conditions.

2.2 Proposed link model formulation

The main idea underlying the proposed model is that in order to describe the link's boundary conditions, we do not need to track the full marginal distributions of UQ and of DQ as in Lu and Osorio (2018), let alone track their full joint distribution as in Osorio and Flötteröd (2015). More specifically, we have identified two specific queue states that are essential to describe these boundary conditions. The first state is $DQ(k) = 0$, which describes whether or not there is vehicular flow downstream ready to depart the link. The second state is $UQ(k) = \ell$, which describes whether or not there is road space available at the upstream end of the link. Intuitively, in a network setting with two links, vehicular flow can be transmitted from the upstream link to the downstream link if the following two conditions hold: (i) there is flow at the upstream link ready to depart to the downstream link (i.e., for the upstream link $DQ(k) > 0$), and (ii) there is space available at the upstream end of the downstream link (i.e., for the downstream link $UQ(k) < \ell$). Thus, for a given link, the proposed model approximates two state probabilities: $P(DQ(k) = 0)$ and $P(UQ(k) = \ell)$.

More formally, for a given time interval k, the expected link inflow is estimated as:

$$q^{in}(k) = \lambda(k)(1 - P(UQ(k) = \ell)). \tag{2}$$

Equation (2) states that vehicles can enter the link as long as there is space available at the upstream end of the link (i.e., $UQ(k) < \ell$), which happens with probability $P(UQ(k) < \ell) = 1 - P(UQ(k) = \ell)$. This equation is the stochastic counterpart of the receiving function of Yperman et al. (2007), see Section 2 of Osorio and Flötteröd (2015) for detailed derivations. More specifically, this equation can be thought of as the expected value of the receiving function. Similarly, the expected link outflow is estimated as:

$$q^{out}(k) = \mu(k)(1 - P(DQ(k) = 0)). \tag{3}$$

Equation (3) states that there are vehicle departures from the link as long as there are vehicles at the downstream end of the link that are ready for departure (i.e., $DQ(k) > 0$), which happens with probability $P(DQ(k) > 0) = 1 - P(DQ(k) = 0)$. Equation (3) is the stochastic counterpart of the sending function of Yperman et al. (2007), see Section 2 of Osorio and Flötteröd (2015) for detailed derivations, and can be interpreted as the expected value of the sending function.

The mixture model of Lu and Osorio (2018) derives the marginal distributions of $UQ(k)$ and of $DQ(k)$ at every time step k. However, the only information needed to compute the dynamics of the link's boundary conditions are the two probabilities $P(UQ(k) = \ell)$ and $P(DQ(k) = 0)$.

In this paper, we propose a model that only keeps track of these two key probabilities over time (i.e., $P(UQ(k) = \ell)$ and $P(DQ(k) = 0)$). It improves model scalability by reducing the dimension of the state space. The proposed model has a state space of dimension two. In other words, its complexity is now constant and no longer depends on the space capacity of the link. Thus, in a network setting, the proposed model linearly scales with the number of links in the network, independently of link attributes such as link lengths. The rest of this section is organized as follows. Section 2.2.1 formulates the model of the link's downstream boundary conditions $P(DQ(k) = 0)$. Section 2.2.2 formulates the model of the link's upstream boundary conditions $P(UQ(k) = \ell)$. Section 2.2.3 summarizes the algorithm for the proposed link model.

2.2.1 Downstream boundary conditions

This section formulates the probabilistic model of the link's downstream boundary condition $P(DQ(k) = 0)$. Recall from Section 2.1 that the service process of the link consists of service of vehicles in DQ, and these service times are independent and identically distributed (i.i.d) exponential random variables. Since the service process of the link is the same as the service process of DQ, we do not need to approximate the service process of DQ. In other words, service times of DQ are i.i.d exponential random variables with exogenous rate $\mu(k)$. Thus, we only need to approximate the arrival process of DQ.

As described in Section 2.1, upon entering the link, all vehicles enter LI, and then enter DQ. Thus, arrivals from DQ consist of departures from LI. No arrivals to DQ should be rejected, lost or blocked due to a space capacity limit of DQ. This leads us to model the downstream boundary condition using expressions from an infinite (space) capacity queueing system (i.e., $M/M/1$), rather than a finite (space) capacity queueing system (also referred to as a loss system) during time interval k. The arrival process of DQ during time interval k is approximated as a Poisson process with endogenous rate $\lambda_{DQ}(k)$. To approximate $\lambda_{DQ}(k)$, recall from Section 2.1 that the arrivals to DQ correspond to flow that leaves the last cell of LI. In Figure 1, this cell is the k^{fwd}th cell, which is denoted LLI. Thus, flow that enters DQ during time interval k corresponds to flow that entered the link during time interval $k - k^{fwd}$. Thus, we approximate the arrival rate to DQ during time interval k as the expected flow to enter the link during time interval $k - k^{fwd}$:

$$\lambda_{DQ}(k) = q^{in}(k - k^{fwd}). \tag{4}$$

For an $M/M/1$ queueing system, an exact closed-form expression for the transient queue-length distribution exists (e.g., Eq. (2.163) of Kleinrock (1975)), however the use of such an expression requires keeping track of the entire queue-length distribution at every time step. Since our aim is to track a single state probability, $P(DQ(k) = 0)$, rather than the full distribution, we do not use the closed-form expression.

The transient behavior of a system has also been studied under the name of relaxation processes, which describe the return of a disturbed system to its equilibrium state. Each relaxation process can be characterized by a relaxation time, which quite generally corresponds to the time taken by a disturbed system to reach equilibrium. For instance, in condensed matter physics, relaxation methods are used to study the process by which molecules in high energy states caused by external perturbation return to the Maxwell-Boltzmann distribution, and the relaxation time is then associated the average time between interactions. This idea has been extensively used to study gas kinetics. The work of Prigogine and Andrews (1960) is among the pioneering work that introduced gas-kinetic models for the analysis of vehicular traffic flow. Extensions of their work include Herman and Lam (1971); Paveri-Fontana (1975); Nelson (1995); Helbing (1997).

In the queueing theory literature, the theoretical study of relaxation times is well studied. Our study differs from past literature in the following ways. We focus on the approximation of the relaxation time (or equivalently its inverse) of a single probability state (i.e., $P(DQ(k) = 0)$ of Eq. (5)), while most papers have studied the relaxation time of the expected queue-length for infinite (space) capacity single-server queueing systems, such as in Newell (1982, Chap. 5) and in Odoni and Roth (1983). Our proposed approach allows for an arbitrary initial state for the queueing system, while most theoretical

relaxation time studies focus on initially empty systems (e.g., Newell (1982, Chapter 5)). There are studies, based on numerical stochastic simulations, that have considered arbitrary deterministic initial states, such as in Odoni and Roth (1983). Moreover, most studies consider an isolated queueing system, whereas our work considers several queueing systems that have coupled dynamics. For instance, the dynamics of DQ and of UQ are coupled due to the dependencies between the link's downstream and upstream boundary conditions.

We introduce the following notation:

$P(DQ(k) = 0)$	probability of $DQ(k)$ being empty at the end of time interval k (which is also the beginning of time interval $k+1$);
$P(DQ_k = 0)$	stationary probability of $DQ = 0$ for time interval k;
$\tau_{DQ}(k)$	inverse of the relaxation time of DQ during time interval k.

We propose the following formulation:

$$P(DQ(k) = 0) = P(DQ_k = 0) + \Big[P(DQ(k-1) = 0) - P(DQ_k = 0)\Big] e^{-\tau_{DQ}(k)\delta}. \tag{5}$$

Equation (5) states that the transient probability $P(DQ(k) = 0)$ at the end of time interval k is approximated as the sum of a stationary probability (term $P(DQ_k = 0)$) and a term that decays exponentially with time. The latter term is a function of the difference between the initial condition of time interval k (term $P(DQ(k-1) = 0)$) and the corresponding stationary probability $P(DQ_k = 0)$. The functional form of Equation (5) is inspired by both the exact closed-form expression of the transient queue-length distribution of an $M/M/1/\ell$ of Morse (1958, Chap. 6, Equation (6.13)) as well as by the approximate expression of the transient spillback probability (also known as the blocking probability) of an $M/M/1/\ell$ of Chong and Osorio (2017, Equation (14a)).

Equation (5) contains two endogenous terms, the stationary probability $P(DQ_k = 0)$ and the convergence speed $\tau_{DQ}(k)$. In queueing theory, $1/\tau_{DQ}(k)$ is known as the relaxation time and measures the time a given performance metric needs to reach its stationary value. Thus, $\tau_{DQ}(k)$ measures the speed at which the given performance metric approaches its stationary value (i.e., a larger $\tau_{DQ}(k)$ corresponds to a higher speed of convergence). To approximate $P(DQ_k = 0)$ and $\tau_{DQ}(k)$, we consider two cases depending on whether or not $\mu(k) = 0$.

Case $\mu(k) = 0$

When $\mu(k) = 0$, then there are no departures from DQ during time interval k. Thus DQ is a pure arrival process with rate $\lambda_{DQ}(k)$. This arises, for instance, in the presence of a red traffic light. The stationary probability of DQ being empty during time interval k is zero:

$$P(DQ_k = 0) = 0. \tag{6}$$

This is because for a pure arrival process, the queue length can only increase over time and hence given sufficient amount of time, the queue will eventually and forever leave the queue state of being empty.

Next, we present the approximation of $\tau_{DQ}(k)$. Let $N_{DQ}(k)$ denote the number of new arrivals that enter DQ during time interval k. The number of vehicles in DQ at the end of time interval k (i.e., $DQ(k)$) is the sum of the number of vehicles in DQ at the beginning of time interval k (i.e., $DQ(k-1)$) and the number of vehicles that entered DQ during time interval k:

$$DQ(k) = DQ(k-1) + N_{DQ}(k). \tag{7}$$

Therefore, $P(DQ(k) = 0)$ can be obtained as follows:

$$\begin{aligned} P(DQ(k) = 0) &= P\Big(\big(DQ(k-1) + N_{DQ}(k)\big) = 0\Big) && (8)\\ &= P\Big(DQ(k-1) = 0, N_{DQ}(k) = 0\Big) && (9)\\ &= P(DQ(k-1) = 0)P(N_{DQ}(k) = 0) && (10)\\ &= P(DQ(k-1) = 0)e^{-\lambda_{DQ}(k)\delta} && (11)\\ &= P(DQ_k = 0) + \Big[P(DQ(k-1) = 0) - P(DQ_k = 0)\Big]e^{-\lambda_{DQ}(k)\delta}. && (12) \end{aligned}$$

Equation (8) is obtained by substituting $DQ(k) = 0$ with the expression of Equation (7). Equation (9) is obtained based on the fact that $DQ(k) = 0$ if and only if (i) $DQ(k-1) = 0$ and (ii) $N_{DQ}(k) = 0$. Because we model DQ as an infinite capacity queue, $DQ(k-1)$ and $N_{DQ}(k)$ are independent. Hence, Equation (10) is obtained.

Since the arrival to DQ during time interval k is a Poisson process with rate $\lambda_{DQ}(k)$, the number of arrivals in a time interval length of δ (i.e., $N_{DQ}(k)$) follows a Poisson distribution with parameter $\lambda_{DQ}(k)\delta$ and $P(N_{DQ}(k) = 0) = e^{-\lambda_{DQ}(k)\delta}$, and Equation (11) is hence obtained.

Equation (12) is obtained from Equation (11) by adding and subtracting the term $P(DQ_k = 0)$ with value 0 (Eq. (6)). Note Equation (12) is of the exact form of Equation (5) with $\tau_{DQ}(k) = \lambda_{DQ}(k)$. Hence, when $\mu(k) = 0$,

$$\tau_{DQ}(k) = \lambda_{DQ}(k). \tag{13}$$

Case $\mu(k) > 0$

When $\mu(k) > 0$, there are both arrivals to and departures from DQ during time interval k and hence DQ is a birth and death process, which behaves like an $M/M/1$ queue as discussed at the start of Section 2.2.1. We now present the formulations starting with that of $P(DQ_k = 0)$. We define the traffic intensity of DQ during time interval k as:

$$\rho_{DQ}(k) = \frac{\lambda_{DQ}(k)}{\mu(k)}. \tag{14}$$

We use the following expression to approximate $P(DQ_k = 0)$:

$$P(DQ_k = 0) = \begin{cases} 1 - \rho_{DQ}(k) & \text{if } \rho_{DQ}(k) < 1 \quad (15a)\\ 0 & \text{otherwise.} \quad (15b) \end{cases}$$

Equation (15a) is obtained from the closed-form expression for the stationary queue-length distribution of an $M/M/1$ queue (e.g., Gross (2008, Chap. 2, Equation (2.9))). However, this expression only holds for $\rho_{DQ}(k) < 1$. In our transient setting, the traffic intensity $\rho_{DQ}(k)$ may temporarily exceed 1. Equation (15b) allows for this and is defined such as to ensure continuity of $P(DQ_k = 0)$ at $\rho_{DQ}(k) = 1$.

Next, we present the approximation of $\tau_{DQ}(k)$. We approximate $\tau_{DQ}(k)$ as follows:

$$\begin{cases} \tau_{DQ}(k) = \tau_{DQ,1} + \tau_{DQ,2} & (16a) \\ \tau_{DQ,1} = \left[1 - \alpha_1 e^{-\rho_{DQ}(k)}\right] \times \left[\dfrac{\mu(k)(1-\rho_{DQ}(k))^2}{(1+\rho_{DQ}(k))}\right] & (16b) \\ \tau_{DQ,2} = \alpha_2 \mu(k) \left| \dfrac{P(DQ(k-1)=0) - P(DQ_k = 0)}{\ell(1 - P(DQ_k = 0))} \right|^{1/5}. & (16c) \end{cases}$$

The terms α_1 and α_2 in Equation (16b) and (16c), respectively, are defined as follows:

$$\alpha_1 = \begin{cases} \alpha_{1,1} & \text{if } P(DQ(k-1)=0) > P(DQ_k = 0) \quad (17a) \\ \alpha_{1,2} & \text{otherwise;} \quad (17b) \end{cases}$$

$$\alpha_2 = \begin{cases} \alpha_{2,1} & \text{if } P(DQ(k-1)=0) > P(DQ_k = 0) \quad (18a) \\ \alpha_{2,2} & \text{otherwise;} \quad (18b) \end{cases}$$

where $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ and $\alpha_{2,2}$ are exogenous scalar coefficients. Equation (16a) approximates $\tau_{DQ}(k)$ as the sum of two relaxation terms: $\tau_{DQ,1}$ (which is defined in Eq. (16b)) and $\tau_{DQ,2}$ (which is defined in Eq. (16c)). We now describe how this formulation is derived. In a nutshell, the first term $\tau_{DQ,1}$, defined by Equation (16b), is formulated based on the relaxation time study of Newell (1982, Chap. 5, Equation (5.6)), while the second term $\tau_{DQ,2}$, defined by Equation (16c), is formulated based on insights from our numerical simulation experiments.

More specifically, Equation (16b) defines $\tau_{DQ,1}$ as the product of two terms within brackets. The expression within the right-side bracket corresponds to the inverse of the relaxation time for the expected queue-length of an $M/M/1$ system of Newell (1982, Chap. 5, Equation (5.6)):

$$\tau = \frac{\mu(1-\rho^2)}{C_S^2 + C_A^2 \rho}, \quad (19)$$

where C_A (resp. C_S) is the coefficient of variation for the inter-arrival (resp. service) time, μ and ρ are the service rate and the traffic intensity, respectively, in the notation of Newell (1982). For an $M/M/-$ system, we have $C_A = C_S = 1$. The expression within the left-side bracket is an adjustment term that is based on the following observations.

- The adjustment term should be unit-free. This is based on the observation of Odoni and Roth (1983) that the relaxation time be scaled in time so that it varies directly with the units of the arrival or service rates. In other words, two identical queueing systems measured in different time units should yield the same value of $\tau_{DQ}(k)\delta$ (in Eq. (5)). Thus, τ_{DQ} should vary directly with the units of the arrival or service rates. Since τ_{DQ} is the sum of $\tau_{DQ,1}$ and $\tau_{DQ,2}$ (Eq. (16a)), both $\tau_{DQ,1}$ and $\tau_{DQ,2}$ should also vary directly with the units of the arrival or service rates. Since the right-side bracket term varies directly with the service rate $\mu(k)$, the left-side bracket term should be unit-free.

- The adjustment term should ensure $\tau_{DQ}(k)$ is well-defined and continuous in the domain $\mu(k) \in [0, +\infty)$. In other words, $\lim_{\mu(k)\to 0} \tau_{DQ}(k)$ needs to equate the value of $\tau_{DQ}(k)$ defined at the limit point $\mu(k) = 0$ by Equation (13), i.e.,

$$\lim_{\mu(k)\to 0} \tau_{DQ}(k) = \lambda_{DQ}(k). \tag{20}$$

  This continuity property holds in our formulation. A detailed derivation of Equation (20) is given in Appendix A.

Equation (16c) defines $\tau_{DQ,2}$. This formulation is based on insights obtained from our numerical studies using a simulation-based implementation of the stochastic link transmission model. The simulations, used as a benchmark in Osorio and Flötteröd (2015), samples individual vehicles, and imposes the forward and backward lags explicitly for each vehicle. We carried out a total of 126 simulation experiments. Each experiment starts with an initial empty state and runs for 300 time units, it has one traffic intensity value for the first 150 time units and another value for the remaining 150 time units. We use the arrow notation $0.5 \rightarrow 0.25$ to denote an experiment with a traffic intensity that changes from 0.5 to 0.25. The experiments consider all combinations of traffic intensity $\lambda/\mu \in \{0.5 \rightarrow 0.25, 0.75 \rightarrow 0.25, 1.25 \rightarrow 0.25, 0.75 \rightarrow 0.5, 1.25 \rightarrow 0.5, 1.25 \rightarrow 0.75\}$, service rate $\mu \in \{0.2, 0.4, 0.6\}$, and space capacity $\ell \in \{10, 20, 30, 40, 60, 80, 100\}$.

Equation (16c) defines $\tau_{DQ,2}$ based on the following observations from these simulation experiments.

- We want the relaxation time (or equivalently its inverse) to depend on the space capacity ℓ. This is because we observe that the inverse of the relaxation time of $P(DQ(k) = 0)$ is inversely related to ℓ. In other words, as ℓ increases, so does the time needed to reach stationarity. Blanc and Van Doorn (1983) have also observed that for a finite capacity queueing system (e.g., $M/M/1/\ell$), the relaxation time of the system is positively correlated with the space capacity. The parameter ℓ is in the denominator of Equation (16c). Thus, as ℓ increases, $\tau_{DQ,2}$ decreases, this leads to an increased time to reach stationarity.

- We want the relaxation time to depend on the distance between the initial state (i.e., initial probability) and the steady state. In our simulation experiments, we observe a positive correlation between the inverse of the relaxation time and the absolute difference between the initial state and the corresponding steady state probability. Odoni and Roth (1983) have also observed that for an isolated $M/M/1$ queueing system and for arbitrary deterministic initial states, the relaxation time of the expected queue length may depend on the distance between the initial state and steady state. The term $\left|\frac{P(DQ(k-1)=0)-P(DQ_k=0)}{(1-P(DQ_k=0))}\right|$ of Equation (16c) represents the normalized absolute difference between the initial state and the steady state. Thus, when the value of this term increases, so does $\tau_{DQ,2}$, and this leads to a higher speed (i.e., lower time) to reach stationarity.

- We want the relaxation time to depend on whether congestion is propagating (i.e., aggravating, building up) or dissipating. We have observed in our simulation experiments that it takes a shorter time to reach stationarity when congestion is building up compared to when it is dissipating. The terms α_1 of $\tau_{DQ,1}$ (i.e., Eq. (16b)) and α_2 of $\tau_{DQ,2}$ (i.e., Eq. (16c)) are introduced

to account for this. They are defined by Equations (17) and (18) as functions of the initial state $P(DQ(k-1)=0)$ and the corresponding steady state $P(DQ_k=0)$. More specifically, Equations (17a) and (18a) consider the case of congestion build up, while Equations (17b) and (18b) consider the case of congestion dissipation. A description of how the exogenous scalar parameters $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ are $\alpha_{2,2}$ are fitted is given in Appendix B. In the experiments presented in this paper, the fitted values are $\alpha_{1,1}=\alpha_{1,2}=0.4$, $\alpha_{2,1}=0.6$ and $\alpha_{2,2}=0$. Note that $\alpha_{2,1}>\alpha_{2,2}$, which means that, all else being equal, $\tau_{DQ}(k)$ is larger when congestion builds up than when congestion dissipates. In other words, it takes a longer time to reach stationarity when congestion is dissipating than when it is building up. As mentioned previously, this is consistent with our simulation experiments.

