Network loading model: a probabilistic, analytical, scalable and traffic-theoretic approach

Jing Lu[a] and Carolina Osorio[a,b]

[a]*Operations Research Center, Massachusetts Institute of Technology, USA*
Email: jl3724@mit.edu
[b]*Department of Civil and Environmental Engineering,*
Massachusetts Institute of Technology, USA
Email: osorioc@mit.edu

August 11, 2018

Abstract

We formulate a traffic theoretic and probabilistic analytical network loading 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. 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. The model is then used to address a city-wide traffic signal control problem. Compared to a benchmark analytical model, the proposed model enhances computational efficiency by two orders of magnitude, while deriving signal plans with similar performance. The proposed model yields signal plans that outperform those obtained from a widely used commercial signal control software.

1 Introduction

The field of traffic flow modeling is shifting from the formulation and use of deterministic models to that of stochastic ones. This is facilitated and motivated by a number of factors, including increased availability of urban mobility data, advanced censoring 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.

Detailed reviews of stochastic traffic flow models 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). 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. For analytical models that add Gaussian noise terms to a specific deterministic model, computational inefficiency can arise due to the need to sample from high-dimensional Gaussian distributions.

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). The work of Heidemann (2001) considers transient (i.e., non-stationary) analysis. 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). This paper 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. The Appendices contain additional equation derivations and numerical validation results.

2 Link model formulation

2.1 Past link model formulations

First we 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). 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 $\hat{\rho}$. Consider a discrete-time model, where k denotes the discrete time interval index. Each time interval is of length δ.

The models 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. $q^{out}(k)$). Upon entrance to the link, the vehicular inflow experiences a delay of L/v time units 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 queue*, LI. This delay process can be viewed as if the flow traveled sequentially through a set of $k^{fwd} = \lceil \frac{L}{v} \rceil$ 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 queue*, DQ. Upon departure from the link, the newly available space is delayed for $L/|w|$ time units 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 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 $k^{bwd} = \lceil \frac{L}{|w|} \rceil$ 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 queue*, UQ, which is defined as:

$$UQ = LI + DQ + LO. \tag{1}$$

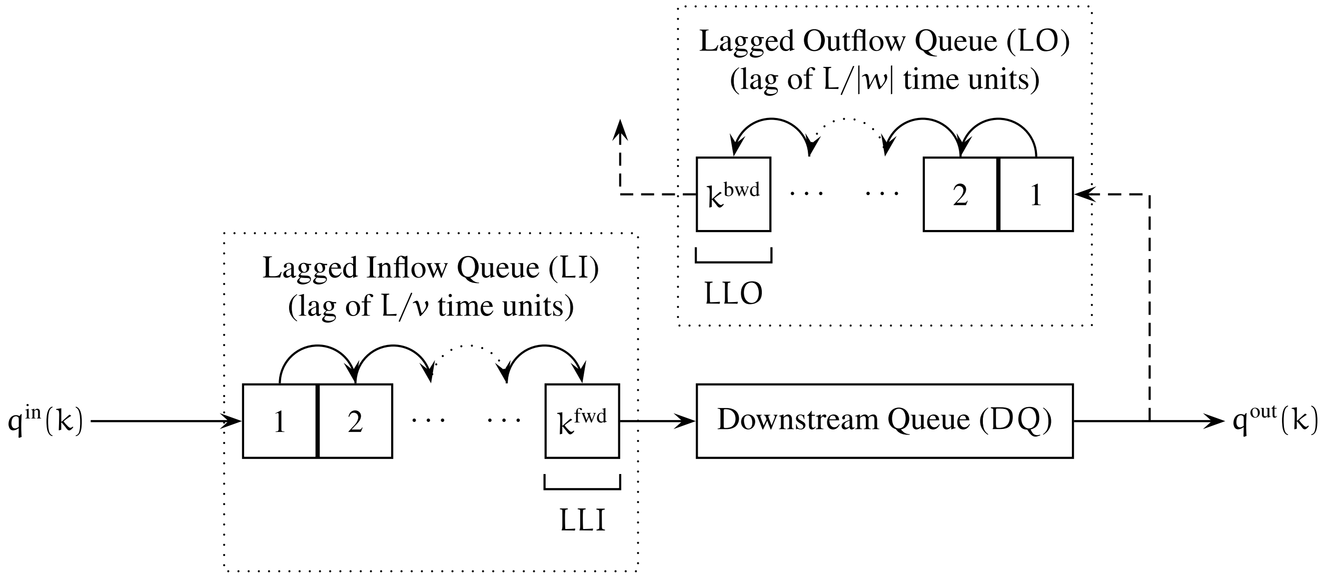


Figure 1: Link dynamics (adapted from Lu and Osorio (2018))

In other words, UQ accounts for all vehicular flow in the link (i.e., $\mathrm{LI}+\mathrm{DQ}$) as well as all "newly available spaces" that are not yet available at the upstream end of the link (i.e., LO). Equivalently, for a link with space capacity ℓ (which is defined as the rounded version of $\hat{\rho}L$), $\ell-\mathrm{UQ}$ represents the vehicular space available at the upstream end of the link.

All queues ($\mathrm{LI}, \mathrm{DQ}, \mathrm{LO}$ and UQ) are time-dependent stochastic processes. Hence, 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 4 queues ($\mathrm{LI}, \mathrm{DQ}, \mathrm{LO}, \mathrm{UQ}$), which are linearly constrained through Eq. (1). Hence, we need only to track 3 of these 4 queues. The model of Osorio and Flötteröd (2015) tracks 3 queues and yields the joint distribution of ($\mathrm{LI}, \mathrm{DQ}, \mathrm{LO}$). For a given link with space capacity ℓ, each queue has a state space defined as $\{0, 1, ..., \ell\}$. Hence, the dimension of the state space of this joint distribution ($\mathrm{LI}, \mathrm{DQ}, \mathrm{LO}$) 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 state space dimension equals to 2. In other words, the model complexity is now independent of the link's space capacity. 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. Hereafter, we use the notation $DQ(k)$ (resp. $LI(k)$, $UQ(k)$, $LO(k)$) to denote the state of DQ (resp. LI, UQ, LO) at the end of time interval k. Hereafter, we use DQ and $DQ(k)$ are used interchangeably.

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 2 specific queue states that are essential to describe these boundary conditions. The first state is $DQ = 0$, which describes whether or not there is vehicular flow downstream ready to depart the link. The second state is $UQ = \ell$, 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 $DQ > 0$), and (ii) there is space available at the upstream end of the downstream link (i.e., for the downstream link $UQ < \ell$). Hence, for a given link, the proposed model approximates only 2 state probabilities: $P(DQ = 0)$ and $P(UQ = \ell)$.

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

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

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 $P(UQ(k) < \ell) = 1 - P(UQ(k) = \ell)$. Similarly, the expected link outflow is defined as:

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

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 $P(DQ(k) > 0) = 1 - P(DQ(k) = 0)$.

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 2. In other words, its complexity is now constant and no longer depends on the space capacity of the link. Hence, 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 $P(DQ(k) = 0)$. We approximate the downstream queue during time interval k, $DQ(k)$, as an $M/M/1/\ell$ queue. More specifically, the arrival process of $DQ(k)$ is approximated as a Poisson process with endogenous rate $\lambda_{DQ}(k)$. The service process of $DQ(k)$ is exactly the service process at the link's downstream end. Hence, service times of $DQ(k)$ are independent and identically distributed exponential random variables with exogenous rate $\mu(k)$. There is no approximation made in the service process of $DQ(k)$.

First, we describe the approximation of the endogenous arrival rate $\lambda_{DQ}(k)$. 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. Hence, flow that enters DQ during time interval k corresponds to flow that entered the link during time interval $k - k^{fwd}$. Hence, we approximate the arrival rate to $DQ(k)$ as the expected flow to enter the link during time interval $k-k^{fwd}$ divided by the time interval length δ:

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

For an $M/M/1/\ell$ queue, an exact closed-form expression for the transient queue-length distribution is given by Morse (1958, Chap. 6, Equation (6.13)). Nonetheless, the use of this expression requires keeping track of the full queue-length distribution. Our aim is to track a single probability, $P(DQ(k) = 0)$, rather than the full distribution. 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 $k + 1$);
$P_k(DQ = 0)$	time-interval specific stationary probability of $DQ = 0$;
$\tau_{DQ}(k)$	inverse of the relaxation time during time interval k.

We propose the following formulation:

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

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 $P_k(DQ = 0)$) and a term that decays exponentially with time. The latter term is the difference between the initial condition of time interval k (term $P(DQ(k-1) = 0)$) and the corresponding stationary probability $P_k(DQ = 0)$. The functional form of Equation (5) is inspired by both the expression of Morse (1958, Chap. 6, Equation (6.13)) as well as by the recent work of Chong and Osorio (2017, Equation (14a)), which models the spillback probability (also known as the blocking probability) of a queue with such a functional form.

Equation (5) contains two endogenous terms, $P_k(DQ = 0)$ and $\tau_{DQ}(k)$. We now present their formulations. We first present the approximation of $P_k(DQ = 0)$. Recall that we approximate $DQ(k)$ as an $M/M/1/\ell$ queue with arrival rate $\lambda_{DQ}(k)$ and service rate $\mu(k)$. For an $M/M/1/\ell$ queueing system, there is a closed-form expression for the stationary queue-length distribution (e.g., Gross (2008, Chap. 2, Equation (2.49))). We use this expression to approximate $P_k(DQ = 0)$:

$$\begin{cases} P_k(DQ = 0) = \dfrac{1 - \rho_{DQ}(k)}{1 - \rho_{DQ}(k)^{\ell+1}} & \text{(6a)} \\ \rho_{DQ}(k) = \lambda_{DQ}(k)/\mu(k). & \text{(6b)} \end{cases}$$

We now present the approximation for $\tau_{DQ}(k)$. In queueing theory, $\tau_{DQ}(k)$ is known as the inverse of the relaxation time. It measures the speed at which a given performance measure reaches its stationary value (i.e., a higher $\tau_{DQ}(k)$ corresponds to a higher speed of convergence to stationary values). We approximate $\tau_{DQ}(k)$ as follows:

$$\begin{cases} \tau_{DQ}(k) = \tau_{DQ} \cdot \left(\dfrac{\alpha_1 |P(DQ(k-1)=0) - P_k(DQ=0)|^{\rho_{DQ}(k)}}{1 + e^{-\alpha_2 \rho_{DQ}(k)}} \right) & \text{(7a)} \\ \tau_{DQ} = \dfrac{\mu(k)(1-\rho_{DQ}(k))^2 \ell + \alpha_3 \mu(k) \rho_{DQ}(k)^2 \sqrt{\ell}}{(1+\rho_{DQ}(k))(\ell+1)}, & \text{(7b)} \end{cases}$$

where α_1, α_2 and α_3 are exogenous scalar coefficients. Equation (7a) approximates $\tau_{DQ}(k)$ as the product of τ_{DQ} (which is defined in Eq. (7b)) and of the term within parenthesis. We now describe how the formulation for each of these two terms is derived.

