
Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# A Probabilistic Traffic-Theoretic Network Loading Model Suitable for Large-Scale Network Analysis

Jing Lu, Carolina Osorio

To cite this article:
Jing Lu, Carolina Osorio (2018) A Probabilistic Traffic-Theoretic Network Loading Model Suitable for Large-Scale Network Analysis. Transportation Science 52(6):1509-1530. https://doi.org/10.1287/trsc.2017.0804

http://pubsonline.informs.org/journal/trsc/

TRANSPORTATION SCIENCE
Vol. 52, No. 6, November–December 2018, pp. 1509–1530
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

# A Probabilistic Traffic-Theoretic Network Loading Model Suitable for Large-Scale Network Analysis

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

[a] Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [b] Department of Civil and Environmental Engineering, School of Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

**Contact:** jl3724@mit.edu, http://orcid.org/0000-0002-9937-8067 (JL); osorioc@mit.edu, http://orcid.org/0000-0003-0979-6052 (CO)

**Received:** September 2, 2016
**Revised:** April 17, 2017; July 10, 2017
**Accepted:** July 21, 2017
**Published Online in Articles in Advance:** June 26, 2018

**https://doi.org/10.1287/trsc.2017.0804**



**Abstract.** This paper formulates an analytical stochastic network loading model. It is a stochastic formulation of the link transmission model (LTM), which itself is an operational formulation of Newell's simplified theory of kinematic waves. The proposed model builds on an existing initial model. It proposes a formulation with enhanced scalability. In particular, compared with the initial model, it has a complexity that is linear rather than cubic in the link's space capacity. This makes it suitable for large-scale network analysis. The model is validated versus a simulation-based implementation of the stochastic LTM. The proposed model yields significant gains in computational efficiency while preserving accuracy. The validation experiments illustrate how computational run times of the proposed model increase linearly with the link's space capacity, while the initial model has an exponential increase in run times. The proposed model yields accurate distributional approximations of the link's boundary conditions. It is used to address a probabilistic formulation of a citywide signal control problem. The model is shown to be robust to the quality of the initial signal plans. It yields signal plans that systematically outperform both initial plans, as well as a plan derived by widely used commercial signal control software. The model is suitable for large-scale network optimization.

**Funding:** The work of J. Lu and C. Osorio was partially supported by the U.S. National Science Foundation [Grant CMMI-1562912].
**Supplemental Material:** The online appendix is available at https://doi.org/10.1287/trsc.2017.0804.

**Keywords:** link transmission model • stochastic network loading model • traffic flow theoretic model

## 1. Introduction

This paper focuses on the formulation of stochastic (i.e., probabilistic) network loading models of road traffic. The vast majority of the literature in the field of traffic flow theory has focused on the development of deterministic traffic models. There has been a recently renewed interest in the development of analytical stochastic models that is, arguably, triggered by both (i) the interest of major transportation agencies around the world in estimating and improving the robustness and reliability of their networks (Transport for London 2010, U.S. Department of Transportation 2008) and (ii) the availability of high-resolution traffic data, which enable the validation of more detailed models.

In a transportation network, there are sources of uncertainty both in supply (e.g., weather) and in demand (e.g., spatial and temporal distribution of travel demand, heterogeneous population of travelers). Studies that have reviewed sources and modeling approaches to demand and supply uncertainty include Sumalee et al. (2011) and Lam, Shao, and Sumalee (2008). For instance, in the field of microscopic travel demand modeling, a variety of probabilistic models have been developed to account for uncertainties in various travel choices such as departure time, mode, route, etc. In the field of macroscopic modeling, the variability (or scatter) in the fundamental diagrams has led the community to develop probabilistic models to better interpret and fit field data. A review of recent approaches to model, or account for, the variability in fundamental diagrams is given in Sumalee et al. (2011) and in Jabari, Zheng, and Liu (2014). For instance, the work of Heidemann (2001) uses a probabilistic nonstationary (i.e., transient) traffic model to interpret hysteresis loops, and the case study in Sumalee et al. (2011) uses a probabilistic model to improve the fit of a fundamental diagram with high scatter. Nonetheless, there is a lack of probabilistic traffic models that are both (i) consistent with mainstream traditional deterministic traffic flow theoretic models and (ii) tractable enough to enable the efficient analysis and optimization of large-scale networks. The main contribution of this paper is to *formulate a probabilistic link model that is both* (i) *consistent with mainstream deterministic traffic flow*

*theory and* (ii) *computationally tractable to enable large-scale network analysis.*

Jabari (2012) and Laval and Chilukuri (2014) provide reviews of stochastic traffic flow theoretic models. Recent formulations include those derived from the variational theory of Daganzo (2005) (e.g., Deng, Lei, and Zhou 2013; Laval and Chilukuri 2014; Laval and Castrillón 2015). The most popular approach to stochastic traffic modeling is the formulation of stochastic cell-transmission models (CTMs; e.g., Boel and Mihaylova 2006, Jabari and Liu 2012, Sumalee et al. 2011). The approach of Boel and Mihaylova (2006) is an example of the most common approach to stochastic CTMs in that it adds Gaussian noise terms to the deterministic formulation. This contributes to model tractability yet does not guarantee expected (i.e., average) traffic dynamics consistent with the CTM dynamics. The implications of this are further discussed in Jabari and Liu (2012). The model of Jabari and Liu (2012) considers stochastic vehicle headways. It allows for a variety of headway distributions and has a fluid limit approximation that is consistent with the CTM. Boel and Mihaylova (2006) and Jabari and Liu (2012) are sampling-based approaches, which can become computationally intensive for large-scale networks. Jabari and Liu (2013) propose a second-order Gaussian approximation of the model of Jabari and Liu (2012) that can be evaluated without sampling. The CTM is a space-discretized approximation of the kinematic wave model (KWM; Lighthill and Whitham 1955, Richards 1956); hence a stochastic CTM formulation does not guarantee consistency with the KWM.

The recent work of Osorio and Flötteröd (2015) extends the model of Osorio, Flötteröd, and Bierlaire (2011) and proposes a link model that is a stochastic formulation of the deterministic link transmission model of Yperman, Tampere, and Immers (2007), which itself is an operational formulation of Newell's simplified theory of kinematic waves (Newell 1993). The model considers an isolated link and derives an analytical description of the transient (i.e., time-dependent) distribution of link's boundary conditions. It yields the joint distribution of the link's upstream and downstream boundary conditions. Hence, it provides a higher-order (i.e., beyond first-order) description of within-link dependencies. The model represents the link as a set of three finite space capacity stochastic queues. For a link with space capacity $l$, the dimension of the state space of the joint distribution is $\frac{1}{6}(l+1)(l^2+2l+6)$. In other words, the model complexity is in the order of $\mathcal{O}(l^3)$.

This paper formulates a link model with a complexity that is linear rather than cubic in the link's space capacity; that is, the proposed model has $\mathcal{O}(l)$ complexity. It is therefore scalable and appropriate for large-scale network analysis. The proposed model is derived from the model of Osorio and Flötteröd (2015). It is therefore a stochastic formulation of Newell's simplified theory of kinematic waves (Newell 1993).

Section 2 formulates the proposed model. The model is validated in Section 3 and used to address a large-scale signal control problem in Section 4. Conclusions and a discussion of ongoing work are presented in Section 5. The online appendices contain additional numerical validation results.

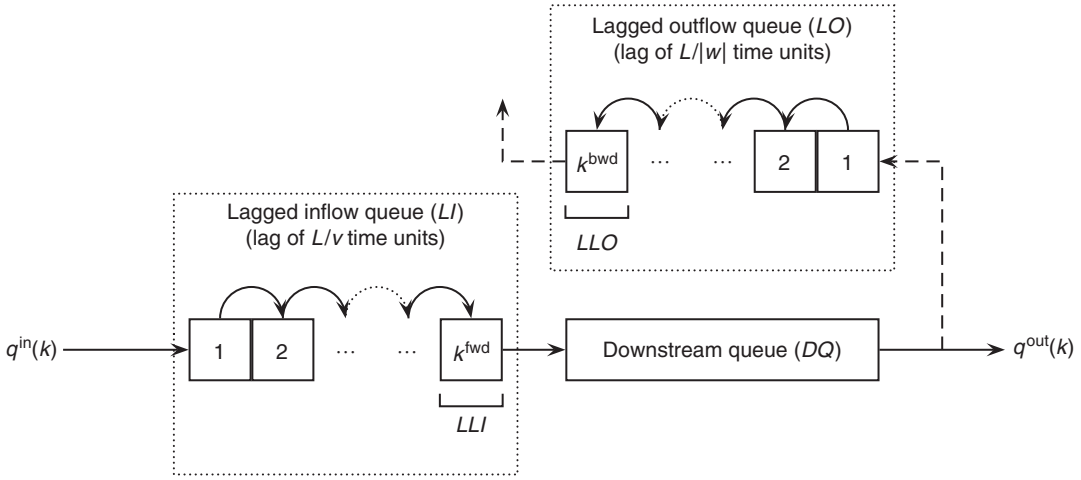
## 2. Link Model Formulation

### 2.1. Multivariate Link Model

We outline here the main ideas of the model of Osorio and Flötteröd (2015). Hereafter, we refer to the Osorio and Flötteröd (2015) model as the *multivariate* (*link*) *model*. For a description of how this model relates to Newell's simplified theory of kinematic waves or to the operational formulation of Yperman, Tampere, and Immers (2007), we refer the reader to Osorio and Flötteröd (2015). Consider a link with a triangular fundamental diagram, free-flow velocity $v$, backward wave speed $w$ (negative), flow capacity $\hat{q}$, jam density $\hat{\rho}$, and link length $L$. The process that vehicular traffic flow goes through within the link is described as follows. Upon entrance into the link, a vehicle is delayed by $L/v$ time units. It is then ready for departure and enters the physical vehicular queue downstream, if one exists. Upon departure from the link, the vehicle experiences an additional delay of $L/|w|$ before the newly available space becomes available upstream of the link. This delay represents the time it takes a kinematic backward wave to traverse the link. The multivariate model is a continuous space discrete-time model, where $L/v$ (respectively, $L/|w|$) is rounded to the integer $k^{\text{fwd}}$ (respectively, $k^{\text{bwd}}$).

This process is summarized in Figure 1. During time interval $k$, the link has an expected inflow (respectively, outflow) denoted by $q^{\text{in}}(k)$ (respectively, $q^{\text{out}}(k)$). The delay incurred upon entrance to the link is represented by the *lagged inflow queue*, denoted by *LI*. In discrete time, *LI* can be thought of as a set of $k^{\text{fwd}}$ cells. One can think of this delay as if the flow traveled sequentially from the first until the $k^{\text{fwd}}$th cell of *LI*. This last cell of *LI* is denoted by *LLI* in Figure 1. This cell configuration of *LI* is a mere representation; the multivariate model describes *LI* aggregately (i.e., it is not decomposed into individual cells). After this delay, the flow enters the *downstream queue*, denoted by *DQ*. The departure of flow from the link triggers two events: the flow departs *DQ* (in a network setting, it would enter a downstream link), and it enters the *lagged outflow queue*, denoted by *LO*. The purpose of *LO* is to capture the kinematic backward wave delay. One can think of this delay as if the newly available space traveled sequentially from the first until the $k^{\text{bwd}}$th cell of *LO*. This last cell of *LO* is denoted by *LLO* in Figure 1. The multivariate link

**Figure 1.** Link Dynamics of the Multivariate Link Model




model accounts for stochasticity in the link's arrival and departure processes. Time-dependent (i.e., inhomogeneous, nonhomogeneous) finite-state birth-death processes are assumed. This leads to stochastic link flows, to stochastic cumulative flows both upstream and downstream of the link, and hence, to a stochastic description of link states.

The multivariate model jointly tracks the dynamics between the three queues $LI$, $DQ$, and $LO$. It also defines the *upstream queue*, $UQ$, as

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

More specifically, the model is a discrete-time model. We denote $LI(t;k)$ (respectively, $DQ(t;k)$, $LO(t;k)$, and $UQ(t;k)$) as the number of vehicles in $LI$ (respectively, $DQ$, $LO$, and $UQ$) at continuous time $t$ within discrete-time interval $k$ of duration $\delta$. The model yields the joint distribution of $P(LI(t;k), DQ(t;k), LO(t;k), UQ(t;k))$. The linear equality (1) implies that this four-dimensional joint distribution can be obtained by tracking three of the four variables. The model implementation of Osorio and Flötteröd (2015) tracks $LI, DQ$, and $LO$. For a given link with space capacity $l$ (which is defined as a rounded version of $\hat{\rho}L$), the state space is defined by $\{(li, dq, lo) \in \{0, \ldots, l\}^3 : li + dq + lo \leq l\}$. The state space dimension is $\frac{1}{6}(l+1)(l^2+2l+6)$.

In this paper, we propose a formulation with a state space dimension that is linear instead of cubic in the space capacity while still providing a detailed representation of the within-link dependencies. This enables its use for the efficient analysis and optimization of large-scale networks.

## 2.2. Univariate Link Models

Hereafter, unless necessary, we drop the time dependency notation and use $LI$ or $LI(k)$ to denote $LI(t;k)$. We do the same for $DQ, LO$, and $UQ$. The main insights of the multivariate model that underlie the newly proposed formulation are the following: $UQ$ provides a detailed description of the link's upstream boundary conditions, while $DQ$ provides a detailed description of the link's downstream boundary conditions. One approach would be to propose a model that jointly describes $(UQ, DQ)$. This would improve model scalability by going from a three- to a two-dimensional state space. The idea considered in this paper goes even further: it proposes a univariate (i.e., one-dimensional) state space, which leads to a scalable formulation. We consider the following two independent univariate models.