2.2.2 Upstream boundary conditions

This section formulates the probabilistic model of the link's upstream boundary conditions $P(UQ(k)=\ell)$. In queueing theory $P(UQ(k)=\ell)$ is known as the blocking probability of UQ. In traffic flow theory it represents the spillback probability of the link.

Recall from Section 2.1 that the arrival process to the link is assumed to be a Poisson process with exogenous rate $\lambda(k)$. Since the arrival process to the link is the same as the arrival process to UQ, the arrival process to UQ during time interval k is also a Poisson process with exogenous rate $\lambda(k)$. Thus, we only need to approximate the service process of UQ.

Recall from Section 2.1 that flow that enters UQ sequentially undergoes the following three phases of service: (i) it is delayed k^{fwd} time intervals (this delay is represented in Figure 1 by LI); (ii) it enters DQ, where it experiences a sojourn (waiting and service) time; (iii) vehicular flow that leaves the link (i.e., leaves DQ) generates newly available road space, which becomes available at the upstream end of the link after a delay of k^{bwd} time intervals (this delay is represented in Figure 1 by LO). Once this space becomes available upstream, the corresponding flow leaves UQ.

Flow departures from UQ correspond to flow departures from the most downstream cell of LO (which is denoted LLO in Figure 1). Let $q^{LLO}(k)$ denote the expected outflow from LLO during time interval k. It corresponds to vehicular flow that left the link during time interval $k-k^{bwd}$, i.e.,

$$q^{LLO}(k)=q^{out}(k-k^{bwd}). \tag{21}$$

To approximate $P(UQ(k)=\ell)$ we consider two cases depending on whether or not $q^{LLO}(k)=0$. Note that at time interval k, $q^{LLO}(k)$ is known since it defined by expected link outflows from past time intervals (Eq. (21)).

Case $q^{LLO}(k)=0$

If $q^{LLO}(k)=0$, then the expected outflow from UQ is also zero during time interval k. This implies that, with probability 1, there are no departures from UQ. Thus, UQ is a pure arrival process. This arises, for instance, in the presence of a red traffic light during time interval $k-k^{bwd}$ so that there is no vehicular flow that left the link from past time intervals (Eq. (21)).

Let $N(k)$ denote the number of attempted new arrivals to the link during time interval k whether or not they successfully enter UQ. Thus, the number of arrivals that successfully entered UQ during time interval k is the minimum of $N(k)$ and the available space left, i.e., $\ell - UQ(k-1)$. Note that the proposed model is a single link model. The vehicles that wish to enter the link while the link is full are considered lost demand. If a model with no losses is desired, an infinite capacity queue can be added to the upstream of the link to hold the vehicles that are waiting to enter the link. In a network setting, as in Section 4, vehicles that have already been loaded onto the network are preserved and are not lost. Only vehicles that would like to start their trip and cannot find available space in their entry link can potentially be lost.

The number of vehicles in UQ at the end of time interval k (i.e., $UQ(k)$) is sum of the number of vehicles in UQ at the beginning of time interval k (i.e., $UQ(k-1)$) and the number of vehicles that successfully entered UQ during time interval k:

$$UQ(k) = UQ(k-1) + \min\Big\{N(k), \ell - UQ(k-1)\Big\}. \tag{22}$$

Therefore, $P(UQ(k) = \ell)$ can be obtained as follows:

$$P(UQ(k) = \ell) = P\Big(\big(UQ(k-1) + \min\{N(k), \ell - UQ(k-1)\}\big) = \ell\Big) \tag{23}$$

$$= \sum_{i=0}^{\ell} P\Big(\min\{N(k), \ell - i\} = \ell - i \Big| UQ(k-1) = i\Big) P\Big(UQ(k-1) = i\Big) \tag{24}$$

$$= \sum_{i=0}^{\ell} P\Big(N(k) \geq \ell - i \Big| UQ(k-1) = i\Big) P\Big(UQ(k-1) = i\Big) \tag{25}$$

$$= \sum_{i=0}^{\ell} P\Big(N(k) \geq \ell - i\Big) P\Big(UQ(k-1) = i\Big) \tag{26}$$

Equation (23) is obtained by substituting $UQ(k) = \ell$ with the expression of Equation (22). Equation (24) is obtained by conditioning on the states of UQ at the beginning of time interval k, or equivalently at the end of time interval k-1, i.e., $UQ(k-1)$. In the conditional probability of Equation (24), the equality $\min\{N(k), \ell - i\} = \ell - i$ holds if and only if $N(k) \geq \ell - i$. Thus, Equation (25) is obtained. Since the process of attempted arrivals does not have any dependence on the initial state of the link, $P\Big(N(k) \geq \ell - i \Big| UQ(k-1) = i\Big) = P\big(N(k) \geq \ell - i\big)$ and Equation (26) is obtained.

Since the arrival process to the link, which is also the arrival process to UQ, is a Poisson process with rate $\lambda(k)$, then $N(k)$, the number of attempted arrivals during time interval k, follows a Poisson distribution with parameter $\lambda(k)\delta$. Thus, $P\big(N(k) \geq \ell - i\big)$ is calculated as follows:

$$P\Big(N(k) \geq \ell - i\Big) = 1 - P\Big(N(k) \leq \ell - i - 1\Big) \tag{27}$$

$$= 1 - e^{-\lambda(k)\delta} \sum_{j=0}^{\ell - i - 1} \frac{(\lambda(k)\delta)^j}{j!}. \tag{28}$$

Equation (26) depends on the full marginal distribution of UQ (i.e., it depends on all terms $P\big(UQ(k-1) = i\big), \forall i \in \{0, \ldots, \ell\}$). However, the proposed model does not track the full distribution of UQ, it

only tracks the scalar probability $P\big(UQ(k-1)=\ell\big)$. Thus, we propose the following approximation for $P\big(UQ(k-1)=i\big)$, $0 \leq i \leq \ell-1$:

$$\begin{cases} P\big(UQ(k-1)=i\big) = \dfrac{1-P\big(UQ(k-1)=\ell\big)}{\sum_{j=0}^{\ell-1} f\big(j, q^{UQ}(k-1)\delta\big)} f\big(i, q^{UQ}(k-1)\delta\big) & \text{(29a)} \\ f\big(i, q^{UQ}(k-1)\delta\big) = \dfrac{(q^{UQ}(k-1)\delta)^i e^{-q^{UQ}(k-1)\delta}}{i!} & \text{(29b)} \\ q^{UQ}(k-1) = \sum_{r=0}^{k-2} q^{in}(r) - \sum_{r=0}^{k-k^{bwd}-2} q^{out}(r). & \text{(29c)} \end{cases}$$

Equation (29b) gives the probability mass function (pmf) of a Poisson distribution with parameter $q^{UQ}(k-1)\delta$. Thus, Equation (29a) is a normalized and finite support (i.e., $\{0, ..., \ell-1\}$) Poisson distribution (with parameter $q^{UQ}(k-1)\delta$). The normalization term (the fraction term) is defined such as to obtain a proper probability distribution, see Appendix D for detailed explanations. Equation (29c) defines the expected flow in $UQ(k-1)$, based on flow conservation principles, as the difference between: (i) aggregated (over time) flow that has entered the link up until the end of time interval $k-2$ (first summation) and (ii) aggregated (over time) vehicular flow that has left the link up until the end of time interval $k-k^{bwd}-2$ (second summation). The second summation accounts for the kinematic backward wave delay.

Case $q^{LLO}(k) > 0$

When $q^{LLO}(k) > 0$, we account for all three service processes that vehicular flow within UQ goes through, which were mentioned at the start of Section 2.2.2. This leads us to approximate UQ as an $M/G/\ell/\ell$ system. Let us detail this. First, since the service time of UQ is the sum of that of these three processes, we assume it to be generally distributed, i.e., it can follow a general distribution with a known mean and variance. For an $-/-/\ell/\ell$ system the service time and the sojourn time are identical because there is no waiting or buffer space. The expected service time, $E[S_{UQ}(k)]$, is given by:

$$E[S_{UQ}(k)] = k^{fwd}\delta + E[S_{DQ}(k-k^{bwd})] + k^{bwd}\delta \tag{30}$$

where $E[S_{DQ}(k-k^{bwd})]$ is the expected sojourn time (service plus waiting time) of the DQ system from the past time interval $k-k^{bwd}$, which corresponds to the vehicular flows in LLO that are ready to leave UQ during time interval k. Second, we approximate UQ as a multi-server, rather than a single-server, queueing system. This is because the flow in LI and in LO is served (or processed) simultaneously, rather than sequentially.

We introduce the following notation.

$P(UQ(k)=\ell)$	probability of $UQ(k)$ being full at the end of time interval k (which is also the beginning of time interval $k+1$);
$P(UQ_k=\ell)$	stationary probability of $UQ=\ell$ for time interval k;
$\tau_{UQ}(k)$	inverse of the relaxation time of UQ during time interval k.

We use the same functional form as for $DQ(k)$ (Eq. (5)) to approximate $P(UQ(k) = \ell)$:

$$P(UQ(k) = \ell) = P(UQ_k = \ell) + \Big[P(UQ(k-1) = \ell) - P(UQ_k = \ell)\Big] e^{-\tau_{UQ}(k)\delta}. \tag{31}$$

Just as for $DQ(k)$ (Eq. (5)), the transient probability $P(UQ(k) = \ell)$ is defined as the sum of a time-interval specific stationary probability (term $P(UQ_k = \ell)$) and a term that decays exponentially with time and accounts for the difference between the initial conditions (i.e., $P(UQ(k-1) = \ell)$) and the corresponding stationary conditions (i.e., $P(UQ_k = \ell)$). Our motivation to use this functional form differs from that of Equation (5). Here we are inspired by Jagerman (1975, Equation (166)) which expresses the transient blocking probability of an $M/M/\ell/\ell$ system as the sum of the corresponding stationary probability (known as the Erlang-B formula, it is presented below in Equation (32a)) and a term that decays exponentially with time. We consider generally distributed, rather than Markovian, service times.

The stationary probability $P(UQ_k = \ell)$ of Equation (31) is approximated as follows:

$$\begin{cases} P(UQ_k = \ell) = \dfrac{\rho_{UQ}(k)^\ell/\ell!}{\sum_{n=0}^{\ell} \rho_{UQ}(k)^n/n!} & \text{(32a)} \\ \rho_{UQ}(k) = \lambda(k)/\mu_{UQ}(k) & \text{(32b)} \\ \mu_{UQ}(k) = 1/(k^{fwd}\delta + E[S_{DQ}(k - k^{bwd})] + k^{bwd}\delta) & \text{(32c)} \\ E[S_{DQ}(k - k^{bwd})] = \dfrac{\ell\rho_{DQ}(k-k^{bwd})^{\ell+1} - (\ell+1)\rho_{DQ}(k-k^{bwd})^\ell + 1}{\mu(k-k^{bwd})(1-\rho_{DQ}(k-k^{bwd})^\ell)(1-\rho_{DQ}(k-k^{bwd}))}. & \text{(32d)} \end{cases}$$

The system $M/G/\ell/\ell$ has been extensively studied and is known as the Erlang loss model. Equation (32a) is the stationary blocking probability for an $M/G/\ell/\ell$. It is known as the Erlang-B formula. It was first derived by Erlang (1917) for an $M/M/\ell/\ell$. Khinchin (1962) later proved that it holds for generally distributed service times with finite expectation. Equation (32b) is the definition of the traffic intensity of the $M/G/\ell/\ell$: it is the ratio of the arrival rate $\lambda(k)$ to the inverse of the expected service time, which is given by the inverse of Equation (32c) or equivalently by Equation (30). Equation (32d) approximates the expected sojourn time of DQ during time interval $k - k^{bwd}$. The expression corresponds to the closed-form expression for the expected sojourn time of an $M/M/1/\ell$ system (e.g., Gross (2008, Chap. 2, Equations (2.48) and (2.51))). Note that Equation (32d) uses expressions from an $M/M/1/\ell$ model for DQ, while Section 2.2.1 uses expressions from an $M/M/1$ model. The use of an $M/M/1/\ell$ model yields an expected sojourn time that: (i) is well defined for all traffic intensity values (i.e., even when $\rho_{DQ}(k - k^{bwd}) \geq 1$), and (ii) is bounded from above. This is not the case for the expected sojourn time of an $M/M/1$ model (i.e., $E[S_{DQ}(k - k^{bwd})] = 1/(\mu(k - k^{bwd}) - \lambda_{DQ}(k - k^{bwd}))$), which assumes a traffic intensity strictly smaller than 1 and goes to infinity as the traffic intensity approaches 1.

Equation (32d) depends on $\rho_{DQ}(k-k^{bwd})$, which is defined by Equation (14). Under the current case where $q^{LLO}(k) > 0$, Equation (14) has a non-zero denominator $\mu(k - k^{bwd})$. We can show this by contradiction as follows. If $\mu(k - k^{bwd}) = 0$, then according to Equation (3), $q^{out}(k - k^{bwd}) = 0$, and according to Equation (21), $q^{LLO}(k) = 0$. This contradicts the assumption of the current case: $q^{LLO}(k) > 0$.

The endogenous parameter $\tau_{UQ}(k)$ of Eq. (31) represents the inverse of the relaxation time. As discussed in Section 2.2.1, it measures the speed of convergence to the stationary value. We approximate

$\tau_{UQ}(k)$ as follows.

$$\begin{cases} \tau_{UQ}(k) = \dfrac{\alpha_3 \ell \mu_{UQ}(k)(1-\rho_{UQ}(k)/\ell)^2}{1+C_{UQ}(k)^2} \times \dfrac{1}{\ell} & \text{(33a)} \\ C_{UQ}(k) = \sqrt{\text{Var}(S_{UQ}(k))}/E[S_{UQ}(k)] & \text{(33b)} \\ \text{Var}(S_{UQ}(k)) = [\ell\rho_{DQ}(k-k^{bwd})^{2\ell+2} - 2\ell\rho_{DQ}(k-k^{bwd})^{2\ell+1} + (\ell+1)\rho_{DQ}(k-k^{bwd})^{2\ell} \\ \qquad - \ell(\ell+1)\rho_{DQ}(k-k^{bwd})^{\ell+2} + 2\ell(\ell+1)\rho_{DQ}(k-k^{bwd})^{\ell+1} - (\ell^2+\ell+2) \\ \quad \times \rho_{DQ}(k-k^{bwd})^{\ell} + 1]/[\mu(k-k^{bwd})^2(1-\rho_{DQ}(k-k^{bwd})^{\ell})^2(1-\rho_{DQ}(k-k^{bwd}))^2], & \text{(33c)} \end{cases}$$

where the term α_3 in Equation (33a) is defined by:

$$\alpha_3 = \begin{cases} \alpha_{3,1} & \text{if } P(UQ(k-1)=\ell) < P(UQ_k = \ell) \quad \text{(34a)} \\ \alpha_{3,2} & \text{otherwise .} \quad \text{(34b)} \end{cases}$$

Studies of the relaxation time of an Erlang loss model are limited and mostly focus on the special case of an $M/M/\ell/\ell$ system (e.g., van Doorn and Zeifman (2009)) and on the asymptotic analysis of the relaxation time of the expected queue length. Our focus is on the relaxation time of the blocking probability of an Erlang loss model with generally distributed service times (i.e., $M/G/\ell/\ell$). To derive an approximation of the relaxation time for such a system, we follow the approach of Roth (1994). Roth assumes that the functional form of the relaxation time of a multi-server $M/M/\ell$ system is the same as that of the single server $M/M/1$ system and replaces the service rate μ with $\ell\mu$. The expression for a single-server $M/M/1$ system is derived by Odoni and Roth (1983). We proceed similarly: we use the same functional form for the relaxation time of an $M/G/1$ system, and scale the service rate by ℓ.

To derive an expression for the relaxation time of an $M/G/1$, we proceed as follows. Roth (1994) studies the relaxation times for infinite-capacity, single-server queueing system that are Markovian (including systems in which inter-arrival and service times can be represented as exponential, Erlangian, hyperexponential, and phase-type random variables) and derives a general inverse relaxation formula for such a system as

$$\tau_{\text{Roth}} = \frac{\mu(1-\lambda/\mu)^2}{C_A^2 + C_S^2}, \tag{35}$$

where C_A (resp. C_S) is the coefficient of variation for the inter-arrival (resp. service) time. We approximate the relaxation time of our $M/G/1$ system as an $M/PH/1$ system. Past work has shown that there always exists a phase-type (PH) distribution to approximate a non-negative general distribution with an arbitrary precision level (Asmussen; 2003). For an $M/-/-$ system, we have $C_A = 1$. We set C_S equal to the coefficient of variation of the service time of UQ during time interval k, denoted $C_{UQ}(k)$ and given by Equation (33b).

We then proceed to obtain the inverse of relaxation time of $M/G/\ell$ from Roth's formula (Eq. (35)) by scaling the service rate by ℓ:

$$\tau_{M/G/\ell} \approx \frac{\ell\mu_{UQ}(k)\big(1-\lambda(k)/\ell\mu_{UQ}(k)\big)^2}{1+C_{UQ}(k)^2} \tag{36}$$

$$= \frac{\ell \mu_{uQ}(k)\left(1 - \rho_{uQ}(k)/\ell\right)^2}{1 + C_{uQ}(k)^2}. \tag{37}$$

Notice that Equation (37) is obtained by substituting in Equation (32b) and it is exactly the left part of Equation (33a).

The system we are interested in has a finite capacity (i.e., $M/G/\ell/\ell$), hence we need to further adjust Equation (37) to take the finite capacity into consideration. We follow the idea of Chong and Osorio (2017), in which the relaxation time of an $M/M/1/\ell$ system is approximated as the product of (i) the relaxation time of an $M/M/1$ system and (ii) the finite capacity of the queue ℓ. Applying the idea above, we obtain the approximation of the inverse relaxation time of an $M/G/\ell/\ell$ system as the product of (i) the inverse relaxation time of an $M/G/\ell$ system (i.e., $\tau_{M/G/\ell}$ of Eq. (37)) and (ii) the inverse of the finite capacity of the queue $1/\ell$. Equation (33a) for $\tau_{uQ}(k)$ of an $M/G/\ell/\ell$ queue is hence obtained.

Equation (33b) defines the coefficient of variation of $S_{uQ}(k)$ as the ratio between its standard deviation and expectation. The expectation of $S_{uQ}(k)$ (i.e., $E[S_{uQ}(k)]$) is given by Equation (30). The variance of $S_{uQ}(k)$ (i.e., $Var(S_{uQ}(k))$) is given by Equation (33c) and it is obtained as follows. Recall that by definition: $S_{uQ}(k) = k^{fwd}\delta + S_{DQ}(k - k^{bwd}) + k^{bwd}\delta$, where $k^{fwd}\delta$ and $k^{bwd}\delta$ are constant delays. Thus, $Var(S_{uQ}(k)) = Var(S_{DQ}(k - k^{bwd}))$. The expression for the variance of $S_{DQ}(k - k^{bwd})$ is derived in Appendix C.

The proposed expression for $\tau_{uQ}(k)$, defined by the system of Equations (33) has the following properties.

- Just as for the expression proposed for $\tau_{DQ}(k)$ (Eq. (16)), the expression for $\tau_{uQ}(k)$ has the same time units as the arrival or service rates. More specifically, Equation (33a) has the same units as $\mu_{uQ}(k)$ (note that $\rho_{uQ}(k)$, $C_{DQ}(k)$, and ℓ are unit-free).

- Davis et al. (1995) study non-stationary Erlang loss models with a special focus on $M_t/PH/\ell/\ell$ systems. They observe that the inverse of the relaxation time decreases, as the variability of the service time increases. In other words, the more variable the service time, the longer it takes to reach stationarity. This holds for the proposed equation. In Equation (33a), $\tau_{uQ}(k)$ is inversely proportional to $C_{uQ}(k)$, which is the coefficient of variation of the service time of uQ during time interval k. Thus, the higher the variability of $S_{uQ}(k)$, the higher $C_{uQ}(k)$, the smaller $\tau_{uQ}(k)$, and thus the longer the time it takes to reach stationarity.