The study of relaxation times in the literature is mostly limited to the relaxation time of the expected queue-length for infinite (space) capacity single-server Markovian queueing systems that start off empty, such as in Odoni and Roth (1983) and in Newell (1982, Chap. 5). The work of Odoni and Roth (1983) reviews relaxation time studies and identifies properties that a closed-form expression for the relaxation time of the expected queue-length should have. A closed-form approximation is also provided by Newell (1982, Chap. 5, Equation (5.6)).

The main differences between past studies and our setting are: (i) we consider a finite, rather than an infinite, space capacity queueing system, (ii) we want to approximate the relaxation time (or equivalently its inverse) of a probability state, rather than that of the expected queue-length, (iii) we consider an arbitrary initial state for the queueing system, while past work consider empty initial states. The term τ_{DQ} of Equation (7a) accounts for difference (i), while the expression for the term within parenthesis of Equation (7a) accounts for differences (ii) and (iii).

The proposed formulation for τ_{DQ} (Eq. (7b)) is based on the following two goals. First, when considering the experimental conditions considered in the literature (e.g., $\ell \to \infty$), the desired properties of the relaxation time are preserved. Second, the formulation is extended to account for our experimental conditions (i.e., finite capacity systems). We now describe this in more detail.

- As the space capacity increases, we retrieve the expression for an infinite space capacity $M/M/1$ system of Newell (1982, Chap. 5, Equation (5.6)), i.e.,:

  $$\lim_{\ell \to \infty} \tau_{DQ} = \frac{\mu(k)(1-\rho_{DQ}(k))^2}{(1+\rho_{DQ}(k))}. \tag{8}$$

  The calculation of this limit is given in Appendix A.

- For infinite space capacity queueing systems, the relaxation time is only defined for $\rho_{DQ}(k) < 1$. For finite space capacity systems, τ_{DQ} should be defined for all non-negative values of traffic intensity, including values greater than 1 (i.e., $\rho_{DQ}(k) \geq 0$). Equation (7b) is well-defined for $\rho_{DQ}(k) \geq 0$ (even if $\rho_{DQ}(k) = 1$), as both the numerator and denominator are nonzero for all $\rho_{DQ}(k) \geq 0$.

- Odoni and Roth (1983) and Newell (1982, Chap. 5) study how the relaxation time varies with levels of congestion for infinite-capacity queueing systems. We follow a similar reasoning and desire τ_{DQ} to first decrease with increasing congestion levels (i.e., it takes a longer time to reach stationarity), but then increase as the congestion level further increases (i.e., it takes a shorter time to reach stationarity). In the limit as $\ell \to \infty$, Equation (7b) becomes the exact formula proposed by Newell (1982, Chap. 5, Equation (5.6)), which indeed satisfies this desired property. Holding all parameters other than $\rho_{DQ}(k)$ fixed, Equation (7b) is of the form $h(x) = (A(1-x)^2 + Bx^2)/C(x+1)$, where $x \geq 0$ and $A, C > 0, B \geq 0$ are parameters. This function first decreases and then increases as x increases. Hence, Equation (7b) preserves this desired property.
- Odoni and Roth (1983) state that the relaxation time should be scaled in time so that it varies directly with the units of the arrival and service rates. For instance, Newell's expression for an $M/M/1$ system (Newell; 1982, Chap. 5, Equation (5.6)) is directly proportional to the service rate $\mu(k)$ and hence it varies directly with the units of service rate. In other words, two identical queueing systems measured in different time units should yield the same value of $\tau_{DQ}(k)\delta$ (of Eq. (5)). Equation (7a) satisfies this property because $\tau_{DQ}(k)$ is the product of a unit-free term (term within parenthesis) and of a term that has the same unit as $\mu(k)$ (i.e., term τ_{DQ} of Eq. (7b) is the sum of two terms that have the same unit as $\mu(k)$).

The expression for the term within parenthesis of Equation (7a) is based on insights derived from simulation experiments. We use a simulation-based implementation of the stochastic link transmission model. This simulator samples individual vehicles, and imposes the forward and backward lags explicitly for each vehicle. It corresponds to the benchmark simulator used in Osorio and Flötteröd (2015). A total of 84 simulation experiments were carried out with all combinations of traffic intensity $\lambda/\mu \in \{0.25, 0.5, 0.75, 1.25\}$, service rate $\mu \in \{0.2, 0.4, 0.6\}$, and space capacity $\ell \in \{10, 20, 30, 40, 60, 80, 100\}$. For all experiments, we identified a positive correlation between $\tau_{DQ}(k)$ and the absolute difference $|P(DQ(k-1) = 0) - P_k(DQ = 0)|$. Based on these observations, we derived the following unit-free expression for the term within parenthesis of Eq. (7a):

$$\frac{\alpha_1 |P(DQ(k-1) = 0) - P_k(DQ = 0)|^{\rho_{DQ}(k)}}{1 + e^{-\alpha_2 \rho_{DQ}(k)}}, \tag{9}$$

where α_1 and α_2 are exogenous scalar coefficients. A description of how the exogenous scalar parameters α_1, α_2 and α_3 are fitted is given in Appendix B.

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 $UQ(k)$. 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 $UQ(k)$, the arrival process to $UQ(k)$ is also a Poisson process with exogenous rate $\lambda(k)$. Hence, we only need to approximate the service process of $UQ(k)$.

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.,:

$$q^{LLO}(k) = q^{out}(k - k^{bwd}). \tag{10}$$

To approximate $P(\mathrm{UQ}(k) = \ell)$ 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 (see Eq. (10)).

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

If $q^{LLO}(k) = 0$, then the expected outflow from $\mathrm{UQ}(k)$ is also zero. This implies that, with probability 1, there are no departures from $\mathrm{UQ}(k)$ (in other words, positive outflow from $\mathrm{UQ}(k)$ occurs with a probability of zero). Hence, $\mathrm{UQ}(k)$ is a pure arrival process.

Let $N(k)$ denote the number of attempted new arrivals 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., $\ell - \mathrm{UQ}(k-1)$. Hence, the number of vehicles in UQ at the end of time interval k (i.e., $\mathrm{UQ}(k)$) is sum of the number of vehicles in UQ at the beginning of time interval k (i.e., $\mathrm{UQ}(k-1)$) and the number of vehicles that successfully entered UQ:

$$\mathrm{UQ}(k) = \mathrm{UQ}(k-1) + \min\{N(k), \ell - \mathrm{UQ}(k-1)\}. \tag{11}$$

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

$$\begin{aligned}
P(\mathrm{UQ}(k) = \ell) &= P((\mathrm{UQ}(k-1) + \min\{N(k), \ell - \mathrm{UQ}(k-1)\}) = \ell) && (12)\\
&= \sum_{i=0}^{\ell} P(\min\{N(k), \ell - i\} = \ell - i | \mathrm{UQ}(k-1) = i) P(\mathrm{UQ}(k-1) = i) && (13)\\
&= \sum_{i=0}^{\ell} P(N(k) \geq \ell - i | \mathrm{UQ}(k-1) = i) P(\mathrm{UQ}(k-1) = i) && (14)\\
&= \sum_{i=0}^{\ell} P(N(k) \geq \ell - i) P(\mathrm{UQ}(k-1) = i) && (15)
\end{aligned}$$

Equation (12) gives $P(\mathrm{UQ}(k) = \ell)$ by substituting in Equation (11). Equation (13) is obtained by conditioning on the states of UQ at the beginning of time interval k (i.e., $\mathrm{UQ}(k-1)$). In the conditional probability of Equation (13), the equality $\min\{N(k), \ell - i\} = \ell - i$ holds if and only if $N(k) \geq \ell - i$. Hence, Equation (14) is obtained. Since the process of attempted arrivals does not have any dependence on the initial state of the system, $P(N(k) \geq \ell - i | UQ(k-1) = i) = P(N(k) \geq \ell - i)$ and Equation (15) is obtained.

Since the arrival process to the link, which is also the arrival process to $UQ(k)$, 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, $P(N(k) \geq \ell - i)$ is calculated as follows:

$$P(N(k) \geq \ell - i) = 1 - P(N(k) \leq \ell - i - 1) = 1 - e^{-\lambda(k)\delta} \sum_{j=0}^{\ell-i-1} \frac{(\lambda(k)\delta)^j}{j!}. \tag{16}$$

Equation (15) depends on the full marginal distribution of UQ (i.e., it depends on all terms $P(UQ(k-1) = i), \forall i \in \{0, \dots, \ell\}$). However, the proposed model does not track the full distribution of UQ, it only tracks the scalar probability $P(UQ(k-1) = \ell)$. Hence, we propose the following approximation for $P(UQ(k-1) = i), 0 \leq i \leq \ell - 1$:

$$\begin{cases} P(UQ(k-1) = i) = \dfrac{1 - P(UQ(k-1) = \ell)}{\sum_{j=0}^{\ell-1} f(j, q^{UQ}(k-1)\delta)} f(i, q^{UQ}(k-1)\delta) & \text{(17a)} \\ f(i, q^{UQ}(k-1)\delta) = \dfrac{(q^{UQ}(k-1)\delta)^i e^{-q^{UQ}(k-1)\delta}}{i!} & \text{(17b)} \\ q^{UQ}(k-1) = \displaystyle\sum_{r=0}^{k-2} q^{in}(r) - \sum_{r=0}^{k-k^{bwd}-2} q^{out}(r). & \text{(17c)} \end{cases}$$