- One model of $UQ$: Its purpose is to accurately capture the link's upstream boundary conditions.
- One model of $DQ$: Its purpose is to accurately capture the link's downstream boundary conditions.

The proposed model is then defined as a mixture of these two independent univariate models. There is significant dependency between the upstream and the downstream boundary conditions of a link, as illustrated by Equation (1). In other words, there is dependency between the dynamics of $UQ$ and of $DQ$. The numerical case studies in Osorio and Flötteröd (2015) analyze this dependency in more detail. The main challenge addressed in this paper is therefore to develop independent univariate models of $UQ$ and of $DQ$ while still capturing the dependency between the link's upstream and downstream boundary conditions.

Consider an isolated link with space capacity $l$, an inhomogeneous Poisson arrival process with exogenous arrival rate $\lambda(k)$ (time is indexed by $k$), and exponentially distributed service times at the downstream end of the link with exogenous downstream bottleneck flow capacity $\mu(k)$. For this isolated link, Section 2.3 (respectively, 2.4) formulates a univariate model that tracks the distribution of $UQ$ (respectively, $DQ$) over

time. Section 2.5 then formulates the proposed mixture model, which combines the *UQ* and the *DQ* models.

### 2.3. Univariate Upstream Queue Model

This section formulates a univariate model of *UQ*. Following the approach in Osorio and Flötteröd (2015), *UQ* is modeled as a birth-death process with a finite state space defined by $uq \in \{0, \ldots, l\}$. For time interval $k$ of duration $\delta$, the transient probability distribution of *UQ* satisfies a system of linear differential equations with a solution defined by (see, for instance, Reibman 1991 for details)

$$P(UQ(t;k)) = P(UQ(0;k))e^{tQ^{UQ}(k)} \quad \forall\, t \in [0, \delta], \tag{2}$$

where $P(UQ(0;k))$ are the initial conditions at the beginning of time interval $k$, and $Q^{UQ}(k)$ is the transition rate matrix of *UQ*.

The initial conditions are given by ensuring continuity at the start of the time interval

$$P(UQ(0;k)) = P(UQ(\delta;k-1)). \tag{3}$$

Let $P(UQ(k))$ denote the *UQ* distribution at the end of time interval $k$; that is, it is a simplified notation for $P(UQ(\delta;k))$. Equations (2) and (3) can be combined to obtain the equation that yields the distribution of *UQ* at the end of the time interval

$$P(UQ(k)) = P(UQ(k-1))e^{\delta Q^{UQ}(k)}. \tag{4}$$

Equation (4) states that to approximate the transient distribution of *UQ*, we need to approximate the transition rate matrix $Q^{UQ}(k)$. Table 1 defines the nondiagonal and nonnull elements of the transition rate matrix. This table considers for an arbitrary initial state $uq$ of *UQ* (displayed in the first column), the feasible instantaneous transitions that can take place to new states (the second column), the rate at which the transitions take place (the third column), and the conditions on the initial states needed for the transitions to be feasible (the fourth column). For *UQ*, there are two types of events that trigger state changes. The first is flow arrival to the link. This is described in the first row of the table. This row states that arrivals to the link lead to an increase in the state of *UQ* (i.e., the new state is $uq + 1$); this occurs with rate $\lambda(k)$ and can occur as long as *UQ* is not full (i.e., there is available space at the upstream end of the link: $uq < l$). The second type of event is flow departure from *UQ*; these departures are described by the second row of the table. They occur at rate $\mu^{UQ}(uq;k)$ and can occur as long as *UQ* is not empty (i.e., $uq > 0$). The diagonal elements of the transition rate matrix, $Q^{UQ}(k)_{ss}$, are derived from the nondiagonal elements by

$$Q^{UQ}(k)_{ss} = -\sum_{j \neq s} Q^{UQ}(k)_{sj}. \tag{5}$$

**Table 1.** Transition Rate Table of *UQ*

| Initial state | New state | Rate | Condition |
|---|---|---|---|
| $uq$ | $uq+1$ | $\lambda(k)$ | $uq < l$ |
| $uq$ | $uq-1$ | $\mu^{UQ}(uq;k)$ | $uq > 0$ |

Table 1 states that the univariate model of *UQ* depends on two rates: (i) $\lambda(k)$, which for an isolated link is an exogenous rate, and (ii) $\mu^{UQ}(uq;k)$. The latter is referred to as the service rate of *UQ*. It is an endogenous rate. We now formulate its approximation.

#### 2.3.1. Service Rate of *UQ*.

Recall from Section 2.1 and Figure 1 that departures from *UQ* correspond to flow that leaves the last cell of *LO*. In Figure 1, this last cell is the $k^{\text{bwd}}$th cell denoted by *LLO*. Therefore, the number of departures from *UQ* during time interval $k$ is a random variable, and it can be expressed as $LLO(k) \mid UQ(k) = uq$. Let $E[T_m(k)]$ denote the expected interdeparture time from *UQ* conditional on there being a total of $m$ departures during time interval $k$. By definition, the service rate is the inverse of expected time between consecutive departures. The service rate of *UQ* conditional on $UQ = uq$, $\mu^{UQ}(uq;k)$, can be approximated as follows:

$$\mu^{UQ}(uq;k) \approx \sum_{m=1}^{uq} \frac{1}{E[T_m]} P(LLO(k) = m \mid UQ(k) = uq). \tag{6}$$

Equation (6) approximates $\mu^{UQ}(uq;k)$ as the mean inverse of expected interdeparture time from *UQ* conditional on there being a total of $m$ departures during time interval $k$. To approximate $E[T_m(k)]$ for $m > 0$, we use the following property. For a Poisson process, given that a total number of $m$ arrivals have occurred during a time interval of duration $\delta$, then the unordered arrival times are independently, uniformly distributed over the time interval of interest (see, for instance, section 2.12.3 of Larson and Odoni 1981). Hence, the expected interarrival time is $\delta/m$. We approximate the departure process of *UQ* as a Poisson process. Therefore, given a total of $m$ departures from *UQ* during time interval $k$, the expected time between consecutive departures is approximated with $\delta/m$. Equation (6) becomes

$$\mu^{UQ}(uq;k) \approx \sum_{m=1}^{uq} \frac{m}{\delta} P(LLO(k) = m \mid UQ(k) = uq) \tag{7}$$

$$= \frac{1}{\delta} \sum_{m=0}^{uq} m P(LLO(k) = m \mid UQ(k) = uq) \tag{8}$$

$$= \frac{1}{\delta} E[LLO(k) \mid UQ(k) = uq], \tag{9}$$

where $E[LLO(k) \mid UQ(k) = uq]$ represents the expected outflow from *UQ* during time interval $k$, given that *UQ* is in state $uq$. The expression for this conditional expectation is derived as follows.

First, assume $UQ$ to be a Poisson process with rate

$$q^{UQ}(k) = \sum_{r=0}^{k-1} q^{\text{in}}(r) - \sum_{r=0}^{k-k^{\text{bwd}}-1} q^{\text{out}}(r), \tag{10}$$

where $q^{\text{in}}(k)$ (respectively, $q^{\text{out}}(k)$) denotes the instantaneous link inflow (respectively, outflow) rates at the end of time interval $k$. As in the multivariate model (as well as in its deterministic counterpart model of Yperman, Tampere, and Immers 2007), we use these instantaneous link inflow and outflow rates to approximate the expected inflow to (respectively, outflow from) the link during time interval $k$. In other words, these instantaneous rates $q^{\text{in}}(k)$ and $q^{\text{out}}(k)$ are held constant throughout the time interval $k$ of duration $\delta$. Equation (10) approximates the rate of the Poisson process $UQ$ as the difference between (i) all flow that has entered the link from time interval 0 until the end of time interval $k-1$ (this is represented by the first summation) and (ii) all flow that has left the link from time interval 0 until the end of time interval $k-k^{\text{bwd}}-1$ (this is represented by the second summation). Recall that $k^{\text{bwd}}$ represents the number of time intervals needed for a kinematic backward wave to traverse the link. Therefore, this second summation accounts for this kinematic backward wave delay by considering all flow that has left the $LO$ queue and, hence, has left $UQ$.

Second, assume $LLO$ and $\{UQ-LLO\}$ to be two independent Poisson processes. This simplifying independence assumption neglects the temporal dependency between $LLO$ and $\{UQ-LLO\}$. The numerical validation results of Section 3 highlight the small effect this has on the final model's accuracy. The rate of $LLO$ is given by

$$q^{LLO}(k) = q^{\text{out}}(k-k^{\text{bwd}}). \tag{11}$$

The term $q^{\text{out}}(k-k^{\text{bwd}})$ represents the expected flow that has left the link during time interval $k-k^{\text{bwd}}$. This leads to $UQ$ being a sum of two independent Poisson processes: $LLO$ and $\{UQ-LLO\}$. Therefore, the conditional distribution of $LLO(k)$ given $\{UQ(k)=uq\}$ is binomial with parameters $(uq, q^{LLO}(k)/q^{UQ}(k))$ (see, for instance, section 2.12.4 of Larson and Odoni 1981). Hence, the expected number of departures from $UQ$ during time interval $k$ is approximated by

$$E[LLO(k) \mid UQ(k)=uq] \approx uq\frac{q^{LLO}(k)}{q^{UQ}(k)}. \tag{12}$$

The accuracy of this approximation depends on the dependency between $LLO$ and $\{UQ-LLO\}$. In particular, we expect it to decrease as the congestion level increases.

In summary, given the rates that fully define the transition rate matrix, $\lambda(k)$ (an exogenous rate for an isolated link) and $\mu^{UQ}(uq;k)$ (given by Equation (9)), the transient probability distribution of $UQ$ is obtained by evaluating Equation (4).

**2.3.2. Expected Link Inflow and Outflow.** Given the univariate model of $UQ$, we now describe how it can be used to compute the expected inflow and expected outflow of the link during time interval $k$. An arrival may enter the link as long as there is space at the upstream end of the link. This happens with probability $P(UQ(k)<l)$. Hence, the expected inflow to the link is

$$q^{\text{in}}(k) = \lambda(k)P(UQ(k)<l). \tag{13}$$

Note that in a full network model (i.e., if we combine the link model with a node model), a vehicle in an upstream link that cannot enter its desired downstream link because it is full would wait at its current location until an available space downstream is allocated to it. In other words, spillbacks occur with probability $P(UQ(k)=l)$. In this paper we consider a single link model; hence vehicles that wish to enter the link while it is full are considered lost demand. If a model with no losses is desired, then an infinite space-capacity queue can be inserted upstream of the link to capture vehicles that are waiting to enter the link.

Similarly, the probability that there are vehicles ready to leave the link is $P(DQ(k)>0)$. Thus, the expected outflow from the link is

$$q^{\text{out}}(k) = \mu(k)P(DQ(k)>0). \tag{14}$$

From Equation (4), we obtain the distribution of $UQ$ at the end of time interval $k$, which allows us to compute $q^{\text{in}}(k)$ through Equation (13). To compute $q^{\text{out}}(k)$, we need to compute $P(DQ(k)>0)$. Nonetheless, in this univariate $UQ$ model, we do not track $DQ$ directly. Let us now describe how it is approximated.

We proceed as above, where we approximate $UQ$ as a sum of independent Poisson processes and approximate the distribution of $DQ$ given $\{UQ=uq\}$ as a binomial with parameters $(uq, q^{DQ}(k)/q^{UQ}(k))$, where

$$q^{DQ}(k) = \sum_{r=0}^{k-k^{\text{fwd}}-1} q^{\text{in}}(r) - \sum_{r=0}^{k-1} q^{\text{out}}(r). \tag{15}$$

Equation (15) considers the expected flow in $DQ$ as the difference between (i) the sum of all of the expected inflows into the link from time 0 to time $k-k^{\text{fwd}}-1$ (i.e., omitting the flows that are still in $LI$) and (ii) the sum of all expected outflows out of the link (i.e., outflow from time 0 to time $k-1$).

We obtain $P(DQ(k)>0)$ as follows:

$$P(DQ(k)>0) = 1 - P(DQ(k)=0) \tag{16}$$

$$= 1 - \sum_{n=0}^{l} P(DQ(k)=0 \mid UQ(k)=n)P(UQ(k)=n) \tag{17}$$

$$\approx 1 - \sum_{n=0}^{l}\left(1-\frac{q^{DQ}(k)}{q^{UQ}(k)}\right)^{n} P(UQ(k)=n), \tag{18}$$

where the binomial probability mass function was used to derive the last expression. This approximation is accurate when the dependencies among $LI$, $DQ$, and $LO$ is weak (e.g., uncongested link).

**2.3.3. Marginal Distribution of $DQ$.** The univariate $UQ$ model can be used to approximate the entire marginal distribution of $DQ$ by proceeding similarly as in the derivation of Equation (18). For all $i \in \{0, \ldots, l\}$,

$$P(DQ(k)=i) = \sum_{n=i}^{l} P(DQ(k)=i \mid UQ(k)=n) P(UQ(k)=n) \quad (19)$$

$$\approx \sum_{n=i}^{l} \binom{n}{i} \left(\frac{q^{DQ}(k)}{q^{UQ}(k)}\right)^{i} \left(1-\frac{q^{DQ}(k)}{q^{UQ}(k)}\right)^{n-i} P(UQ(k)=n), \quad (20)$$

where $\binom{n}{i}$ denotes the binomial coefficient. Equation (20) is obtained by approximating $P(DQ(k) \mid UQ(k) = n)$ as a binomial distribution with parameters $(uq, q^{DQ}(k)/q^{UQ}(k))$.