- Just as for $\tau_{DQ}(k)$ (Equation (16)), the relaxation time depends on whether congestion is building up or dissipating. Thus, Equation (34) defines α_3 as a function of whether congestion is building up (Equation (34a)) or dissipating (Equation (34b)). We use the same 126 simulation experiments described in Section 2.2.1 to fit the exogenous scalar coefficients $\alpha_{3,1}$ and $\alpha_{3,2}$ of Equation (34). A description of how these coefficients are estimated is given in Appendix E. The fitted values are $\alpha_{3,1} = 25$ and $\alpha_{3,2} = 7.5$. Note that $\alpha_{3,1} > \alpha_{3,2}$, which means that when traffic dissipates it takes a longer time to reach stationary than when traffic builds up. As mentioned previously, this is consistent with observations from our own numerical experiments.

2.2.3 Algorithm

Algorithm 1 summarizes the proposed model. Steps 1 through 5 are initialization steps. Step 6 is carried out iteratively, it yields for each time interval the two key probabilities: $P(DQ(k) = 0)$ and $P(UQ(k) = \ell)$, as well as expected link inflow (i.e., $q^{in}(k)$) and expected link outflow (i.e., $q^{out}(k)$). More specifically, steps 6(a)-6(e) implement the model of $P(DQ(k) = 0)$ of Section 2.2.1, while step 6(f) implements the model of $P(UQ(k) = \ell)$ of Section 2.2.2. Note that all function evaluations in step 6 can be done sequentially and no simultaneous evaluation of a system of equations is required. This makes our algorithm computationally efficient.

3 Validation

In this section, we evaluate the accuracy and the computational efficiency of the proposed model. We compare its performance to that of two stochastic simulators: (i) the stochastic link transmission model (LTM) simulator (Section 3.1), and (ii) the microscopic traffic simulator Aimsun (TSS; 2014) (Section 3.2). The stochastic LTM simulator is a discrete-event implementation of a stochastic formulation of the deterministic link transmission model (LTM) (Yperman et al.; 2007). It assumes an inhomogeneous Poisson arrival process at the upstream end of the link and a stochastic departure process at the downstream end of the link. The simulator samples individual vehicles, it implements the exact forward and backward lags. The vehicles at the downstream end of the link are served following a first-come first-serve rule. The service times are independent and identically distributed exponential random variables. The simulator was used as a benchmark in the validation experiments in Osorio and Flötteröd (2015) and in Lu and Osorio (2018). The microscopic simulator Aimsun (TSS; 2014) is a commercial traffic simulator that relies on disaggregate car-following and lane-changing models for individual vehicles.

3.1 Validation versus a stochastic link transmission model simulator

We benchmark the performance of the proposed model versus that of the multivariate model (Osorio and Flötteröd; 2015) and of the mixture model (Lu and Osorio; 2018). The analytical approximations provided by each of the three analytical models are compared to simulation-based estimates obtained from 10^6 simulation replications of the stochastic LTM simulator.

We consider a link with parameters defined in Table 1 for all experiments conducted in Section 3.1. First, we consider two experiments with time-varying demand and evaluate the ability of the proposed model to approximate the link's upstream and downstream boundary conditions. The space capacity of the link ℓ, is fixed at 10 for both experiments. The link is initially empty. Experiment 1 considers a case where traffic conditions change from uncongested to highly-congested and then to congested. More specifically, it considers an arrival rate, $\lambda(k)$, of 0.1 veh/sec during the interval $[0, 125]$ seconds, of 0.5 veh/sec during the interval $[125, 175]$ seconds and of 0.3 veh/sec during the interval $[175, 300]$ seconds. Experiment 2 considers a case where traffic conditions change from congested to uncongested and then to highly-congested. It considers an arrival rate of 0.3 veh/sec during the interval

Algorithm 1 Link model algorithm

1. set exogenous link parameters $\hat{\rho}, v, w, \ell$ and the duration of each time interval δ
2. compute the forward and backward lags: $k^{fwd} = \lceil \frac{\ell}{\hat{\rho} v \delta} \rceil$ and $k^{fwd} = \lceil \frac{\ell}{\hat{\rho} |w| \delta} \rceil$
3. set, for each time interval, the exogenous arrival rates and service rates $\lambda(k)$ and $\mu(k)$ for $\forall\, k = 1, 2, ...$
4. set initial link conditions: $q^{in}(0)$, $q^{out}(0)$, $q^{UQ}(0)$, $P(UQ(0) = \ell)$ and $P(DQ(0) = 0)$
5. set $q^{in}(k) = 0$ and $q^{out}(k) = 0$ for $k < 0$
6. repeat the following for time intervals $k = 1, 2, ...$
   (a) compute $\lambda_{DQ}(k)$ and $q^{LLO}(k)$ according to Eq. (4) and Eq. (21), respectively
   (b) if $\mu(k) = 0$:
      i. compute $P(DQ_k = 0)$ according to Eq. (6)
      ii. compute $\tau_{DQ}(k)$ according to Eq. (13)

      else:
      i. compute $\rho_{DQ}(k)$ according to Eq. (14)
      ii. compute $P(DQ_k = 0)$ according to Eq. (15)
      iii. compute α_1 and α_2 according to Eq. (17) and Eq. (18), respectively
      iv. compute $\tau_{DQ,1}$ and $\tau_{DQ,2}$ according to Eq. (16b) and Eq. (16c), respectively
      v. compute $\tau_{DQ}(k)$ according to the system of Eq. (16a)
   (c) compute $P(DQ(k) = 0)$ according to Eq. (5)
   (d) if $q^{LLO}(k) = 0$:
      i. compute $q^{UQ}(k-1)$ according to Eq. (29c) if $k > 1$
      ii. compute $f(i, q^{UQ}(k-1)\delta)\ \ \forall i \in \{0, 1, ..., \ell - 1\}$ according to Eq. (29b)
      iii. compute $P(N(k) \geq \ell - i)\ \ \forall i \in \{0, 1, ..., \ell\}$ and compute $P(UQ(k-1) = i)\ \ \forall i \in \{0, 1, ..., \ell - 1\}$ according to Eq. (28) and Eq. (29a), respectively
      iv. compute $P(UQ(k) = \ell)$ according to Eq. (26)

      else:
      i. compute $E[S_{DQ}(k)]$ and $Var(S_{UQ}(k))$ according to Eq. (32d) and Eq. (33c), respectively
      ii. compute $E[S_{UQ}(k)]$ and $\mu_{UQ}(k)$ according to Eq. (30) and Eq. (32c), respectively
      iii. compute $\rho_{UQ}(k)$ and $C_{UQ}(k)$ according to Eq. (32b) and Eq. (33b), respectively
      iv. compute $P(UQ_k = \ell)$ according to Eq. (32a)
      v. compute α_3 according to Eq. (34)
      vi. compute $\tau_{UQ}(k)$ according to Eq. (33a)
      vii. compute $P(UQ(k) = \ell)$ according to Eq. (31)
   (e) compute $q^{in}(k)$ and $q^{out}(k)$ according to Eq. (2) and (3), respectively

Parameter	Value
v	0.01 km/sec
w	-0.005 km/sec
$\hat{\rho}$	200 veh/km
$\hat{q}$	2400 veh/h = 0.67 veh/sec
δ	1 sec
$\mu(k)$	1440 veh/h = 0.4 veh/sec
$\lambda(k)$	varies by experiment
ℓ, L, k^{fwd}, k^{bwd}	varies by experiment

Table 1: Link parameters

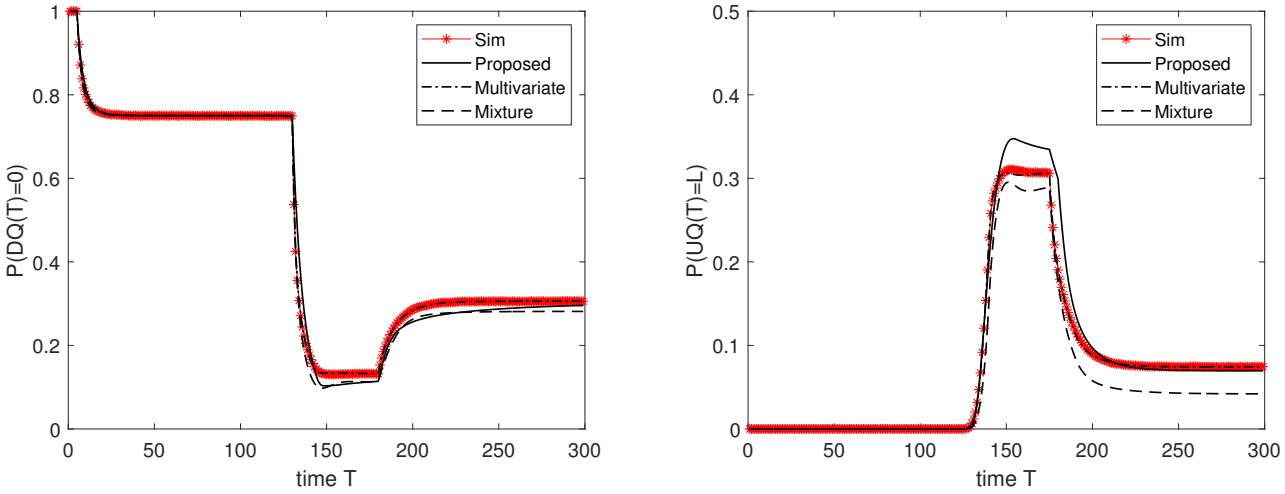


Figure 2: Experiment 1: impact of the temporal variation of demand on the link's upstream and downstream boundary conditions

$[0, 100]$ seconds, of 0.1 veh/sec during the interval $[100, 200]$ seconds and of 0.5 veh/sec during the interval $[200, 300]$ seconds.

Figure 2 considers experiment 1. The left (resp. right) plot considers $P(DQ(T) = 0)$ (resp. $P(UQ(T) = \ell)$). For each plot, the x-axis displays the integer time T in seconds and the y-axis displays the corresponding probability. The simulated estimates are displayed as a red line with asterisks, those of the proposed model are the black solid line, those of the the multivariate model are the black dot-dashed line, and those of the mixture model are the black dashed line. The simulated estimates are displayed with 95% confidence intervals, which are barely visible.

Recall that in experiment 1, there is a sharp increase in demand at time $T = 125$ seconds and a sharp decrease at time $T = 175$ seconds. All analytical models yield similar temporal trends for both $P(DQ(T) = 0)$ and $P(UQ(T) = \ell)$. More specifically, as congestion increases, we expect $P(DQ(T) = 0)$ to decrease and $P(UQ(T) = \ell)$ to increase. Similarly, as congestion decreases, we expect $P(DQ(T) = 0)$ to increase and $P(UQ(T) = \ell)$ to decrease. All models exhibit these trends. They all capture the sharp decrease and increase trends of the simulator. The multivariate model yields the most accurate approximation. The proposed model tends to overestimate the stationary spillback

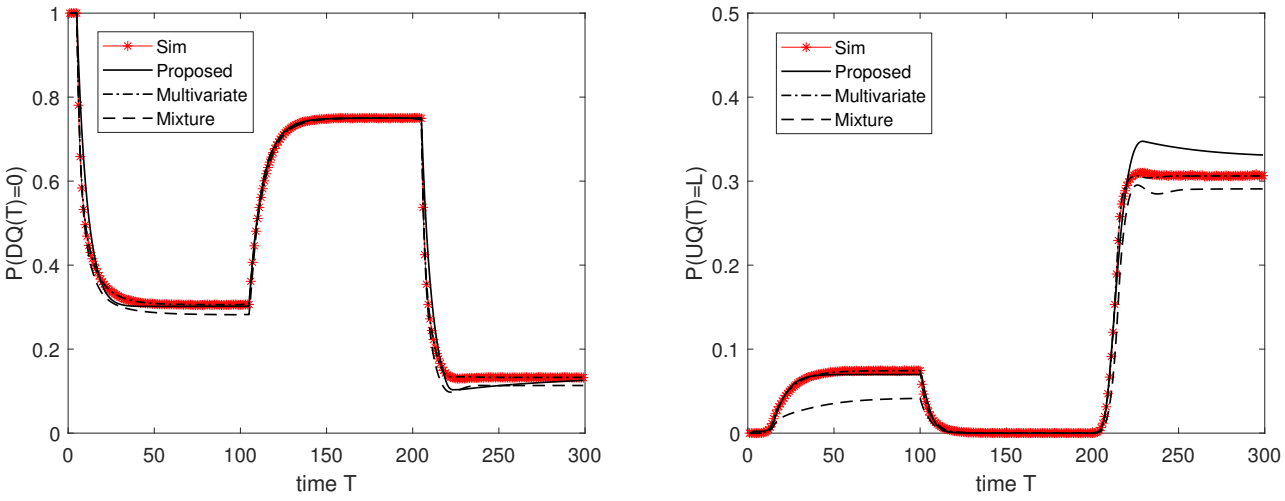


Figure 3: Experiment 2: impact of the temporal variation of demand on the link's upstream and downstream conditions

probability during the highly congested regime, whereas the mixture model tends to underestimate the stationary spillback probability during both the congested and the highly congested regimes. Overall, the proposed model yields a good approximation of both probabilities $P(DQ(T) = 0)$ and $P(UQ(T) = \ell)$.

The results of experiment 2 are displayed in Figure 3. These plots have the same layout as those of Figure 2. The simulated estimates are displayed with 95% confidence intervals, which are barely visible. Recall that experiment 2 considers a sharp decrease in demand at $T = 100$ seconds and a sharp increase at $T = 200$ seconds. The left plot shows an increase in $P(DQ(T) = 0)$ after $T = 100$ seconds and a decrease after $T = 200$ seconds. The right plot shows that the spillback probability ($P(UQ(T) = \ell)$) decreases after $T = 100$ seconds and increases after $T = 200$ seconds. All analytical models capture these sharp changes in probability mass. The proposed model yields a less accurate approximation of the stationary spillback probability during the highly congested regime, while the mixture model yields a less accurate approximation for both the congested and the highly congested regimes. Overall, all three models approximate well the dynamics of the link's boundary conditions for sudden and significant changes in congestion levels.

Next, we benchmark the accuracy of the proposed model over a set of 21 experiments. All experiments start with an empty link and have a duration of $T_F = 300$ seconds. For each experiment, the arrival rate changes at 150 seconds. The experiments consider all combinations of the following time-varying arrival rates ($\lambda \in \{0.2 \rightarrow 0.1, 0.3 \rightarrow 0.1, 0.3 \rightarrow 0.2\}$ veh/sec) and space capacities ($\ell \in \{10, 20, 30, 40, 60, 80, 100\}$). The space capacity values considered correspond to link lengths $L \in \{50, 100, 150, 200, 300, 400, 500\}$ (in meters), forward lags $k^{fwd} \in \{5, 10, 15, 20, 30, 40, 50\}$ and backward lags $k^{bwd} \in \{10, 20, 30, 40, 60, 80, 100\}$. For all experiments the arrival rates are such that congestion builds up during the first 150 seconds and the link reaches a stationary regime, then the arrival rate decreases, congestion dissipates and the link reaches a less congested stationary regime. For each experiment and each model, we set a maximum computation runtime of 40 hours. Model evaluations that have not concluded within the 40 hours are terminated.

The error metric used to evaluate the accuracy of a given analytical model is the average absolute difference between the simulated estimate and the analytical approximation:

$$\bar{e}_{DQ} = \frac{1}{T_F} \sum_{T=1}^{T_F} |P_A(DQ(T) = 0) - P_S(DQ(T) = 0)| \tag{38}$$

$$\bar{e}_{UQ} = \frac{1}{T_F} \sum_{T=1}^{T_F} |P_A(UQ(T) = \ell) - P_S(UQ(T) = \ell)|, \tag{39}$$

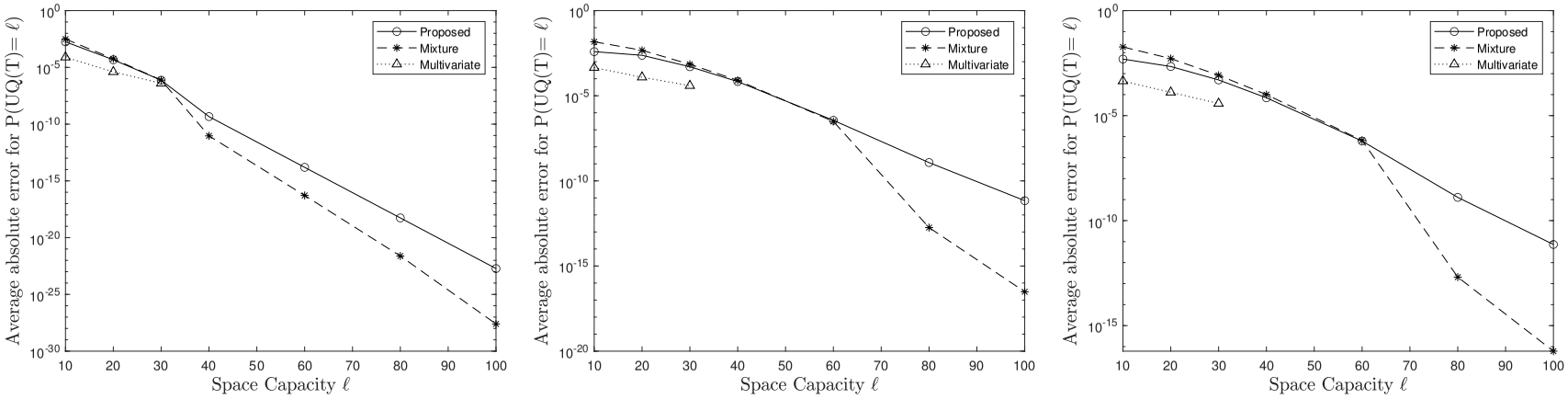
where P_A denotes the probability approximated by an analytical model (proposed, mixture or multivariate) and P_S denotes the simulated estimate.

Figure 4 displays the average absolute difference for the 21 experiments. The top (resp. bottom) three plots consider the spillback probability $P(UQ(T) = \ell)$ (resp. $P(DQ(T) = 0)$). The first, second and third column of plots consider the experiments with arrival rate $0.2 \rightarrow 0.1$ veh/sec, $0.3 \rightarrow 0.1$ veh/sec and $0.3 \rightarrow 0.2$ veh/sec, respectively. Each plot compares the three models: the proposed model (circles), the mixture model (asterisks) and the multivariate model (triangles). The x-axis displays the space capacity (i.e., ℓ) and the y-axis displays the average absolute difference (i.e., $\bar{e}_{UQ}$ or $\bar{e}_{DQ}$). The top three plots, which consider the spillback probability, have a logarithmic-scaled y-axis. For the experiments with space capacity greater than 30 (i.e., $\ell > 30$), the multivariate model does not conclude within 40 hours, thus these runs are terminated and are not displayed in the plots.

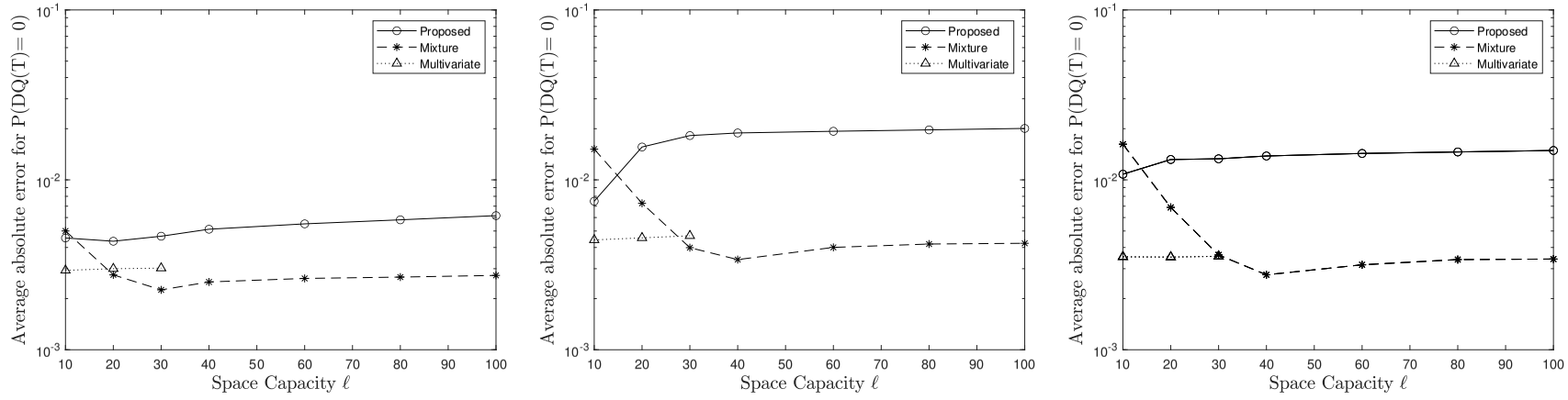
The main insights from Figure 4 are as follows. For most experiments, the multivariate model gives the lowest errors for both $P(UQ(T) = \ell)$ and $P(DQ(T) = 0)$, followed by the mixture model. For the spillback probabilities (i.e., top three plots), both the proposed model and the mixture model have errors that decrease exponentially as the space capacity ℓ increases. As the space capacity increases, the error in the approximation of $P(DQ(T) = 0)$ (bottom three plots) for the proposed model preserves the same order of magnitude. The numerical values of the errors displayed in Figure 4 are provided in Appendix F (Tables 6 and 7). The average (over the 21 experiments) $\bar{e}_{UQ}$ and $\bar{e}_{DQ}$ of the proposed model are 0.0008 and 0.0119, respectively, whereas those of the mixture model are 0.0023 and 0.0073. Compared to the mixture model, the proposed model, on average, gains accuracy in approximating the upstream boundary conditions and loses accuracy in approximating the downstream boundary conditions.