Equation (17b) gives the probability mass function (pmf) of a Poisson distribution with parameter $q^{UQ}(k-1)\delta$. Equation (17a) is a normalized and finite support ($\{0, ..., \ell - 1\}$) Poisson distribution (with parameter $q^{UQ}(k-1)\delta$). The normalization term (the fraction term) is defined such that:

$$\sum_{i=0}^{\ell} P(UQ(k-1) = i) = P(UQ(k-1) = \ell) + \sum_{i=0}^{\ell-1} \frac{1 - P(UQ(k-1) = \ell)}{\sum_{j=0}^{\ell-1} f(j, q^{UQ}(k-1)\delta)} f(i, q^{UQ}(k-1)\delta) \tag{18}$$

$$= P(UQ(k-1) = \ell) + \frac{1 - P(UQ(k-1) = \ell)}{\sum_{j=0}^{\ell-1} f(j, q^{UQ}(k-1)\delta)} \sum_{i=0}^{\ell-1} f(i, q^{UQ}(k-1)\delta) \tag{19}$$

$$= P(UQ(k-1) = \ell) + 1 - P(UQ(k-1) = \ell) = 1 \tag{20}$$

Equation (17c) defines the expected flow in $UQ(k-1)$ 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 flow within UQ goes through, which were mentioned at the start of Section 2.2.2. This leads us to approximate $UQ(k)$ as an $M/G/\ell/\ell$ system. Let us detail this. Denote $S_{DQ}(k)$ the sojourn time of $DQ(k)$ system. First, since the service time of $UQ(k)$ is the sum of that of these three processes, we assume it to be generally distributed. The expected service time, $E[S_{UQ}(k)]$, is given by:

$$E[S_{UQ}(k)] = k^{fwd} + E[S_{DQ}(k)] + k^{bwd} \tag{21}$$

where $E[S_{DQ}(k)]$ is the expected sojourn time of the $DQ(k)$ system. Second, we approximate $UQ(k)$ 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 $k+1$);
$P_k(UQ = \ell)$	time-interval specific stationary probability of $UQ = \ell$;
$\tau_{UQ}(k)$	inverse of the relaxation time during time interval k.

If $q^{LLO}(k) > 0$, then we use the same functional form as for $DQ(k)$ (Eq. (5)) to approximate $P(UQ(k) = \ell)$:

$$P(UQ(k) = \ell) = P_k(UQ = \ell) + \Big[P(UQ(k-1) = \ell) - P_k(UQ = \ell)\Big] e^{-\tau_{UQ}(k)\delta}. \tag{22}$$

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 $P_k(UQ = \ell)$) and a term that decays exponentially with time and accounts for the difference between the initial conditions ($P(UQ(k-1) = \ell)$) and the corresponding stationary probability ($P_k(UQ = \ell)$).

The stationary probability $P_k(UQ = \ell)$ of Equation (22) is approximated as follows:

$$\begin{cases} P_k(UQ = \ell) = \dfrac{\rho_{UQ}(k)^\ell/\ell!}{\sum_{n=0}^{\ell} \rho_{UQ}(k)^n/n!} & \text{(23a)} \\ \rho_{UQ}(k) = \lambda(k)(k^{fwd} + E[S_{DQ}(k)] + k^{bwd}) & \text{(23b)} \\ E[S_{DQ}(k)] = \dfrac{\ell\rho_{DQ}(k)^{\ell+1} - (\ell+1)\rho_{DQ}(k)^\ell + 1}{\mu(k)(1-\rho_{DQ}(k)^\ell)(1-\rho_{DQ}(k))}. & \text{(23c)} \end{cases}$$

The system $M/G/\ell/\ell$ has been extensively studied and is known as the Erlang loss model. Studies of the transient blocking probability of an $M/G/\ell/\ell$ system include Jagerman (1974, 1975) and Davis et al. (1995). Jagerman (1975, Equation (166)) expresses the transient blocking probability of $M/M/\ell/\ell$ systems as the sum of the corresponding stationary probability (known as the Erlang-B formula, it is presented below) and a term that decays exponentially with time. We consider generally distributed, rather than Markovian, service times. Nonetheless, we use a similar functional form for the transient probability.

Equation (23a) 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 (23b) 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 Equation (21). Equation (23c) approximates the expected sojourn time (waiting plus service time) of $DQ(k)$. As discussed in Section 2.2.1, $DQ(k)$ is modeled as an $M/M/1/\ell$ queue with arrival rate $\lambda_{DQ}(k)$ and service rate $\mu(k)$. For such a queueing system, there is a closed-form expression for the expected sojourn time (e.g., Gross (2008, Chap. 2, Equations (2.48) and (2.51))), which yields Equation (23c).

The endogenous parameter $\tau_{UQ}(k)$ of Equation (22) 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 propose the following formulation for $\tau_{UQ}(k)$:

$$\begin{cases} \tau_{UQ}(k) = \alpha_4 \dfrac{\mu(k)}{C_{DQ}(k)\ell^2} + \alpha_5 \dfrac{\mu(k)}{\ell^2} & \text{(24a)} \\ C_{DQ}(k) = \sqrt{\mathrm{Var}(S_{DQ}(k))}/E[S_{DQ}(k)] & \text{(24b)} \\ \mathrm{Var}(S_{DQ}(k)) = [\ell\rho_{DQ}(k)^{2\ell+2} - 2\ell\rho_{DQ}(k)^{2\ell+1} + (\ell+1)\rho_{DQ}(k)^{2\ell} - \ell(\ell+1)\rho_{DQ}(k)^{\ell+2} & \\ \quad + 2\ell(\ell+1)\rho_{DQ}(k)^{\ell+1} - (\ell^2+\ell+2)\rho_{DQ}(k)^{\ell} + 1]/[\mu(k)^2(1-\rho_{DQ}(k)^{\ell})^2(1-\rho_{DQ}(k))^2], & \text{(24c)} \end{cases}$$

where α_4 and α_5 are exogenous scalar coefficients.

Studies of the relaxation time of an $M/G/\ell/\ell$ system are limited. Equations (24a) and (24b) are inspired from the work of Davis et al. (1995). Davis et al. (1995) study non-stationary Erlang loss models with a special focus on $M_t/PH/n/n$ systems. They observe that the inverse of 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. Recall that the service time of $UQ(k)$ (denoted $S_{UQ}(k)$) is the sum of three parts: two constant delays in LI and LO (i.e., k^{fwd} and k^{bwd}) and the sojourn time in $DQ(k)$ (denoted $S_{DQ}(k)$). Hence, the variability of the service time of $UQ(k)$ only comes from the variability of the sojourn time of $DQ(k)$. We use the coefficient of variance of $S_{DQ}(k)$, which is unit-free, as a measure of the variability of $S_{UQ}(k)$. The coefficient of variance of $S_{DQ}(k)$ (denoted $C_{DQ}(k)$) is defined by Equation (24b). In Equation (24a), $\tau_{UQ}(k)$ is inversely proportional to $C_{DQ}(k)$. This implies that the larger the variability of $S_{UQ}(k)$, the larger $C_{DQ}(k)$, the smaller $\tau_{UQ}(k)$, and thus the longer it takes to reach stationarity.

The variable $C_{DQ}(k)$ (of Eq. (24b)) is defined as a function of the expectation of $S_{DQ}(k)$, which is given by Equation (23c), and of the variance of $S_{DQ}(k)$, which is given by Equation (24c). The latter is obtained as follows. Recall from Section 2.2.1 that $DQ(k)$ is approximated as an $M/M/1/\ell$ system, the closed form expression for the variance of the sojourn time of such a system is derived in Appendix C and is given by Equation (24c).

Similar to our reasoning for $\tau_{DQ}(k)$, the expression for $\tau_{UQ}(k)$ should vary directly with the time units of the arrival and service rates. Equation (24a) contains two additive terms, both of which are directly proportional to $\mu(k)$ (note that $C_{DQ}(k)$ and ℓ are unit-free). Hence, $\tau_{UQ}(k)$ satisfies this property.

We use the same 84 simulation experiments described in Section 2.2.1 to fit the two exogenous scalar coefficients α_4 and α_5 of Equation (24a). A description of how these coefficients were estimated is given in Appendix D.

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 ($q^{in}(k)$) and expected link outflow ($q^{out}(k)$). All function evaluations 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 benchmark its performance versus the multivariate model of Osorio and Flötteröd (2015) and versus the mixture model of Lu and Osorio (2018). The analytical approximations provided by each of the three analytical models are compared to simulation-based estimates obtained from the stochastic simulator that was used as a benchmark in the validation experiments in Osorio and Flötteröd (2015) and in Lu and Osorio (2018).

The simulator is a discrete-event implementation of the link transmission model (LTM) (Yperman et al.; 2007). More specifically, it implements a stochastic LTM with 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 simulated estimates are obtained from 10^6 simulation replications.

The validation experiments are those used in Lu and Osorio (2018). We consider a link with parameters defined in Table 1. First, we consider two experiments with time-varying demand and evaluate the ability of the proposed model to approximate the 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 (i.e., $\mu(k) < \lambda(k)$) and then to congested. More specifically, it considers an arrival rate, $\lambda(k)$, of 0.1 veh/sec during the time $[0, 125]$ seconds, of 0.5 veh/sec for $[125, 175]$ seconds and of 0.3 veh/sec for $[175, 300]$ 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 time $[0, 100]$ seconds, of 0.1 veh/sec for $[100, 200]$ seconds and of 0.5 veh/sec for $[200, 300]$ 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

