A probabilistic traffic-theoretic network loading model suitable for large-scale network analysis

Jing Lu
Operations Research Center
Massachusetts Institute of Technology, USA
Email: jl3724@mit.edu

Carolina Osorio
Department of Civil and Environmental Engineering
Operations Research Center
Massachusetts Institute of Technology, USA

July 10, 2017

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 upon the initial model of Osorio and Flötteröd (2015). It proposes a formulation with enhanced scalability. In particular, compared to 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 runtimes of the proposed model increase linearly with the link's space capacity, while the initial model has an exponential increase in runtimes. The proposed model yields accurate distributional approximations of the link's boundary conditions. It is used to address a probabilistic formulation of a city-wide 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 a widely used commercial signal control software. The model is suitable for large-scale network optimization.

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 which 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; Department of Transportation; 2008); (ii) the availability of high resolution traffic data, which enables 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). Recent studies that review sources and modeling approaches to demand and supply uncertainty include Sumalee et al. (2011); Lam et al. (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 et al. (2014). For instance, the work of Heidemann (2001) uses a probabilistic non-stationary (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) is 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 et al. (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; Sumalee et al.; 2011; Jabari and Liu; 2012). The approach of Boel and Mihaylova (2006) is an example of the most common approach to stochastic CTM models 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 Witham (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 et al. (2011) and proposes

a link model that 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 an isolated link and derives an analytical description of the transient (i.e., time-dependent) distribution of link 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 $\frac{1}{6}(\ell+1)(\ell^2+2\ell+6)$. In other words, the model complexity is in the order of $\mathcal{O}(\ell^3)$.

This paper formulates a link model with a complexity that is linear, rather than cubic, in the link's space capacity, i.e., the proposed model has $\mathcal{O}(\ell)$ 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 (Section 3) and used to address a large-scale signal control problem (Section 4). Conclusions and a discussion of ongoing work are presented in Section 5. The 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 et al. (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 to the link, it 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, there is 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 (resp. $L/|w|$) is rounded to the integer k^{fwd} (resp. k^{bwd}).

This process is summarized in Figure 1. During time interval k, the link has an expected inflow (resp. outflow) denoted $q^{in}(k)$ (resp. $q^{out}(k)$). The delay incurred upon entrance to the link is represented by the *lagged inflow queue*, denoted LI. In discrete time, LI can be thought of as a set of k^{fwd} cells. One can think of this delay as if the flow traveled sequentially from the first until the k^{fwd}th cell of LI. This last cell of LI is denoted 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 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 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^{bwd}th cell of LO. This last cell of LO is denoted LLO in Figure 1. The multivariate link model accounts for stochasticity in the link's arrival and departure processes. Time-dependent (i.e., inhomogeneous, non-homogeneous) 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.

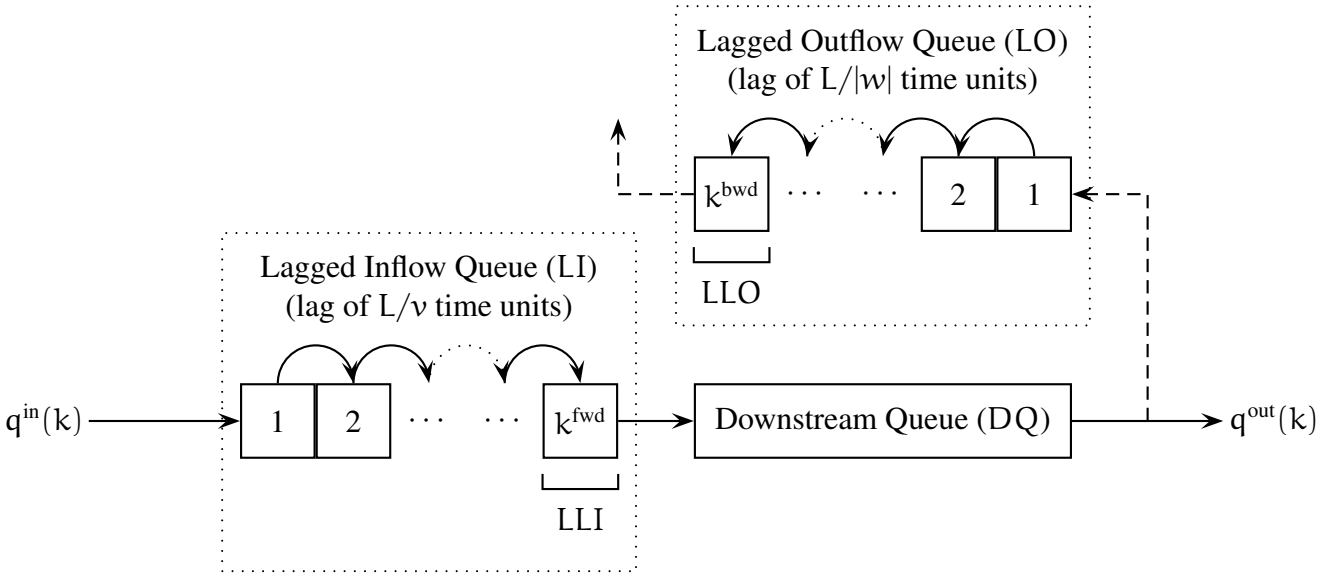


Figure 1: Link dynamics of the multivariate link model

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 $LI(t;k)$ (resp. $DQ(t;k)$, $LO(t;k)$, $UQ(t;k)$) as the number of vehicles in LI (resp. DQ,LO,UQ) at continuous time t within discrete time interval k of duration δ. 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 $\{(li, dq, lo) \in [0,\ell]^3 : li + dq + lo \leq \ell\}$. The state space dimension is $\frac{1}{6}(\ell+1)(\ell^2+2\ell+6)$.

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 underly 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 (resp. 2.4) formulates a univariate model that tracks the distribution of UQ (resp. 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 $\{uq \in [0, \ell]\}$. For time interval k of duration δ, the transient probability distribution of UQ satisfies a system of linear differential equations with 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, i.e., 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:

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

Equations (4) states that in order to approximate the transient distribution of UQ, we need to approximate the transition rate matrix $Q^{UQ}(k)$. Table 1 defines the non-diagonal and non-null elements of the transition rate matrix. This table considers for an arbitrary initial state uq of UQ (displayed in column 1), the feasible instantaneous transitions that can take place to new states (column 2), the rate at which the transitions take place (column 3), and the conditions on the initial states needed for the transitions to be feasible (column 4). 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: $uq < \ell$). The second type of event are flow departures from UQ, these 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 non-diagonal elements by:

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

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^{bwd}th cell denoted LLO. Therefore, the number of departures from UQ during time interval k is a random variable and it can be expressed as $LLO(k)|UQ(k) = uq$. Let $E[T_m(k)]$ denote the expected inter-departure time from UQ conditional on there being a total of m departures during time interval k. By definition, 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:

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

initial state	new state	rate	condition
uq	$uq+1$	$\lambda(k)$	$uq < \ell$
uq	$uq-1$	$\mu^{UQ}(uq;k)$	$uq > 0$

Table 1: Transition rate table of UQ.

Equation (6) approximates $\mu^{UQ}(uq;k)$ as the mean inverse of expected inter-departure time from UQ conditional on there being a total of m departures during time interval k. In order 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 δ, then the unordered arrival times are independently, uniformly distributed over the time interval of interest (cf., for instance, Section 2.12.3 of Larson and Odoni (1981)). Hence, the expected inter-arrival time is δ/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 δ/m. Equation (6) becomes:

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

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

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

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:

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

where $q^{in}(k)$ (resp. $q^{out}(k)$) denotes the instantaneous link inflow (resp. outflow) rates at the end of time interval k. As in the multivariate model (as well as in its deterministic counterpart model of Yperman et al. (2007)), we use these instantaneous link inflow and outflow rates to approximate the expected inflow to (resp. outflow from) the link during time interval k. In other words, these instantaneous rates $q^{in}(k)$ and $q^{out}(k)$ are held constant throughout the time interval k of duration δ. 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 $k - k^{bwd} - 1$ (this is represented by the second summation). Recall that k^{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:

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

The term $q^{out}(k - k^{bwd})$ represents the expected flow that has left the link during time interval $k - k^{bwd}$. 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 $(uq, q^{LLO}(k)/q^{UQ}(k))$ (cf., 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 with:

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

The accuracy of this approximation depends on the dependency between LLO and $\{UQ - LLO\}$. In particular, we expect it to decrease as 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 $P(UQ(k) < \ell)$. Hence, the expected inflow to the link is:

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

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 $P(UQ(k) = \ell)$. 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:

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

From Equation (4), we obtain the distribution of UQ at the end of time interval k, which allows us to compute $q^{in}(k)$ through Equation (13). In order to compute $q^{out}(k)$, 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:

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

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 $k - k^{fwd} - 1$ (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:

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

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

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

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 $i \in [0, \ell]$:

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

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

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.