**2.3.4. Algorithm.** Algorithm 1 summarizes the numerical evaluation of the $UQ$ model. In the algorithm, we omit the computation of the marginal distribution of $DQ$ at each time interval $k$. However, all the parameters in Equation (20) are stored, and thus the distribution of $DQ$ for any time interval $k$ can be computed if needed.

**Algorithm 1** (Algorithm of the univariate upstream queue model)

1. set exogenous parameters $\hat{\rho}, v, w, l$, and $\delta$
2. set arrival and service rate over time $\lambda(k)$ and $\mu(k)$ for $\forall\, k = 1, 2, \ldots$
3. compute $k^{\text{fwd}} = \lceil l/(\hat{\rho}v) \rceil$ and $k^{\text{fwd}} = \lceil l/(\hat{\rho}\lvert w \rvert) \rceil$
4. set exogenous initial link conditions: $q^{\text{in}}(0)$, $q^{\text{out}}(0)$, $P(UQ(0))$, $q^{UQ}(0)$, $q^{LLO}(0)$, and $q^{DQ}(0)$
5. set $q^{\text{in}}(r) = 0$ and $q^{\text{out}}(r) = 0$ for $r < 0$
6. repeat the following for time intervals $k = 1, 2, \ldots$
   (a) compute $q^{UQ}(k)$, $q^{LLO}(k)$, and $q^{DQ}(k)$ according to Equations (10), (11), and (15), respectively
   (b) for $uq = 0, 1, \ldots, l$, compute $\mu^{UQ}(uq; k)$ according to Equation (9)
   (c) form the transition rate matrix $Q^{UQ}(k)$ defined in Table 1
   (d) compute $P(UQ(k))$ according to Equation (4)
   (e) compute $P(DQ(k) > 0)$ according to Equation (18)
   (f) compute $q^{\text{in}}(k)$ and $q^{\text{out}}(k)$ according to Equations (13) and (14)

## 2.4. Univariate Downstream Queue Model

The approach to formulate the univariate $DQ$ model is similar to that used for the univariate $UQ$ model of Section 2.3. We model $DQ$ as a birth-death process with finite state space defined by $dq \in \{0, \ldots, l\}$. Just as for the $UQ$ model, the transient distribution of $DQ$, $P(DQ(k))$, satisfies an equation of the form (4)

**Table 2.** Transition Rate Table of $DQ$

| Initial state | New state | Rate | Condition |
|---|---|---|---|
| $dq$ | $dq+1$ | $\lambda^{DQ}(k)$ | $dq < l$ |
| $dq$ | $dq-1$ | $\mu(k)$ | $dq > 0$ |

with initial conditions $P(DQ(k-1))$ and transition rate matrix $Q^{DQ}(k)$. The nondiagonal and nonnull elements of the transition rate matrix of $DQ$, $Q^{DQ}(k)$ are given in Table 2. The first row of Table 2 describes the event of arrivals to $DQ$—that is, flow that transitions from $LI$ to $DQ$ (see Figure 1). The second row describes the event of flow departing $DQ$ (i.e., departing the link). The corresponding rate $\mu(k)$ is the downstream bottleneck flow capacity and is considered exogenous for an isolated link. The diagonal elements of the transition rate matrix are computed following equations as in (5).

Table 2 indicates that the transition rate matrix of $DQ$ is defined by two rates: (i) an endogenous arrival rate $\lambda^{DQ}(k)$ and (ii) an exogenous service rate $\mu(k)$. This table is simpler than that of the $UQ$ model (see Table 1) because both rates are state independent (i.e., neither depends on the state $dq$). In Table 1, the service rate of $UQ$ is state dependent; that is, $\mu^{UQ}(uq; k)$ depends on $uq$.

For a finite capacity birth-death process with state-independent rates, Morse (1958, equation (6.13), chapter 6) provides a closed-form expression to Equation (4), which avoids the need to numerically evaluate the matrix exponential. For time interval $k$ of length $\delta$, the $DQ$ distribution at the end of time interval $k$, $P(DQ(k))$, is given by

$$\begin{cases} P(DQ(k)=n) = \sum_{m=0}^{l} P(DQ(k-1)=m) P_n^m(\delta) \\ \qquad \text{for } n = 0, \ldots, l, \quad (21a) \\ P_n^m(\delta) = P_n(k) + \dfrac{2\rho(k)^{(n-m)/2}}{l+1} \sum_{s=1}^{l} \dfrac{\mu(k)}{\gamma_s(k)} \\ \cdot \left[\sin\left(\dfrac{sm\pi}{l+1}\right) - \sqrt{\rho(k)}\sin\left(\dfrac{s(m+1)\pi}{l+1}\right)\right] \\ \cdot \left[\sin\left(\dfrac{sn\pi}{l+1}\right) - \sqrt{\rho(k)}\sin\left(\dfrac{s(n+1)\pi}{l+1}\right)\right] e^{-\gamma_s(k)\delta}, \quad (21b) \\ \gamma_s(k) = \lambda^{DQ}(k) + \mu(k) - 2\sqrt{\lambda^{DQ}(k)\mu(k)}\cos\left(\dfrac{s\pi}{l+1}\right) \\ \qquad \text{for } s = 1, 2, \ldots, l, \quad (21c) \\ P_n(k) = \left(\dfrac{1-\rho(k)}{1-\rho(k)^{l+1}}\right)\rho(k)^n, \quad (21d) \\ \rho(k) = \dfrac{\lambda^{DQ}(k)}{\mu(k)}. \quad (21e) \end{cases}$$

Equation (21a) states that the distribution of $DQ$ at the end of time interval $k$ can be obtained by a convex combination of distributions $P_n^m(\delta)$, each of which is

defined in Equation (21b) as the sum of (i) the stationary probability of being in state $n$, which is denoted by $P_n(k)$ and defined by Equation (21d), and (ii) a time-dependent term with exponential decay. The exponential decay is parameterized by $\gamma_s(k)$, which is defined by Equation (21c) and is referred to in the queuing literature as the inverse of the relaxation time. In summary, the distribution of $DQ$ is given by System of Equations (21), which depends on two rates: (i) an exogenous service rate $\mu(k)$ and (ii) an endogenous arrival rate $\lambda^{DQ}(k)$. We now describe how we approximate this endogenous arrival rate.

**2.4.1. Arrival Rate of $DQ$.** The distribution of $DQ$ at the end of time interval $k$ is given by the System of Equations (21), which depends on the endogenous rate, $\lambda^{DQ}(k)$. To approximate this rate, we observe that for a queue with finite space capacity $l$ and arrival rate $\lambda$, the expected inflow to the queue is given by $\lambda P(N < l)$, where $N$ represents the number of vehicles in the queue. We use this property to obtain the following expression for the arrival rate to $DQ$:

$$\lambda^{DQ}(k) \approx \frac{q^{\text{in}}(k - k^{\text{fwd}})}{P(DQ(k-1) < l)}. \tag{22}$$

The numerator $q^{\text{in}}(k - k^{\text{fwd}})$ represents the expected inflow to the link during time interval $k - k^{\text{fwd}}$; that is, this is the flow that is expected to leave the last cell of $LI$ (denoted by $LLI$ in Figure 1) and enter $DQ$ during time interval $k$. The denominator $P(DQ(k-1) < l)$ is based on the $DQ$ distribution at the end of time interval $k-1$, which is the $DQ$ distribution at the beginning of time interval $k$.

**2.4.2. Expected Link Inflow and Outflow.** Given the univariate model of $DQ$, we now describe how it can be used to compute the expected inflow and expected outflow of the link during time interval $k$. Recall their definition given in (13) and (14). The System of Equations (21) yields the marginal distribution of $DQ$; hence the expected outflow $q^{\text{out}}(k)$ (defined by Equation (14)) can be directly computed.

To compute the expected inflow $q^{\text{in}}(k)$ (defined by Equation (13)), we need $P(UQ(k) < l)$. Nonetheless, in this univariate $DQ$ model, we do not track $UQ$ directly. Let us now describe how it is approximated.

We express $P(UQ(k) < l)$ as a function of the conditional distribution of $\{UQ - DQ\}$ given $DQ$:

$$P(UQ(k) < l) = 1 - P(UQ(k) = l) \tag{23}$$

$$= 1 - \sum_{n=0}^{l} P(UQ(k) = l \mid DQ(k) = n) P(DQ(k) = n) \tag{24}$$

$$= 1 - \sum_{n=0}^{l} P(UQ(k) - DQ(k) = l - n \mid DQ(k) = n) \cdot P(DQ(k) = n) \tag{25}$$

$$\approx 1 - \sum_{n=0}^{l} p_1(k)^{l-n} P(DQ(k) = n). \tag{26}$$

Equation (25) is obtained from (24) by observing that $P(UQ(k) = l \mid DQ(k) = n)$ equals $P(UQ(k) - DQ(k) = l - n \mid DQ(k) = n)$. Equation (26) is obtained by approximating the conditional distribution of $\{UQ - DQ\}$ given $\{DQ = n\}$ with a binomial distribution with parameters $(l - n, p_1(k))$.

The first parameter of this distribution $l - n$ is derived by observing that the random variable $\{UQ - DQ\}$ given $\{DQ = n\}$ can only take values in $\{0, \ldots, l - n\}$. Let us elaborate this. Equation (1) implies $UQ - DQ \geq 0$. Additionally, by definition, $UQ \leq l$. Thus, conditional on $DQ = n$, we have $UQ - DQ \leq l - n$.

Let us now approximate the second parameter of this binomial distribution, $p_1(k)$:

$$E[UQ(k)] = E[DQ(k)] + E[UQ(k) - DQ(k)] \tag{27}$$

$$= E[DQ(k)] + E[E[UQ(k) - DQ(k) \mid DQ(k)]] \tag{28}$$

$$= E[DQ(k)] + \sum_{n=0}^{l} E[UQ(k) - DQ(k) \mid DQ(k) = n] \cdot P(DQ(k) = n) \tag{29}$$

$$\approx E[DQ(k)] + \sum_{n=0}^{l} (l - n) p_1(k) P(DQ(k) = n) \tag{30}$$

$$= E[DQ(k)] + p_1(k)(l - E[DQ(k)]). \tag{31}$$

Equation (27) is obtained by adding and subtracting $E[DQ(k)]$ on the right-hand side. The law of total expectation is used in (28) and rewritten in more detail in (29). Since $\{UQ - DQ\}$ conditional on $\{DQ = n\}$ is approximated as a binomial distribution with parameters $(l - n, p_1(k))$, then $E[UQ - DQ \mid DQ = n]$ equals $(l - n)p_1(k)$, which leads to (30). The summation is simplified to obtain (31), which itself can be rearranged to obtain the approximation for $p_1(k)$

$$p_1(k) \approx \frac{E[UQ(k)] - E[DQ(k)]}{l - E[DQ(k)]}. \tag{32}$$

To evaluate Equation (32), $E[DQ(k)]$ can be computed from the marginal distribution of $DQ$ (system of Equations (21)) as

$$E[DQ(k)] = \sum_{n=0}^{l} n P(DQ(k) = n), \tag{33}$$

and $E[UQ(k)]$ can be obtained from the approximation of $UQ$ as a Poisson process with rate defined by Equation (10), and thus

$$E[UQ(k)] \approx q^{UQ}(k) \cdot \delta = \left( \sum_{r=0}^{k-1} q^{\text{in}}(r) - \sum_{r=0}^{k-k^{\text{bwd}}-1} q^{\text{out}}(r) \right) \cdot \delta. \tag{34}$$

In summary, $P(UQ(k) < l)$ is approximated by Equation (26), with $p_1(k)$ given by Equation (32) and $P(DQ(k) = n)$ given by the System of Equations (21).

**2.4.3. Marginal Distribution of *UQ*.** The univariate *DQ* model can be used to approximate the entire marginal distribution of *UQ* by proceeding similarly as in the derivation of Equation (26). For all $i \in \{0, \ldots, l\}$,

$$P(UQ(k) = i)$$

$$= \sum_{n=0}^{i} P(UQ(k) = i \mid DQ(k) = n) P(DQ(k) = n) \quad (35)$$

$$= \sum_{n=0}^{i} P(UQ(k) - DQ(k) = i - n \mid DQ(k) = n) \cdot P(DQ(k) = n) \quad (36)$$

$$\approx \sum_{n=0}^{i} \binom{l-n}{i-n} p_1(k)^{i-n} (1 - p_1(k))^{l-i} P(DQ(k) = n), \quad (37)$$

where $P(DQ(k) = n)$ is given by the System of Equations (21), and $p_1(k)$ is given by Equation (32). Equation (37) is obtained by approximating $P(UQ(k) - DQ(k) \mid DQ(k) = n)$ as a binomial distribution with parameters $(l - n, p_1(k))$.

**2.4.4. Algorithm.** Algorithm 2 summarizes the numerical evaluation of the *DQ* model. In the algorithm, we omit the computation of the marginal distribution of *UQ* at each time interval $k$. However, all the parameters in Equation (37) are stored, and thus the distribution of *UQ* for any time interval $k$ can be computed if needed.