Algorithm 1 Link model algorithm

```
1. set exogenous link parameters $\hat{\rho}, \nu, w, \ell$ and the duration of each time interval $\delta$
2. compute the forward and backward lags: $k^{fwd} = \lceil \frac{\ell}{\hat{\rho}\nu} \rceil$ and $k^{fwd} = \lceil \frac{\ell}{\hat{\rho}|w|} \rceil$
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: $q^{in}(0)$, $q^{out}(0)$, $q^{UQ}(0)$, $q^{LLO}(0)$, $P(UQ(0) = \ell)$ 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. (10), respectively
   (b) compute $\rho_{DQ}(k)$ according to Eq. (6b)
   (c) compute $P_k(DQ = 0)$ and $\tau_{DQ}$ according to Eq. (6a) and Eq. (7b), respectively
   (d) compute $\tau_{DQ}(k)$ according to Eq. (7a)
   (e) compute $P(DQ(k) = 0)$ according to Eq. (5)
   (f) if $q^{LLO}(k) = 0$:
         i. compute $q^{UQ}(k-1)$ according to Eq. (17c)
        ii. compute $f(i, q^{UQ}(k-1)\delta)\ \forall i \in \{0, 1, ..., \ell - 1\}$ according to Eq. (17b)
       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. (16) and Eq. (17a), respectively
        iv. compute $P(UQ(k) = \ell)$ according to Eq.  (15)
       else:
         i. compute $E[S_{DQ}(k)]$ and $Var(S_{DQ}(k))$ according to Eq. (23c) and Eq. (24c), respectively
        ii. compute $\rho_{UQ}(k)$ and $C_{DQ}(k)$ according to Eq. (23b) and Eq. (24b), respectively
       iii. compute $P_k(UQ = \ell)$ and $\tau_{UQ}(k)$ according to Eq. (23a) and Eq. (24a), respectively
        iv. compute $P(UQ(k) = \ell)$ according to Eq.  (22)
   (g) 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

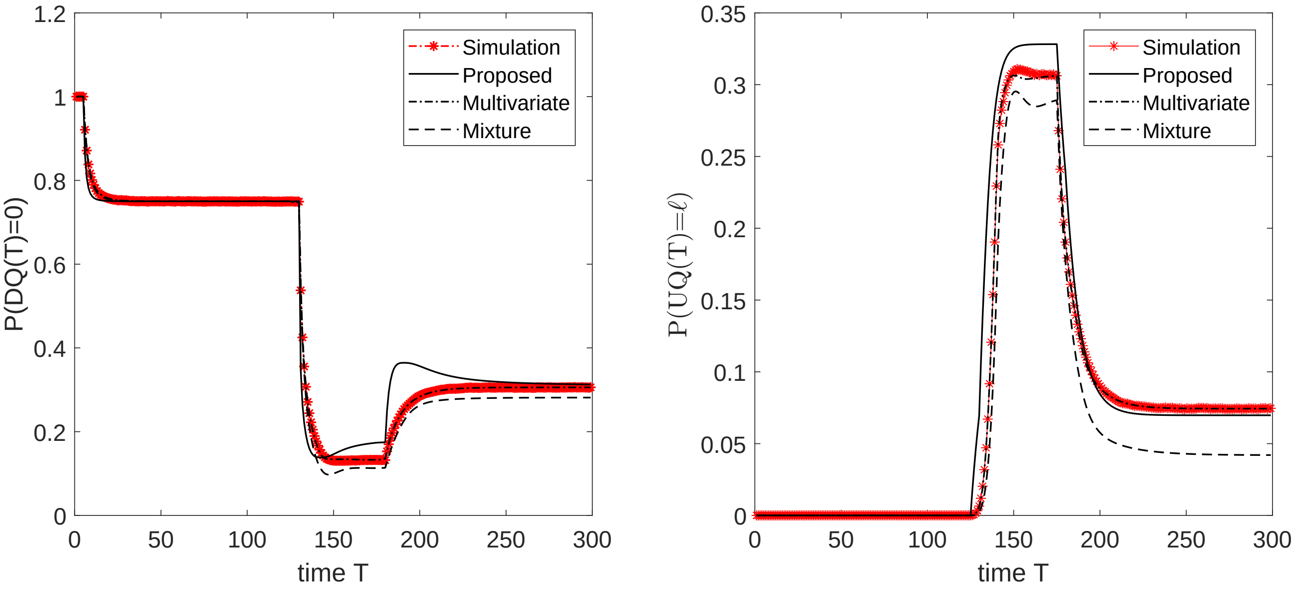


Figure 2: Experiment 1: impact of the temporal variation of demand on the link's upstream and downstream boundary conditions

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. The changes in $P(DQ(T) = 0)$ and $P(UQ(T) = \ell)$ are visible for all models. 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 largest deviation from the simulated estimates of $P(DQ(T) = 0)$ comes from the proposed model during the time interval $[125, 175]$. For $P(UQ(T) = \ell)$, the largest deviation comes from the mixture model during the time interval $[175, 300]$. Overall, the proposed model yields a good approximation of both probabilities $P(DQ(T) = 0)$ and $P(UQ(T) = \ell)$.

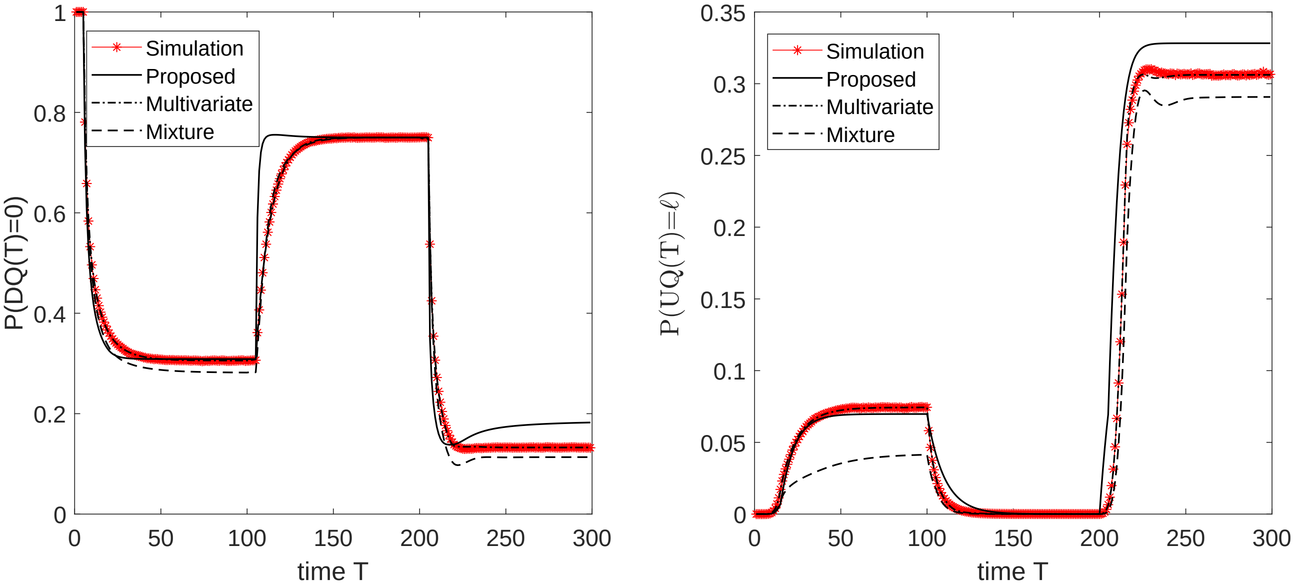


Figure 3: Experiment 2: impact of the temporal variation of demand on the link's upstream and downstream conditions

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. Once again, the multivariate model is the most accurate. The mixture model yields a less accurate approximation of the stationary value of the spillback probability during the congested and the highly congested regimes. The proposed model yields a less accurate approximation of both the stationary probability that DQ is empty and the stationary spillback probability during the highly congested regime. 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, which consider all combinations of the following arrival rates ($\lambda \in \{0.1, 0.2, 0.3\}$ 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 $k^{fwd} \in \{5, 10, 15, 20, 30, 40, 50\}$ (in seconds) and backward lags $k^{bwd} \in \{10, 20, 30, 40, 60, 80, 100\}$ (in seconds). All experiments start with an empty link and have a duration of 250 seconds. 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:

$$\bar{e}_{DQ} = \frac{1}{250}\sum_{T=1}^{250} |P_A(DQ(T) = 0) - P_S(DQ(T) = 0)| \tag{25}$$

$$\bar{e}_{UQ} = \frac{1}{250}\sum_{T=1}^{250} |P_A(UQ(T) = \ell) - P_S(UQ(T) = \ell)|, \tag{26}$$

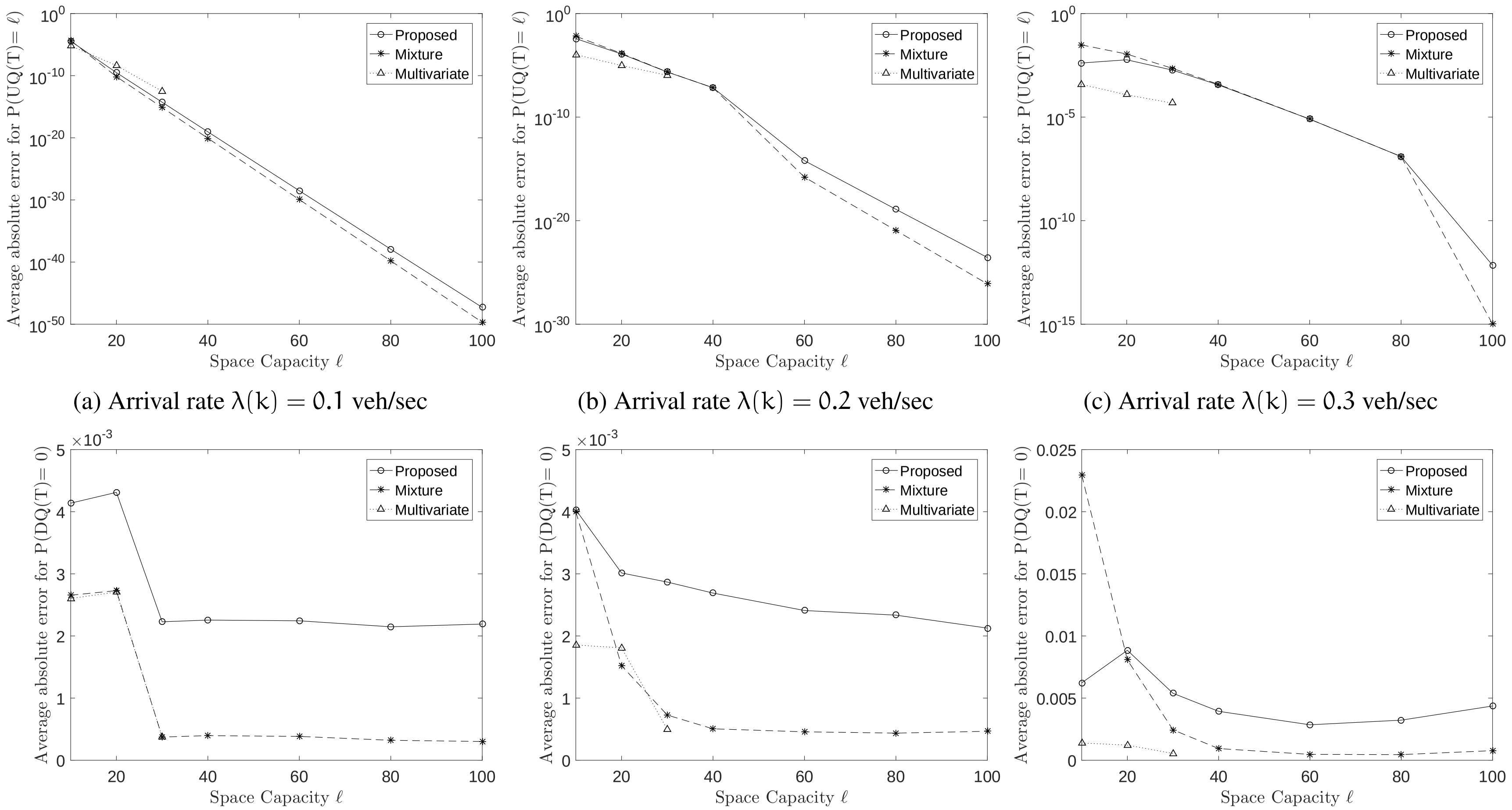
where P_A denotes the probability approximated by an analytical model (proposed, mixture or multivariate) and P_S denotes the simulated estimate.

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 0.1 veh/sec, 0.2 veh/sec and 0.3 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 the y-axis in logarithmic scale. For the experiments with space capacity greater than 30 ($\ell > 30$), the multivariate model does not conclude within 40 hours, hence 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 all analytical models, holding the space capacity ℓ constant, experiments with larger arrival rate have larger errors. Holding the arrival rate constant, experiments with smaller ℓ have larger errors. 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) first decreases and then remains around a low value for both the proposed model and the mixture model. The numerical values of the errors displayed in Figure 4 are provided in Appendix E (Tables 3 and 4). The average (over the 21 experiments) $\bar{e}_{UQ}$ and $\bar{e}_{DQ}$ of the proposed model are 0.0007 and 0.0035, respectively, whereas those of the mixture model are 0.0024 and 0.0025. 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. The proposed model performs well for both time-varying and constant demand 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 arrival rate 0.1 veh/sec, 0.2 veh/sec and 0.3 veh/sec, respectively. 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. As illustrated in Figure 5, regardless of the arrival rates, the runtime of the multivariate model increases