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 $\{dq \in [0, \ell]\}$. Just as for the UQ model, the transient distribution of DQ, $P(DQ(k))$, satisfies an equation of the form (4) with initial conditions $P(DQ(k-1))$ and transition rate matrix $Q^{DQ}(k)$. The

Algorithm 1 Algorithm of the univariate upstream queue model

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

non-diagonal and non-null 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, i.e., 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 (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, i.e., $\mu^{UQ}(uq; k)$ depends on uq.

For a finite capacity birth-death process with state-independent rates, Morse (1958, Equation (6.13), Chap. 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 δ, DQ distribution at the end of time

initial state	new state	rate	condition
dq	$dq + 1$	$\lambda^{DQ}(k)$	$dq < \ell$
dq	$dq - 1$	$\mu(k)$	$dq > 0$

Table 2: Transition rate table of DQ.

interval k, $P(DQ(k))$, is given by:

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

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 $P_n(k)$ 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 in the queuing literature as the inverse of the relaxation time. In summary, the distribution of DQ is given by (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, $\lambda^{DQ}(k)$. In order to approximate this rate, we observe that for a queue with finite space capacity ℓ and arrival rate λ, the expected inflow to the queue is given by: $\lambda P(N < \ell)$, 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:

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

The numerator $q^{in}(k-k^{fwd})$ represents the expected inflow to the link during time interval $k-k^{fwd}$, i.e., this is the flow that is expected to leave the last cell of LI (denoted LLI in Figure 1) and enter DQ during time interval k. The denominator $P(DQ(k-1) < \ell)$ 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.

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

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

$$\begin{aligned} P(UQ(k) < \ell) &= 1 - P(UQ(k) = \ell) && (23)\\ &= 1 - \sum_{n=0}^{\ell} P(UQ(k) = \ell \mid DQ(k) = n) P(DQ(k) = n) && (24)\\ &= 1 - \sum_{n=0}^{\ell} P(UQ(k) - DQ(k) = \ell - n \mid DQ(k) = n) P(DQ(k) = n) && (25)\\ &\approx 1 - \sum_{n=0}^{\ell} p_1(k)^{\ell-n} P(DQ(k) = n). && (26) \end{aligned}$$

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

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

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

$$\begin{aligned} E[UQ(k)] &= E[DQ(k)] + E[UQ(k) - DQ(k)] && (27)\\ &= E[DQ(k)] + E[E[UQ(k) - DQ(k) | DQ(k)]] && (28)\\ &= E[DQ(k)] + \sum_{n=0}^{\ell} E[UQ(k) - DQ(k) \mid DQ(k) = n] P(DQ(k) = n) && (29)\\ &\approx E[DQ(k)] + \sum_{n=0}^{\ell} (\ell - n) p_1(k) P(DQ(k) = n) && (30)\\ &= E[DQ(k)] + p_1(k)(\ell - E[DQ(k)]). && (31) \end{aligned}$$

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 $(\ell - n, p_1(k))$, then $E[UQ - DQ \mid DQ = n]$ equals $(\ell - n)p_1(k)$, which leads to (30). The summation is simplified to obtain (31), which itself can be rearranged to obtain the approximation for $p_1(k)$:

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

In order to evaluate Equation (32): $E[DQ(k)]$ can be computed from the marginal distribution of DQ (Equations (21)) as:

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

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

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

In summary, $P(UQ(k) < \ell)$ 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 $i \in [0, \ell]$:

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

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

$$\approx \sum_{n=0}^{i} \binom{\ell - n}{i - n} p_1(k)^{i-n} (1 - p_1(k))^{\ell - i} P(DQ(k) = n). \tag{37}$$

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 $(\ell - n, p_1(k))$.

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

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

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

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 (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:

$$\tilde{w}(\ell, \mu, k^{fwd}) = e^{-\frac{\ell^2}{70\mu k^{fwd}}}. \tag{40}$$

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 $\ell \in \{5, 10, 15, \ldots, 100\}$; $\rho = \lambda/\mu \in \{0.25, 0.5, 0.75\}$; $\mu \in \{0.2, 0.4, 0.6\}$. A more detailed description of the derivation of weight parameter, $\tilde{w}$, is given in Appendix A.

For the mixture model, the expected inflow and outflow, i.e. $q^{in}(k)$ and $q^{out}(k)$, are obtained according to Equations (13) and (14) where $P(UQ(k) < \ell)$ 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 runtime.

3 Validation

In this section we validate the model. We evaluate and compare both in terms of computational runtime and accuracy. First, we compare the computational runtimes 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 veh/sec for all experiments. The experiments consider different arrival rates and link lengths (and hence, different space capacities, forward

Algorithm 3 Algorithm of the mixture model

1. set exogenous parameters $\hat{\rho}, v, w, \ell$ and δ
2. set arrival and service rate over time $\lambda(k)$ and $\mu(k)$ for $\forall\, k = 1, 2, ...$
3. compute $k^{fwd} = \lceil \frac{\ell}{\hat{\rho} v} \rceil$ and $k^{fwd} = \lceil \frac{\ell}{\hat{\rho}\lvert w \rvert} \rceil$
4. compute $\tilde{w}$ according to Eq. (40)
5. set exogenous initial link conditions: $q^{in}(0)$, $q^{out}(0)$, $P(UQ(0))$, $P(DQ(0))$, $q^{UQ}(0)$, $q^{LLO}(0)$, $q^{LLI}(0)$ and $q^{DQ}(0)$
6. set $q^{in}(r) = 0$ and $q^{out}(r) = 0$ for $r < 0$
7. run step 6 of algorithm 1, this yields $P^{UQ}(UQ(k))$ for all $k = 1, 2...$
8. run step 6 of algorithm 2, this yields $P^{DQ}(DQ(k))$ for all $k = 1, 2...$
9. for any time interval k,
   (a) compute $P^{UQ}(DQ(k))$ according to Eq. (20)
   (b) compute $P^{DQ}(UQ(k))$ according to Eq. (37)
   (c) compute $P(UQ(k))$ according to Eq. (38)
   (d) compute $P(DQ(k))$ according to Eq. (39)
   (e) compute $q^{in}(k)$ and $q^{out}(k)$ according to Eq. (13) and (14)

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	0.1 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 3: Link Parameters

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 ($\ell \in \{10, 20, 30, 40, 60, 80, 100\}$). 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 $k^{fwd} \in \{5, 10, 15, 20, 30, 40, 50\}$ (in seconds) and backward lags $k^{bwd} \in \{10, 20, 30, 40, 60, 80, 100\}$ (in seconds). Each experiment starts with an empty link at time zero 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 runtimes of the mixture model (circles) and of the multivariate model (asterisks). The x-axis considers the space capacity values ℓ. The y-axis displays the average computational runtime (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 runtime for evaluating an experiment is set to be 40 hours. If an experiment has not concluded within 40 hours, it is terminated. For $\ell = 30$ the average runtime of the multivariate model is already 2366 minutes ($\approx$ 39 hours). Hence, for experiments where $\ell > 30$, the multivariate model is not evaluated. Figure 2 illustrates that the runtime of the multivariate model increases exponentially with ℓ, while for the mixture model the increase appears linear. For the mixture model, the average runtime over all 21 experiments is 0.05 minutes. The maximum average runtime is obtained for $\ell = 100$ and is 0.11 minutes. Thus, compared to 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. In order 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, inter-departure 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, $\ell = 10$. 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 $[175, 300]$

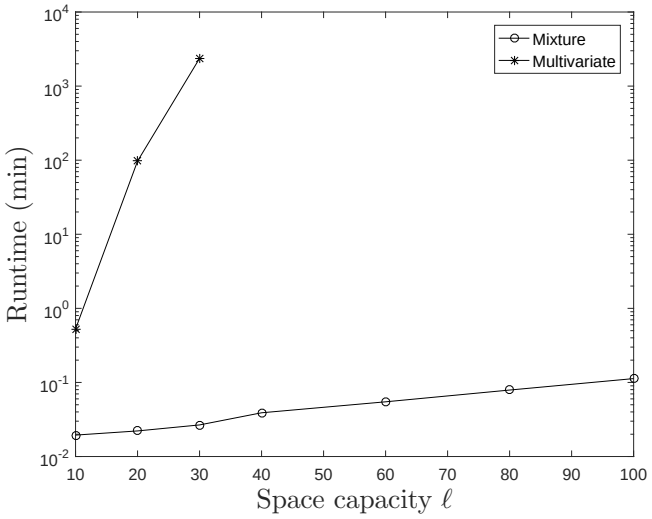


Figure 2: Model runtime comparison

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 (resp. DQ) have higher probabilities. After time $T = 175$, states with higher values of UQ (resp. 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 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 $P(UQ(T) = \ell)$. For instance, in the right-most plot of the second row of Figure 3(a), the spillback probability is non-zero (i.e., $P(UQ(T) = \ell) > 0$).