**Algorithm 2** (Algorithm of the univariate downstream queue model)

1. set exogenous parameters $\hat{\rho}, v, w, l$, and $\delta$
2. set arrival and service rate over time $\lambda(k)$ and $\mu(k)$ for $\forall\, k = 1, 2, \ldots$
3. compute $k^{\text{fwd}} = \lceil l/(\hat{\rho} v) \rceil$ and $k^{\text{fwd}} = \lceil l/(\hat{\rho} \lvert w \rvert) \rceil$
4. set exogenous initial link conditions: $q^{\text{in}}(0)$, $q^{\text{out}}(0)$, $P(DQ(0))$, $q^{UQ}(0)$, and $q^{LLI}(0)$
5. set $q^{\text{in}}(r) = 0$ and $q^{\text{out}}(r) = 0$ for $r < 0$
6. repeat the following for time intervals $k = 1, 2, \ldots$
   (a) compute $q^{UQ}(k)$ according to Equation (10)
   (b) compute $\lambda^{DQ}(k)$ according to Equation (22)
   (c) compute $P(DQ(k))$ according to the System of Equations (21)
   (d) compute $E[UQ(k)]$ according to Equation (34)
   (e) compute $E[DQ(k)]$ according to Equation (33)
   (f) compute $p_1(k)$ according to Equation (32)
   (g) compute $P(UQ(k) < l)$ according to Equation (26)
   (h) compute $q^{\text{in}}(k)$ and $q^{\text{out}}(k)$ according to Equations (13) and (14)

## 2.5. Mixture Model

Recall that, by design, the role of *UQ* is to capture the link's upstream boundary conditions, while that of *DQ* is to capture the link's downstream boundary conditions. To capture both the link's upstream and downstream boundary conditions, while ensuring a model suitable for large-scale network analysis, we propose a link model that is a mixture of the univariate *UQ* model (formulated in Section 2.3) and of the univariate *DQ* model (formulated in Section 2.4). The proposed model is given by

$$P(UQ(k)) = \tilde{w} P^{UQ}(UQ(k)) + (1 - \tilde{w}) P^{DQ}(UQ(k)), \quad (38)$$

$$P(DQ(k)) = \tilde{w} P^{UQ}(DQ(k)) + (1 - \tilde{w}) P^{DQ}(DQ(k)), \quad (39)$$

where the following notation is used:

- $P^{UQ}(UQ(k))$ *UQ* distribution from the *UQ* model (Equation (4)),
- $P^{DQ}(UQ(k))$ *UQ* distribution from the *DQ* model (Equation (37)),
- $P^{UQ}(DQ(k))$ *DQ* distribution from the *UQ* model (Equation (20)),
- $P^{DQ}(DQ(k))$ *DQ* distribution from the *DQ* model (System of Equations (21)).

An analytical expression for the weight parameter, $\tilde{w}$, is derived through insights obtained from a variety of numerical experiments. Its expression is given by

$$\tilde{w}(l, \mu, k^{\text{fwd}}) = e^{-l^2/(70 \mu k^{\text{fwd}})}. \quad (40)$$

The experiments compared the performance of the proposed mixture model to that of a discrete-event simulation model used in Osorio and Flötteröd (2015), which implements the stochastic link transmission model. It samples individual vehicles. The forward and backward lags are explicitly implemented on each vehicle. A total of 180 experiments were conducted considering combinations of $l \in \{5, 10, 15, \ldots, 100\}$; $\rho = \lambda/\mu \in \{0.25, 0.5, 0.75\}$; and $\mu \in \{0.2, 0.4, 0.6\}$. A more detailed description of the derivation of weight parameter, $\tilde{w}$, is given in Online Appendix A.

For the mixture model, the expected inflow and outflow—that is, $q^{\text{in}}(k)$ and $q^{\text{out}}(k)$—are obtained according to Equations (13) and (14), where $P(UQ(k) < l)$ and $P(DQ(k) > 0)$ are given by (38) and (39), respectively. Algorithm 3 summarizes the mixture model approach. Notice that steps 7 and 8 in the algorithm can be run simultaneously and independently to further enhance the run time.

**Algorithm 3** (Algorithm of the mixture model)

1. set exogenous parameters $\hat{\rho}, v, w, l$, and $\delta$
2. set arrival and service rate over time $\lambda(k)$ and $\mu(k)$ for $\forall\, k = 1, 2, \ldots$
3. compute $k^{\text{fwd}} = \lceil l/(\hat{\rho} v) \rceil$ and $k^{\text{fwd}} = \lceil l/(\hat{\rho} \lvert w \rvert) \rceil$
4. compute $\tilde{w}$ according to Equation (40)
5. set exogenous initial link conditions: $q^{\text{in}}(0)$, $q^{\text{out}}(0)$, $P(UQ(0))$, $P(DQ(0))$, $q^{UQ}(0)$, $q^{LLO}(0)$, $q^{LLI}(0)$, and $q^{DQ}(0)$
6. set $q^{\text{in}}(r) = 0$ and $q^{\text{out}}(r) = 0$ for $r < 0$
7. run step 6 of Algorithm 1, this yields $P^{UQ}(UQ(k))$

for all $k = 1, 2 \ldots$

8. run step 6 of Algorithm 2, this yields $P^{DQ}(DQ(k))$ for all $k = 1, 2 \ldots$
9. for any time interval $k$,
   (a) compute $P^{UQ}(DQ(k))$ according to Equation (20)
   (b) compute $P^{DQ}(UQ(k))$ according to Equation (37)
   (c) compute $P(UQ(k))$ according to Equation (38)
   (d) compute $P(DQ(k))$ according to Equation (39)
   (e) compute $q^{\text{in}}(k)$ and $q^{\text{out}}(k)$ according to Equations (13) and (14)

## 3. Validation

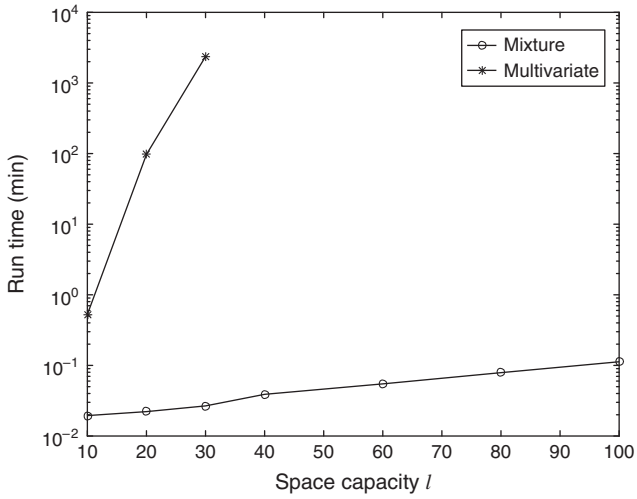
In this section we validate the model. We evaluate and compare both in terms of computational run time and accuracy. First, we compare the computational run times of the proposed model to those of the multivariate model (Osorio and Flötteröd 2015). We consider a single-lane link with parameters shown in Table 3. The link configuration is the same as that used in Osorio and Flötteröd (2015) except for the service rate. The service rate of the link is fixed at 0.4 vehicles per second (veh/sec) for all experiments. The experiments consider different arrival rates and link lengths (and hence, different space capacities, forward lags, and backward lags). We consider a set of three different arrival rates ($\lambda \in \{0.1, 0.2, 0.3\}$ veh/sec) and seven different space capacities ($l \in \{10, 20, 30, 40, 60, 80, 100\}$). The combination of these values leads to a total of 21 experiments. The considered space capacity values correspond to link lengths $L \in \{50, 100, 150, 200, 300, 400, 500\}$ (in meters), forward lags $k^{\text{fwd}} \in \{5, 10, 15, 20, 30, 40, 50\}$ (in seconds), and backward lags $k^{\text{bwd}} \in \{10, 20, 30, 40, 60, 80, 100\}$ (in seconds). Each experiment starts with an empty link at time 0 and runs for 250 seconds, at which point the link is ensured to have reached a stationary regime. All experiments are carried out on a standard laptop machine with Intel Core i7-4700HQ CPU running at 2.40 GHz.

Figure 2 compares the run times of the mixture model (circles) and of the multivariate model (asterisks). The $x$ axis considers the space capacity values $l$.

**Table 3.** Link Parameters

| Parameter | Value |
|---|---|
| $v$ | 0.01 km/sec |
| $w$ | −0.005 km/sec |
| $\hat{\rho}$ | 200 veh/km |
| $\hat{q}$ | 2,400 veh/h = 0.67 veh/sec |
| $\delta$ | 0.1 sec |
| $\mu(k)$ | 1,440 veh/h = 0.4 veh/sec |
| $\lambda(k)$ | Varies by experiment |
| $l$, L, $k^{\text{fwd}}$, $k^{\text{bwd}}$ | Varies by experiment |

**Figure 2.** Model Run Time Comparison




The $y$ axis displays the average computational run time (in minutes). The average is computed over the three experiments with three different arrival rate values. The $y$ axis is plotted on a logarithmic scale. The maximum run time for evaluating an experiment is set to be 40 hours. If an experiment has not concluded within 40 hours, it is terminated. For $l = 30$, the average run time of the multivariate model is already 2,366 minutes (≈ 39 hours). Hence, for experiments where $l > 30$, the multivariate model is not evaluated. Figure 2 illustrates that the run time of the multivariate model increases exponentially with $l$, while for the mixture model, the increase appears linear. For the mixture model, the average run time over all 21 experiments is 0.05 minutes. The maximum average run time is obtained for $l = 100$ and is 0.11 minutes. Thus, compared with the multivariate model, the mixture model achieves significant improvements in computational complexity both theoretically and numerically.

We now compare the multivariate model and the mixture model in terms of their accuracy. To evaluate the accuracy of each of these analytical models, we use a discrete-event simulator of the stochastic link transmission model. The simulator is the same as that used for validation in Osorio and Flötteröd (2015). It samples individual vehicles and implements for each vehicle exact forward and backward lags. The arrival process is a Poisson process. For vehicles at the downstream end of the link, interdeparture times are independent and identically distributed exponential random variables. The simulated estimates are obtained from $10^6$ replications.

First, we consider two experiments with temporal variations in demand and evaluate the ability of the analytical models to approximate the transient distributions of $UQ$ and of $DQ$. For both experiments, $l = 10$. Experiment 1 has an arrival rate of 0.1 veh/sec during time $[0, 125]$ seconds, an arrival rate of 0.5 veh/sec

during time [125, 175] seconds, and an arrival rate of 0.3 veh/sec during time [175, 300] seconds. This experiment corresponds to step changes from uncongested to highly congested (i.e., $\lambda(k) > \mu(k)$) and then to congested traffic conditions. Experiment 2 considers first an arrival rate of 0.3 veh/sec during time [0, 100] seconds, of 0.1 veh/sec during time [100, 200] seconds, and then of 0.5 veh/sec during time [200, 300] seconds. This experiment corresponds to step changes from congested to uncongested and then to highly congested traffic conditions. The two experiments are designed such that during the highly congested period (where $\lambda(k) > \mu(k)$), the period is not long enough in experiment 1 for the transient distribution to converge to its stationary counterpart, while in experiment 2, it is a long enough period.

Figure 3 considers experiment 1. Each plot of Figure 3(a) considers a given time $T$ (in seconds) and displays the distribution of $UQ$, $P(UQ(T))$, at time $T$ as proposed by the mixture model (red squares), the multivariate model (blue diamonds), and the simulated estimates (black crosses). The different plots consider different times: $T \in \{1, 30, 60, 90, 120, 150, 180, 210, 240, 270\}$ seconds. Similarly, each plot of Figure 3(b) displays the distribution of $DQ$, $P(DQ(T))$, at time $T$. The simulated estimates are displayed with 95% confidence intervals. These are barely visible.

Recall that for this experiment, there is a sharp increase in demand at time $T = 125$ sec and a sharp decrease at time $T = 175$ sec. The changes in the distributions of $UQ$ and $DQ$ after time $T = 125$ seconds and $T = 175$ seconds are visible for all models. During time [125, 175], states with higher values of $UQ$ (respectively, $DQ$) have higher probabilities. After time $T = 175$, states with higher values of $UQ$ (respectively, $DQ$) have comparably lower probabilities. Figures 3(a) and 3(b) show that the dynamics of the simulator are well approximated by both the mixture and the multivariate models. Additionally, both analytical models converge, both before $T = 125$ seconds and after $T = 175$ seconds, to stationary distributions that approximate well the simulated distribution.

The left and right plots of Figure 3(c) display, respectively, $E[UQ(T)]$ and $E[DQ(T)]$ as a function of time $T$. The sharp increase in expectation after time $T = 125$ seconds and the sharp decrease after time $T = 175$ seconds are well approximated by both analytical models. The stationary values before $T = 125$ seconds and after $T = 175$ seconds are also well approximated.

Note also that for all three models considered here (mixture, multivariate, and simulator) their arrival process and their departure process are stochastic. Hence, spillback may occur even when $\mu(k) > \lambda(k)$. More specifically, the spillback probability is given by $P(UQ(T) = l)$. For instance, in the rightmost plot of the second row of Figure 3(a), the spillback probability is nonzero (i.e., $P(UQ(T) = l) > 0$).

Experiment 2 considers a sharp decrease in demand at $T = 100$ seconds and a sharp increase in demand at $T = 200$. Figures 4(a) and 4(b) display, respectively, the distributions of $UQ$ and of $DQ$ as a function of time (i.e., $P(UQ(T))$ and $P(DQ(T))$). In this experiment, we observe a shift in probability mass to states with smaller values of $UQ$ and $DQ$ during time [100, 200] seconds and a shift in probability mass to states with larger values of $UQ$ and $DQ$ after time $T = 200$ seconds. In this experiment, both analytical models converge to the stationary distribution after each change in demand. The conclusions here are the same as for the previous experiment: both the stationary and the transient distributions are well approximated by the analytical models. The time-dependent expectations $E[UQ(T)]$ and $E[DQ(T)]$ are displayed in Figure 4(c). Again, the dynamics are well captured by both analytical models. In summary, for Experiments 1 and 2, the approximations of both the mixture and the multivariate models are good. The transient and the stationary distributions are well approximated by both models.