(a) Arrival rate $\lambda(k) = 0.1$ veh/sec

(b) Arrival rate $\lambda(k) = 0.2$ veh/sec

(c) Arrival rate $\lambda(k) = 0.3$ veh/sec

(d) Arrival rate $\lambda(k) = 0.1$ veh/sec

(e) Arrival rate $\lambda(k) = 0.2$ veh/sec

(f) Arrival rate $\lambda(k) = 0.3$ veh/sec

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

exponentially with ℓ, that of the mixture model increases linearly, while for the proposed model, the computational runtime appears constant over all experiments. The average runtime over the 21 experiments of the proposed model is 0.26 seconds, whereas that of the mixture model is 3.03 seconds. The average runtime is improved by one order of magnitude.

In summary, for all experiments with both constant or time-varying demand, 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 significant and increases as the space capacity increases.

4 Case study

In this section, we evaluate and benchmark the computational efficiency of the proposed model with a traffic signal control 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 presents the numerical results and evaluates the computational runtime of the proposed model compared to the mixture model. It also compares the derived signal plans with the signal plan proposed by a widely used commercial software.

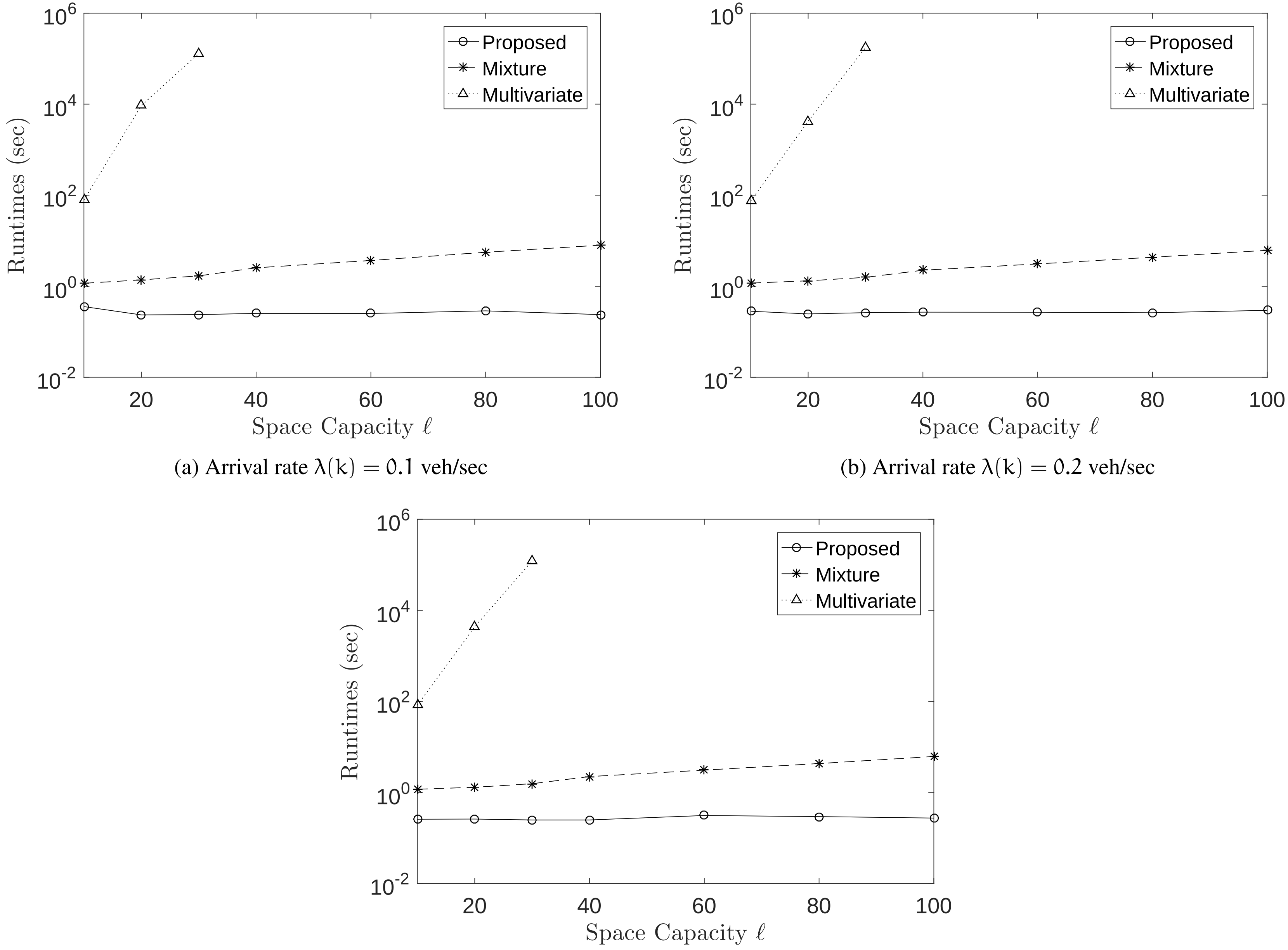
4.1 City-scale signal control

The Lausanne network consists of 603 links, 902 lanes and 231 intersections. We consider a fixed signal control problem in which we determine the signal plans of 17 intersections distributed throughout the network. 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 different intersections. All other traditional control variables (e.g., cycle times and offsets) are fixed. This lead 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 $\mathcal{L}$.

The problem is formulated as follows:

$$\min_{x} f(x) = \frac{1}{\tilde{T}N_1} \sum_{i \in \mathcal{L}} \sum_{\hat{t}=1}^{\tilde{T}} P(UQ_i(\hat{t};x) = \ell_i) \tag{27}$$



(a) Arrival rate $\lambda(k) = 0.1$ veh/sec

(b) Arrival rate $\lambda(k) = 0.2$ veh/sec

(c) Arrival rate $\lambda(k) = 0.3$ veh/sec

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

subject to

$$\sum_{j\in\mathcal{P}_D(d)} x(j) = b_d, \quad \forall d \in \mathcal{D} \tag{28}$$

$$x \geq x_{LB}. \tag{29}$$

The decision vector, x, is the green splits of the signal controlled lanes. Constraint (28) ensures that, for every intersection, the sum of the green times equals the available cycle time. Constraint (29) sets lower bounds, which are set to 4 seconds in this case study. $P(UQ_i(\hat{t};x) = \ell_i)$ denotes the spillback probability of lane i at integer time $\hat{t}$ under signal plan x. 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. For other implementation details, we refer the reader to Section 4.1 of Lu and Osorio (2018). 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). 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

For each model, Problem (27)-(29) is solved considering four different initial points. The initial points are drawn uniformly randomly from the feasible region (Equations (28)-(29)). The uniform sampling is conducted using the code of Stafford (2006).

For the proposed model, 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 mixture model, the algorithms do not converge within the time limit. Table 2 compares the average computation time (in minutes) per algorithmic iteration. Each column of Table 2 corresponds to a different initial point. For the proposed model, the average runtime per iteration is in the order of 1 minute, while for the mixture model it is in the order of 2.4 hours (i.e., 146 minutes). The proposed model reduces the runtime per iteration by 2 orders of magnitude.

Initial point	1	2	3	4
Mixture model	144.98	146.14	144.37	149.38
Proposed model	1.31	1.31	1.31	1.30

Table 2: Average runtime (in min) per iteration of the signal control optimization algorithm

We now compare the performance of the derived signal plans. To evaluate the performance of a given signal plan, we use a microscopic traffic simulation model of Lausanne (Dumont and Bert; 2006), which is calibrated for the evening peak period demand. It is implemented in the Aimsun software (TSS; 2014). 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. (27)), 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.

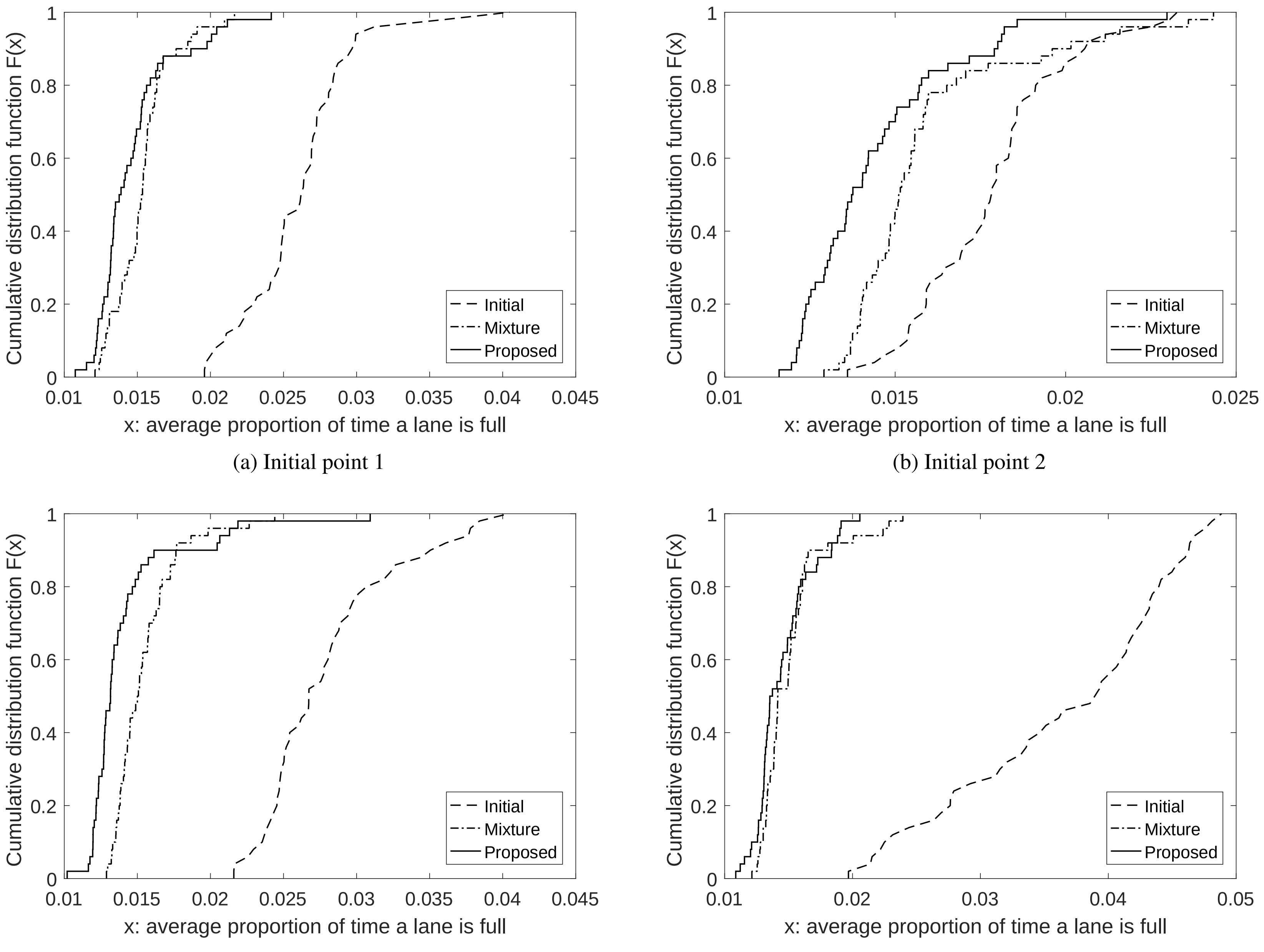