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.

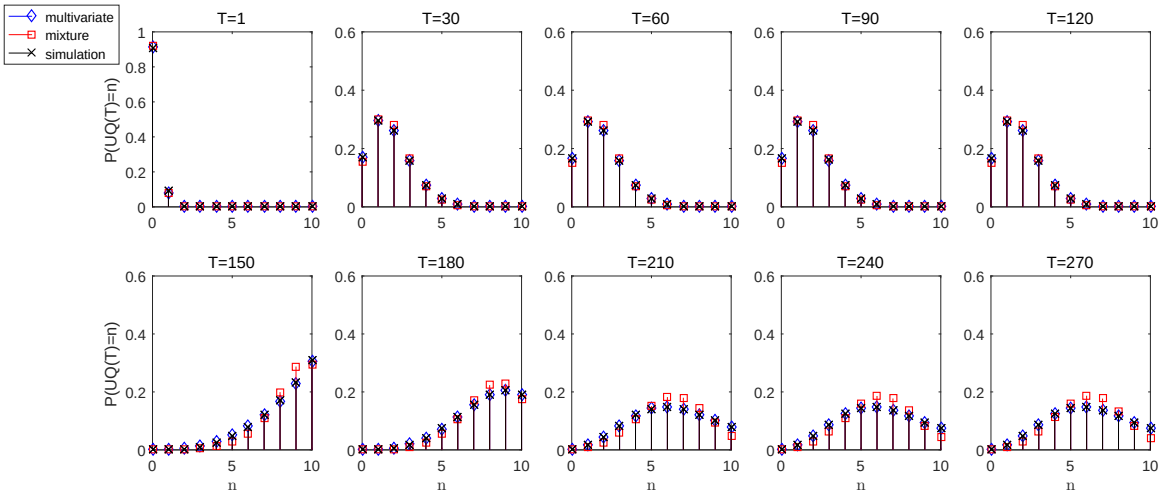
We now evaluate the accuracy of the mixture model over a larger set of experiments. We consider the 21 experiments mentioned above. The main goal is to evaluate the loss of accuracy of the mixture model compared to the (less scalable but more accurate) multivariate model. In order to evaluate the accuracy 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:

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

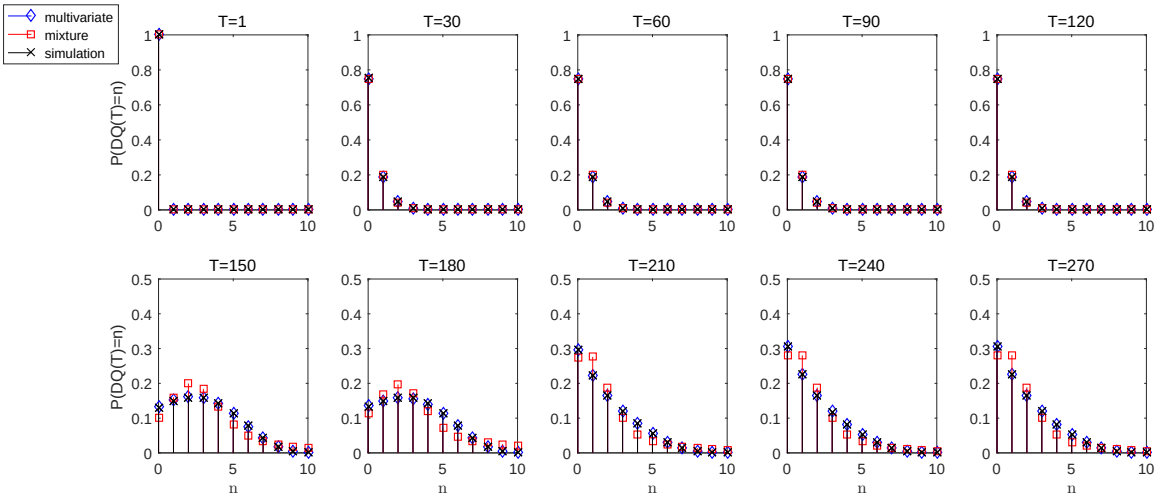
$$D(P_1 \parallel P_2) = \sum_i P_1(i) \log \frac{P_1(i)}{P_2(i)}, \tag{42}$$

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, i.e.: $\frac{1}{250}\sum_{T=1}^{250} JSD(P_1(T) \parallel P_2(T))$ where $P_1(T)$ and $P_2(T)$ are the distributions evaluated at time T.

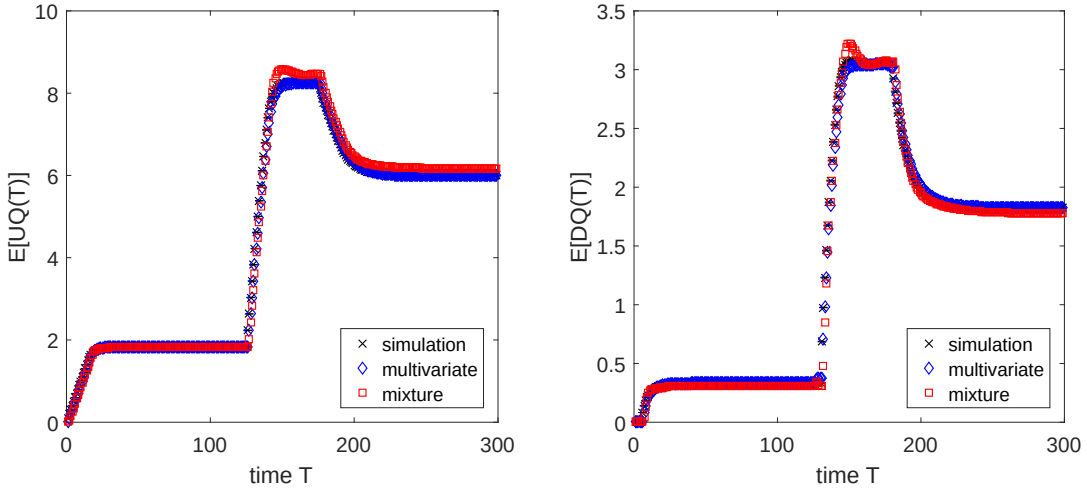
Since the main goal is to evaluate the accuracy loss of the mixture model compared to the multivariate model, we will compare the time-average JSD values of the mixture model (i.e., the time-average JSD distance between the distribution approximated by the mixture model and the simulated distribution) 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). In



(a) Distribution of UQ over time, P(UQ(T))

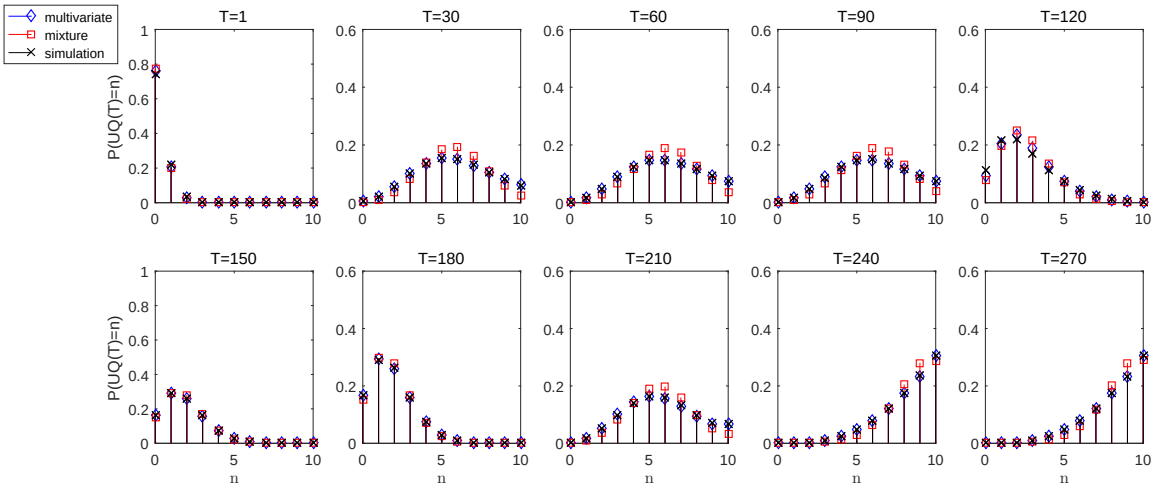


(b) Distribution of DQ over time, P(DQ(T))