We now evaluate the accuracy of the mixture model over a larger set of experiments. We consider the 21 experiments mentioned above. The main goal is to evaluate the loss of accuracy of the mixture model compared with the (less scalable but more accurate) multivariate model. To evaluate the accuracy of a given distribution ($UQ$ or $DQ$), we evaluate its distance to the distribution estimated via simulation with the stochastic LTM simulator described previously and used for validation in Osorio and Flötteröd (2015). Recall that this simulator is an exact implementation of the stochastic LTM. The distance between an analytical distribution (mixture or multivariate) and the simulated distribution is evaluated with the Jensen–Shannon divergence (JSD) metric (Endres and Schindelin 2003). For a pair of distributions $P_1$ and $P_2$, the JSD metric is defined by

$$\mathrm{JSD}(P_1 \parallel P_2) = \tfrac{1}{2} D(P_1 \parallel M) + \tfrac{1}{2} D(P_2 \parallel M), \tag{41}$$

$$D(P_1 \parallel P_2) = \sum_i P_1(i) \log \frac{P_1(i)}{P_2(i)}, \tag{42}$$

where $D(P_1 \parallel P_2)$ is the Kullback–Leibler divergence (KLD) (Kullback and Leibler 1951) and $M = \frac{1}{2}(P_1 + P_2)$. Unlike the KLD, the JSD is both symmetric and upper bounded by 1. The lower the JSD value, the smaller the distance between the two distributions (i.e., the higher the accuracy). We define the time-average JSD over the entire time period (i.e., 250 seconds) as the temporal mean of the JSD values; that is, $(1/250) \sum_{T=1}^{250} \mathrm{JSD}(P_1(T) \parallel P_2(T))$, where $P_1(T)$ and $P_2(T)$ are the distributions evaluated at time $T$.

**Figure 3.** (Color online) Experiment 1: Impact of the Temporal Variation of Demand on the Distributions, as Well as the Expected Values, of *UQ* and of *DQ*

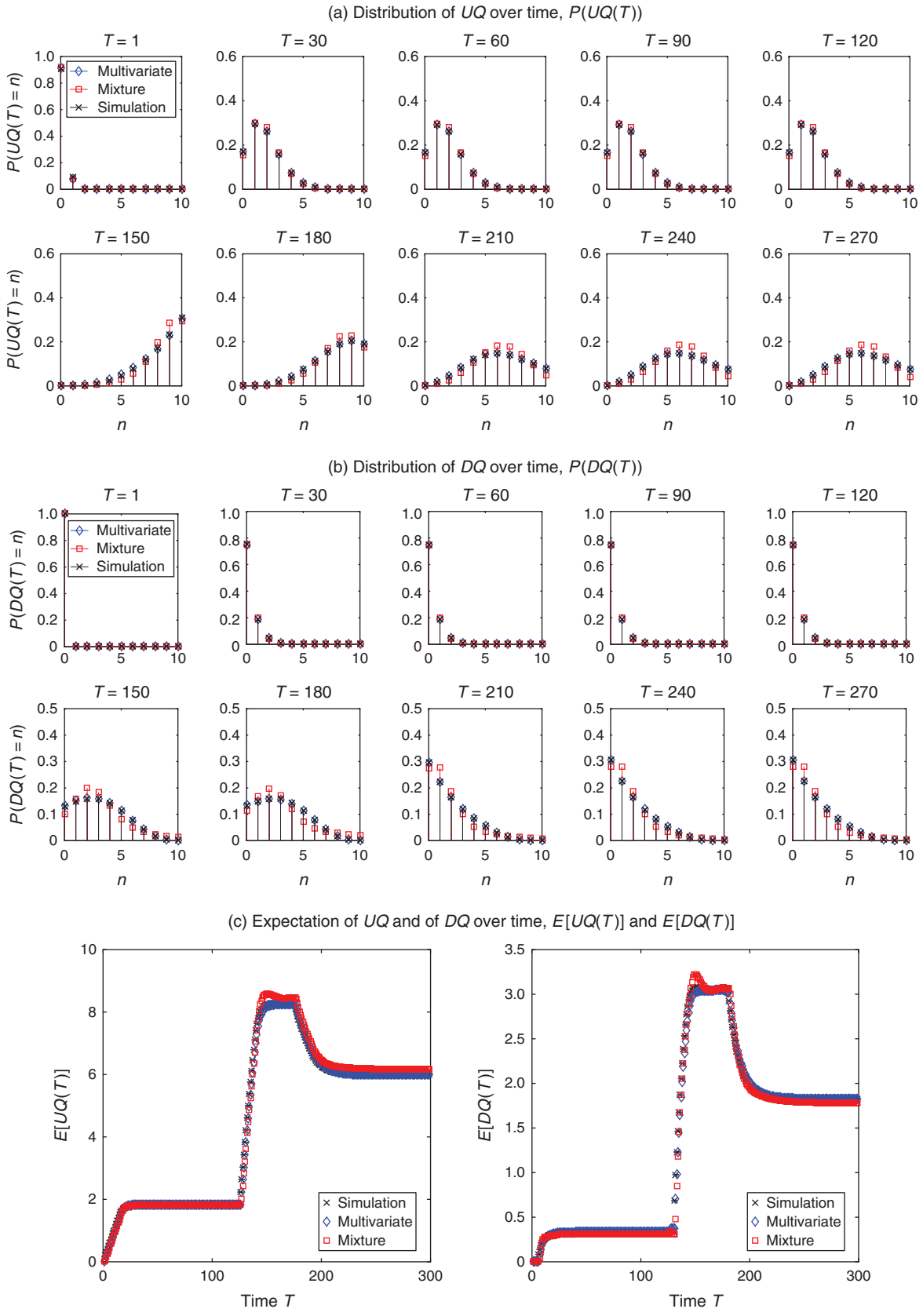
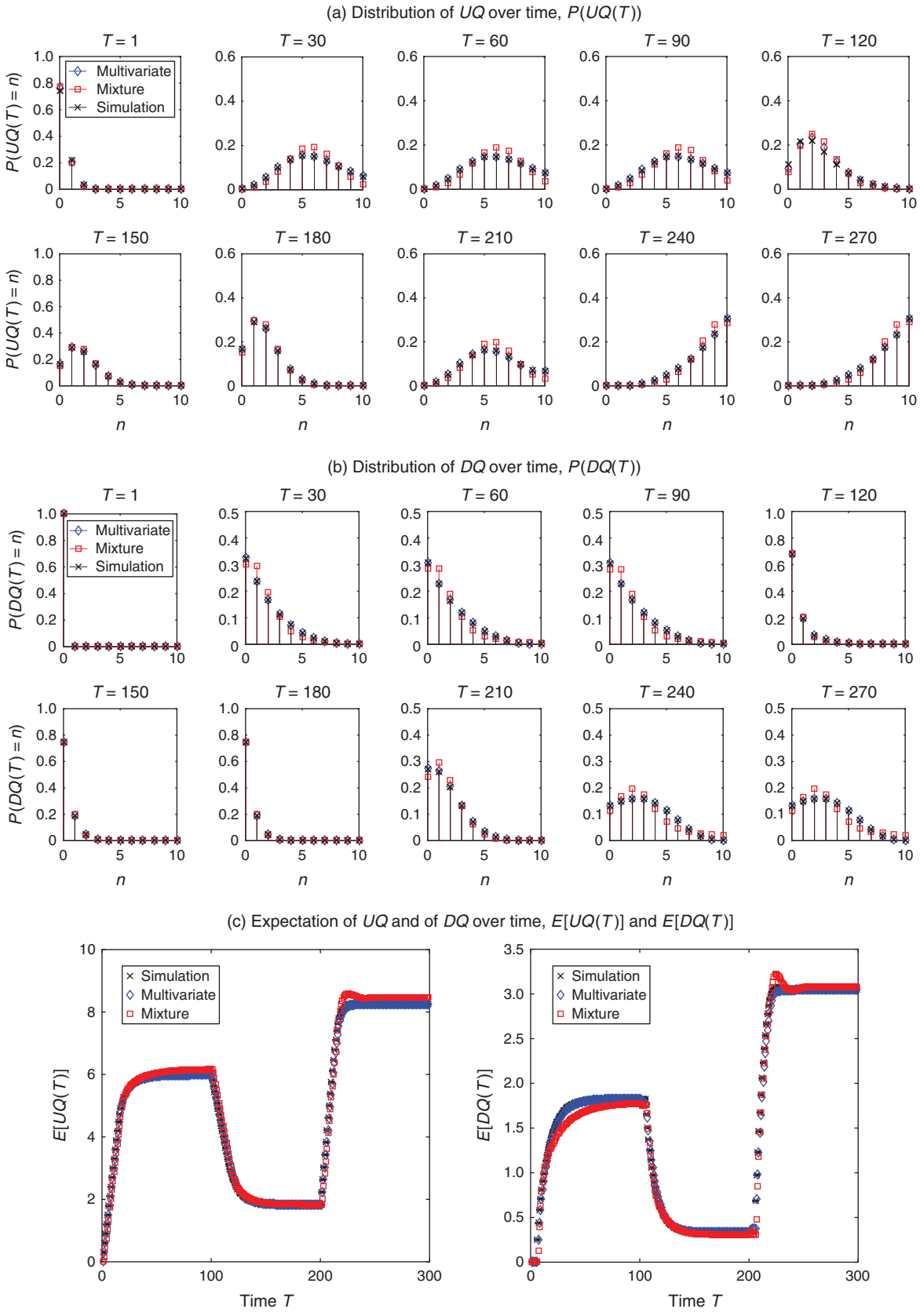


**Figure 4.** (Color online) Experiment 2: Impact of the Temporal Variation of Demand on the Distributions, as Well as the Expected Values, of *UQ* and of *DQ*



Since the main goal is to evaluate the accuracy loss of the mixture model compared with the multivariate model, we will compare the time-average JSD values of the mixture model (i.e., the time-average JSD distance between the distribution approximated by the mixture model and the simulated distribution) and the time-average JSD values of the multivariate model (i.e., the time-average JSD distance between the distribution approximated by the multivariate model and the simulated distribution). To guide us in the interpretation of the magnitude of the JSD metric, we provide three additional models to compare the proposed model with (i) the deterministic LTM (denoted by *DetDet*, which stands for deterministic arrivals and deterministic departures); (ii) a simulation-based instance of the LTM with deterministic arrivals and independent exponentially distributed interdeparture times (denoted by *DetExp*), and (iii) a simulation-based instance of the LTM with independent, exponentially distributed interarrival times and deterministic interdeparture times (denoted by *ExpDet*). Since *DetDet* is a deterministic traffic model, for a given experiment and a given time, it generates a unique link state (i.e., the distribution has all the probability mass concentrated in a single state). For the simulation-based models, the distributional estimates are obtained from $10^6$ simulation replications. In summary, for a given experiment (of the 21 experiments), a given model (mixture, multivariate, *DetDet*, *DetExp*, and *ExpDet*), and a given distribution ($UQ$ or $DQ$), we evaluate its distance to the simulated distribution using the time-average JSD metric.

As described above, the simulator consists of the deterministic LTM yet with a probabilistic arrival process and a probabilistic departure process. Hence, the underlying distributions (of $UQ$ and of $DQ$) it yields are expected to differ from those of the purely deterministic LTM. Thus, the time-average JSD values of *DetDet* can be interpreted as the effect of extending the LTM with a given probabilistic arrival process and a given probabilistic departure process. Similarly, the time-average JSD values of *ExpDet* (respectively, *DetExp*) can be interpreted as the effect of extending the LTM with a given probabilistic arrival (respectively, departure) process.

Figure 5 displays the time-average JSD values for the 21 experiments described above. The top (respectively, bottom) row plots consider the $UQ$ (respectively, $DQ$) distribution. The first column of plots considers the experiments with arrival rate $\lambda(k) = 0.1$ veh/sec. The second and third columns consider arrival rate values of 0.2 and 0.3 veh/sec, respectively. Each plot compares five models: the mixture model (circles), the multivariate model (asterisks), *DetDet* (square), *ExpDet* (triangle), and *DetExp* (cross). Each plot displays the time-average JSD metric ($y$ axis) as a function of the space capacity ($x$ axis). Recall that for the multivariate model, the run times for the experiments with $l > 30$ exceed 40 hours and are hence not computed. Figure 6 considers a zoomed-in version of Figure 5. It displays only the mixture, the multivariate, and the *ExpDet* models, which are those with the lowest error values (i.e., their curves mostly overlap along the $x$ axis in Figure 5).

For all plots of Figure 5, the time-average JSD values of *DetDet* and *DetExp* are significantly higher than those of the other models. In particular, the curves of the three other models (mixture, multivariate, and *ExpDet*) are barely visible along the $x$ axis. Figure 6 presents in more detail the curves of these three models. For $P(UQ(T))$ (i.e., top row plots), the time-average JSD values of the mixture model are higher than those of the multivariate and of *ExpDet*. Yet the values remain very small. For $P(DQ(T))$ (i.e., bottom row plots), the time-average JSD values of the *ExpDet* model are higher than those of the mixture and of the multivariate model. For space capacities $l \geq 30$, the curve of the mixture model overlaps with the $x$ axis; it is barely visible. This indicates very high accuracy. Recall also that for $l > 30$, the computation time for the mixture model exceeds 40 hours and is hence not evaluated. Overall, these experiments indicate that the loss of accuracy of the mixture model compared with the multivariate model is not significant. The numerical time-average JSD values displayed in Figure 5 are provided, for all experiments, in Tables 1 and 2 of Online Appendix B.