Each plot of Figure 6 considers a different initial signal plan and plots three cdf curves: one for the initial signal plan (dashed line), one for the solution derived by the proposed model (solid line), and one for the solution derived by the mixture model (dot-dashed line). The x-axis displays the 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. Hence, the more a cdf curve is shifted to the left, the better the performance of the corresponding signal plan. For all 4 plots (Fig. 6(a)-6(d)), the cdf of the derived signal plans, from both the mixture model and from the proposed model, are to the left of the initial signal plan. Hence, both models identify signal plans that outperform the initial signal plans. This holds for all initial signal plans. Paired-sample t-tests at a significance level of 5% are carried out to test for differences (in the average proportion of time a lane is full) between the signal plans derived by the proposed model and by the mixture model. For Figures 6(b) and 6(c), the plan derived by the proposed model has statistically improved performance compared to the plan derived by the mixture model. For Figures 6(a) and 6(d) both models yield plans with statistically similar performance.

Figures 7 compares the performance of the signal plans in terms of the average lane queue-length (in vehicles). As before, it compares the cdf curves of the different signal plans. It has a similar layout as Figure 6. 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 lane queue-lengths). All four plots in Figure 7 indicate that all derived signal plans outperform their corresponding initial signal plans. The derived signal plans from both analytical models have similar performance.

Figure 8 compares the performance of the signal plans in terms of the average trip travel times (in minutes). For all initial points, the corresponding derived solutions yield lower average trip travel times compared to the initial points. The signal plans derived by the proposed model and by the mixture model have similar performance.

We compare the performance of the derived signal plans with that of a signal plan proposed by the widely used commercial software Synchro (Trafficware; 2011). Synchro is a signal control optimization software based on a deterministic macroscopic traffic model, it does not solve the same optimization problem (27)-(29). For details on how the Synchro signal plan is derived, we refer the reader to Section 5.3 of Osorio and Chong (2015). Figures 9, 10 and 11 display, respectively, the three performance metrics: average proportion of time a lane is full, average lane queue-length and average trip travel time. Each figure contains 9 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 from 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.

In summary, compared to the mixture model, the proposed model improves the runtimes by 2 orders of magnitude, on average, and yields signal plans with improved or similar performance. This case



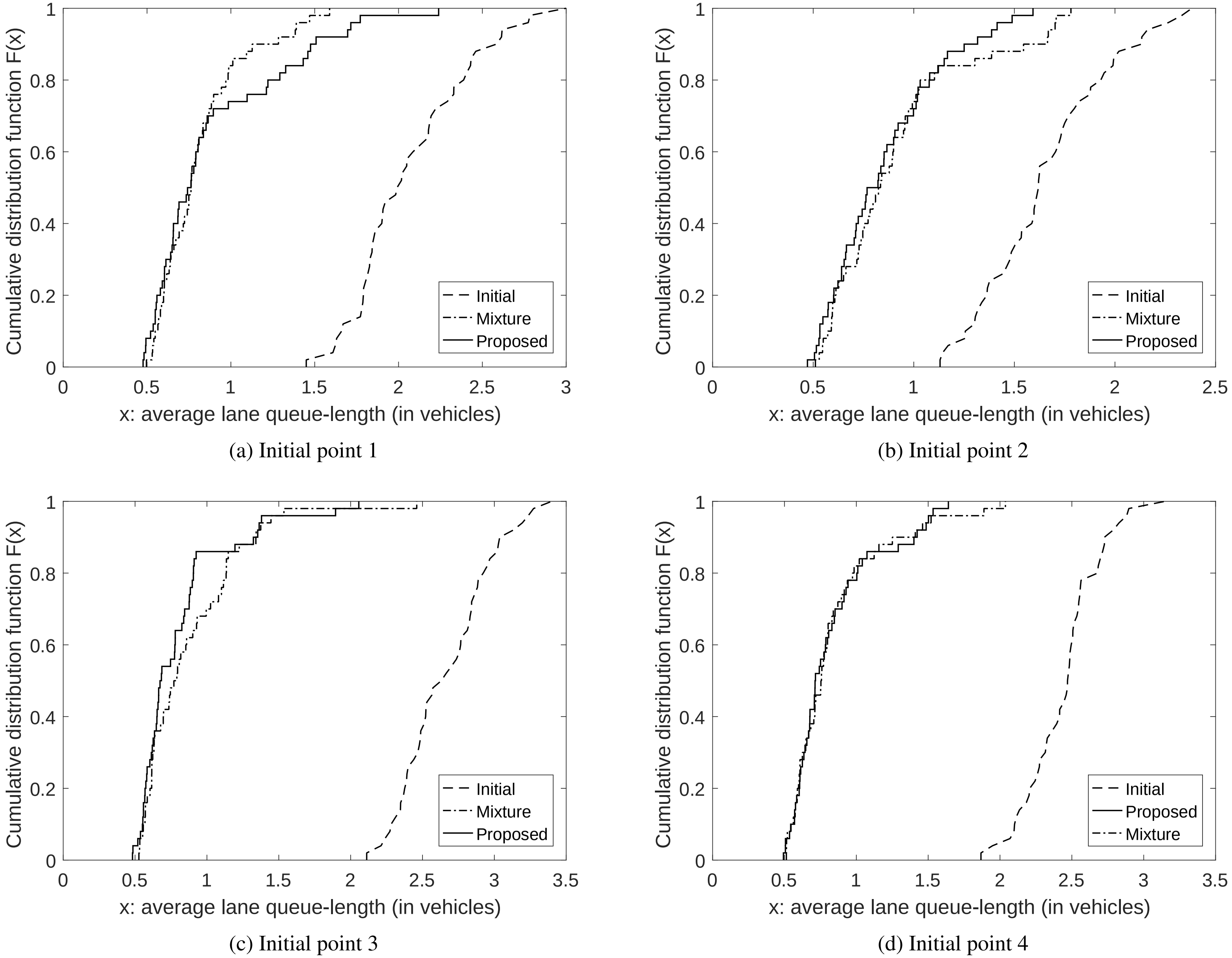
(a) Initial point 1

(b) Initial point 2

(c) Initial point 3

(d) Initial point 4

Figure 6: Cumulative distribution functions of the average proportion of time a lane is full, considering different initial signal plans



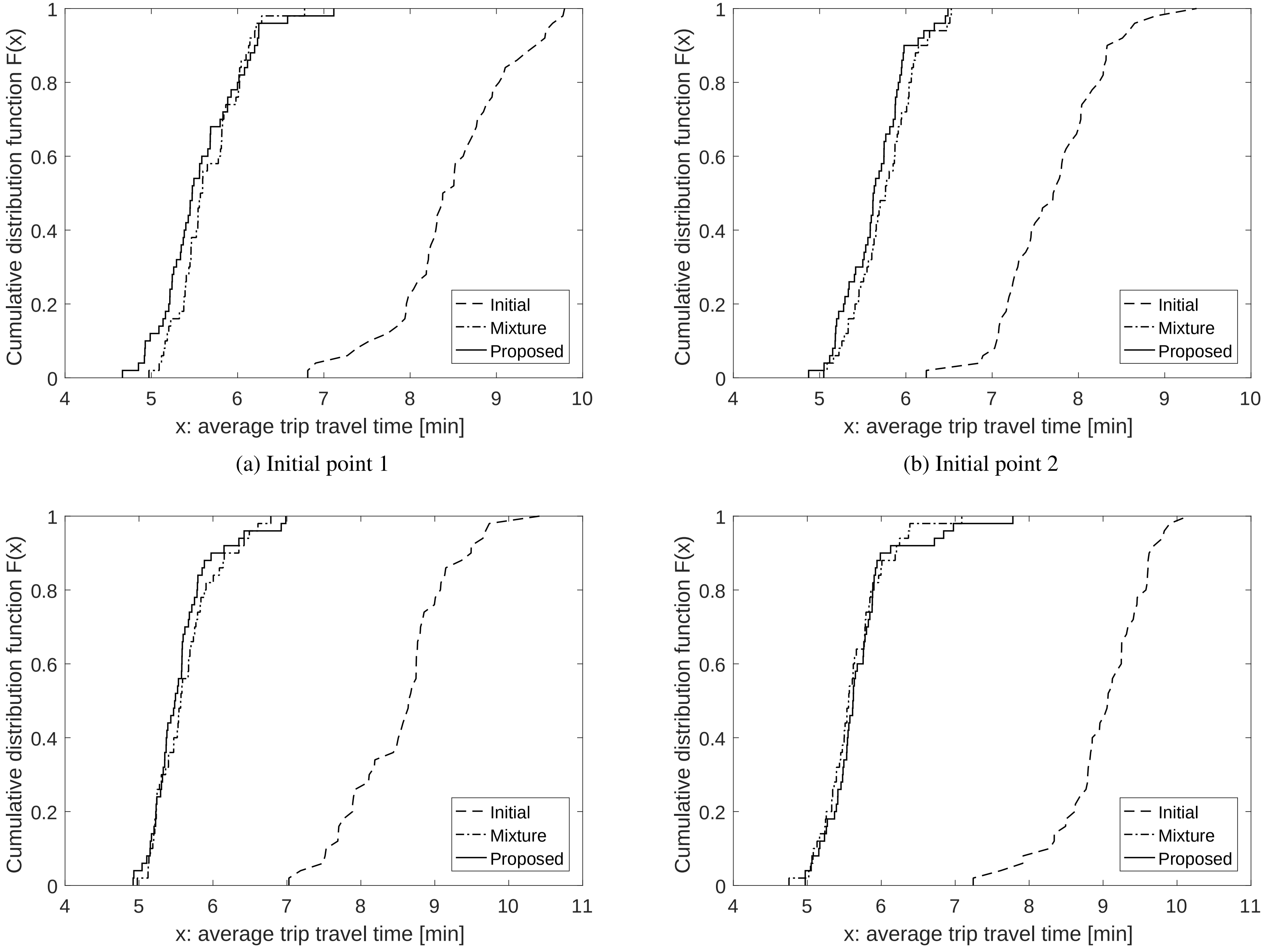
(a) Initial point 1

(b) Initial point 2

(c) Initial point 3

(d) Initial point 4

Figure 7: Cumulative distribution functions of the average lane queue-length considering different initial signal plans



(a) Initial point 1

(b) Initial point 2

(c) Initial point 3

(d) Initial point 4

Figure 8: Cumulative distribution functions of the average trip travel times considering different initial signal plans

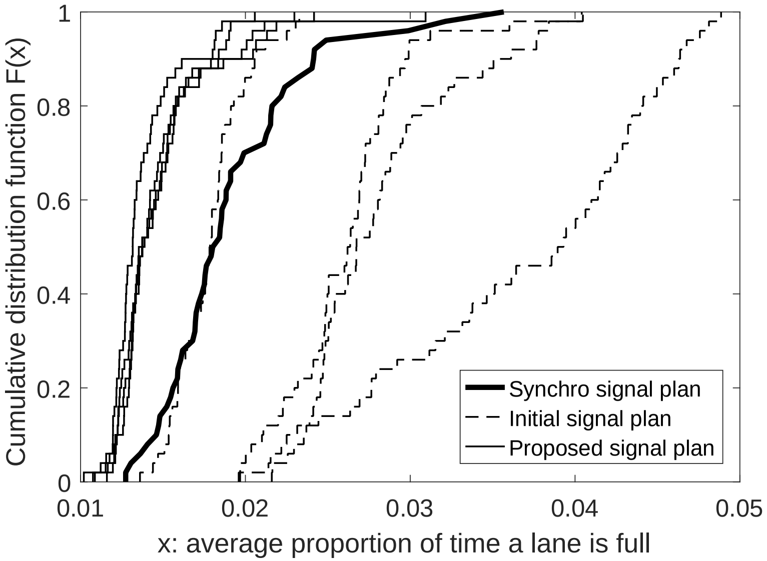


Figure 9: Cumulative distribution functions of the average proportion of time a lane is full

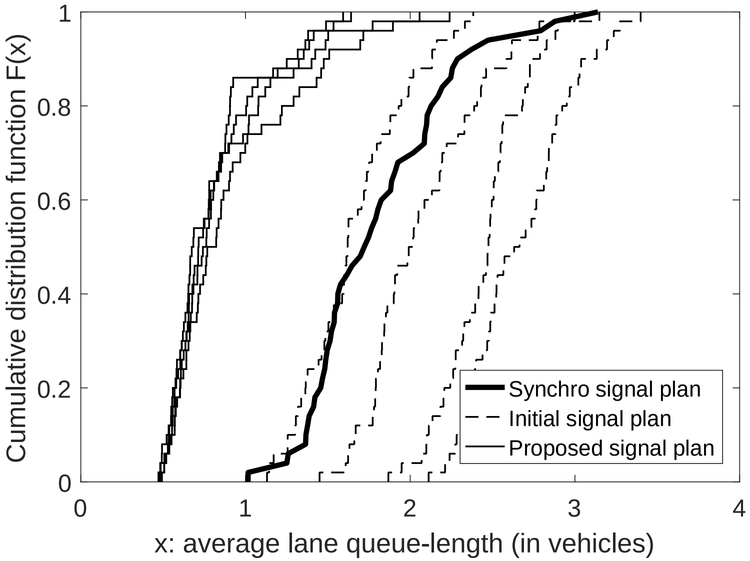


Figure 10: Cumulative distribution functions of the average lane queue-length

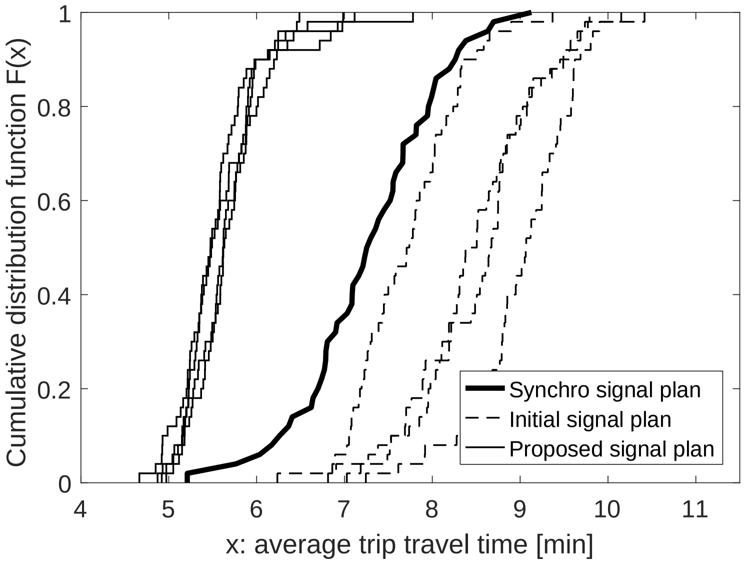


Figure 11: Cumulative distribution functions of the average trip travel time

study illustrates the scalability and efficiency of the proposed model. It 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 key 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) scales cubically and linearly, respectively, with the link's space capacity, the proposed model has a constant dimension of 2. Hence, 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. 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). 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 plans 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 2 orders of magnitude.

Future work 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 provides 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 part of ongoing work. Second, scalable and computationally efficient network model formulations are required. Consider a network of 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).

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.