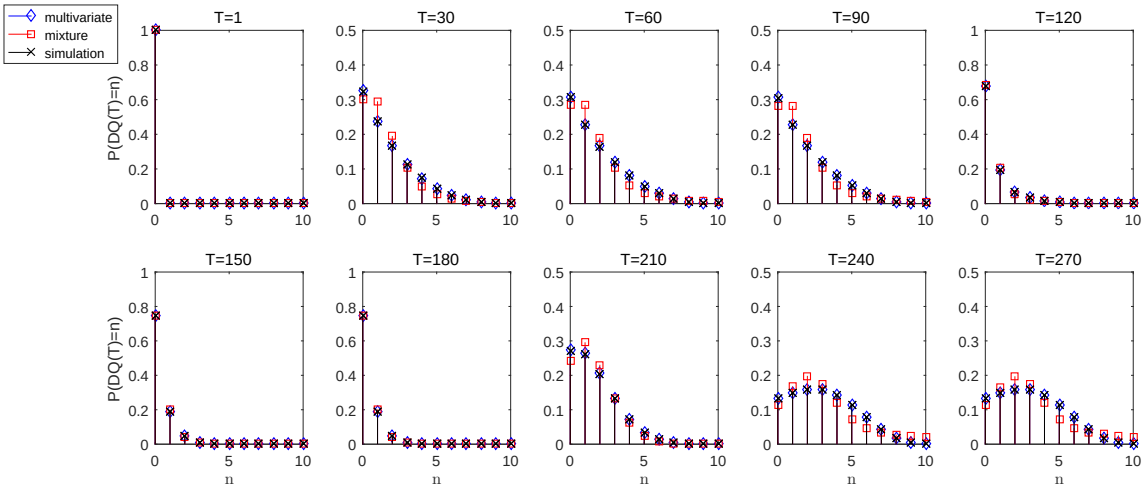


(c) Expectation of UQ and of DQ over time, E[UQ(T)] and E[DQ(T)]

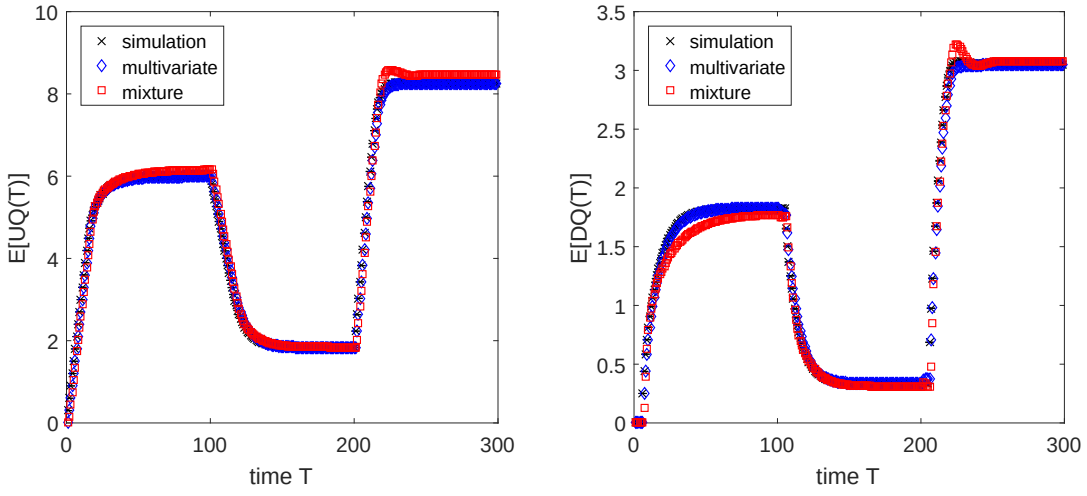
Figure 3: Experiment 1: impact of the temporal variation of demand on the distributions, as well as the expected values, of UQ and of DQ



(a) Distribution of UQ over time, P(UQ(T))



(b) Distribution of DQ over time, P(DQ(T))



(c) Expectation of UQ and of DQ over time, E[UQ(T)] and E[DQ(T)]

Figure 4: Experiment 2: impact of the temporal variation of demand on the distributions, as well as the expected values, of UQ and of DQ

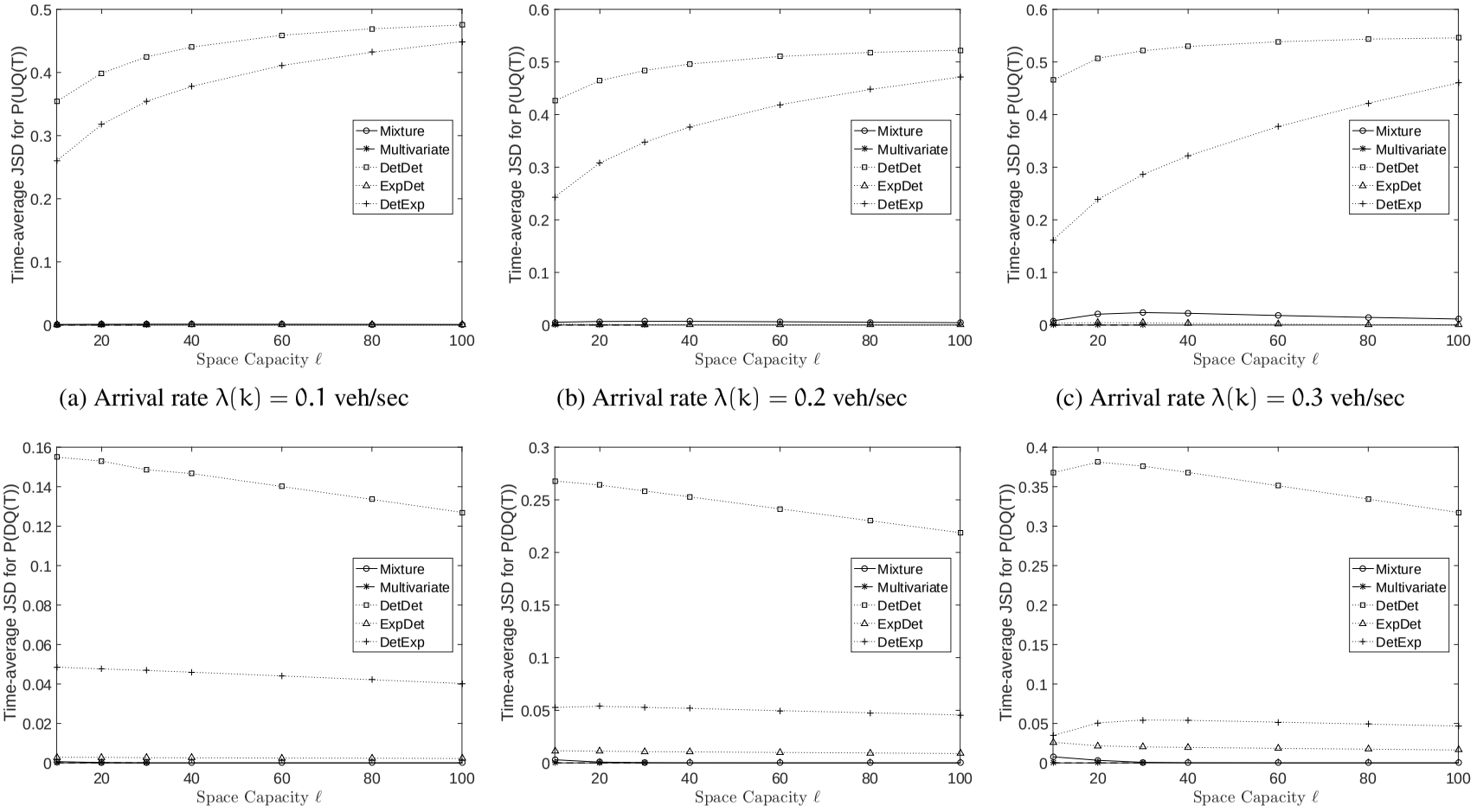
order to guide us in the interpretation of the magnitude of the JSD metric, we provide three additional models to compare the proposed model with: (i) the deterministic LTM (denoted DetDet, which stands for deterministic arrivals and deterministic departures), (ii) a simulation-based instance of the LTM with deterministic arrivals and independent exponentially distributed inter-departure times (denoted DetExp), (iii) a simulation-based instance of the LTM with independent exponentially distributed inter-arrival times and deterministic inter-departure times (denoted 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 (out 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 (resp. DetExp) can be interpreted as the effect of extending the LTM with a given probabilistic arrival (resp. departure) process.