Overall, the multivariate model has the highest accuracy, yet is computationally inefficient for large space capacity values. The proposed model and the mixture model have comparable accuracy. They both perform well for all experiments.

We now compare the multivariate model, the mixture model and the proposed model in terms of computational runtime. Figure 5 compares the runtimes for the 21 experiments. Figure 5(a), 5(b) and 5(c) consider the experiments with time-varying arrival rate $0.2 \rightarrow 0.1$ veh/sec, $0.3 \rightarrow 0.1$ veh/sec and $0.3 \rightarrow 0.2$ veh/sec, respectively. Figure 5(d) plots the average runtime over all three time-varying arrival rate experiments. For each plot, the x-axis displays the space capacity ℓ and the y-axis displays the computational runtime (in seconds). The y-axis is plotted on a logarithmic scale. Each plot considers runtimes of the three models: proposed (circles), mixture (asterisks) and multivariate (triangles). Since the multivariate model does not conclude within 40 hours for experiments with $\ell > 30$, they are not evaluated. For all three arrival rate experiments (i.e., for Figures 5(a), 5(b)



(a) Arrival rate $\lambda(k) = 0.2 \rightarrow 0.1$ veh/sec (b) Arrival rate $\lambda(k) = 0.3 \rightarrow 0.1$ veh/sec (c) Arrival rate $\lambda(k) = 0.3 \rightarrow 0.2$ veh/sec



(d) Arrival rate $\lambda(k) = 0.2 \rightarrow 0.1$ veh/sec (e) Arrival rate $\lambda(k) = 0.3 \rightarrow 0.1$ veh/sec (f) Arrival rate $\lambda(k) = 0.3 \rightarrow 0.2$ veh/sec

Figure 4: Comparison of the average absolute errors for the 21 experiments with time-varying demand

and 5(c)) the following trends hold: the runtime of the multivariate model increases exponentially with ℓ, that of the mixture model increases linearly, while for the proposed model, the computational runtime appears constant. The average runtime over the 21 experiments of the proposed model is 0.09 seconds, whereas that of the mixture model is 0.49 seconds. The average runtime is improved by one order of magnitude.

In summary, for all experiments, the proposed model performs comparably with the mixture and multivariate model in describing the dynamics of the link's boundary conditions. The gain in computational runtime is of at least one order of magnitude and increases as the space capacity increases.

3.2 Validation versus a microscopic traffic simulator

We benchmark the proposed model versus the microscopic traffic simulator Aimsun (TSS; 2014). We consider a signal-controlled single-lane link, depicted in Figure 6, with parameters shown in Table 2. There are two detectors on the link that provide vehicle count data every second. We study the traffic dynamics between the entrance and the exit detectors of the link. The detector denoted entrance detector (resp. exit detector) captures the link's upstream (resp. downstream) boundary conditions. The performance metrics that we consider are the expected inflow and the expected outflow, in vehicles per second, and are computed as the average, over 2 seconds, vehicle count. We consider four experiments with different demand scenarios. The first three experiments have a constant arrival rate over the simulation period. The arrival rates rates are 0.1 veh/sec, 0.2 veh/sec and 0.3 veh/sec, respectively. The fourth experiment has a time-varying arrival rate that mimics arrival patterns from an upstream signal controlled intersection. More specifically, the arrival rate [veh/sec] within every signal cycle (i.e., every 60 seconds) is given by:

$$\lambda(k) = \begin{cases} 0.3 & \text{for } 0 \leq k \leq 40 \text{ sec} \\ 0 & \text{for } 40 < k \leq 60 \text{ sec.} \end{cases} \tag{40}$$

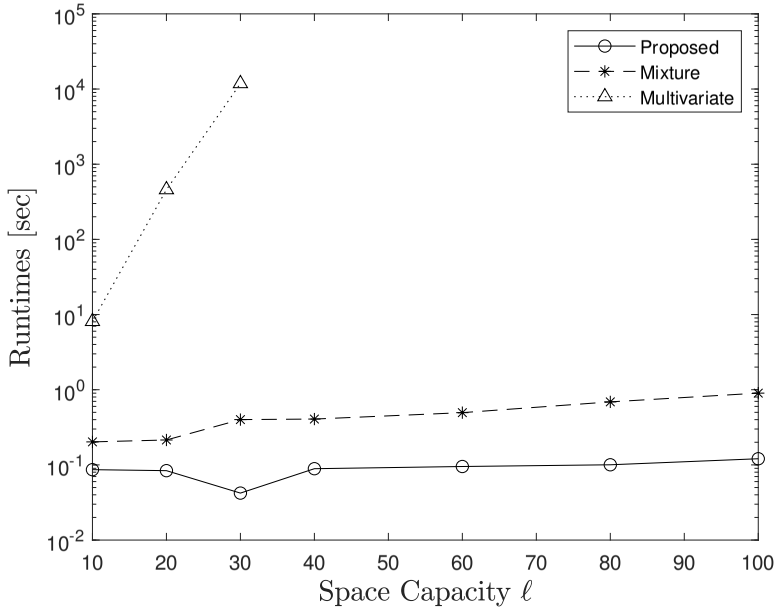
For all experiments, we start with an empty link with a warm-up period of 5 seconds, the free-flow travel time for the arrivals to first reach the entrance detector and the performance metrics are estimated from 200 simulation replications.

We use the link model to describe the state of the roadway that is between the two detectors. The corresponding link parameters used in the proposed model are given in Table 3. All experiments initiate with an empty link and have a duration of 250 seconds. The time-varying service rate [veh/sec] of the link used in the proposed model within every signal cycle (i.e., every 60 seconds) is given by:

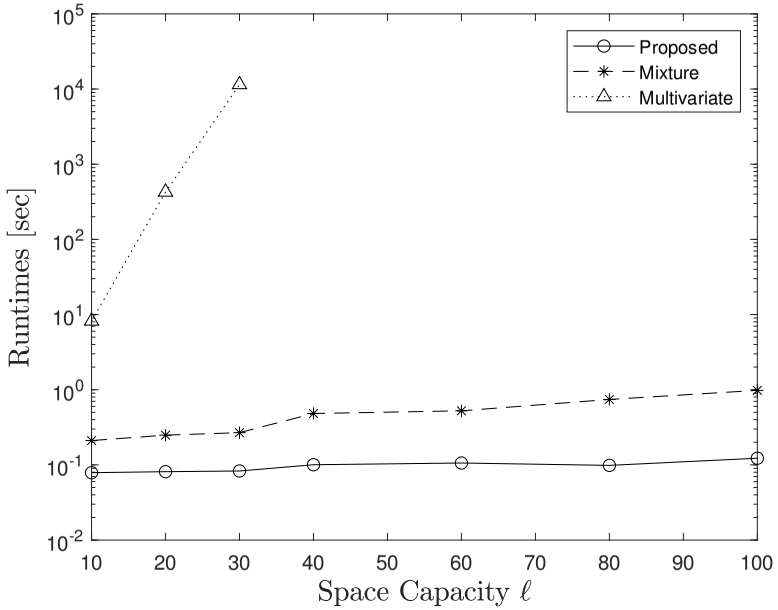
$$\mu(k) = \begin{cases} 0.67 & \text{for } 0 \leq k \leq 48 \text{ sec} \\ 0 & \text{for } 48 < k \leq 60 \text{ sec.} \end{cases} \tag{41}$$

The arrival rates used in the proposed model for the four experiments are set to the same values as in the simulator, i.e., $\lambda(k) = 0.1$ veh/sec, $\lambda(k) = 0.2$ veh/sec, $\lambda(k) = 0.3$ veh/sec for the first three experiments, respectively, and for the fourth experiment, $\lambda(k)$ within every signal cycle is given by Equation (40).

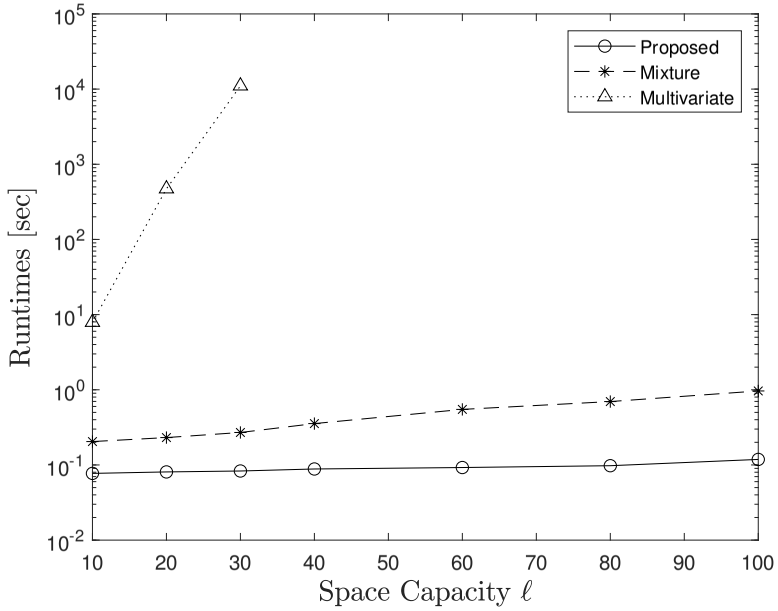
Figures 7-10 compare estimates of the expected inflow (computed at the entrance detector) and of the expected outflow (computed at the exit detector) obtained from the simulator to those obtained



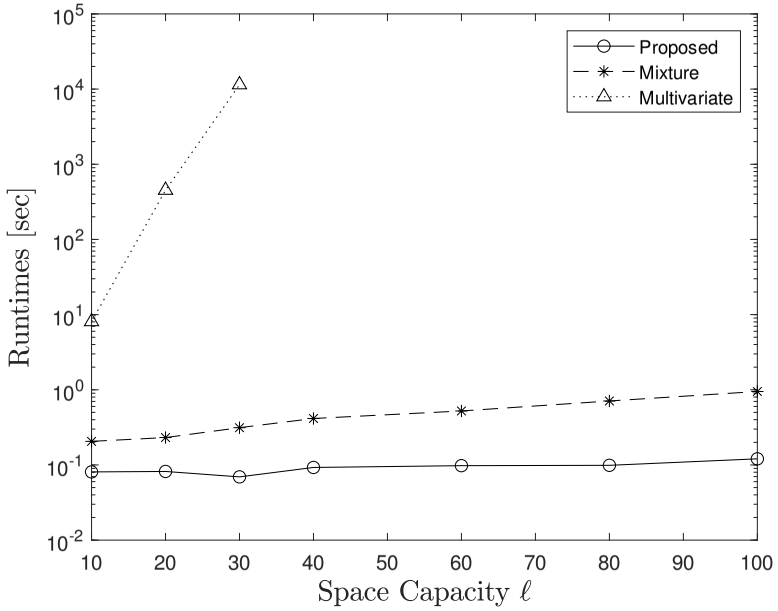
(a) Runtime for time-varying arrival rate $\lambda(k) = 0.2 \rightarrow 0.1$ veh/sec



(b) Runtime for time-varying arrival rate $\lambda(k) = 0.3 \rightarrow 0.1$ veh/sec



(c) Runtime for time-varying arrival rate $\lambda(k) = 0.3 \rightarrow 0.2$ veh/sec



(d) Average runtime over all three time-varying arrival rates

Figure 5: Comparison of the computational runtimes for the 21 experiments with time-varying demand

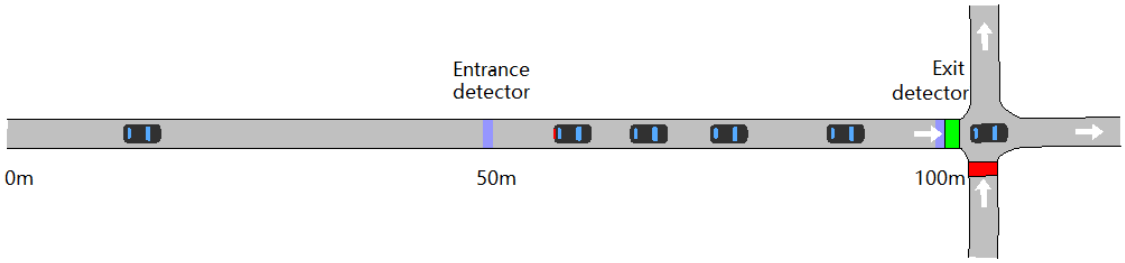


Figure 6: Microscopic simulation model of a single-lane link

Parameter	Value
Link length	100 m
Entrance detector position	50 m
Exit detector position	100 m
Flow capacity	2400 veh/h = 0.67 veh/sec
Maximum speed	36 km/h = 10 m/sec
Downstream control type	fixed-time signal plan
Cycle time	60 sec
Green phase	48 sec
Simulation length	250 sec
Replications number	200
Detection interval	1 sec
Arrival rate $\lambda(k)$	varies by experiment

Table 2: Link parameters used in the microscopic simulator

Parameter	Value
v	0.01 km/sec
w	-0.005 km/sec
$\hat{\rho}$	200 veh/km
$\hat{q}$	2400 veh/h = 0.67 veh/sec
δ	1 sec
ℓ	10
L	50 m
k^{fwd}	5
k^{bwd}	10
$\lambda(k)$	varies by experiment
$\mu(k)$	varies within signal cycle given by Eq. (41)

Table 3: Link parameters used in the proposed model

from the proposed analytical model. For each figure, the left (resp. right) plot depicts the expected inflow (resp. outflow). Each plot displays, as dotted lines, 95% confidence intervals of the simulated estimates. The proposed model results are displayed as a solid line with circles. For all plots, two color bars indicating the upstream arrival and downstream traffic light pattern are displayed. For the arrival pattern bars, the yellow part indicates the period of time when the arrival rate is nonzero whereas the white part indicates the period of time with a zero arrival rate. Note that the bar for upstream arrival pattern in the right plot is shifted by 5 seconds, which is the free-flow travel time of traffic from the upstream to the downstream end of the link.

Figure 7 considers the experiment with a constant arrival rate of 0.1 veh/sec (i.e., the whole arrival pattern bar is yellow). The expected inflow (left plot) remains constant around 0.1 veh/sec for the whole experiment. This indicates that the demand is low enough such that the queue-length does not exceed the entrance detector and thus does not constrain the arrivals to it. The expect outflow (right plot) shows the zero outflow during the red phases of the traffic signal. The outflow pattern during the first traffic signal cycle differs from the others because the link starts off empty and there is only a warm-up period of 5 seconds for the arrivals to first reach the entrance detector. So when the first green phase starts, there is no vehicular queue. For all other cycles, there is a queue of vehicles waiting to leave the link when the green phase starts. The outflow pattern shows how the queue gradually dissipates (i.e., the outflow gradually decreases).

Figures 8 and 9 display the results for the experiments with constant arrival rates of 0.2 veh/sec and 0.3 veh/sec, respectively. The expected outflow patterns (right plots of each figure) are similar to those of the right plot of Figure 7. The left plots of Figures 8 and 9 indicate that as demand increases, the expected inflow may temporarily decrease due to the vehicular queue extending beyond the entrance detector. This phenomenon is also captured by the proposed analytical model.

Figure 10 considers the experiment with time-varying demand. The left and right plot display the expected inflow and outflow pattern, respectively. For the arrival pattern bars, the yellow parts indicate the period of time when the arrival rate is 0.3 veh/sec and white when arrival rate is zero. The expected inflow (left plot) oscillates between 0.3 and 0 veh/sec as expected. The expect outflow (right plot) shows the zero outflow during the red phases of the traffic signal. The outflow pattern during the first traffic signal cycle differs from the others because the link starts off empty and there is only a warm-up period of 5 seconds for the arrivals to first reach the entrance detector. So when the first green phase starts, there is no vehicular queue. For all other cycles, since the downstream traffic light turns red before the arrival pattern goes to zero, there is a residual queue of vehicles left on the link that is ready to depart as soon as the downstream traffic light turns green. This departure of the residual queue is reflected in the quick spikes of outflow that occur shortly after the traffic light turns green. This phenomenon is well captured by the proposed model.

We now consider an experiment with platoon arrivals. This experiment serves to evaluate the ability of the proposed model, and in particular its use of a time-varying Poisson arrival process, to approximate the link's boundary conditions under platoon arrival patterns. We consider two tandem single-lane links as shown in Figure 11. Both links are controlled by a fixed-time signal plan with a 60 second cycle time. The cycle is composed of a single green phase followed by a single red phase. The offset is set to zero, i.e., both intersections start their green phase at the same time. The duration of the green phase for the upstream (resp. downstream) intersection is 40 (resp. 48) seconds. The maximum speed of both links as well as all other link attributes are the same as in Table 2. Arrivals from the upstream

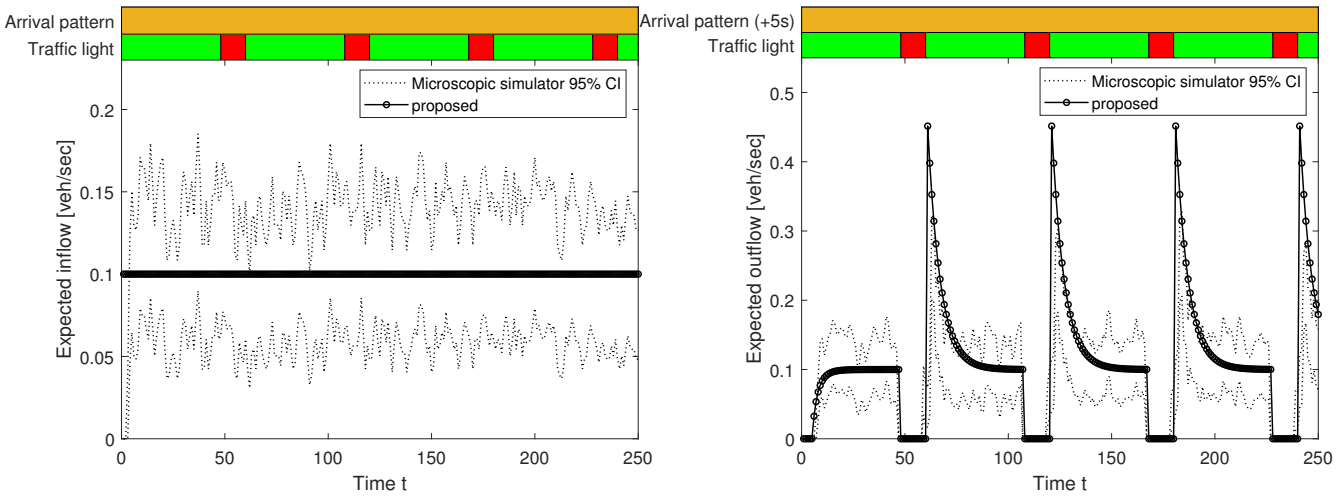


Figure 7: Comparison of the expected inflow and outflow for the experiment with arrival rate $\lambda = 0.1$ veh/sec

link to the downstream link form a platoon created by the upstream signalized intersection.