Appendices

A Calculation of $\lim_{\ell\to\infty} \tau_{DQ}$ of Equation (8)

In this appendix, we derive the limit of τ_{DQ} (Equation (7b)) as ℓ goes to infinity.

$$\lim_{\ell\to\infty} \tau_{DQ} = \lim_{\ell\to\infty} \frac{\mu(k)(1-\rho(k))^2\ell + 1.5\lambda_{DQ}(k)\rho(k)\sqrt{\ell}}{(1+\rho(k))(\ell+1)} \tag{30}$$

$$= \lim_{\ell\to\infty} \frac{\mu(k)(1-\rho(k))^2\ell}{(1+\rho(k))(\ell+1)} + \lim_{\ell\to\infty} \frac{1.5\lambda(k)\rho(k)\sqrt{\ell}}{(1+\rho(k))(\ell+1)} \tag{31}$$

$$= \frac{\mu(k)(1-\rho(k))^2}{(1+\rho(k))} \lim_{\ell\to\infty} \frac{\ell}{\ell+1} + \frac{1.5\lambda(k)\rho(k)}{(1+\rho(k))} \lim_{\ell\to\infty} \frac{\sqrt{\ell}}{\ell+1} \tag{32}$$

$$= \frac{\mu(k)(1-\rho(k))^2}{(1+\rho(k))} \cdot 1 + \frac{1.5\lambda(k)\rho(k)}{(1+\rho(k))} \cdot 0 \tag{33}$$

$$= \frac{\mu(k)(1-\rho(k))^2}{(1+\rho(k))}. \tag{34}$$

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

This appendix describes the procedure to fit the exogenous coefficients (α_i for $i = 1, 2, 3$) of Equation (7). A total of 84 simulation experiments considering combinations of $\rho = \lambda/\mu \in \{0.25, 0.5, 0.75, 1.25\}$; $\mu \in \{0.2, 0.4, 0.6\}$; $\ell \in \{10, 20, 30, 40, 60, 80, 100\}$ are conducted. Each experiment considers a duration of 250 seconds. The simulator yields an estimate of $P(DQ(k) = 0)$, denoted $P_S(DQ(k) = 0)$, for all $k = 1, ..., 250$. The coefficients α_i (for $i = 1, 2, 3$) are fit such as to minimize, over all 84 experiments, the error function given by Equation (25) and rewritten here:

$$\bar{e}_{DQ} = \frac{1}{250} \sum_{T=1}^{250} |P_A(DQ(T) = 0) - P_S(DQ(T) = 0)|, \tag{35}$$

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,

- $P(UQ(k) = \ell)$ is obtained from the simulator;
- $\lambda_{DQ}(k)$ is obtained from the simulator;
- $\tau_{DQ}(k)$ is obtained from Equation (7), which depends on scalar parameters α_i, $i = 1, 2, 3$.

In other words, perfect information about the link's upstream boundary conditions is assumed in the calculation of $P_A(DQ(k) = 0)$. Hence, $P_A(DQ(k) = 0)$ only depends on the choice of α_i, $i = 1, 2, 3$ and thus the error function $\bar{e}_{DQ}$ only depends on α_i, $i = 1, 2, 3$. The scalars are estimated jointly and the numerical values obtained are: $\alpha_1 = 12$, $\alpha_2 = 1.5$ and $\alpha_3 = 1.5$.

C Variance of the sojourn time of $\mathrm{DQ}(k)$

In this section, we derive the expression for the variance of the sojourn time of $\mathrm{DQ}(k)$. Recall that $\mathrm{DQ}(k)$ is approximated as an $M/M/1/\ell$ queue with arrival rate $\lambda_{\mathrm{DQ}}(k)$ and service rate $\mu(k)$. To make the notation simpler, hereafter the time index k is dropped. Let $\rho = \lambda_{\mathrm{DQ}}/\mu$.

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

$$f_{S_{\mathrm{DQ}}}(t) = \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \frac{P(\mathrm{DQ}=n)}{1-P(\mathrm{DQ}=\ell)} \tag{36}$$

where $P(\mathrm{DQ}=n)$ is the steady state probability of DQ.

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

$$\begin{align}
E[S_{\mathrm{DQ}}^2] &= \int_0^\infty t^2 f_{S_{\mathrm{DQ}}}(t)dt \tag{37}\\
&= \int_0^\infty t^2 \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \frac{P(\mathrm{DQ}=n)}{1-P(\mathrm{DQ}=\ell)} dt \tag{38}\\
&= \int_0^\infty t^2 \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \frac{\left(\frac{1-\rho}{1-\rho^{\ell+1}}\right)\rho^n}{1-\left(\frac{1-\rho}{1-\rho^{\ell+1}}\right)\rho^\ell} dt \tag{39}\\
&= \frac{(1-\rho)}{1-\rho^\ell} \int_0^\infty t^2 \sum_{n=0}^{\ell-1} \mu \frac{(\mu t)^n}{n!} e^{-\mu t} \rho^n dt \tag{40}\\
&= \frac{(1-\rho)}{1-\rho^\ell} \sum_{n=0}^{\ell-1} \rho^n \int_0^\infty \frac{t(\mu t)^{n+1}}{n!} e^{-\mu t} dt \tag{41}\\
&= \frac{(1-\rho)}{1-\rho^\ell} \sum_{n=0}^{\ell-1} \rho^n \frac{\Gamma(n+3)}{\mu^2 n!} \tag{42}\\
&= \frac{(1-\rho)}{(1-\rho^\ell)\mu^2} \sum_{n=0}^{\ell-1} \rho^n (n+2)(n+1) \tag{43}\\
&= \frac{-\ell(\ell+1)\rho^{\ell+2} + 2\ell(\ell+2)\rho^{\ell+1} - (\ell+1)(\ell+2)\rho^\ell + 2}{\mu^2(1-\rho^\ell)(1-\rho)^2} \tag{44}
\end{align}$$

Equation (39) is obtained from Equation (38) 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:

$$P(\mathrm{DQ}=n) = \left(\frac{1-\rho}{1-\rho^{\ell+1}}\right)\rho^n, \ \ \forall n \in \{0, ..., \ell\}. \tag{45}$$

The expected value of the sojourn time of DQ is given by (see Eq. (23c)):

$$E[S_{DQ}] = \frac{\ell\rho^{\ell+1} - (\ell+1)\rho^{\ell} + 1}{\mu(1-\rho^{\ell})(1-\rho)} \tag{46}$$

Hence the variance of the sojourn time of DQ is given by:

$$\begin{aligned} Var(S_{DQ}) &= E[S_{DQ}^2] - E[S_{DQ}]^2 \qquad (47)\\ &= \frac{\ell\rho^{2\ell+2} - 2\ell\rho^{2\ell+1} + (\ell+1)\rho^{2\ell} - \ell(\ell+1)\rho^{\ell+2} + 2\ell(\ell+1)\rho^{\ell+1} - (\ell^2+\ell+2)\rho^{\ell} + 1}{\mu^2(1-\rho^{\ell})^2(1-\rho)^2} \qquad (48)\end{aligned}$$

D Estimation of the scalar coefficients in $\tau_{UQ}(k)$