Figure 5 displays the time-average JSD values for the 21 experiments described above. The top (resp. bottom) row plots consider the UQ (resp. DQ) distribution. The first column of plots considers the experiments with arrival rate $\lambda(k) = 0.1$ veh/sec. The second and third column consider arrival rate values of 0.2 and 0.3 veh/sec, respectively. Each plot compares 5 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 runtimes for the experiments with $\ell > 30$ 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 $\ell \geq 30$, the curve of the mixture model overlaps with the x-axis, it is barely visible. This indicates very high accuracy. Recall also that for $\ell > 30$, 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 5 and 6 of Appendix B.

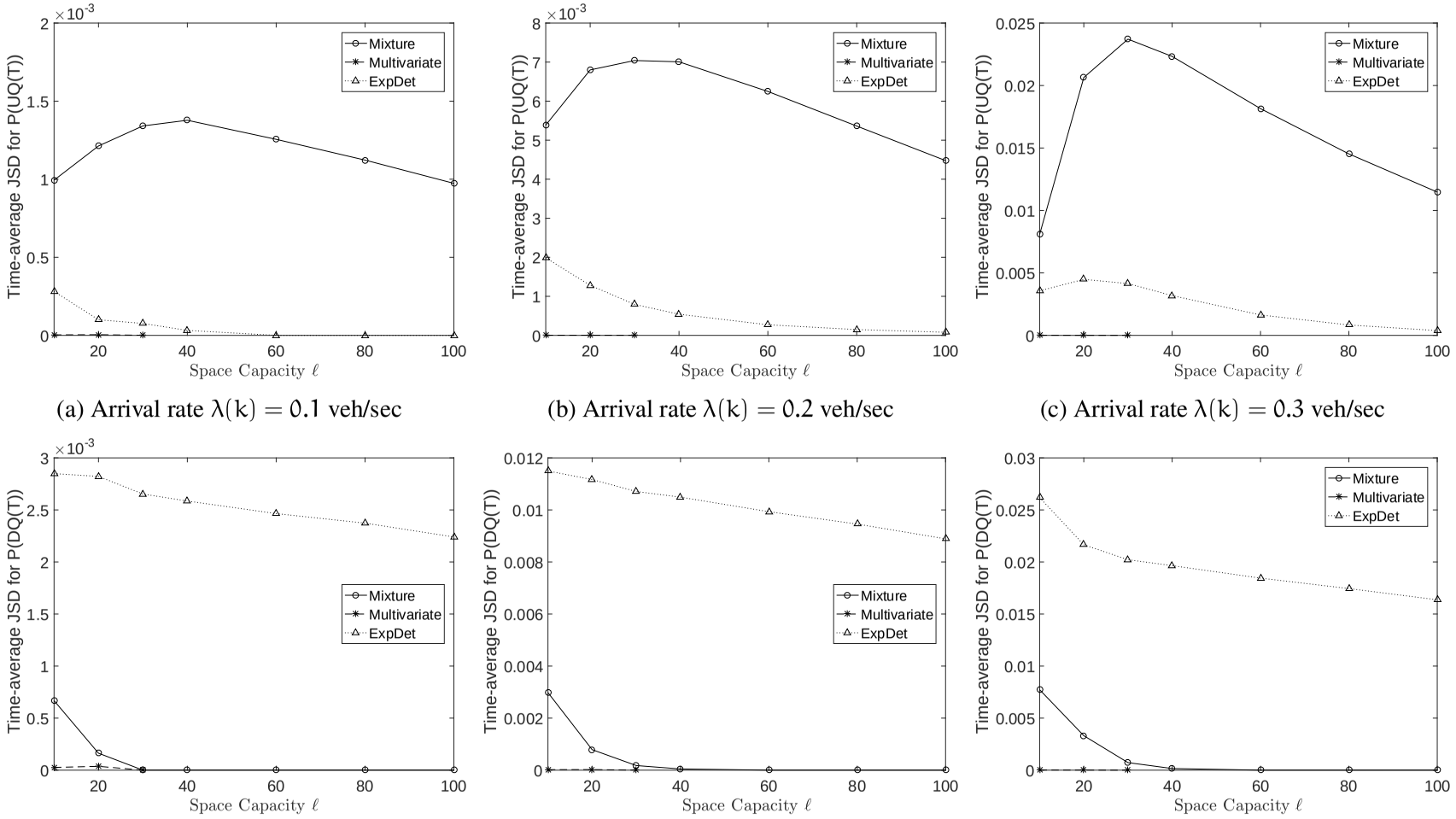
In summary, for experiments with both constant and time-varying demand, the mixture model performs comparably with the multivariate model, while being significantly faster to evaluate. The gain



(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 5: Comparison of the JSD values for the 21 experiments with time-independent demand



(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 6: Comparison of the JSD values for the 21 experiments with time-independent demand (zoomed-in results)

in computational runtime 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 a 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 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 pre-timed plan, is an off-line pre-determined 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 5:00-5:30pm 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, i.e., 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;
$\mathcal{D}$	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	number of lanes, i.e., cardinality of $\mathcal{L}$.

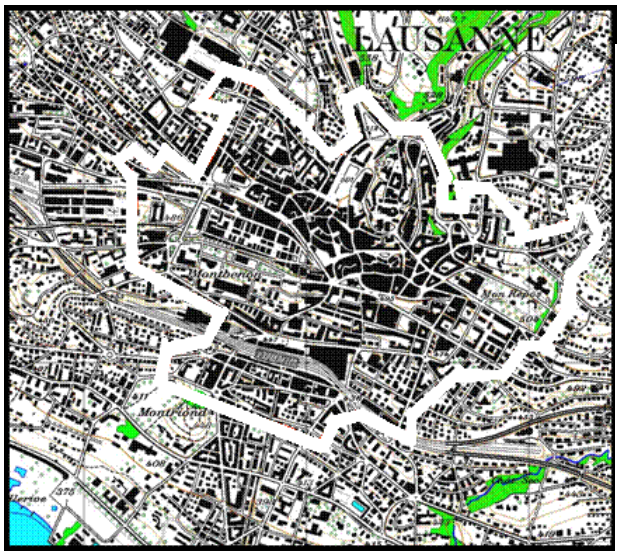


Figure 7: Lausanne city road network (adapted from Dumont and Bert (2006))

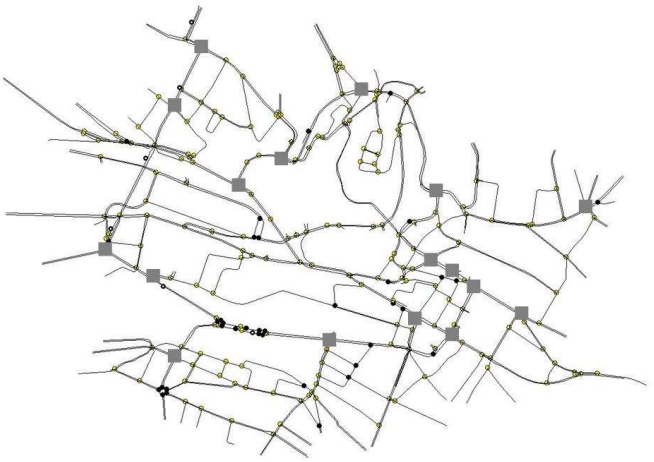


Figure 8: Lausanne network model

The problem is formulated as follows:

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

subject to

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

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

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 $P(UQ_i(\hat{t}; x) = \ell_i)$, 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 (cf. 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 underlying lanes. More specifically, for a signal controlled lane i, its flow capacity is given by:

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

where s represents the saturation flow, e_i represents the ratio of fixed green time to cycle time of signalized lane i, $\mathcal{P}_I(i)$ represents the set of endogenous signal phases of lane i and $\tilde{\mathcal{L}}$ 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. In order 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, i.e., 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 $\lambda_i(k)$, is computed, prior to optimization, by solving the following linear system of equations:

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

where γ_i denotes an external arrival rate (i.e., rate of trips that start at lane i), p_{ji} is a turning probability from lane j to lane i. Both γ_i and p_{ji} are exogenous and time-independent, hence λ is also exogenous and time-independent. Equation (47) states that the arrival rate of lane i is the sum of the external arrival rate γ_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

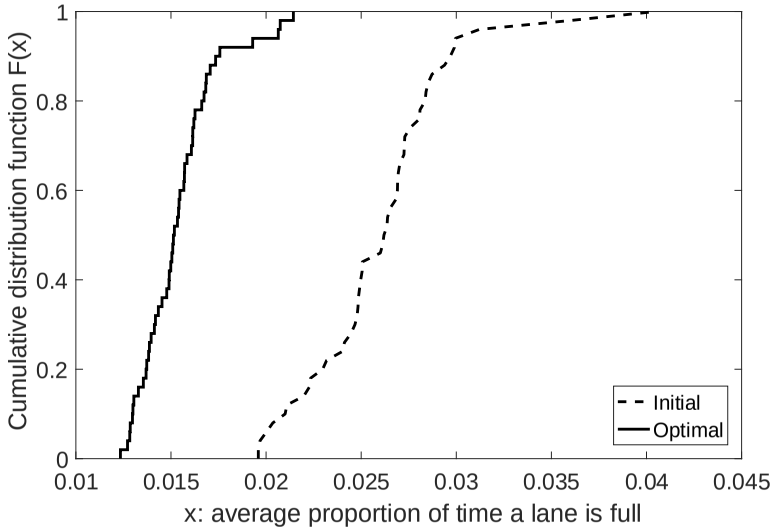
We solve Problem (43)-(45) considering four different initial points. Each point is drawn uniformly randomly from the feasible space (Equations (44)-(45)). The uniform sampling is conducted using the code of Stafford (2006). The use of four different initial points leads to four optimal solutions. In order 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 (TSS; 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 (5:00-5:30pm) 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.

Parameter	Value
$\tilde{T}$	30 one-minute intervals
N	902 lanes
δ	0.1 sec
x_{LB}	4 sec
v	50 km/h
w	-15 km/h
$\hat{\rho}$	200 veh/km
s	1800 veh/h
μ	varies by signal plans
λ	calculated from Equation (47)
ℓ, γ, p_{ij}, e_i, b_d	exogenous values obtained from Osorio (2010, Chap. 4)
k^{fwd}	$k^{fwd} = \lceil \frac{\ell}{\hat{\rho} v} \rceil$
k^{bwd}	$k^{bwd} = \lceil \frac{\ell}{\hat{\rho} \lvert w \rvert} \rceil$

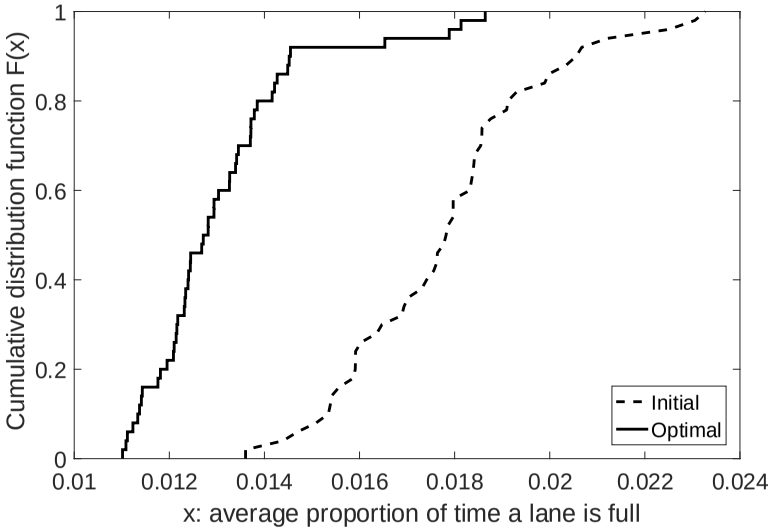
Table 4: Parameters for Lausanne case study

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 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, the dashed curves represent that of the optimal signal plans of Problem (43)-(45). As shown in plots 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 lower average proportion of time a lane is full.

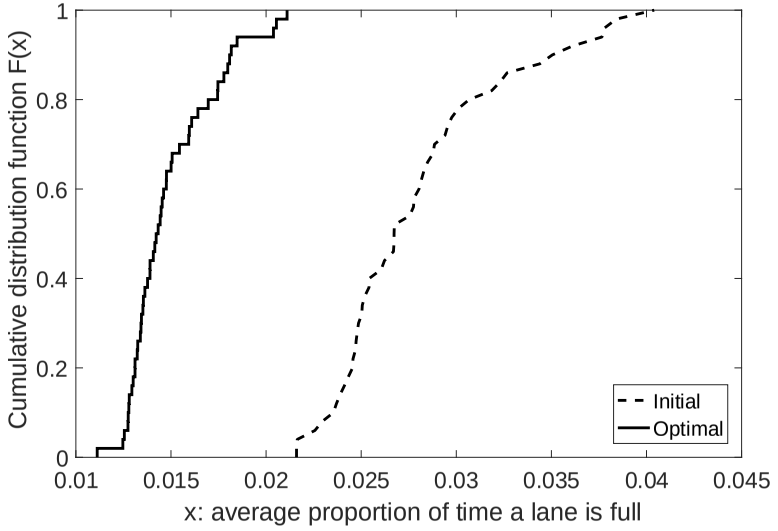
Figures 10 and 11 have similar figure structure as Figure 9. Figure 10 analyses the performance of the signal plans in terms of the average lane queue-length (in vehicles). This average is computed over time and over lanes. The x-axis displays the average lane queue-length. For a given value of x, the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