Just as in our previous experiments, the downstream link has the entrance and the exit detectors, which are used to estimate expected inflows and outflows. We study the traffic dynamics between the entrance and the exit detectors of the downstream link. Since node modeling is not the focus of this paper, we use a third detector to estimate the downstream link's arrival rate $\lambda(k)$ of the analytical model. More specifically, $\lambda(k)$ is estimated by the flow observed at the detector labeled detector 1 in Figure 11 and shifted in time by the free-flow travel time between detector 1 and the entrance detector (5 seconds). The parameters of the proposed analytical model are displayed in Table 4. We start with an empty link with a warm-up period of 10 seconds, the free-flow time for the arrivals to first reach the entrance detector, and the performance metrics are estimated from 200 simulation replications.

The left (resp. right) plot of Figure 12 shows the expected inflow (resp. outflow). For both plots, the analytical approximations of both the expected inflow and of the expected outflow fall within the 95% confidence intervals of the simulated estimates. This shows the ability of the proposed model to capture the impact of platoon arrivals on the links' boundary conditions.

In summary, the comparisons to a microscopic traffic simulator indicate that the proposed model can accurately approximate the links' boundary conditions for realistic traffic situations such as signalized links and platoon arrival patterns.

4 Optimization case study

In this section, we evaluate and benchmark the computational efficiency of the proposed model. We use it to tackle a signal control optimization problem for the Swiss city of Lausanne. The signal control problem considered is the same as that studied in Lu and Osorio (2018). Section 4.1 formulates the problem. Section 4.2 compares the performance of the proposed model with that of three

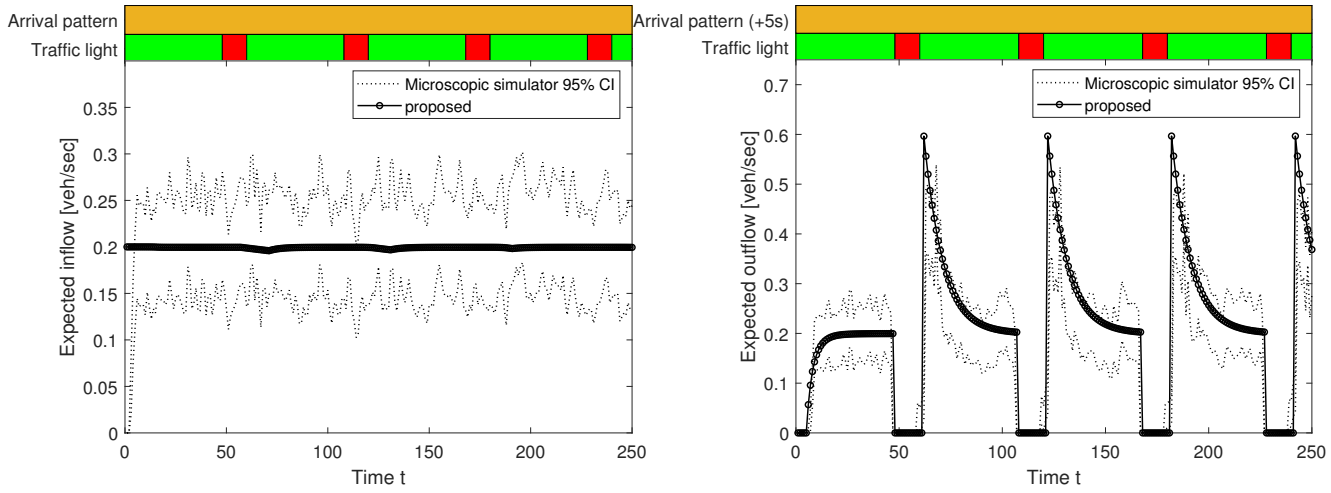


Figure 8: Comparison of the expected inflow and outflow for the experiment with arrival rate $\lambda = 0.2$ veh/sec

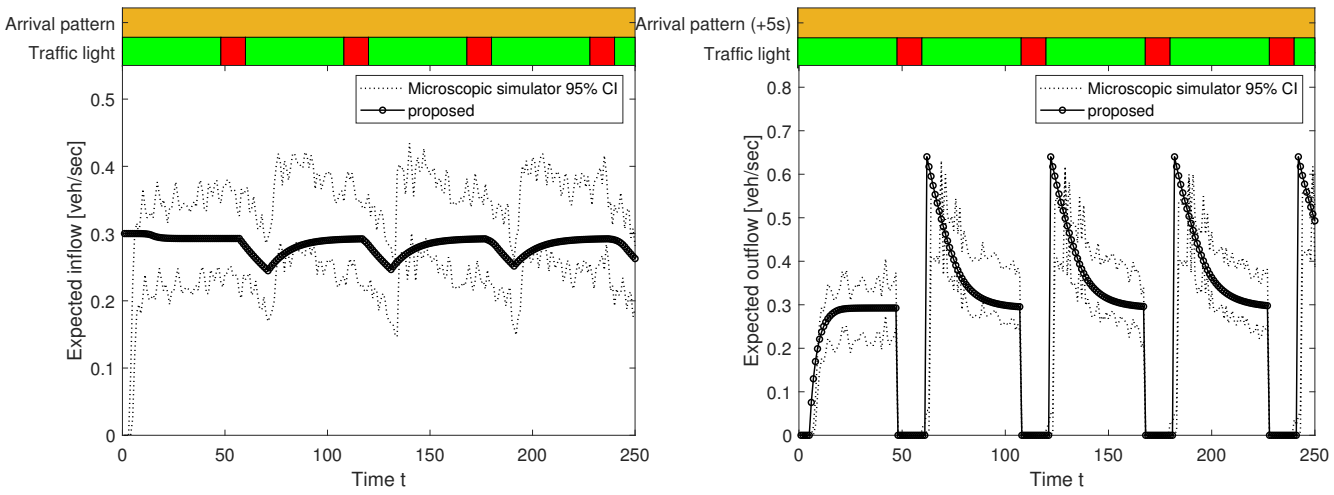


Figure 9: Comparison of the expected inflow and outflow for the experiment with arrival rate $\lambda = 0.3$ veh/sec

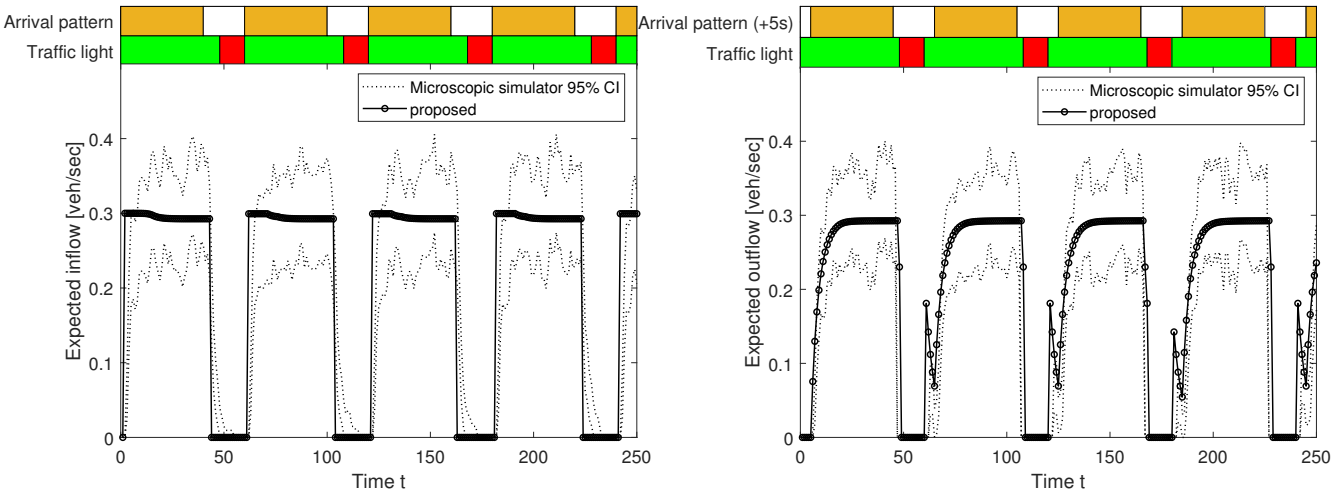


Figure 10: Comparison of the expected inflow and outflow for the experiment with alternating arrival rate between 0.3 veh/sec and 0 veh/sec

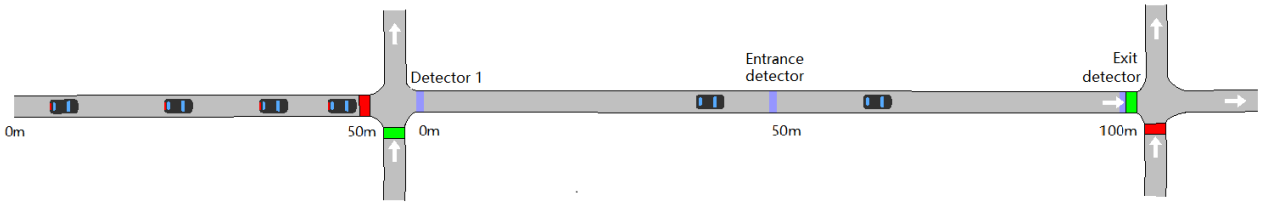


Figure 11: Microscopic simulation model for platoon arrival experiments

Parameter	Value
v	0.01 km/sec
w	-0.005 km/sec
$\hat{\rho}$	200 veh/km
$\hat{q}$	2400 veh/h = 0.67 veh/sec
δ	1 sec
ℓ	10
L	50 m
k^{fwd}	5
k^{bwd}	10
$\lambda(k)$	estimates from detector 1
$\mu(k)$	varies within signal cycle given by Eq. (41)

Table 4: Link parameters used in the proposed model

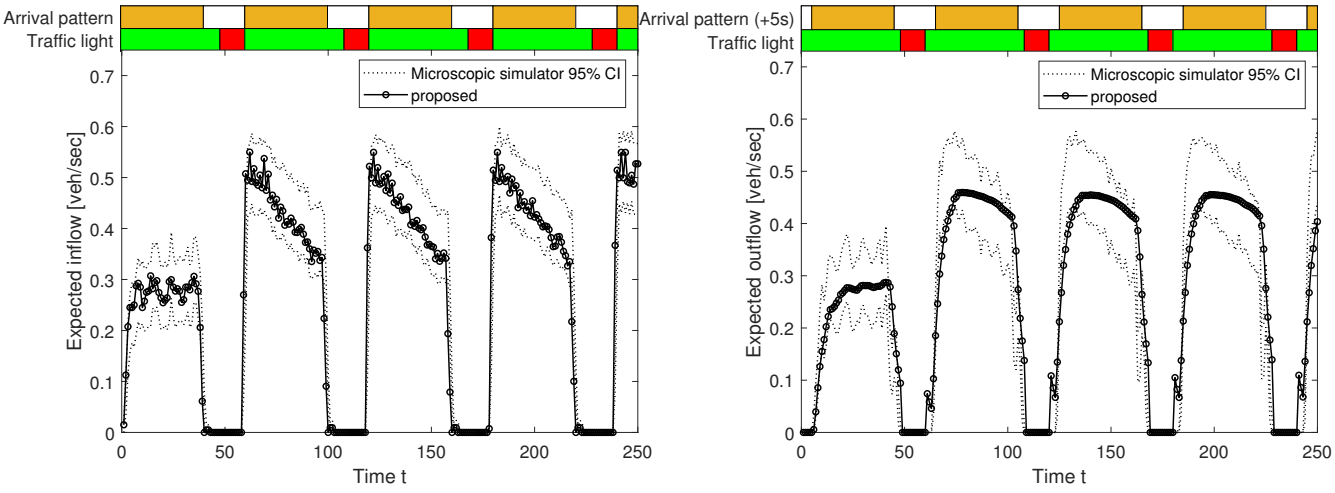


Figure 12: Comparison of the expected inflow and outflow for the tandem link experiment

other approaches: (i) the mixture model Lu and Osorio (2018), (ii) the deterministic intelligent link transmission model (ILTM) (Himpe et al.; 2016), and (iii) a widely used commercial signal control software Synchro (Trafficware; 2011). The ILTM is a network loading model that combines an efficient link transmission model (Himpe et al.; 2016) with the node model of Tampère et al. (2011).

4.1 City-scale signal control

The Lausanne network consists of 652 links, 902 lanes and 231 intersections. The network model of the stochastic microscopic simulator is shown in Figure 13. We consider a fixed-time signal control problem in which we determine the signal plans of 17 intersections distributed throughout the network (displayed as squares in Figure 13). The signal plans of the 17 intersections are determined jointly. The problem is a fixed-time signal control problem for the evening peak period 5:00-5:30pm. The decision variables are the green splits of the signal phases of the 17 intersections. All other control variables (e.g., cycle times and offsets) are fixed. This leads to a total of 99 endogenous signal phase variables (i.e., the decision vector is of dimension of 99). We use the following notation.

b_d	ratio of available cycle time to total cycle time for intersection d;
x	vector of green splits;
$x(j)$	green split of signal phase j;
x_{LB}	vector of lower bounds for green splits;
$\mathcal{D}$	set of intersection indices;
$\mathcal{P}_D(d)$	set of endogenous signal phase indices of intersection d;
$\mathcal{L}$	set of all lanes;
$\tilde{T}$	total number of one-minute time intervals;
N_1	number of lanes, i.e., cardinality of $\mathcal{L}$.

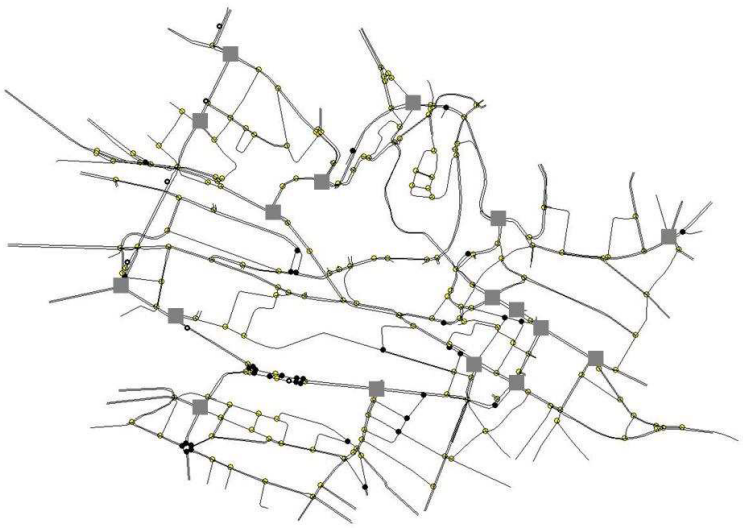


Figure 13: Lausanne network model

The problem is formulated as follows:

$$\min_{x} f(x) = \frac{1}{\tilde{T} N_l} \sum_{i \in \mathcal{L}} \sum_{\hat{t}=1}^{\tilde{T}} P(UQ_i(\hat{t}; x) = \ell_i) \tag{42}$$

subject to

$$\sum_{j \in \mathcal{P}_D(d)} x(j) = b_d, \quad \forall d \in \mathcal{D} \tag{43}$$

$$x \geq x_{LB}. \tag{44}$$

The decision vector, x, is the green splits of the signal controlled lanes. Constraint (43) ensures that, for every intersection, the sum of the green times equals the available cycle time. Constraint (44) sets lower bounds, which are set to 4 seconds in this case study. The spillback probability of lane i at integer time $\hat{t}$ under signal plan x is denoted by $P(UQ_i(\hat{t}; x) = \ell_i)$. Therefore, the objective function is the average (over space and over time) spillback probability. The goal is to find a signal plan that minimizes the spatial and temporal occurrence of spillbacks.

To evaluate the performance of the signal plans proposed by the various methods, we use a microscopic model of the Lausanne network (Dumont and Bert; 2006), which is calibrated for the evening peak period demand. When evaluating the performance of a signal plan, the simulator accounts for how route choices and link demand can vary with signal plan changes. Route choice probabilities depend mainly on route travel times. More specifically, we use the default multinomial logit route choice model of Aimsun (TSS; 2014) that is based on the route travel time.

For all the three analytical models (i.e., the proposed model and the mixture model and the deterministic ILTM), each lane in the Lausanne network is modeled as an individual link and the decision variables, which are the green splits of the signal plans, determine the service rate of each lane. The

service rates are computed according to Eq. (46) of Lu and Osorio (2018). The probabilistic analytical models (i.e., the proposed model and the mixture model) are link models. In order to use them for network optimization, they assume demand for each link is exogenous and describe the within-link dynamics, they do not describe the across-link (i.e., node) dynamics. We follow the approach of Lu and Osorio (2018) to compute the exogenous lane demands, which depends on (i) the external arrival rates to the network and (ii) the exogenous turning probabilities obtained from Osorio (2010, Chapter 4). The deterministic ILTM is a complete network loading model that embeds the well-established node model of Tampère et al. (2011), and hence the lane demands are endogenous. The required exogenous inputs for the algorithm are (i) network travel demand (i.e., OD matrix) and (ii) route choice (e.g., turning fractions). We use the same values for the turning fractions as for the other analytical probabilistic models. The above problem is solved with the proposed model and with the mixture model using the *interior-point* algorithm of the *fmincon* routine of Matlab (MATLAB; 2016).

Since the ILTM is a deterministic model and it is not intended to approximate the spillback probabilities that define the objective function (42). Thus, we use two alternate objective functions for ILTM:

1. The average, over time and over the network lanes, proportion of a lane that is occupied by vehicles:

$$f(x) = \frac{1}{\tilde{T} N_l} \sum_{i \in \mathcal{L}} \sum_{\hat{t}=1}^{\tilde{T}} \left(c_i^{up}(\hat{t}; x) - c_i^{down}(\hat{t}; x) \right) / \ell_i, \tag{45}$$

   where $c_i^{up}(\hat{t}; x)$ (resp. $c_i^{down}(\hat{t}; x)$) is the cumulative number of vehicles that cross the upstream (resp. downstream) end of link i up until time $\hat{t}$ given signal plan x, which is the output metric produced by the deterministic ILTM model. This objective function is between 0 and 1. It indicates increasing levels of congestion, and increasing occurrence of spillbacks.

2. The average, over time and over the network lanes, link travel time:

$$f(x) = \frac{1}{\tilde{T} N_l} \sum_{i \in \mathcal{L}} \sum_{\hat{t}=1}^{\tilde{T}} d_i(\hat{t}; x), \tag{46}$$

   where $d_i(\hat{t}; x)$ is the travel time of lane i at time $\hat{t}$ given signal plan x, which is an output metric produced by the deterministic ILTM. This is a common choice of objective function for deterministic models. For fixed demand, as this objective function increases, so do spillback probabilities.

We also benchmark the proposed model with a signal plan derived by the widely used commercial signal control software Synchro (Trafficware; 2011). The Synchro software is based on a deterministic macroscopic traffic model, it does not solve the same optimization problem (42)-(44). For details on how the Synchro signal plan is derived, we refer the reader to Section 5.3 of Osorio and Chong (2015). To illustrate the congestion levels considered in this case study, we have included in Appendix G congestion statistics of the network under they Synchro plan. Based on traffic conditions defined in Dowling et al. (2008, Class IV in Exhibit 2 Urban Street Level of Service (LOS)), 27% of the network under the Synchro signal plan are still operating near or above capacity.

Initial point	1	2	3	4
Mixture	145.0	146.1	144.4	149.4
ILTM	10.6	10.7	10.7	10.8
Proposed	1.3	1.3	1.3	1.3

Table 5: Average runtime (in min) per iteration of the signal control optimization algorithm

For all methods, the maximum runtime is set to 24 hours. If the algorithm does not converge to a local optimal solution within the time limit, the algorithm is terminated and the current iterate is used as the final solution.

4.2 Numerical results

Problem (42)-(44) for the proposed and for the mixture model, and Problems (45) and (46) with constraints (43) and (44) for ILTM are solved considering four different initial points. The initial points are drawn uniformly randomly from the feasible region (Equations (43)-(44)) using the sampling code of Stafford (2006).