In summary, for experiments with both constant and time-varying demand, the mixture model performs comparably with the multivariate model while being significantly faster to evaluate. The gain in computational run time increases with the space capacity. In particular, for medium-dimensional state spaces (i.e., medium-sized links), the evaluation of the mixture model remains instantaneous (i.e., in the order of seconds), while that of the multivariate model increases exponentially.

## 4. Network Analysis

In this section, the proposed mixture model is used to address a traffic signal control problem for the city of Lausanne, Switzerland. Section 4.1 formulates the problem and describes the case study. Section 4.2 presents the numerical results, and Section 4.3 compares the performance of the resulting signal plans to that of a signal plan derived by commercial signal control software.

### 4.1. City-Scale Signal Control

We consider the city of Lausanne, Switzerland. The city map is shown in Figure 7, and the area of consideration is delimited in white. The network model of a stochastic microscopic simulator is displayed in

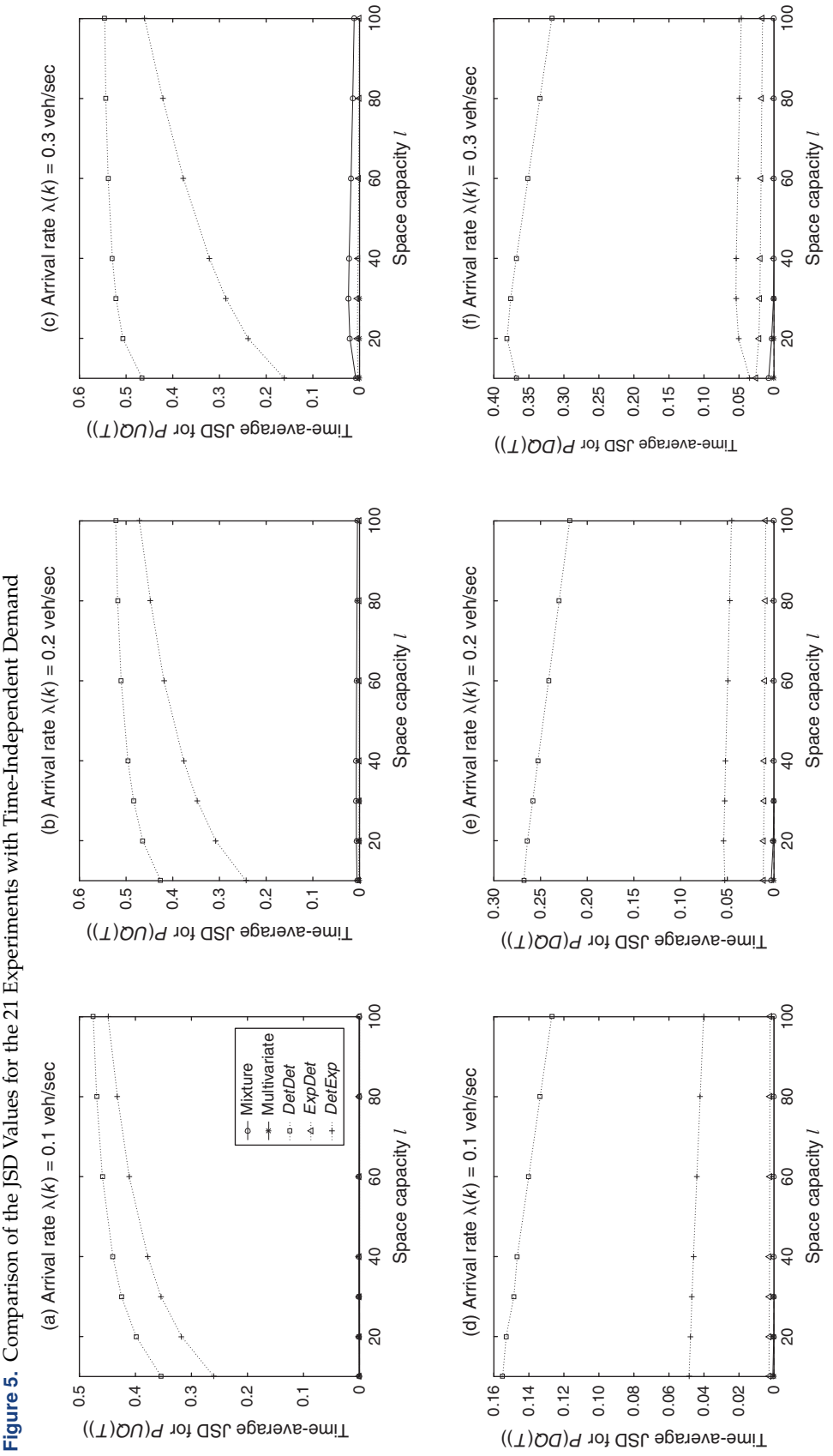



**Figure 5.** Comparison of the JSD Values for the 21 Experiments with Time-Independent Demand

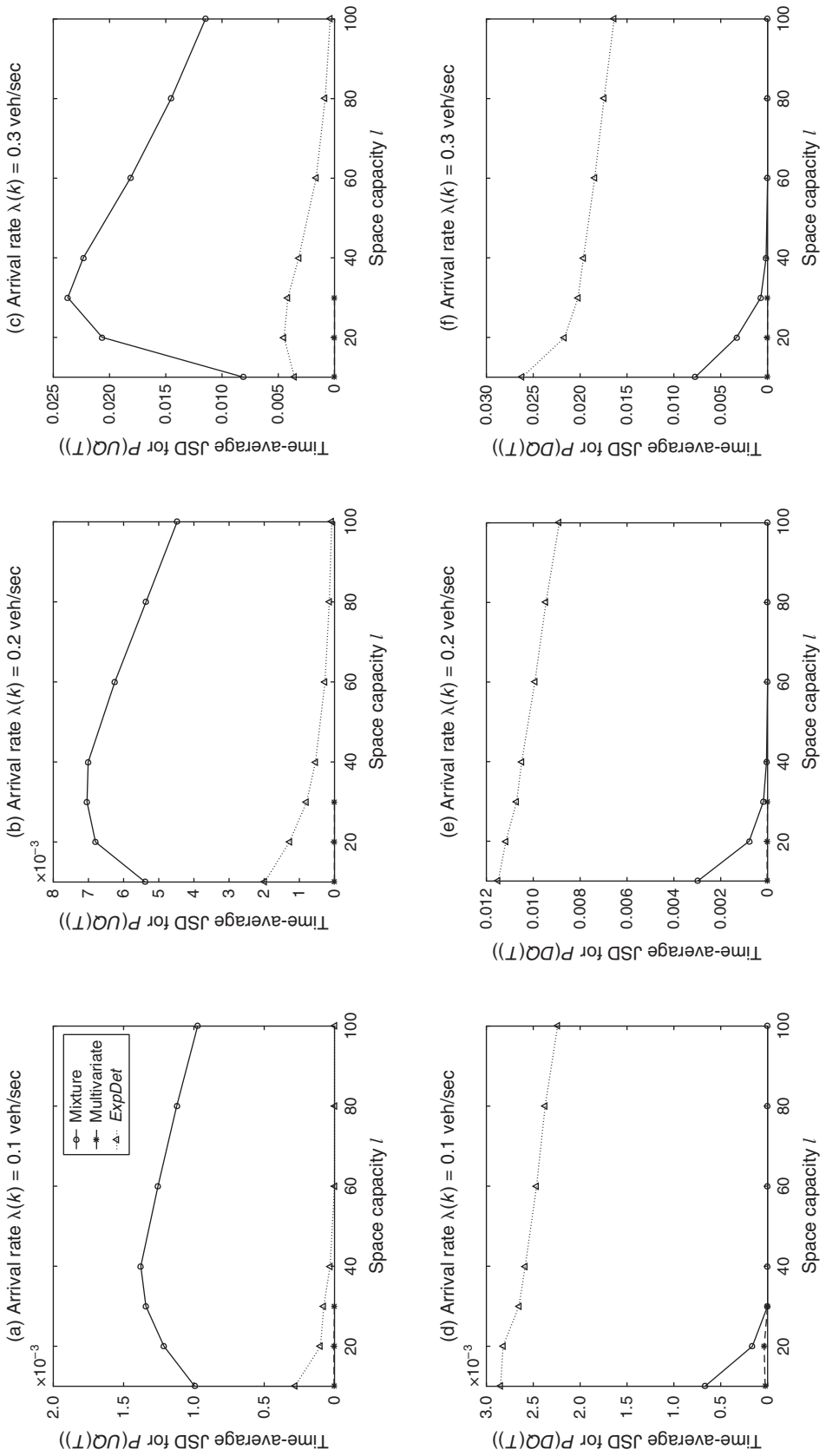
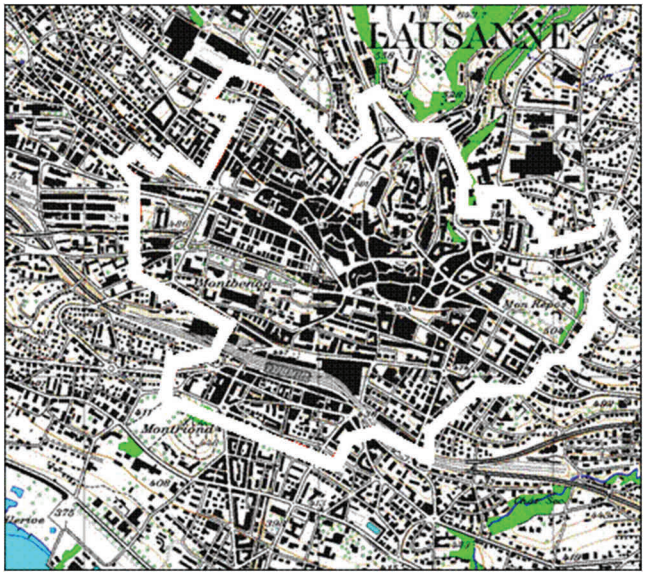



**Figure 6.** Comparison of the JSD Values for the 21 Experiments with Time-Independent Demand (Zoomed-In Results)

**Figure 7.** (Color online) Lausanne City Road Network




*Source.* Adapted from Dumont and Bert (2006).

Figure 8. The network consists of 603 links, 902 lanes, and 231 intersections. We consider a problem where we determine the signal plans of 17 intersections distributed throughout the city. These 17 intersections are depicted as squares in Figure 8. We consider a fixed-time signal control problem. For a review of traffic signal control terminology and formulations, see Appendix A of Osorio (2010). A fixed-time signal plan, also called time-of-day or pretimed plan, is an off-line predetermined plan that is periodical during a specific time of day (e.g., evening peak). Fixed-time plans are appropriate for networks with sparse or unreliable real-time data. They are also commonly used by major cities with high and uniformly distributed congestion levels, such as New York City (Osorio et al. 2015).

We consider a fixed-time signal control problem for the 5:00–5:30 P.M. evening peak. The signal plans of the 17 intersections are determined jointly. The decision variables are the green splits (i.e., normalized green times) of the phases of the different intersections. All other traditional control variables (e.g., cycle times, offsets, stage structure) are assumed fixed. This leads to a total of 99 endogenous signal phase variables; that is, the dimension of the decision vector is 99.

To formulate the problem, we introduce 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.
- $\mathscr{D}$ Set of intersection indices.
- $\mathscr{P}_D(d)$ Set of endogenous signal phase indices of intersection $d$.

**Figure 8.** (Color online) Lausanne Network Model

- $\mathscr{L}$ Set of all lanes.
- $\tilde{T}$ Total number of one-minute time intervals.
- $N$ Number of lanes (i.e., cardinality of $\mathscr{L}$).

The problem is formulated as follows:

$$\min_x f(x) = \frac{1}{\tilde{T}N} \sum_{i \in \mathscr{L}} \sum_{\hat{t}=1}^{\tilde{T}} P(UQ_i(\hat{t}; x) = l_i) \tag{43}$$

subject to

$$\sum_{j \in \mathscr{P}_D(d)} x(j) = b_d, \quad \forall\, d \in \mathscr{D}, \tag{44}$$

$$x \geq x_{LB}. \tag{45}$$

The decision vector, $x$, denotes the green splits of the signal controlled lanes. The linear equality constraints (44) ensure that, for each intersection, the sum of green times equals the available cycle time. Constraint (45) ensures lower bounds for the green splits. This objective function averages, over time and over all lanes, the spillback probability of each lane. This spillback probability is represented by $P(UQ_i(\hat{t}; x) = l_i)$, which denotes the probability of $UQ$ being full at integer time $\hat{t}$ under signal plan $x$. This problem formulation minimizes the spatial and temporal occurrence of spillbacks.

The above signal control problem has a probabilistic formulation, which is naturally addressed with probabilistic traffic models. Given the high computation times of the multivariate model (see Section 3), the above problem is only solved with the proposed mixture model.