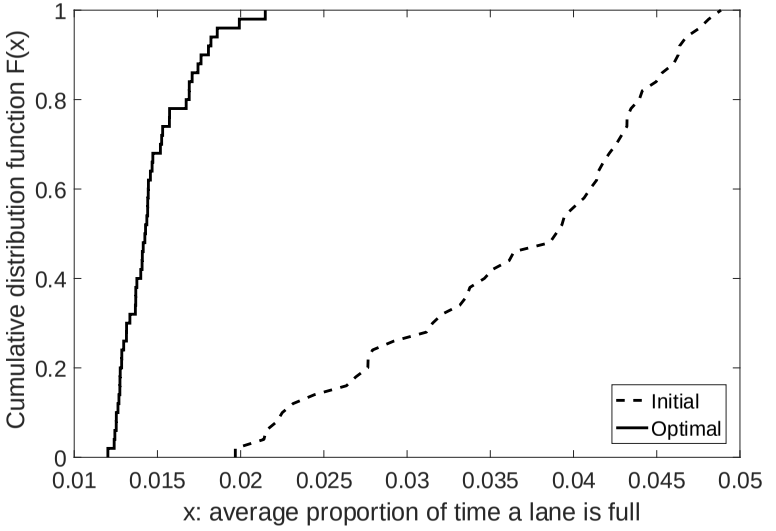
(a) Initial point 1



(b) Initial point 2

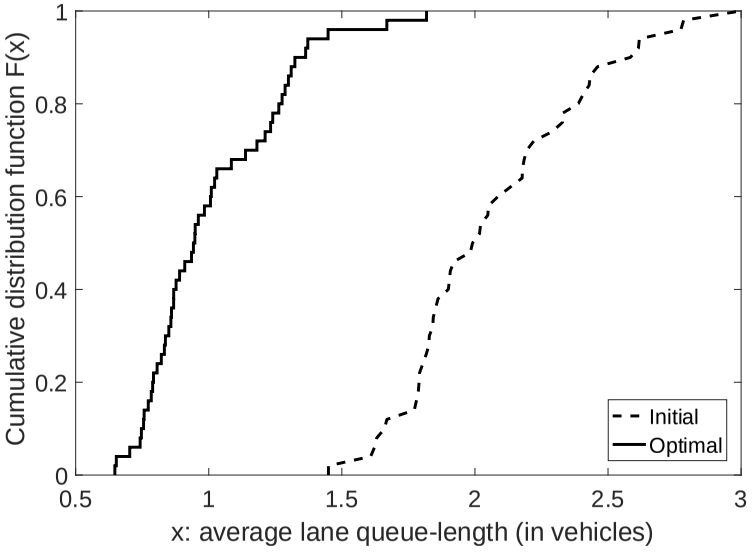


(c) Initial point 3

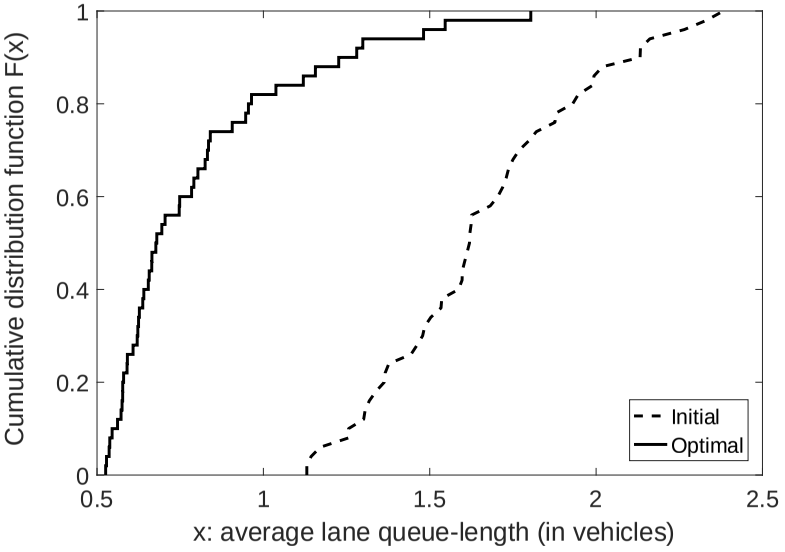


(d) Initial point 4

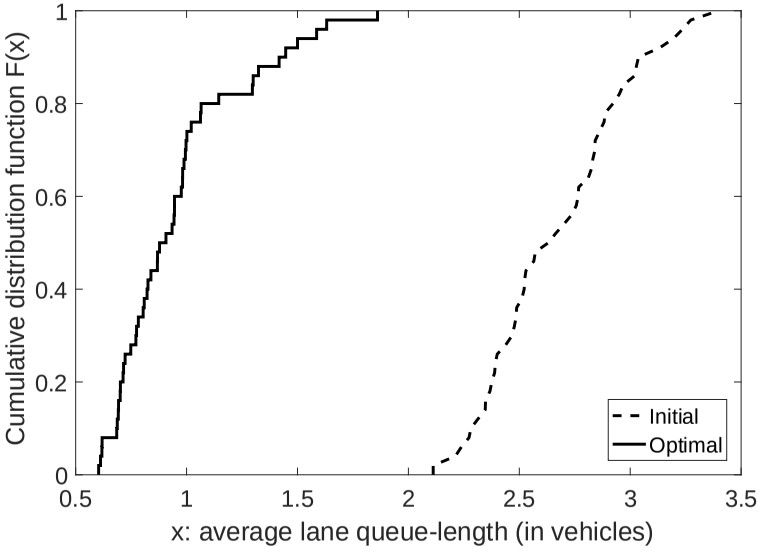
Figure 9: Cumulative distribution functions of the average proportion of time a lane is full considering different initial signal plans



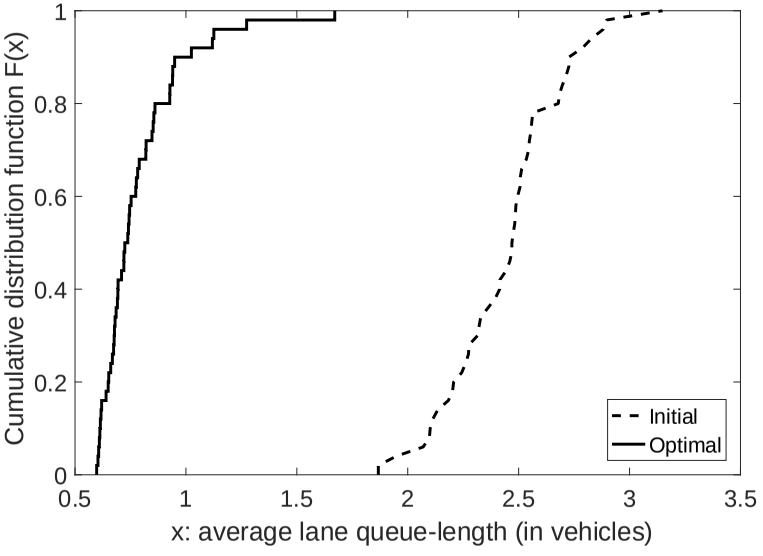
(a) Initial point 1



(b) Initial point 2

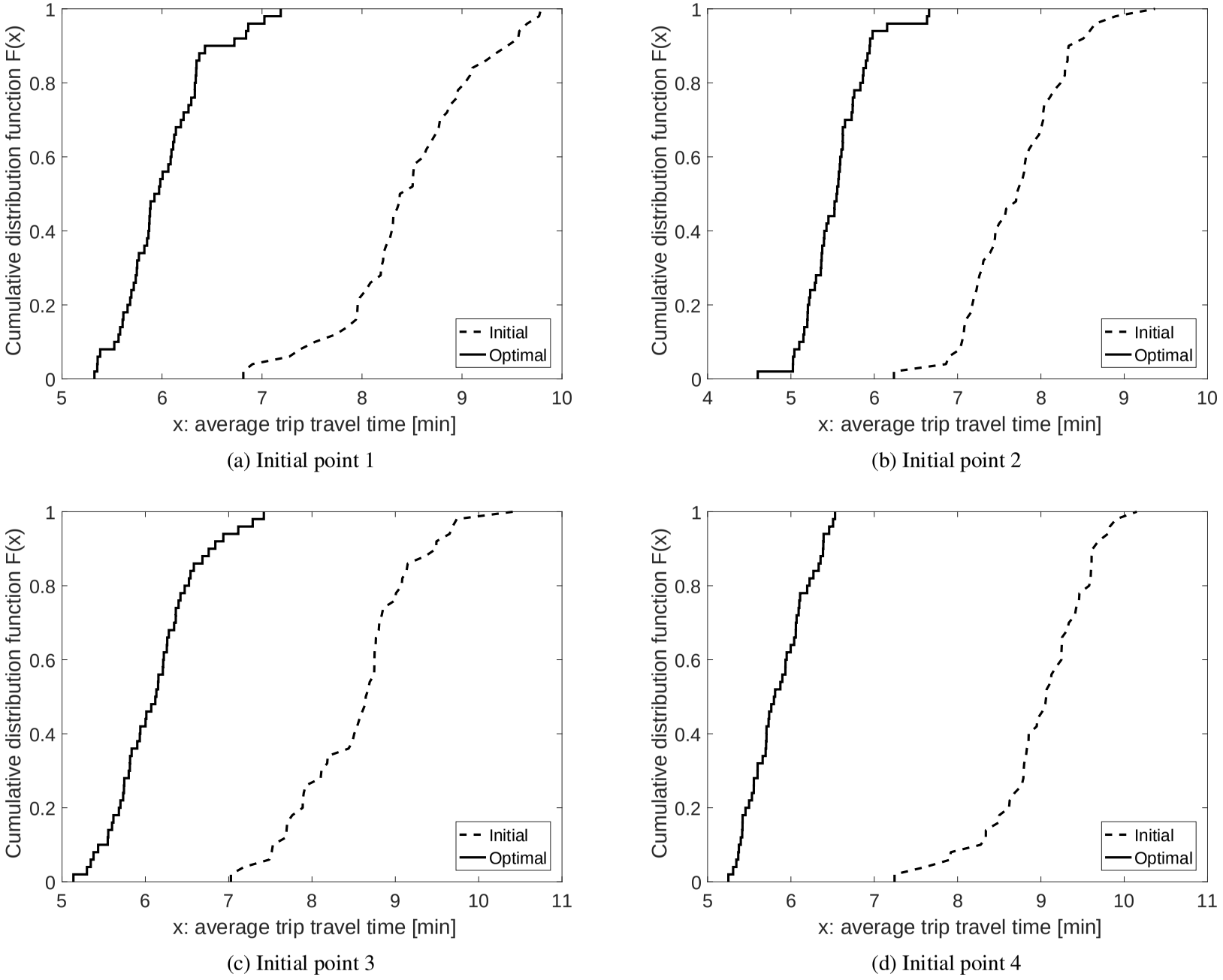


(c) Initial point 3



(d) Initial point 4

Figure 10: 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 11: Cumulative distribution functions of the average trip travel times considering different initial signal plans

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 a 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, 13 and 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 9 cdf curves. The four dashed (resp. solid thin) curves correspond to the four initial (resp. 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 left-most 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. For two metrics, average proportion of time a lane is full and average lane queue-length, the Synchro plan outperforms 3 of the 4 initial plans and performs similarly to the fourth plan. For the average trip travel time metric, the Synchro plan outperforms all 4 initial points.