For the proposed model and for ILTM, all four optimization runs (i.e., one for each initial point) conclude within the time limit. Actually, they all finish within 2.5 hours for the proposed model and within 5.5 hours for ILTM. For the mixture model, the algorithms do not converge within the time limit. Table 5 compares the average computation time (in minutes) per algorithmic iteration. Each column of Table 5 corresponds to a different initial point. Since we consider two different objective functions for ILTM, Table 5 displays the average computation time, averaged over the two optimization problems. Table 5 indicates the computation time of all models does not vary significantly across initial points. For the proposed model, the average runtime per iteration is in the order of 1 minute. For ILTM, it is in the order of 10 minutes, and for the mixture model it is in the order of 2.4 hours (i.e., 146 minutes). Thus, the proposed model improves the runtime by 1 order of magnitude compared to ILTM and by 2 orders of magnitude compared to the mixture model. Note however, that, unlike the probabilistic models (proposed or mixture) ILTM is a full network model that embeds an endogenous node model. This added realism comes, of course, with an increase in the computational runtime.

We now compare the performance of the derived signal plans. For a given signal plan, we embed it within the microscopic simulation software, and evaluate 50 simulation replications. Each replication consists of a warm-up period of 15 minutes followed by a simulation period of 30 minutes. For each simulation replication, we estimate the objective function (Eq. (42)), which is the average (over lanes) proportion of time (over 30 minutes) a lane is full. For each signal plan, we construct a cumulative distribution function (cdf) of these 50 objective function observations.

Each plot of Figure 14 considers a different initial signal plan and plots five cdf curves: one for the initial signal plan (dashed line), one for the solution derived by the proposed model (solid line), one for the solution derived by the mixture model (dot-dashed line) and two for the solutions derived by the deterministic ILTM with different objective functions (dotted lines): ILTM-1 (black circle-dotted line) represents the result for ILTM with objective function (45), while ILTM-2 (red cross-dotted line) for ILTM with objective function (46). The x-axis displays the objective function realizations (i.e., average (over all the lanes in the network) proportion of time a lane is full). The y-axis displays the

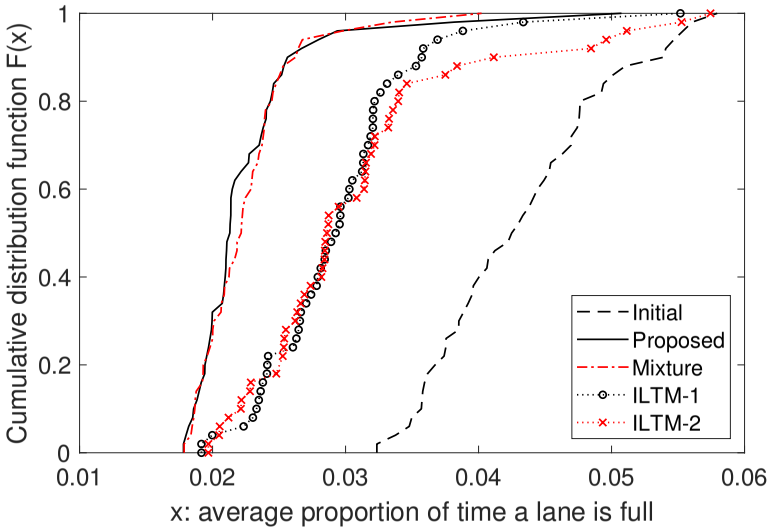
proportion of the 50 simulation replications that have objective function realizations smaller than x. Thus, the more a cdf curve is shifted to the left, the better the performance of the corresponding signal plan.

For all four plots (Fig. 14(a)-14(d)), the cdf curves of the derived signal plans, from the mixture model, from the proposed model and from ILTM, are to the left of the initial signal plan. Thus, all models identify signal plans that outperform the initial signal plans. For all four plots, the cdf curves of the derived signal plans from the stochastic models (i.e, the proposed and mixture model) are to the left of the derived signal plans from ILTM. Thus, the stochastic models can identify signal plans that outperform the signal plans derived by ILTM (with both objective functions). The average (over the derived signal plans) objective function realizations of the proposed model is 2% better than that of the mixture model and 25% better than that of the ILTM.

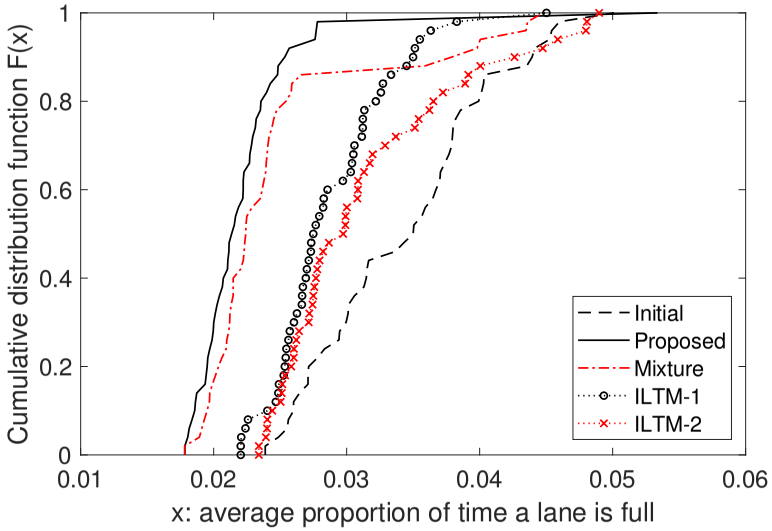
Figure 15 compares the performance of the signal plans in terms of the average (over all lanes in the network) proportion of the lane that is occupied by vehicles, which corresponds to the first objective function used by ILTM (i.e., objective function (45) and method ILTM-1). This figure has a similar layout as Figure 14. The figure compares the cdf curves of the different signal plans. As before, the more a cdf curve is shifted to the left, the better its performance (i.e., the higher the proportion of simulation replications, out of the 50, that have low average proportion of the lane that is occupied by vehicles). All four plots in Figure 15 indicate that all derived signal plans outperform their corresponding initial signal plans. The derived signal plans from both probabilistic models (mixture and proposed) have similar performance and they outperform the signal plans derived by the deterministic ILTM with both objective functions. The average (over the derived signal plans) performance, in terms of the average proportion of the lane that is occupied by vehicles, of the proposed model is 1% worse than that of the mixture model and 10% better than that of the ILTM.

Figure 16 compares the performance of the signal plans in terms of the average trip travel time, which corresponds to the second objective function used by ILTM (i.e., objective function (46) and method ILTM-2). For all initial points, all models yield signal plans that outperform the initial points. The signal plans derived by the proposed model and by the mixture model have similar performance and they outperform the signal plans derived by the deterministic ILTM with both objective functions. The average (over the derived signal plans) performance, in terms of the average trip travel time, of the proposed model is 0.5% worse than that of the mixture model and 10% better than that of the ILTM.

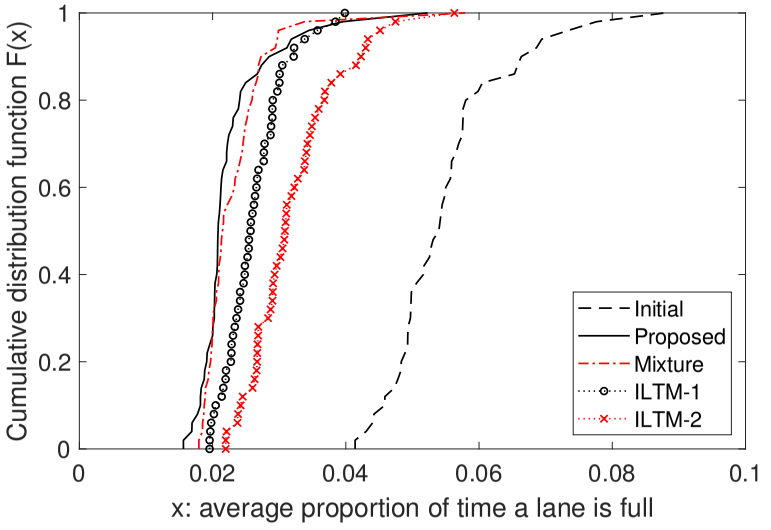
We compare the performance of the signal plans derived by the propose model with that of a signal plan derived by the widely used commercial signal control software Synchro (Trafficware; 2011). Figures 17, 18 and 19 display, respectively, the three performance metrics of interest: average proportion of time a lane is full, average proportion of the lane that is occupied by vehicles and average trip travel time. Each figure contains nine cdf curves: four black dashed lines for the four initial points, four solid thin lines for the four solutions derived by the proposed model and one solid thick line for the signal plan proposed by Synchro. For all figures, the left-most curves are the ones corresponding to the signal plans derived by the proposed model. In other words, for all three performance metrics, the signal plans derived by the proposed model outperform both the initial signal plans and the signal plan derived by Synchro. For each figure, the performance of the four initial points varies significantly, while that of the proposed signal plans is similar. In other words, the proposed method is robust to the quality of the initial points.



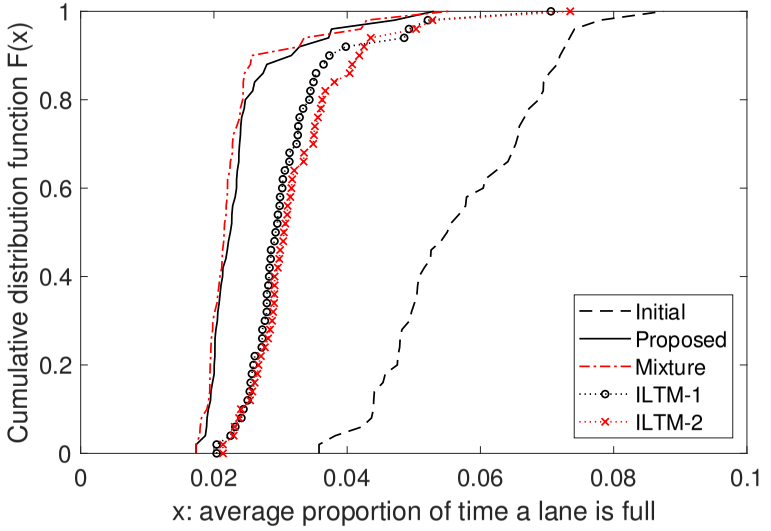
(a) Initial point 1



(b) Initial point 2

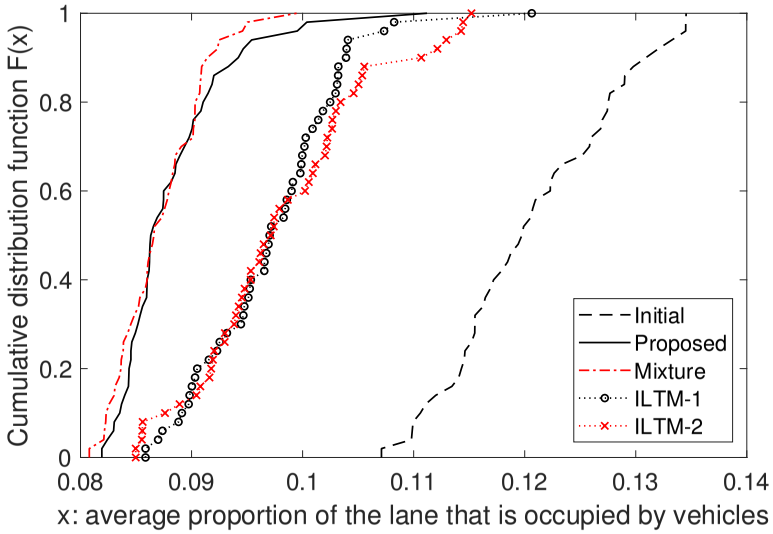


(c) Initial point 3

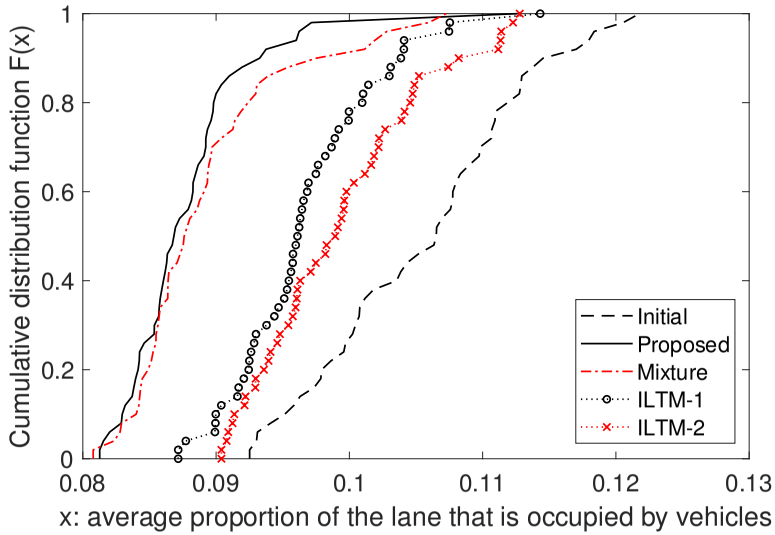


(d) Initial point 4

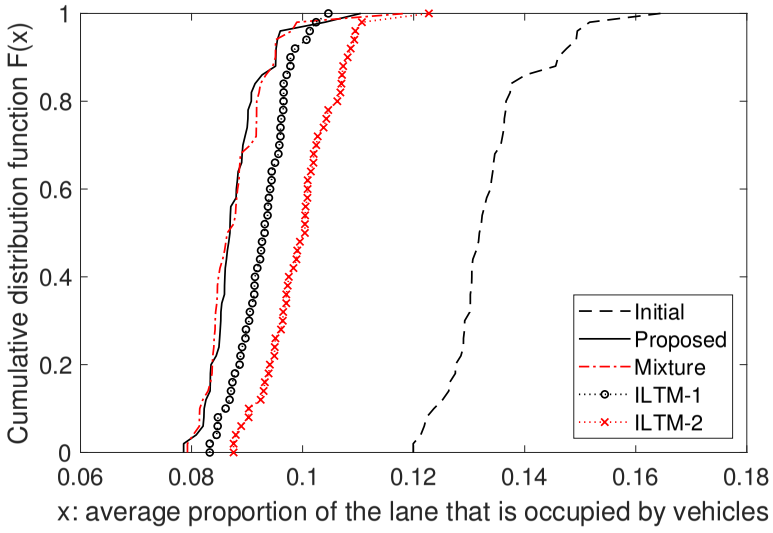
Figure 14: Cumulative distribution functions of the average, over all lanes in the network, proportion of time a lane is full considering different initial signal plans



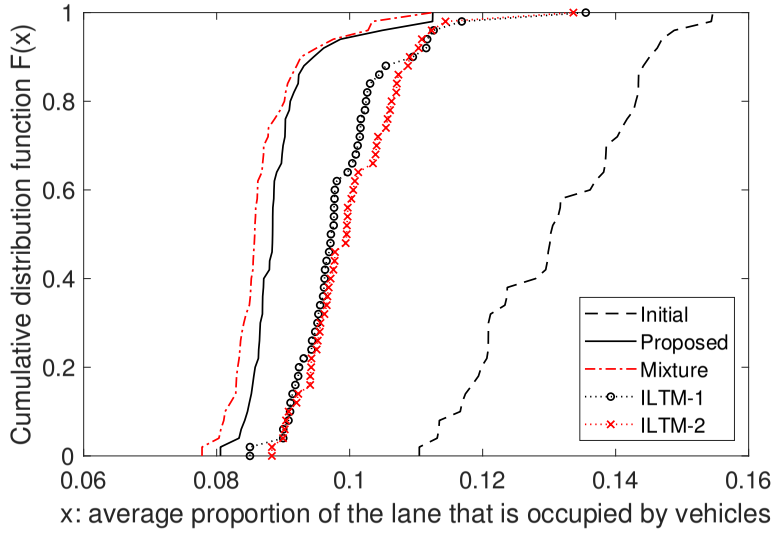
(a) Initial point 1



(b) Initial point 2

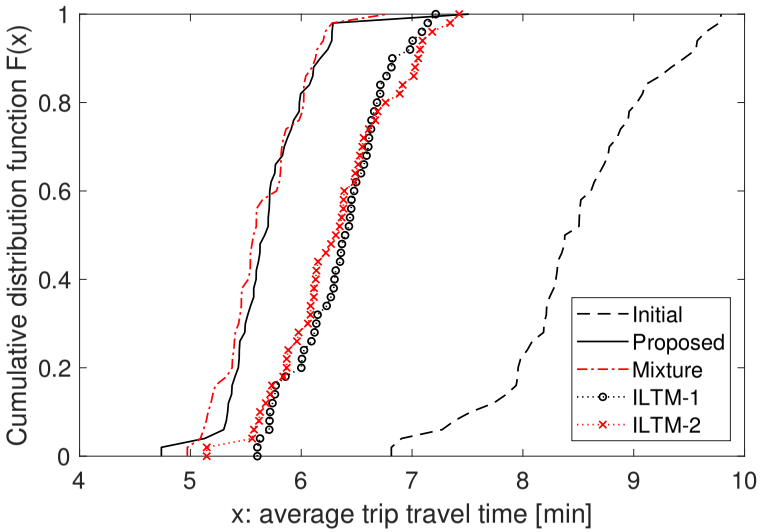


(c) Initial point 3

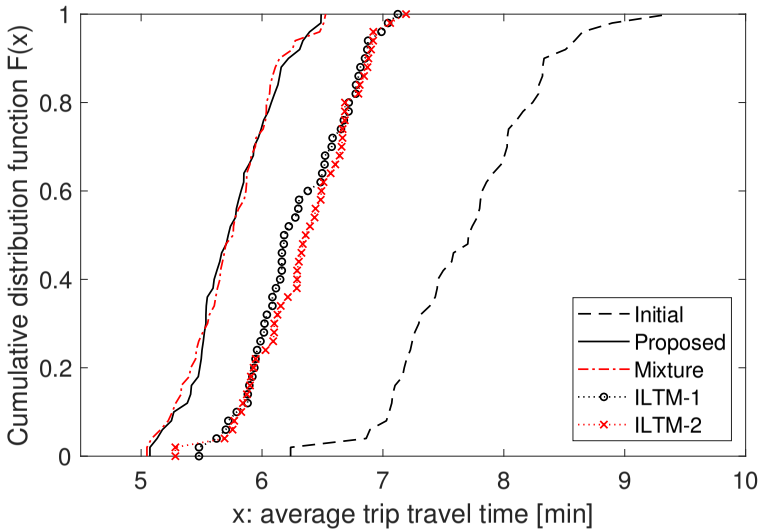


(d) Initial point 4

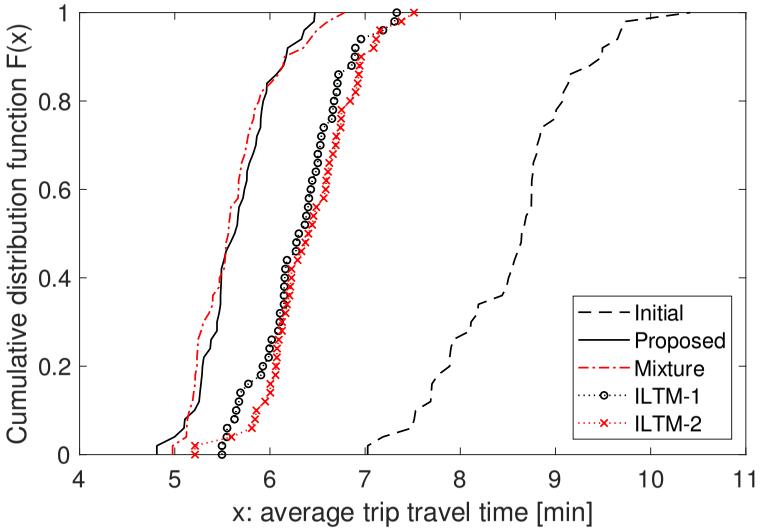
Figure 15: Cumulative distribution functions of the average proportion of the lane that is occupied by vehicles considering different initial signal plans



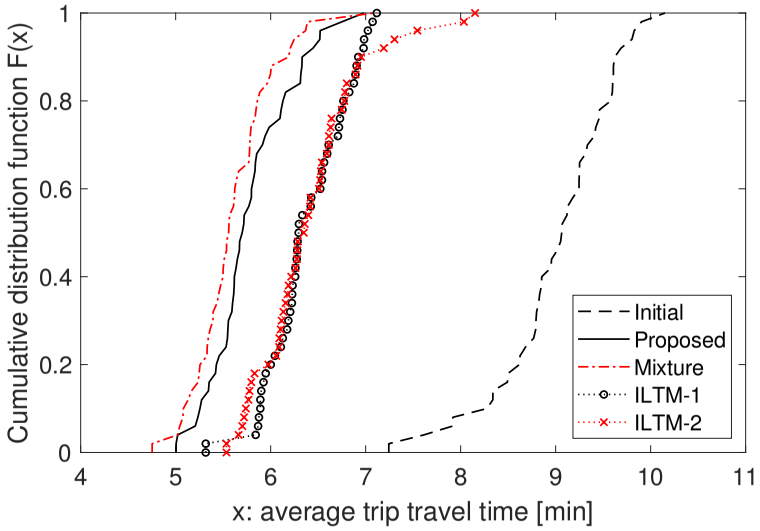
(a) Initial point 1



(b) Initial point 2



(c) Initial point 3



(d) Initial point 4

Figure 16: Cumulative distribution functions of the average trip travel time considering different initial signal plans