### Implementation Details

The values of the main exogenous parameters of the mixture model are displayed in Table 4. The decision variables of this problem (the green splits of the signal plans) determine the downstream flow capacity of the

**Table 4.** Parameters for Lausanne Case Study

| Parameter | Value |
|---|---|
| $\tilde{T}$ | 30 one-minute intervals |
| $N$ | 902 lanes |
| $\delta$ | 0.1 sec |
| $x_{LB}$ | 4 sec |
| $v$ | 50 km/h |
| $w$ | −15 km/h |
| $\hat{\rho}$ | 200 veh/km |
| $s$ | 1,800 veh/h |
| $\mu$ | Varies by signal plans |
| $\lambda$ | Calculated from Equation (47) |
| $l, \gamma, p_{ij}, e_i, b_d$ | Exogenous values obtained from Osorio (2010, chapter 4) |
| $k^{\text{fwd}}$ | $k^{\text{fwd}} = \lceil l/(\hat{\rho}v) \rceil$ |
| $k^{\text{bwd}}$ | $k^{\text{bwd}} = \lceil l/(\hat{\rho}\lvert w \rvert) \rceil$ |

underlying lanes. More specifically, for a signal controlled lane $i$, its flow capacity is given by

$$\mu_i - \sum_{j \in \mathscr{P}_I(i)} x(j)s = e_i s, \quad \forall\, i \in \tilde{\mathscr{L}}, \tag{46}$$

where $s$ represents the saturation flow, $e_i$ represents the ratio of fixed green time to cycle time of signalized lane $i$, $\mathscr{P}_I(i)$ represents the set of endogenous signal phases of lane $i$, and $\tilde{\mathscr{L}}$ denotes the set of signal controlled lanes.

This paper formulates a link model. It can be coupled with a probabilistic node model to formulate a full network model. As is discussed in Section 5, the formulation of probabilistic traffic-theoretic node models is part of ongoing work. To limit this case study to the use of the link model (rather than link and node models), we assume link demand to be exogenous; that is, it does not vary with signal plans. Hence, the mixture model is used to design signal plans that improve within-link traffic dynamics. Across-link dynamics—or more generally, changes in traffic assignment—are not accounted for in this formulation. The results of this case study show that even with the use of such simplifying assumptions (e.g., the lack of an endogenous node model), the link model identifies signal plans with good network-wide performance.

The exogenous arrival rate (or demand rate) for lane $i$ at time interval $k$, denoted by $\lambda_i(k)$, is computed, prior to optimization, by solving the following linear system of equations:

$$\lambda_i(k) = \gamma_i + \sum_j p_{ji}\lambda_j(k), \quad \forall\, i \in \mathscr{L}, \tag{47}$$

where $\gamma_i$ denotes an external arrival rate (i.e., rate of trips that start at lane $i$), and $p_{ji}$ is a turning probability from lane $j$ to lane $i$. Both $\gamma_i$ and $p_{ji}$ are exogenous and time independent; hence $\lambda$ is also exogenous and time independent. Equation (47) states that the arrival rate of lane $i$ is the sum of the external arrival rate $\gamma_i$ to lane $i$ and of the demand that arises from upstream lanes. Problem (43)–(45) is solved using the *active-set* algorithm of the *fmincon* routine of MATLAB (MATLAB 2016).

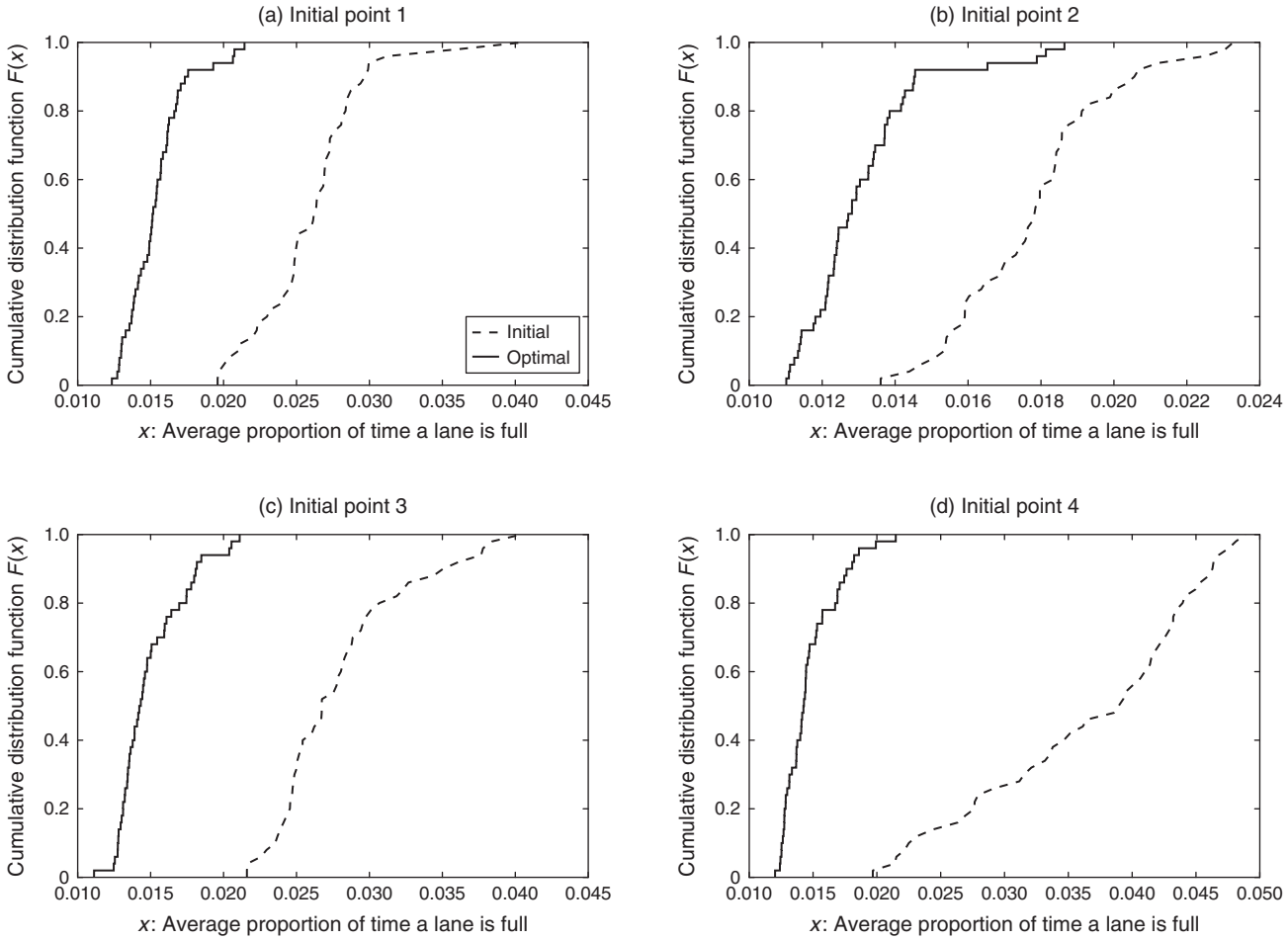
### 4.2. Numerical Analysis

We solve problem (43)–(45) considering four different initial points. Each point is drawn uniformly randomly from the feasible space (Equations (44) and (45)). The uniform sampling is conducted using the code of Stafford (2006). The use of four different initial points leads to four optimal solutions. To evaluate the performance of the various signal plans (initial and optimal), we use a microscopic traffic simulation model of Lausanne (Dumont and Bert 2006), which is calibrated for the evening peak period demand and implemented with the Aimsun simulator (Transport Simulation Systems 2014). Each signal plan is embedded within the simulator; 50 simulation replications are run. We then compare the cumulative distribution (obtained over these 50 replications) of the main network performance measures. Each simulation replication consists of a 15-minute warm-up period, followed by a 30-minute (5:00–5:30 P.M.) simulation period. For a given simulation replication, the objective function (43) is estimated as the average (over all lanes) proportion of time a lane is full.

Each plot of Figure 9 considers one random initial point. Each plot displays two cumulative distribution curves: one for the initial signal plan and one for the optimal plan of problem (43)–(45). Each curve is the cumulative distribution function (cdf) of the average proportion of time a lane is full. More specifically, the $x$ axis displays the average proportion of time a lane is full. For a given value of $x$, the $y$ axis displays the proportion of simulation replications (out of 50) that have the average proportion of time a lane is full smaller than $x$. Therefore, the more a cdf curve is shifted to the left, the better the performance of the corresponding signal plan. The solid curves correspond to the cdf of the initial signal plans, and the dashed curves represent that of the optimal signal plans of problem (43)–(45). As shown in Figures 10(a)–10(d), all the cdf curves of the optimal signal plan are to the left of the corresponding initial plan. In other words, the model yields solutions that have a lower average proportion of time a lane is full.

Figures 10 and 11 have a similar figure structure as Figure 9. Figure 10 analyzes the performance of the signal plans in terms of the average lane queue length (in vehicles). This average is computed over time and over lanes. The $x$ axis displays the average lane queue length. For a given value of $x$, the $y$ axis displays the proportion of simulation replications (out of 50) that have an average lane queue length smaller than $x$. As before, the more these curves are shifted to the left,

**Figure 9.** Cumulative Distribution Functions of the Average Proportion of Time a Lane Is Full, Considering Different Initial Signal Plans



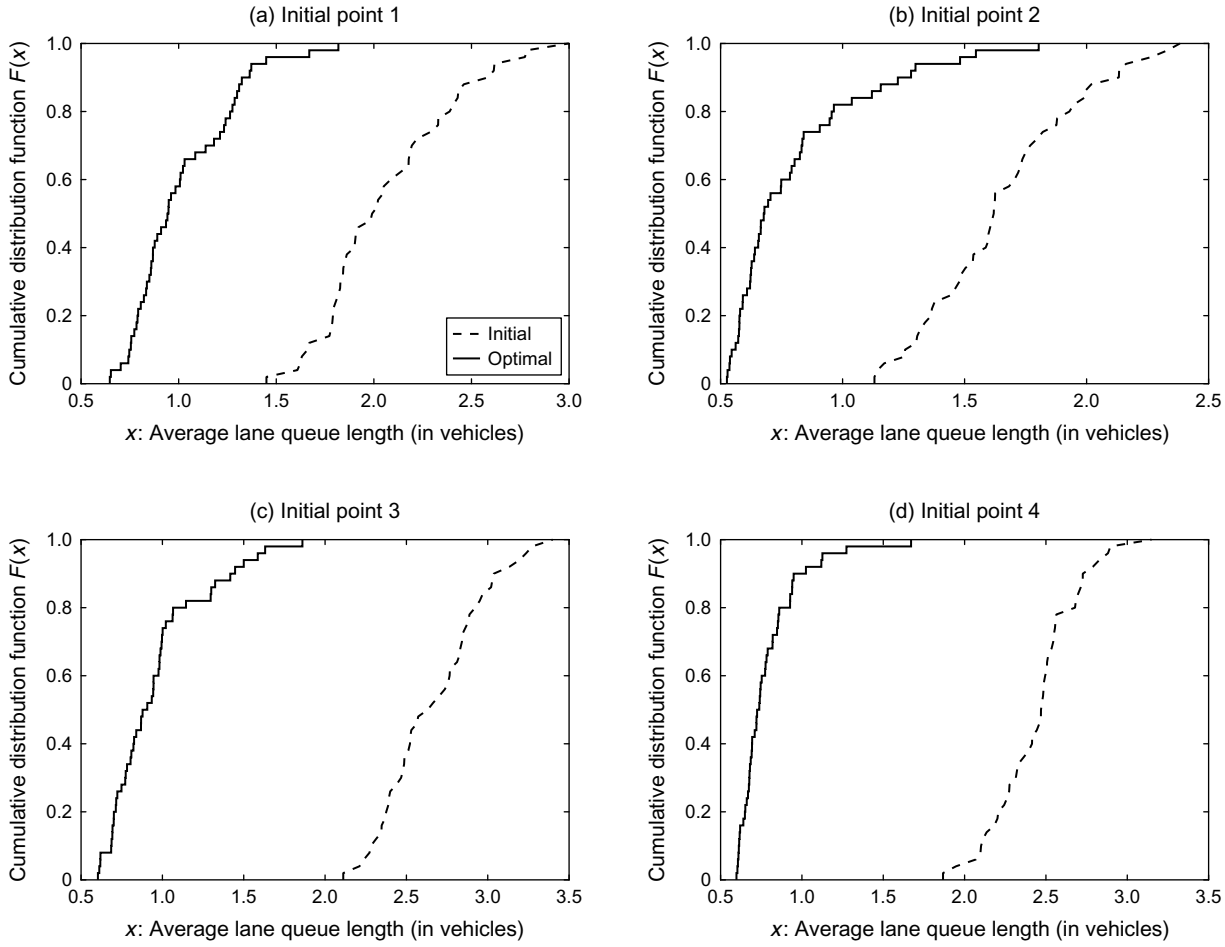

the better the performance of the corresponding signal plans. The four plots of Figure 10 indicate that, for all initial points, the proposed optimal signal plans yields a lower average lane queue length. Figure 11 analyzes the performance of the signal plans in terms of the average trip travel times (in minutes). The $x$ axis displays the average trip travel time. For a given value of $x$, the $y$ axis displays the proportion of simulation replications (out of 50) that have average trip travel times smaller than $x$. For all initial points, the proposed optimal signal plans yield lower average trip travel times.

### 4.3. Comparison to Signal Plans Derived by Commercial Signal Control Software

In this section, we compare the performance of the optimal signal plans with that of a signal plan obtained from widely used commercial signal control software (Synchro Trafficware 2011). For details on how the signal plan for the city of Lausanne is obtained from Synchro, we refer the reader to Section 5.3 of Osorio and Chong (2015). Note that Synchro, which is a signal control optimization software based on a deterministic macroscopic traffic model, does not solve problem (43)–(45).

Figures 12–14 consider the same performance metrics as before: average proportion of time a lane is full, average lane queue length, and average trip travel time. Each figure displays nine cdf curves. The four dashed (respectively, solid thin) curves correspond to the four initial (respectively, optimal) points of the previous analysis. The solid thick curve corresponds to the signal plan proposed by Synchro. Recall that for each figure, the more a cdf curve is shifted to the left, the better the performance of the corresponding signal plan. For all three figures, the four leftmost curves are the four plans proposed by the mixture model. In other words, for all three performance metrics, the proposed plans outperform all initial plans as well as the Synchro plan. These figures also show that, for all three metrics, the performance of the initial plans varies significantly, while the performance of the proposed signal plans is very similar. This illustrates the robustness of the proposed model to the quality of the initial points.