This section describes the procedure to fit the exogenous coefficients α_4 and α_5 of Equation (24a). The same set of 84 simulation experiments as described in Appendix B are used. The coefficients α_i for $i = 4, 5$ are fit by minimizing, over all 84 experiments, the following error function given by Equation (26) and rewritten here:

$$\bar{e}_{UQ} = \frac{1}{250}\sum_{T=1}^{250} |P_A(UQ(T) = \ell) - P_S(UQ(T) = \ell)|, \tag{49}$$

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,

- $P(DQ(k) = 0)$ is obtained from the simulator;
- $q^{UQ}(k)$ and $q^{LLO}(k)$ are obtained from the simulator;
- $\tau_{UQ}(k)$ is obtained from Equation (24), which depends on scalar parameters α_i, $i = 4, 5$.

In other words, perfect information about the link's downstream boundary conditions is assumed in the calculation of $P_A(UQ(T) = \ell)$. Hence, $P_A(UQ(T) = \ell)$ only depends on the choice of α_i, $i = 4, 5$ and thus the error function $\bar{e}_{UQ}$ depends only on α_i, $i = 4, 5$. The scalars are estimated jointly and the numerical values obtained are: $\alpha_4 = 0.1$ and $\alpha_5 = 25$.

E Tables of mean absolute differences

Tables 3 and 4 display, respectively, the mean absolute error of the link's upstream and downstream boundary conditions.

Experiment		$\bar{e}_{\mathrm{UQ}}$		
$\lambda(k)$	ℓ	Mixture	Multivariate	Proposed
	10	$4.60e-5$	$6.50e-6$	$3.67e-5$
	20	$7.07e-11$	$4.29e-9$	$3.11e-10$
	30	$7.58e-16$	$3.17e-13$	$5.39e-15$
0.1	40	$8.54e-21$	NaN	$9.20e-20$
	60	$1.14e-30$	NaN	$3.01e-29$
	80	$1.55e-40$	NaN	$1.18e-38$
	100	$2.05e-50$	NaN	$5.09e-48$
	10	$6.44e-3$	$1.07e-4$	$3.79e-3$
	20	$1.41e-4$	$9.56e-6$	$1.23e-4$
	30	$2.32e-6$	$1.08e-6$	$2.26e-6$
0.2	40	$6.82e-8$	NaN	$6.86e-8$
	60	$1.59e-16$	NaN	$6.17e-15$
	80	$1.12e-21$	NaN	$1.24e-19$
	100	$7.78e-27$	NaN	$2.75e-24$
	10	$3.11e-2$	$3.85e-4$	$4.08e-3$
	20	$1.11e-2$	$1.19e-4$	$5.85e-3$
	30	$2.27e-3$	$4.73e-5$	$1.86e-3$
0.3	40	$3.85e-4$	NaN	$3.62e-4$
	60	$8.04e-6$	NaN	$7.98e-6$
	80	$1.25e-7$	NaN	$1.25e-7$
	100	$1.01e-15$	NaN	$7.02e-13$

Table 3: Mean absolute difference $\bar{e}_{\mathrm{UQ}}$ of $P(\mathrm{UQ}(k) = \ell)$. The value NaN denotes cases where the evaluation of the multivariate model exceeded the limit of 40 hours.

References

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

Experiment		$\bar{e}_{DQ}$		
$\lambda(k)$	ℓ	Mixture	Multivariate	Proposed
0.1	10	$0.27e-2$	$0.26e-2$	$0.41e-2$
	20	$0.28e-2$	$0.27e-2$	$0.43e-2$
	30	$0.04e-2$	$0.04e-2$	$0.22e-2$
	40	$0.04e-2$	NaN	$0.23e-2$
	60	$0.04e-2$	NaN	$0.22e-2$
	80	$0.03e-2$	NaN	$0.21e-2$
	100	$0.03e-2$	NaN	$0.22e-2$
0.2	10	$0.40e-2$	$0.19e-2$	$0.40e-2$
	20	$0.15e-2$	$0.18e-2$	$0.30e-2$
	30	$0.07e-2$	$0.05e-2$	$0.29e-2$
	40	$0.05e-2$	NaN	$0.27e-2$
	60	$0.05e-2$	NaN	$0.24e-2$
	80	$0.04e-2$	NaN	$0.23e-2$
	100	$0.05e-2$	NaN	$0.21e-2$
0.3	10	$2.30e-2$	$0.14e-2$	$0.62e-2$
	20	$0.81e-2$	$0.12e-2$	$0.88e-2$
	30	$0.24e-2$	$0.05e-2$	$0.54e-2$
	40	$0.09e-2$	NaN	$0.39e-2$
	60	$0.05e-2$	NaN	$0.29e-2$
	80	$0.05e-2$	NaN	$0.32e-2$
	100	$0.08e-2$	NaN	$0.44e-2$

Table 4: Mean absolute difference $\bar{e}_{DQ}$ of $P(DQ(k) = \ell)$. The value NaN denotes cases where the evaluation of the multivariate model exceeded the limit of 40 hours.

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

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: Methodological* **102**: 180–209.

Gross, D. (2008). *Fundamentals of queueing theory*, John Wiley & Sons, New York, U.S.

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.

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: Methodological* **68**: 205–223.

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

Jagerman, D. L. (1974). Some properties of the Erlang loss function, *Bell System Technical Journal* **53**(3): 525–551.

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

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.

Lu, J. and Osorio, C. (2018). A probabilistic traffic-theoretic network loading model suitable for large-scale network analysis, *Forthcoming in Transportation Science* .
Available at: http://web.mit.edu/osorioc/www/papers/luOso17.pdf .

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.

Newell, C. (1982). *Applications of queueing theory*, Chapman and Hall, New York, USA.

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. (2002). A simplified car-following theory: a lower order model, *Transportation Research Part B: Methodological* **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.

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.

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. Accessed July 20, 2018.
URL: *https://pdfs.semanticscholar.org/848f/a1f48ad9d3edb24b05667f15cfc633eb8f69.pdf*

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).

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.

Ypperman, 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.