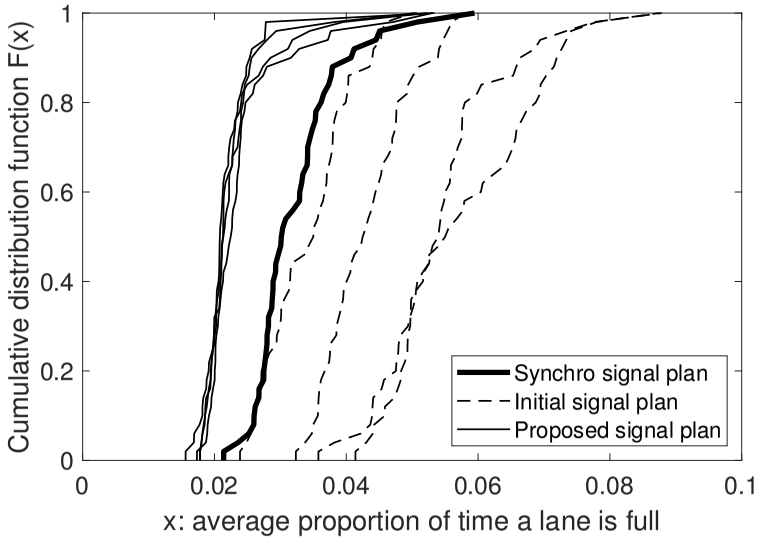


Figure 17: Cumulative distribution functions of the average proportion of time a lane is full

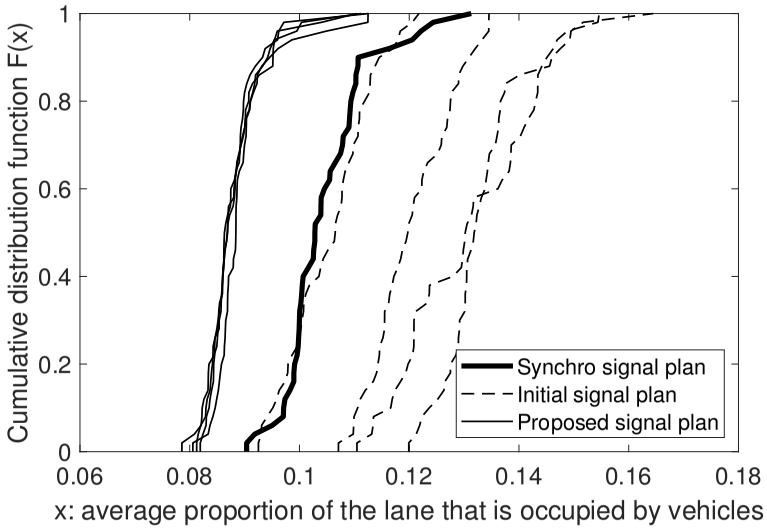


Figure 18: Cumulative distribution functions of the average proportion of the lane that is occupied by vehicles

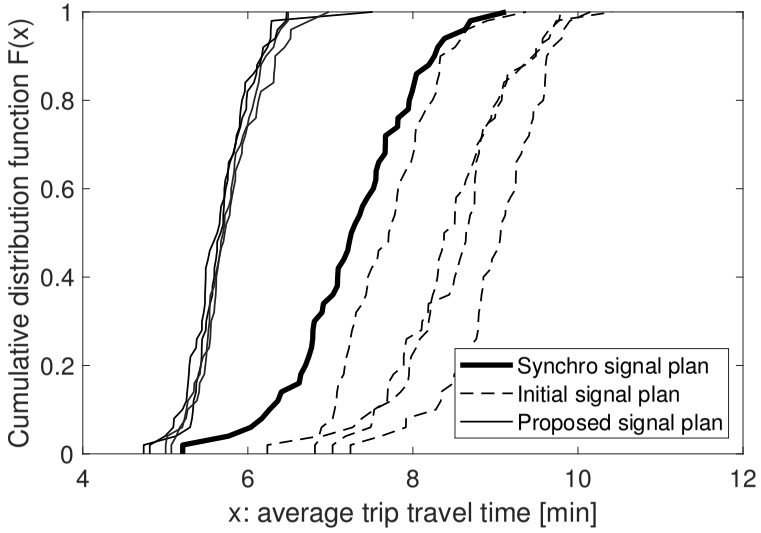


Figure 19: Cumulative distribution functions of the average trip travel time

In summary, compared to the mixture model, the proposed model improves the average runtime by 2 orders of magnitude, and yields signal plans with either improved or similar performance. Compared to the deterministic ILTM, both the proposed and the mixture model are able to find signal plans with improved performance. This case study illustrates the scalability and computational efficiency of the proposed model. The proposed model is suitable for large-scale network analysis and optimization.

5 Conclusion

This paper formulates an analytical probabilistic stochastic model that is scalable and suitable for large-scale network optimization. The main idea of the proposed model is to describe the link's boundary conditions with only two scalar probabilities instead of tracking the full marginal, or full joint, distributions. More specifically, while the dimension of the state space of the models of Osorio and Flötteröd (2015) and of Lu and Osorio (2018) scale cubically and linearly, respectively, with the link's space capacity, the proposed model has a constant dimension of two. Thus, it scales independently of the link attributes such as the link's space capacity. This makes it suitable for large-scale network analysis and optimization. The model is validated versus stochastic simulation results from a simulation-based implementation of a stochastic link transmission model. The model's accuracy is comparable to that of Osorio and Flötteröd (2015) and of Lu and Osorio (2018), while being more computationally efficient by at least two orders of magnitude. The model is also validated versus a microscopic traffic simulator, and it can accurately approximate the link's boundary conditions for realistic traffic situations such as platoon-like arrivals. The proposed model is then used to address a signal control problem for the city of Lausanne (Switzerland). The derived solutions are benchmarked with those derived by the mixture model of Lu and Osorio (2018) and the ILTM of Himpe et al. (2016). The derived signal plans from both the proposed model and the mixture model have similar performance, considering various performance metrics. They both outperform the initial signal plans, the signal plans derived by ILTM and a signal plan proposed by a widely used commercial software. Compared to the model of Lu and Osorio (2018), the proposed model reduces computational runtime by two orders of magnitude.

Although the gain in computational efficiency is substantial, it does come at the cost of information loss. More specifically, this work builds upon past probabilistic link models that yield either the full joint distribution or the full marginal distributions of the respective queues. Instead, the present work yields two scalar probabilities over time. For applications where a more detailed description of the link's boundary conditions is needed, past probabilistic link model formulations or even deterministic link model formulations may be an appropriate choice.

Of interest for future work is the study of how the relaxation time formulation proposed in this work contributes to describe non-equilibrium traffic dynamics. Future work also focuses on the formulation of scalable stochastic network models, the goal is to be able to recover the joint distribution of a path or a network. First, there is a need to formulate scalable probabilistic node models that are consistent with their deterministic counterparts. Osorio et al. (2011) includes a two-link probabilistic node model that accounts for the dependencies of the links' boundary conditions across a node. It yields the joint distribution of the downstream boundary conditions of the upstream link and the upstream boundary conditions of the downstream link. The extension of this formulation to nodes with multiple

incoming and outgoing links is of interest for future work. Second, scalable and computationally efficient network model formulations are required. Consider a network with n links, directly coupling the proposed link model with the node model of Osorio et al. (2011) would yield a complexity of $\mathcal{O}(2^n)$, which is not scalable. Possible techniques to achieve scalability include network decomposition (Flötteröd and Osorio; 2017) and aggregation-disaggregation (Osorio and Yamani; 2017; Osorio and Wang; 2017). Similar to the proposed work, the stationary work of Osorio and Wang (2017) and the transient work of Osorio and Yamani (2017) identify the empty state and the full state as the two critical states to track to describe boundary conditions. However, those models consider a general queueing-theoretic perspective without vehicular traffic-theoretic considerations.

6 Acknowledgment

The work of J. Lu and C. Osorio is partially supported by the U.S. National Science Foundation under Grant No. 1562912. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The work of J. Lu is also partially supported by the UPS Doctoral fellowship. The authors thank Prof. Gunnar Flötteröd (KTH) for providing access to the simulator. The authors thank Prof. Chris Tampère and Dr. Willem Himpe for discussions about the deterministic ILTM and for providing the access to the model code.

Appendices

A Derivation of Equation (20)

In this section, we derive the result of Equation (20), which is the limit of the limit of $\tau_{DQ}(k)$. Equation (16) defined $\tau_{DQ}(k)$. We first calculate the limits of $\tau_{DQ,1}$ (given by Eq. (16b)) and $\tau_{DQ,2}$ (given by Eq. (16c)) independently as follows:

$$\lim_{\mu(k)\to 0} \tau_{DQ,1} = \lim_{\mu(k)\to 0} (1 - \alpha_1 e^{-\rho_{DQ}(k)}) \times \frac{\mu(k)(1-\rho_{DQ}(k))^2}{(1+\rho_{DQ}(k))} \tag{47}$$

$$= \lim_{\mu(k)\to 0} (1 - \alpha_1 e^{-\rho_{DQ}(k)}) \times \lim_{\mu(k)\to 0} \frac{\mu(k)(1-\rho_{DQ}(k))^2}{(1+\rho_{DQ}(k))} \tag{48}$$

$$= \lim_{\mu(k)\to 0} (1 - \alpha_1 e^{-\lambda_{DQ}(k)/\mu(k)}) \times \lim_{\mu(k)\to 0} \frac{\mu(k)^2(1-\rho_{DQ}(k))^2}{\mu(k)(1+\rho_{DQ}(k))} \tag{49}$$

$$= (1-0) \times \lim_{\mu(k)\to 0} \frac{(\mu(k) - \mu(k)\rho_{DQ}(k))^2}{(\mu(k) + \mu(k)\rho_{DQ}(k))} \tag{50}$$

$$= 1 \times \lim_{\mu(k)\to 0} \frac{(\mu(k) - \lambda_{DQ}(k))^2}{(\mu(k) + \lambda_{DQ}(k))} \tag{51}$$

$$= \frac{(0 - \lambda_{DQ}(k))^2}{(0 + \lambda_{DQ}(k))} \tag{52}$$

$$= \lambda_{DQ}(k) \tag{53}$$

$$\lim_{\mu(k)\to 0} \tau_{DQ,2} = \lim_{\mu(k)\to 0} \alpha_2 \mu(k) \left| \frac{P(DQ(k-1)=0) - P(DQ_k=0)}{\ell(1 - P(DQ_k=0))} \right|^{1/5} \tag{54}$$

$$= \alpha_2 \lim_{\mu(k)\to 0} \left| \frac{P(DQ(k-1)=0) - P(DQ_k=0)}{\ell(1 - P(DQ_k=0))} \right|^{1/5} \lim_{\mu(k)\to 0} \mu(k) \tag{55}$$

$$= \alpha_2 \left| \frac{P(DQ(k-1)=0) - \lim_{\mu(k)\to 0} P(DQ_k=0)}{\ell(1 - \lim_{\mu(k)\to 0} P(DQ_k=0))} \right|^{1/5} * 0 \tag{56}$$

$$= \alpha_2 \left| \frac{P(DQ(k-1)=0) - 0}{\ell(1-0)} \right|^{1/5} * 0 \tag{57}$$

$$= 0 \tag{58}$$

Equation (56) is obtained by taking the limit operation to both the numerator and denominator for the term within the absolute operator (i.e., $\frac{P(DQ(k-1)=0)-P(DQ_k=0)}{\ell(1-P(DQ_k=0))}$). Note that $P(DQ(k-1)=0)$ is the initial state of DQ at time step k and hence it is independent of the service rate during time step k (i.e., $\mu(k)$). Moreover, when $\mu(k)$ approaches zero and $\rho_{DQ}(k)$ goes to infinity (Eq. (14)), $P(DQ_k=0)=0$ according to Equation (15b). Equation (56) is hence obtained. Combining the results of Equations (53) and (58), we obtain:

$$\lim_{\mu(k)\to 0} \tau_{DQ}(k) = \lim_{\mu(k)\to 0} \tau_{DQ,1} + \lim_{\mu(k)\to 0} \tau_{DQ,2} \tag{59}$$

$$= \lambda_{DQ}(k) + 0 \tag{60}$$

$$= \lambda_{DQ}(k) \tag{61}$$

B Estimation of the scalar coefficients in $\tau_{DQ}(k)$ (Equation (16))

This appendix describes the procedure to fit the exogenous coefficients ($\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ and $\alpha_{2,2}$) of Equation (16). A total of 126 simulation experiments were carried out. Each experiment starts off empty and runs for $T_F = 300$ time units, it has one traffic intensity value for the first 150 time units and another value for the remaining 150 time units. We use the arrow notation $0.5 \to 0.25$ to denote an experiment with a traffic intensity that changes from 0.5 to 0.25. The experiments consider all combinations of traffic intensity $\lambda/\mu \in \{0.5 \to 0.25, 0.75 \to 0.25, 1.25 \to 0.25, 0.75 \to 0.5, 1.25 \to 0.5, 1.25 \to 0.75\}$, service rate $\mu \in \{0.2, 0.4, 0.6\}$, and space capacity $\ell \in \{10, 20, 30, 40, 60, 80, 100\}$. The simulator yields an estimate of $P(DQ(k)=0)$, denoted $P_S(DQ(k)=0)$, for all $k = 1, ..., T_F$. The coefficients $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ and $\alpha_{2,2}$ are fit such as to minimize, over all 126 experiments, the error function given by Equation (38) and rewritten here:

$$\bar{e}_{DQ} = \frac{1}{T_F} \sum_{T=1}^{T_F} |P_A(DQ(T)=0) - P_S(DQ(T)=0)|, \tag{62}$$

where $P_S(DQ(T) = 0)$ is the estimate from the simulator and $P_A(DQ(T) = 0)$ is the analytical approximation obtained from Algorithm 1 with the following adjustments. At every time step k,

- In step 6(a) of Algorithm 1, $\lambda_{DQ}(k)$ is estimated from the simulator instead of calculating from Eq. (4).
- In step 6(d) of Algorithm 1, $\tau_{DQ}(k)$ is obtained from Equation (16), which depends on scalar parameters $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ and $\alpha_{2,2}$.
- In step 6(f) of Algorithm 1, $P(UQ(k) = \ell)$ is estimated from the simulator instead of from calculating through all the sub-steps.

In other words, perfect information about the link's upstream boundary conditions is assumed in the calculation of $P_A(DQ(k) = 0)$. Thus, $P_A(DQ(k) = 0)$ only depends on the choice of $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ and $\alpha_{2,2}$, thus the error function $\bar{e}_{DQ}$ only depends on $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$ and $\alpha_{2,2}$. The scalars are estimated jointly and the numerical values obtained are $\alpha_{1,1} = 0.4$, $\alpha_{1,2} = 0.4$, $\alpha_{2,1} = 0.6$ and $\alpha_{2,2} = 0$.

C Derivation of Equation (33c)

In this appendix, we derive the expression for the variance of the sojourn time of DQ during time interval $k - k^{bwd}$. Recall that we use the sojourn time of an $M/M/1/\ell$ queue with arrival rate $\lambda_{DQ}(k - k^{bwd})$ and service rate $\mu(k - k^{bwd})$ to approximate the sojourn time of DQ during time interval $k - k^{bwd}$. To make the notation simpler, hereafter the time index $k - k^{bwd}$ is dropped. Let $\rho = \lambda_{DQ}/\mu$.

The probability density function of the sojourn time of an $M/M/1/\ell$ queue, denoted $f_{S_{DQ}}(t)$, is given by Sztrik (2012, Chap. 2.4, Page 34):

$$f_{S_{DQ}}(t) = \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \frac{P(DQ = n)}{1 - P(DQ = \ell)} \tag{63}$$

where $P(DQ = n)$ is the steady state probability of DQ.

We use this probability density function expression to compute $E[S_{DQ}^2]$ as follows.

$$E[S_{DQ}^2] = \int_0^\infty t^2 f_{S_{DQ}}(t) dt \tag{64}$$

$$= \int_0^\infty t^2 \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \frac{P(DQ = n)}{1 - P(DQ = \ell)} dt \tag{65}$$

$$= \int_0^\infty t^2 \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \frac{\left(\frac{1-\rho}{1-\rho^{\ell+1}}\right) \rho^n}{1 - \left(\frac{1-\rho}{1-\rho^{\ell+1}}\right) \rho^\ell} dt \tag{66}$$

$$= \frac{(1-\rho)}{1-\rho^{\ell}} \int_0^{\infty} t^2 \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \rho^n dt \tag{67}$$

$$= \frac{(1-\rho)}{1-\rho^{\ell}} \sum_{n=0}^{\ell-1} \rho^n \int_0^{\infty} \frac{t(\mu t)^{n+1}}{n!} e^{-\mu t} dt \tag{68}$$

$$= \frac{(1-\rho)}{1-\rho^{\ell}} \sum_{n=0}^{\ell-1} \rho^n \frac{\Gamma(n+3)}{\mu^2 n!} \tag{69}$$

$$= \frac{(1-\rho)}{(1-\rho^{\ell})\mu^2} \sum_{n=0}^{\ell-1} \rho^n (n+2)(n+1) \tag{70}$$

$$= \frac{-\ell(\ell+1)\rho^{\ell+2} + 2\ell(\ell+2)\rho^{\ell+1} - (\ell+1)(\ell+2)\rho^{\ell} + 2}{\mu^2(1-\rho^{\ell})(1-\rho)^2} \tag{71}$$

Equation (66) is obtained from Equation (65) by substituting the closed-form expression of the steady state probability distribution of an M/M/1/ℓ system (see Gross (2008, Chap. 2, Equation (2.49))), which is given by:

$$P(DQ = n) = \left(\frac{1-\rho}{1-\rho^{\ell+1}}\right) \rho^n, \quad \forall n \in \{0, ..., \ell\}. \tag{72}$$

Equations (67) and (68) are obtained by pulling out the common terms inside of the integration that are independent of t. Equations (69) and (70) are derived by substituting in $\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x} dx = (n-1)!$ for n a positive integer. Finally, Equation (71) is obtained by calculating the finite power series and collecting common terms.

The expected value of the sojourn time of DQ is given by (see Eq. (32d)):

$$E[S_{DQ}] = \frac{\ell\rho^{\ell+1} - (\ell+1)\rho^{\ell} + 1}{\mu(1-\rho^{\ell})(1-\rho)} \tag{73}$$

Thus, the variance of the sojourn time of DQ is given by:

$$Var(S_{DQ}) = E[S_{DQ}^2] - E[S_{DQ}]^2 \tag{74}$$

$$= \frac{\ell\rho^{2\ell+2} - 2\ell\rho^{2\ell+1} + (\ell+1)\rho^{2\ell} - \ell(\ell+1)\rho^{\ell+2} + 2\ell(\ell+1)\rho^{\ell+1} - (\ell^2+\ell+2)\rho^{\ell} + 1}{\mu^2(1-\rho^{\ell})^2(1-\rho)^2} \tag{75}$$

D Derivation of the normalization term of Equation (29a)

This appendix discusses the normalization term (the fraction term) in Equation (29a). This term is defined such as to obtain a proper probability distribution of $P\big(UQ(k-1) = i\big)$ for $i \in \{0, ..., \ell\}$. More specifically:

$$\sum_{i=0}^{\ell} P\big(UQ(k-1) = i\big) = P\big(UQ(k-1) = \ell\big) + \sum_{i=0}^{\ell-1} P\big(UQ(k-1) = i\big) \tag{76}$$

$$= P\big(UQ(k-1)=\ell\big) + \sum_{i=0}^{\ell-1} \frac{1-P\big(UQ(k-1)=\ell\big)}{\sum_{j=0}^{\ell-1} f\big(j, q^{UQ}(k-1)\delta\big)} f\big(i, q^{UQ}(k-1)\delta\big) \tag{77}$$

$$= P\big(UQ(k-1)=\ell\big) + \frac{1-P\big(UQ(k-1)=\ell\big)}{\sum_{j=0}^{\ell-1} f\big(j, q^{UQ}(k-1)\delta\big)} \sum_{i=0}^{\ell-1} f\big(i, q^{UQ}(k-1)\delta\big) \tag{78}$$

$$= P\big(UQ(k-1)=\ell\big) + 1 - P\big(UQ(k-1)=\ell\big) \tag{79}$$

$$= 1 \tag{80}$$

Equation (77) is obtained by substituting in $P\big(UQ(k-1)=i\big)$ defined in Equation (29a). Note that the term $\frac{1-P\left(UQ(k-1)=\ell\right)}{\sum_{j=0}^{\ell-1} f\left(j,q^{UQ}(k-1)\delta\right)}$ in the summation does not depend on i, and hence it is a constant term that can be pulled outside the summation. Therefore Equation (78) is obtained. Now, the denominator in the fraction term (i.e., $\sum_{j=0}^{\ell-1} f\big(j, q^{UQ}(k-1)\delta\big)$) is exactly the same as the summation term it is multiplied by, and hence these terms cancel each other out and Equation (79) is obtained.