**Figure 10.** Cumulative Distribution Functions of the Average Lane Queue Length, Considering Different Initial Signal Plans




For two metrics, the average proportion of time a lane is full and the average lane queue length, the Synchro plan outperforms three of the four initial plans and performs similarly to the fourth plan. For the average trip travel time metric, the Synchro plan outperforms all four initial points.

## 5. Conclusions

This paper formulates an analytical stochastic link model that is both computationally tractable and consistent with the kinetic theory of traffic flow. The model is validated versus stochastic simulation results, using a simulator of the stochastic link transmission model. Compared with the model of Osorio and Flötteröd (2015), the proposed model has a complexity that is linear in the link space capacity, rather than cubic. This leads to significant gains in computational run times. Both models provide an accurate approximation of the distribution of the link's boundary conditions. The proposed model is used to address a signal control problem for the city of Lausanne. It yields signal plans that systematically outperform initial random plans for various performance metrics. The experiments illustrate the robustness of the model to the quality of the initial points. The proposed plans also outperform a signal plan derived by widely used commercial signal control software.

Ongoing work formulates scalable probabilistic network models. There are two main challenges to be addressed. First, there is a need to formulate probabilistic and scalable node models. The probabilistic model of Osorio, Flötteröd, and Bierlaire (2011) includes a two-link node model that provides a higher-order description of the across-node dependencies. It yields the joint distribution of the boundary conditions that each link adjacent to a node provides to the node—that is, the joint distribution of the upstream link's downstream boundary conditions and the downstream link's upstream boundary conditions. The extension of this formulation to nodes with multiple upstream and downstream links is part of ongoing work. Second, there is a need to formulate scalable network models. For a network with $n$ links, each with space capacity $l$, directly coupling the proposed link model with the node model of Osorio, Flötteröd, and Bierlaire (2011) would yield a model complexity

**Figure 11.** Cumulative Distribution Functions of the Average Trip Travel Times, Considering Different Initial Signal Plans

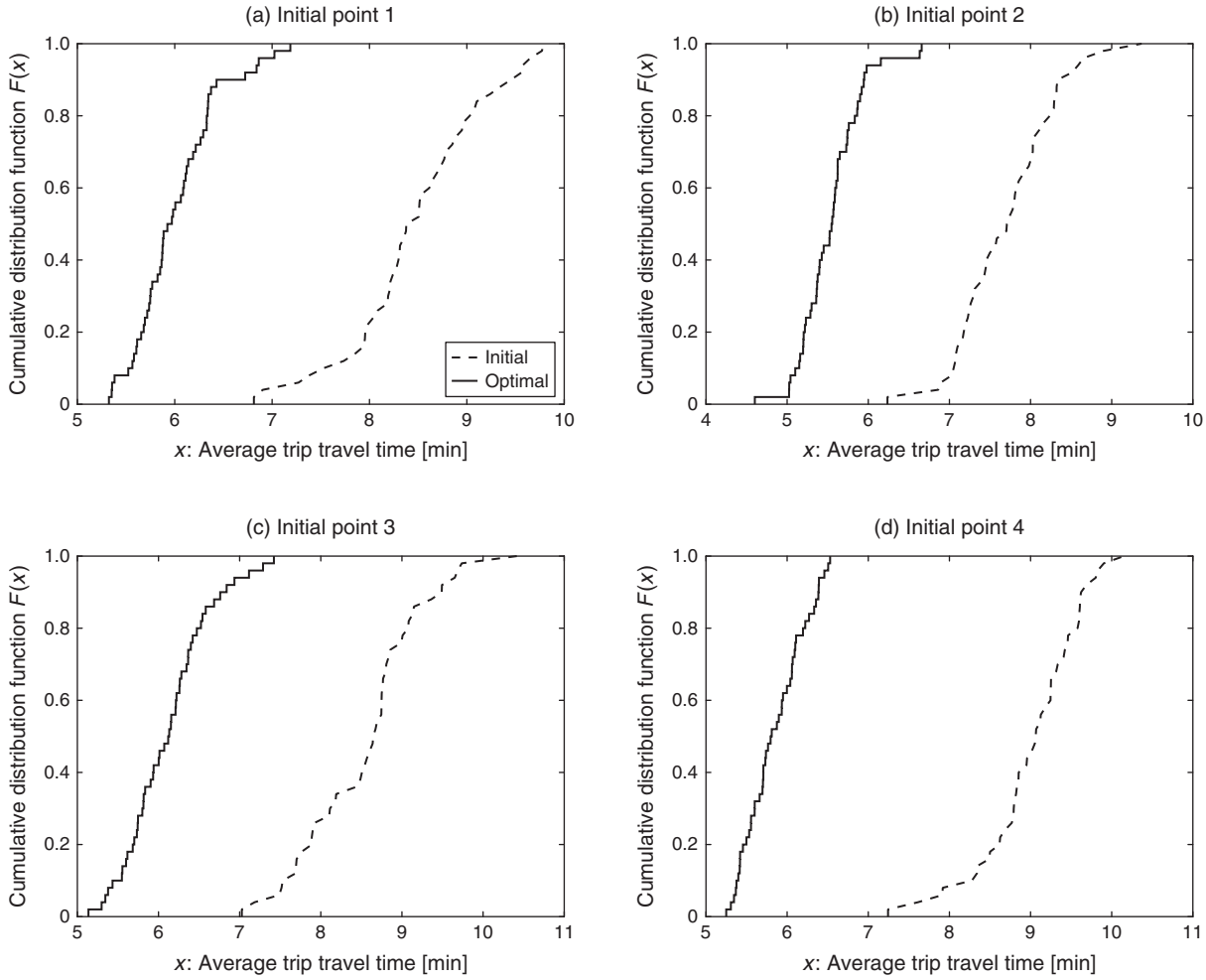



**Figure 12.** Cumulative Distribution Functions of the Average Proportion of Time a Lane Is Full

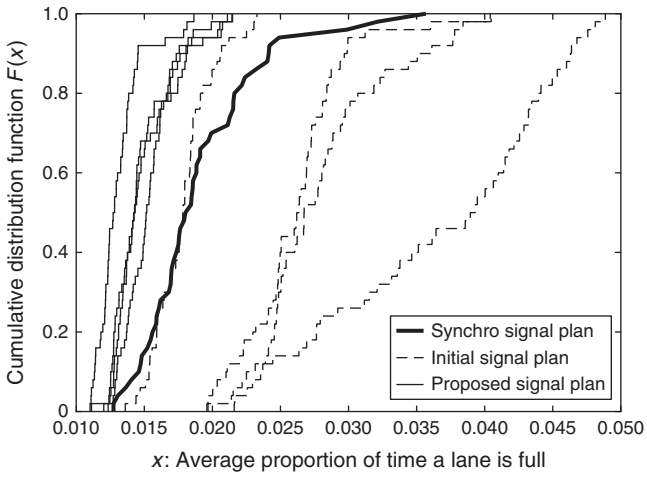
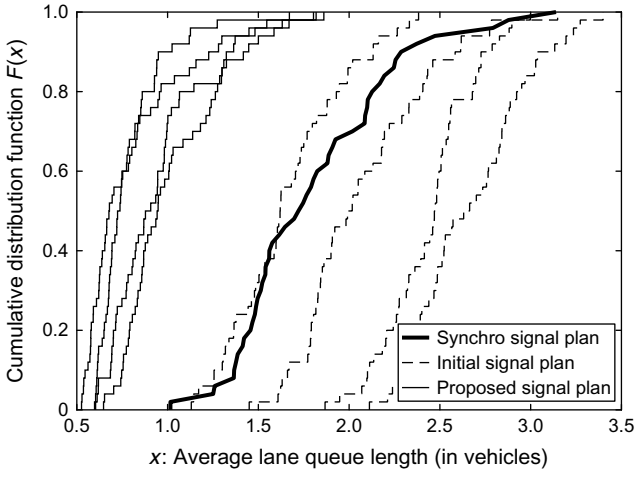
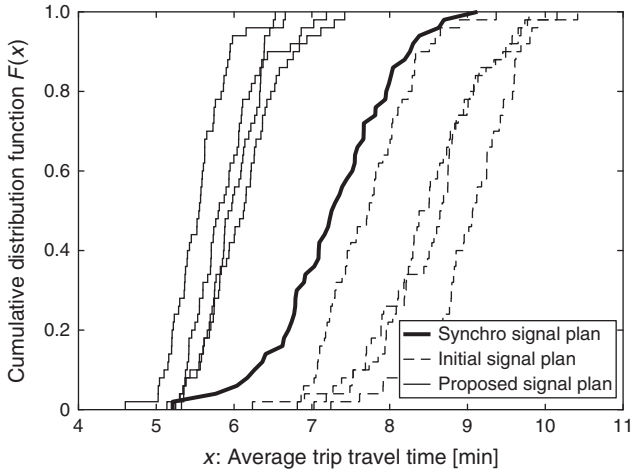



**Figure 13.** Cumulative Distribution Functions of the Average Lane Queue Length




in the order of $\mathcal{O}(l^n)$. Such a model is inappropriate for large-scale network analysis. Ongoing work investigates two research directions. First, we study the use of network decomposition techniques. For instance, combining the link and node models with the technique of Flötteröd and Osorio (2017) would lead to a network

**Figure 14.** Cumulative Distribution Functions of the Average Trip Travel Time




model with complexity $\mathcal{O}(sl^r)$, where $s$ is the number of intersections and $r$ is the maximum number of links adjacent to an intersection. Second, we study the use of aggregation–disaggregation techniques that address the curse of dimensionality by providing an aggregate description of network states (Osorio and Wang 2017, Osorio and Yamani 2017).

## Acknowledgments

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 authors thank Gunnar Flötteröd (KTH Royal Institute of Technology) for providing access to the simulator.

## References

Boel R, Mihaylova L (2006) A compositional stochastic model for real time freeway traffic simulation. *Transportation Res. Part B: Methodological* 40(4):319–334.

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

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

Dumont AG, Bert E (2006) Simulation de l'agglomération lausannoise, SIMLO. Report, Laboratoire des voies de circulation, ENAC, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. http://web.mit.edu/osorioc/www/papers/dumont06BertRapport.pdf.

Endres DM, Schindelin JE (2003) A new metric for probability distributions. *IEEE Trans. Inform. Theory* 49(7):1858–1860.

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

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

Jabari SE (2012) A stochastic model of macroscopic traffic flow: Theoretical foundations. Doctoral thesis, University of Minnesota, Minneapolis.

Jabari SE, Liu HX (2012) A stochastic model of traffic flow: Theoretical foundations. *Transportation Res. Part B: Methodological* 46(1):156–174.

Jabari SE, Liu HX (2013) A stochastic model of traffic flow: Gaussian approximation and estimation. *Transportation Res. Part B: Methodological* 47:15–41.

Jabari SE, Zheng J, Liu HX (2014) A probabilistic stationary speed-density relation based on Newell's simplified car-following model. *Transportation Res. Part B: Methodological* 68:205–223.

Kullback S, Leibler RA (1951) On information and sufficiency. *Ann. Math. Statist.* 22(1):79–86.

Lam WH, Shao H, Sumalee A (2008) Modeling impacts of adverse weather conditions on a road network with uncertainties in demand and supply. *Transportation Res. Part B: Methodological* 42(10):890–910.

Larson RC, Odoni AR (1981) *Urban Operations Research* (Prentice-Hall, Englewood Cliffs, NJ).

Laval JA, Castrillón F (2015) Stochastic approximations for the macroscopic fundamental diagram of urban networks. *Transportation Res. Part B: Methodological* 81:904–916.

Laval JA, Chilukuri BR (2014) The distribution of congestion on a class of stochastic kinematic wave models. *Transportation Sci.* 48(2):217–224.

Lighthill M, Whitham J (1955) On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc. London A* 229(1178):317–345.

MATLAB (2016) *Optimization Toolbox: User's Guide* (*R2016a*) (MathWorks, Natick, MA).

Morse PM (1958) *Queues, Inventories and Maintenance: The Analysis of Operational Systems with Variable Demand and Supply* (John Wiley & Sons, New York).

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

Osorio C (2010) Mitigating network congestion: Analytical models, optimization methods and their applications. Doctoral thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

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

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

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

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

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

Osorio C, Chen X, Gao J, Talas M, Marsico M (2015) On the control of highly congested urban networks with intricate traffic patterns: A New York City case study. Technical report, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge. http://web.mit.edu/osorioc/www/papers/osoChenNYCDOTOfflineSO.pdf.

Reibman A (1991) A splitting technique for Markov chain transient solution. Stewart WJ, ed. *Numerical Solution of Markov Chains* (Marcel Dekker, New York), 373–400.

Richards PI (1956) Shock waves on highways. *Oper. Res.* 4(1):42–51.

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

Sumalee A, Zhong RX, Pan TL, Szeto WY (2011) Stochastic cell transmission model (SCTM): A stochastic dynamic traffic model for traffic state surveillance and assignment. *Transportation Res. Part B: Methodological* 45(3):507–533.

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

Transport for London (2010) Traffic modelling guidelines: TfL traffic manager and network performance best practice, version 3.0. Technical report, Transport for London, London.

Transport Simulation Systems (2014) *AIMSUN 8.1 Microsimulator User's Manual* (Transport Simulation Systems, Barcelona).

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

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