E Estimation of the scalar coefficients in $\tau_{UQ}(k)$ of Equation (33)

This appendix describes the procedure to fit the exogenous coefficients $\alpha_{3,1}$ and $\alpha_{3,2}$ of Equation (33). The same set of 126 simulation experiments as described in Appendix B are used. The coefficients $\alpha_{3,1}$ and $\alpha_{3,2}$ are fit by minimizing, over all 126 experiments, the following error function given by Equation (39) and rewritten here:

$$\bar{e}_{UQ} = \frac{1}{T_F} \sum_{T=1}^{T_F} |P_A(UQ(T)=\ell) - P_S(UQ(T)=\ell)|, \tag{81}$$

where $P_S(UQ(T)=\ell)$ is the estimate from the simulator and $P_A(UQ(T)=\ell)$ is the analytical approximation, which is obtained from Algorithm 1 with the following adjustments. At every time step k,

- In step 6(a) of Algorithm 1, $q^{LLO}(k)$ is estimated from the simulator instead of calculating from Eq. (21).
- In step 6(e) of Algorithm 1, $P(DQ(k)=0)$ is estimated from the simulator instead of calculating from Eq. (5).
- In step 6(f) of Algorithm 1, $q^{UQ}(k-1)$ is estimated from the simulator instead of calculating from Eq. (29c) if $q^{LLO}(k)=0$. Otherwise, $\tau_{UQ}(k)$ is obtained from Equation (33), which depends on scalar parameters $\alpha_{3,1}$ and $\alpha_{3,2}$.

In other words, perfect information about the link's downstream boundary conditions is assumed in the calculation of $P_A(UQ(T)=\ell)$. Thus, $P_A(UQ(T)=\ell)$ only depends on the choice of $\alpha_{3,1}$ and $\alpha_{3,2}$ and thus the error function $\bar{e}_{UQ}$ depends only on $\alpha_{3,1}$ and $\alpha_{3,2}$. The scalars are estimated jointly and the numerical values obtained are $\alpha_{3,1}=25$ and $\alpha_{3,2}=7.5$.

F Detailed results for the mean absolute differences of Figure 4

This appendix provides the detailed results for the mean absolute differences of Figure 4. More specifically, Table 6 provides the mean absolute difference $\bar{e}_{UQ}$ values of $P(UQ(k) = \ell)$ plotted in Figures 4(a), 4(b) and 4(c) and Table 7 the mean absolute difference $\bar{e}_{DQ}$ values of $P(DQ(k) = 0)$ plotted in Figures 4(d), 4(e) and 4(f).

Experiment		$\bar{e}_{UQ}$		
$\lambda(k)$	ℓ	Mixture	Multivariate	Proposed
	10	$3.11e-3$	$7.80e-5$	$1.80e-3$
	20	$5.75e-5$	$4.02e-6$	$4.68e-5$
	30	$7.57e-7$	$3.92e-7$	$7.41e-7$
$0.2 \to 0.1$	40	$9.25e-12$	NaN	$4.60e-10$
	60	$5.20e-17$	NaN	$1.56e-14$
	80	$2.41e-22$	NaN	$5.45e-19$
	100	$2.49e-28$	NaN	$1.89e-23$
	10	$1.48e-2$	$4.50e-4$	$3.90e-3$
	20	$4.67e-3$	$1.23e-4$	$2.38e-3$
	30	$7.03e-4$	$3.90e-5$	$5.08e-4$
$0.3 \to 0.1$	40	$8.02e-5$	NaN	$6.80e-5$
	60	$3.11e-7$	NaN	$3.62e-7$
	80	$1.82e-13$	NaN	$1.19e-9$
	100	$3.08e-17$	NaN	$6.91e-12$
	10	$1.87e-2$	$4.44e-4$	$4.92e-3$
	20	$5.19e-3$	$1.29e-4$	$2.18e-3$
	30	$8.36e-4$	$3.79e-5$	$4.99e-4$
$0.3 \to 0.2$	40	$1.00e-4$	NaN	$7.05e-5$
	60	$6.59e-7$	NaN	$6.15e-7$
	80	$2.04e-13$	NaN	$1.28e-9$
	100	$6.07e-17$	NaN	$7.31e-12$

Table 6: Mean absolute difference $\bar{e}_{UQ}$ of $P(UQ(k) = \ell)$. The value NaN denotes cases where the evaluation of the multivariate model exceeded the limit of 40 hours. (Starting empty with time-varying demand over time)

Experiment		$\bar{e}_{DQ}$		
$\lambda(k)$	ℓ	Mixture	Multivariate	Proposed
	10	$5.00e-3$	$2.93e-3$	$4.55e-3$
	20	$2.76e-3$	$3.00e-3$	$4.35e-3$
	30	$2.25e-3$	$3.02e-3$	$4.66e-3$
$0.2 \to 0.1$	40	$2.51e-3$	NaN	$5.12e-3$
	60	$2.63e-3$	NaN	$5.51e-3$
	80	$2.68e-3$	NaN	$5.82e-3$
	100	$2.74e-3$	NaN	$6.16e-3$
	10	$1.52e-2$	$0.44e-2$	$0.75e-2$
	20	$0.73e-2$	$0.46e-2$	$1.56e-2$
	30	$0.40e-2$	$0.47e-2$	$1.82e-2$
$0.3 \to 0.1$	40	$0.33e-2$	NaN	$1.89e-2$
	60	$0.40e-2$	NaN	$1.93e-2$
	80	$0.42e-2$	NaN	$1.97e-2$
	100	$0.42e-2$	NaN	$2.01e-2$
	10	$1.62e-2$	$0.35e-2$	$1.08e-2$
	20	$0.69e-2$	$0.35e-2$	$1.32e-2$
	30	$0.36e-2$	$0.36e-2$	$1.33e-2$
$0.3 \to 0.2$	40	$0.27e-2$	NaN	$1.38e-2$
	60	$0.32e-2$	NaN	$1.43e-2$
	80	$0.34e-2$	NaN	$1.46e-2$
	100	$0.34e-2$	NaN	$1.49e-2$

Table 7: Mean absolute difference $\bar{e}_{DQ}$ of $P(DQ(k) = 0)$. The value NaN denotes cases where the evaluation of the multivariate model exceeded the limit of 40 hours. (Starting empty with time-varying demand over time)

G Congestion levels of the Lausanne network under the Synchro signal plan

Table 8 defines, for urban roads, four congestion levels or traffic conditions as a function of the underlying speed. These definitions are taken from Dowling et al. (2008, Class IV in Exhibit 2 Urban Street Level of Service (LOS)). Four traffic conditions are defined: free-flow, stable, high-density and breakdown. We consider the Synchro signal plan as the benchmark and estimate for each lane its average speed and classify its traffic condition.

Traffic condition	Level of service (LOS)	Short description	Speed [km/h]
Free-flow	A-B	free flow or reasonably free flow	> 30
Stable	C	stable flow, at or near free flow	$20 - 30$
High-density	D-E	approaching unstable flow, operating at capacity	$11 - 20$
Breakdown	F	forced or breakdown flow	≤ 11

Table 8: Urban street level of service

Each column of Table 9 corresponds to a different traffic condition, as defined in Table 8. The first row of Table 9 displays the breakdown of the lanes according to their traffic conditions, under the Synchro signal plan. For each traffic condition (i.e., each column), it indicates the absolute number of lanes with this traffic condition and, in parenthesis, the percentage of lanes with this traffic condition. There are a total of 676 (i.e., $558 + 118$) out of the total 922 lanes that have either free-flow or stable traffic conditions. This constitutes 73% of the lanes in the network. The remaining 27% of the lanes have either high-density or breakdown traffic condition, i.e., these lanes are operating near or above capacity.

Traffic conditions	Free-flow	Stable	High-density	Breakdown
Number (%) of lanes	558 (60.5%)	118 (12.8%)	149 (16.2%)	97 (10.5%)
Number of lanes with no change	450	85	121	63
Number of lanes with improvement	84	23	15	17
Number of lanes with deterioration	24	10	13	17

Table 9: The number of lanes in different traffic conditions and categories.

References

Asmussen, S. (2003). *Applied probability and queues*, Vol. 2, Springer.

Blanc, J. P. C. and Van Doorn, E. (1983). Relaxation times for queueing systems, *In: de Bakker, J.W., Hazewinkel, M., Lenstra, J.K. (eds.) Mathematics and Computer Science, Proceedings of the CWI symposium*, Vol. 1, CWI Monograph, North-Holland, Amsterdam, pp. 139–162.

Boel, R. and Mihaylova, L. (2006). A compositional stochastic model for real time freeway traffic simulation, *Transportation Research Part B* **40**: 319–334.

Calvert, S., Taale, H., Snelder, M. and Hoogendoorn, S. (2012). Probability in traffic: a challenge for modelling, *4th International Symposium on Dynamic Traffic Assignment (DTA), Massachusetts, USA*.

Chen, X., Li, L. and Shi, Q. (2015). *Stochastic Evolutions of Dynamic Traffic Flow*, Springer, Berlin Heidelberg.

Chong, L. and Osorio, C. (2017). A simulation-based optimization algorithm for dynamic large-scale urban transportation problems, *Transportation Science* **52**(3): 637–656.

Daganzo, C. (2005). A variational formulation of kinematic waves: basic theory and complex boundary conditions, *Transportation Research Part B* **39**(2): 187–196.

Daganzo, C. F. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory, *Transportation Research Part B* **28**(4): 269–287.

Davis, J. L., Massey, W. A. and Whitt, W. (1995). Sensitivity to the service-time distribution in the nonstationary Erlang loss model, *Management Science* **41**(6): 1107–1116.

Deng, W., Lei, H. and Zhou, X. (2013). Traffic state estimation and uncertainty quantification based on heterogeneous data sources: a three detector approach, *Transportation Research Part B* **57**: 132 – 157.

Dowling, R., Flannery, A., Landis, B., Petritsch, T., Rouphail, N. and Ryus, P. (2008). Multimodal level of service analysis for urban streets, *Technical Report No. 616*, National Cooperative Highway Research Program NCHRP, Washington, DC. pp. 5–6.
URL: *https://doi.org/10.17226/14175*

Dumont, A. G. and Bert, E. (2006). *Simulation de l'agglomération Lausannoise SIMULO*, Laboratoire des voies de circulation, ENAC, Ecole Polytechnique Fédérale de Lausanne.
URL: *Available at: http://web.mit.edu/osorioc/www/papers/dumont06BertRapport.pdf*

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, *Post Office Electrical Engineer's Journal* **10**(1917-1918): 189–197.

Flötteröd, G. and Osorio, C. (2017). Stochastic network link transmission model, *Transportation Research Part B* **102**: 180–209.

Gross, D. (2008). *Fundamentals of queueing theory*, John Wiley & Sons, New York, U.S., chapter 2, pp. 49–103.

Heidemann, D. (1991). Queue length and waiting time distributions at priority intersections, *Transportation Research Part B* **25**(4): 163–174.

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

Heidemann, D. (2001). A queueing theory model of nonstationary traffic flow, *Transportation Science* **35**(4): 405–412.

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

Helbing, D. (1997). Modeling multi-lane traffic flow with queuing effects, *Physica A: Statistical Mechanics and its Applications* **242**(1-2): 175–194.

Helbing, D. (2001). Traffic and related self-driven many-particle systems, *Reviews of modern physics* **73**(4): 1067.

Herman, R. and Lam, T. (1971). On the mean speed in the "Boltzmann-Like" traffic theory: Analytical derivation, *Transportation Science* **5**(3): 314–327.

Himpe, W., Corthout, R. and Tampère, M. C. (2016). An efficient iterative link transmission model, *Transportation Research Part B* **92**: 170–190.

Hoogendoorn, S. P. and Bovy, P. H. (2001). Generic gas-kinetic traffic systems modeling with applications to vehicular traffic flow, *Transportation Research Part B* **35**(4): 317–336.

Jabari, S. E. (2012). *A stochastic model of macroscopic traffic flow: Theoretical foundations*, PhD thesis, University of Minnesota.

Jabari, S. E. and Liu, H. X. (2012). A stochastic model of traffic flow: Theoretical foundations, *Transportation Research Part B* **46**(1): 156–174.

Jabari, S. E., Zheng, J. and Liu, H. X. (2014). A probabilistic stationary speed–density relation based on Newell's simplified car-following model, *Transportation Research Part B* **68**: 205–223.

Jagerman, D. (1975). Nonstationary blocking in telephone traffic, *Bell Labs Technical Journal* **54**(3): 625–661.

Khinchin, A. Y. (1962). Erlang's formulas in the theory of mass service, *Theory of Probability & Its Applications* **7**(3): 320–325.

Kleinrock, L. (1975). *Queueing Systems Volume 1:Theory*, Wiley-Interscience, New York, NY, USA, chapter 2, p. 77.

Laval, J. A. and Castrillón, F. (2015). Stochastic approximations for the macroscopic fundamental diagram of urban networks, *Transportation Research Procedia, Papers selected for the International Symposium of Transportation and Traffic Theory (ISTTT)*, Vol. 7, pp. 615–630.

Laval, J. A. and Chilukuri, B. R. (2014). The distribution of congestion on a class of stochastic kinematic wave models, *Transportation Science* **48**(2): 217–224.

Lighthill, M. and Whitham, G. (1955). On kinematic waves. I: Flood movement in long rivers, II: a theory of traffic flow on long crowded roads, *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 229, The Royal Society, pp. 281–345.

Lu, J. and Osorio, C. (2018). A probabilistic traffic-theoretic network loading model suitable for large-scale network analysis, *Transportation Science* **52**(6): 1509–1530.

MATLAB (2016). *Optimization Toolbox: User's Guide (R2016a)*, The Mathworks, Inc., Natick, Massachusetts.

Morse, P. (1958). *Queues, inventories and maintenance; the analysis of operational systems with variable demand and supply*, Wiley, New York, USA, chapter 6, pp. 59–67.

Nelson, P. (1995). A kinetic model of vehicular traffic and its associated bimodal equilibrium solutions, *Transport Theory and Statistical Physics* **24**(1-3): 383–409.

Newell, C. (1982). *Applications of queueing theory*, Chapman and Hall, New York, USA, chapter 3, pp. 143–175.

Newell, G. (1993). A simplified theory of kinematic waves in highway traffic, part I: general theory, *Transportation Research Part B* **27**(4): 281–287.

Newell, G. F. (1961). Nonlinear effects in the dynamics of car following, *Operations Research* **9**(2): 209–229.

Newell, G. F. (2002). A simplified car-following theory: a lower order model, *Transportation Research Part B* **36**(3): 195–205.

Odoni, A. R. and Roth, E. (1983). An empirical investigation of the transient behavior of stationary queueing systems, *Operations Research* **31**(3): 432–455.

Olszewski, P. S. (1994). Modeling probability distribution of delay at signalized intersections, *Journal of Advanced Transportation* **28**(3): 253–274.

Osorio, C. (2010). *Mitigating network congestion: analytical models, optimization methods and their applications*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne.

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

Osorio, C. and Flötteröd, G. (2015). Capturing dependency among link boundaries in a stochastic dynamic network loading model, *Transportation Science* **49**(2): 420–431.

Osorio, C., Flötteröd, G. and Bierlaire, M. (2011). Dynamic network loading: a stochastic differentiable model that derives link state distributions, *Transportation Research Part B* **45**(9): 1410–1423.

Osorio, C. and Wang, C. (2017). On the analytical approximation of joint aggregate queue-length distributions for traffic networks: a stationary finite capacity Markovian network approach, *Transportation Research Part B* **95**: 305–339.

Osorio, C. and Yamani, J. (2017). Analytical and scalable analysis of transient tandem Markovian finite capacity queueing networks, *Transportation Science* **51**(3): 823–840.

Paveri-Fontana, S. (1975). On Boltzmann-like treatments for traffic flow: a critical review of the basic model and an alternative proposal for dilute traffic analysis, *Transportation Research* **9**(4): 225–235.

Prigogine, I. and Andrews, F. C. (1960). A Boltzmann-like approach for traffic flow, *Operations Research* **8**(6): 789–797.

Richards, P. I. (1956). Shock waves on the highway, *Operations Research* **4**(1): 42–51.

Roth, E. (1994). The relaxation time heuristic for the initial transient problem in M/M/k queueing systems, *European Journal of Operational Research* **72**(2): 376–386.

Sayegh, A. S., Connors, R. D. and Tate, J. E. (2017). Uncertainty propagation from the cell transmission traffic flow model to emission predictions: a data-driven approach, *Transportation Science* **52**(6): 1327–1346.

Stafford, R. (2006). Random vectors with fixed sum. Accessed June 1, 2015.
URL: *Http://www.mathworks.com/matlabcentral/fileexchange/9700*

Sumalee, A., Zhong, R. X., Pan, T. L. and Szeto, W. Y. (2011). Stochastic Cell Transmission Model (SCTM): a stochastic dynamic traffic model for traffic state surveillance and assignment, *Transportation Research Part B* **45**(3): 507–533.

Sztrik, J. (2012). *Basic queueing theory*, University of Debrecen, Debrecen, Hungary, chapter 2.4, pp. 32–37. Accessed July 20, 2018.
URL: *https://pdfs.semanticscholar.org/848f/a1f48ad9d3edb24b05667f15cfc633eb8f69.pdf*

Tampère, C., Corthout, R., Cattrysse, D. and Immers, L. (2011). A generic class of first-order node models for dynamic macroscopic simulations of traffic flows, *Transportation Research Part B* **45**(1): 289–309.

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

Transport for London (2010). Traffic modelling guidelines. version 3.0, *Technical report*, Transport for London (TfL).

Treiber, M., Hennecke, A. and Helbing, D. (1999). Derivation, properties, and simulation of a gas-kinetic-based, nonlocal traffic model, *Physical Review E* **59**(1): 239.

TSS (2014). *AIMSUN 8.1 Microsimulator Users Manual*, Transport Simulation System.

U.S. Department of Transportation (2008). Transportation vision for 2030, *Technical report*, U.S. Department of Transportation (DOT), Research and Innovative Technology Administration.

van Doorn, E. A. and Zeifman, A. I. (2009). On the speed of convergence to stationarity of the Erlang loss system, *Queueing Systems* **63**(1-4): 241.

van Zuylen, H. J. and Viti, F. (2003). Uncertainty and the dynamics of queues at controlled intersections, *International Federation of Automatic Control (IFAC) in Transportation Systems Proceedings* **36**(14): 43–48.

Viti, F. and Van Zuylen, H. J. (2010). Probabilistic models for queues at fixed control signals, *Transportation Research Part B* **44**(1): 120–135.

Yperman, I. (2007). *The link transmission model for dynamic network loading*, PhD thesis, Katholieke Universiteit Leuven.

Yperman, I., Tampere, C. and Immers, B. (2007). A kinematic wave dynamic network loading model including intersection delays, *Transportation Research Board 86th Annual Meeting*, Washington DC, USA.

Zheng, F., Jabari, S. E., Liu, H. X. and Lin, D. (2018). Traffic state estimation using stochastic Lagrangian dynamics, *Transportation Research Part B* **115**: 143–165.