5 Conclusions

This paper formulates an analytical stochastic link model that is both computationally tractable and is 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 to 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 runtimes. 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 a 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 et al. (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, i.e., 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

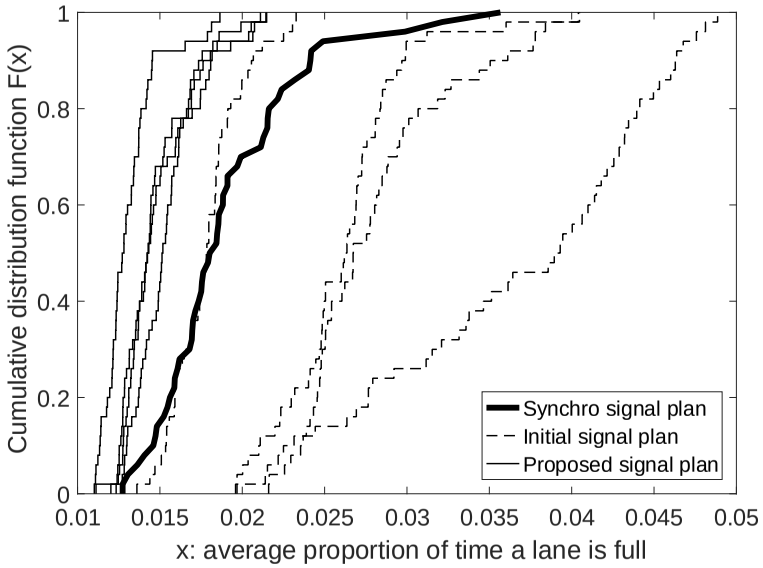


Figure 12: Cumulative distribution functions of the average proportion of time a lane is full

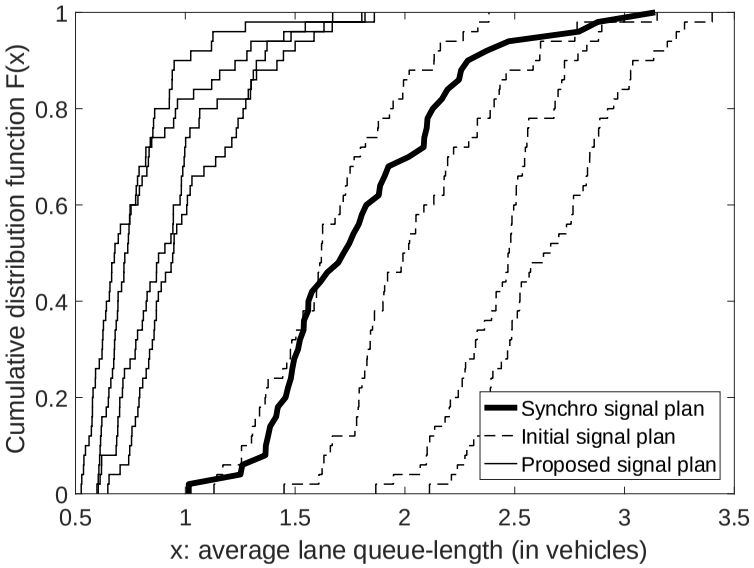


Figure 13: Cumulative distribution functions of the average lane queue-length

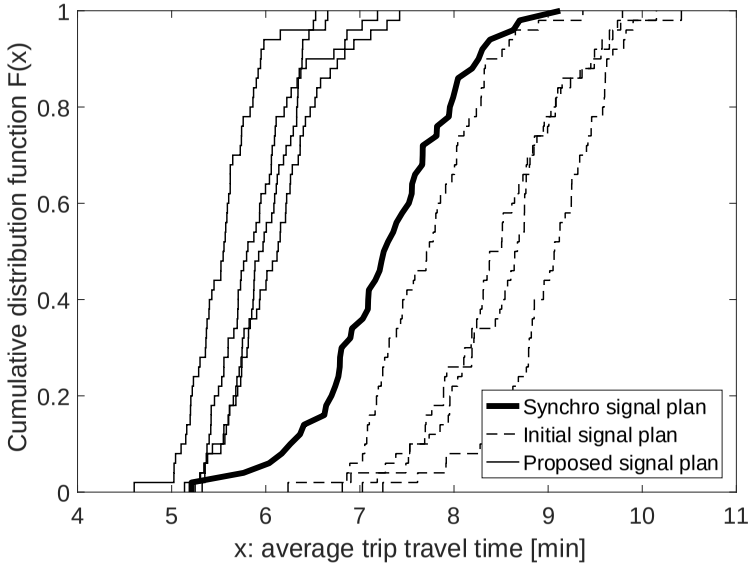


Figure 14: Cumulative distribution functions of the average trip travel time

of Osorio et al. (2011) would yield a model complexity in the order of $\mathcal{O}(\ell^n)$. 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 (2014) would lead to a network model with complexity $\mathcal{O}(s\ell^r)$, 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 Yamani; Forthcoming; Osorio and Wang; 2017).

6 Acknowledgment

The work of J. Lu and C. Osorio are 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 authors thank Prof. Gunnar Flötteröd (KTH) for providing the access to the simulator.

Appendices

A Estimation of the weight parameter $\tilde{w}$

This section describes the procedure followed to formulate, and to fit the coefficients of, the weight parameter of Equation (40). Recall that the goal of the mixture model is to accurately approximate the upstream and the downstream boundary conditions of the link. In other words, it should yield an accurate approximation of the distribution of UQ and of DQ. We consider a single isolated link and conduct a total of 180 experiments with varied combinations of the space capacity ($\ell \in \{5, 10, 15, \ldots, 100\}$), the traffic intensity ($\rho = \lambda/\mu \in \{0.25, 0.5, 0.75\}$) and the service rate (or downstream flow capacity) ($\mu \in \{0.2, 0.4, 0.6\}$). Each experiment considers a time period of duration 250 seconds. For each experiment we compare the approximation of the UQ and of the DQ distributions, over time T, to the distributions estimated via stochastic simulation with a discrete-event simulator of the stochastic link transmissions model. Based on the results of these experiments, we first observed that the parameters that most impact the quality of the approximation are ℓ, μ and k^{fwd}. This lead us to formulate the following expression for the weight parameter:

$$\tilde{w}(\ell, \mu, k^{fwd}; \beta) = e^{-\frac{\ell^2}{\beta \mu k^{fwd}}}, \tag{48}$$

where β is a scalar coefficient. The coefficient is fit such as to minimize, over all 180 experiments, the following error function:

$$\frac{1}{2}\left[\frac{1}{250}\sum_{T=1}^{250} JSD(P_1^{UQ}(T) \parallel P_2^{UQ}(T))\right] + \frac{1}{2}\left[\frac{1}{250}\sum_{T=1}^{250} JSD(P_1^{DQ}(T) \parallel P_2^{DQ}(T))\right], \tag{49}$$

where $P_1^{UQ}(T)$ (resp. $P_1^{DQ}(T)$) is the UQ (resp. DQ) distribution obtained from the mixture model at time T and $P_2^{UQ}(T)$ (resp. $P_2^{DQ}(T)$) is the UQ (resp. DQ) distribution estimated via stochastic simulation at time T. The two summations of (49) consider the error in the UQ distributions and in the DQ distributions, respectively. This leads to $\beta = 70$. This results in the final weight parameter expression defined in Equation (40).

B Tables of time-average JSD metric

Tables 5 and 6 display, respectively, the time-average JSD metric of the UQ and DQ distributions.

Experiment		Time-average JSD of the UQ distribution				
$\lambda(k)$	ℓ	Mixture	Multivariate	DetDet	DetExp	ExpDet
0.1	10	0.0010	0.0000	0.3539	0.2600	0.0003
	20	0.0012	0.0000	0.3988	0.3177	0.0001
	30	0.0013	0.0000	0.4248	0.3543	0.0001
	40	0.0014	NaN	0.4402	0.3779	0.0000
	60	0.0013	NaN	0.4590	0.4111	0.0000
	80	0.0011	NaN	0.4692	0.4320	0.0000
	100	0.0010	NaN	0.4753	0.4489	0.0000
0.2	10	0.0054	0.0000	0.4261	0.2434	0.0020
	20	0.0068	0.0000	0.4644	0.3080	0.0013
	30	0.0070	0.0000	0.4839	0.3476	0.0008
	40	0.0070	NaN	0.4961	0.3767	0.0005
	60	0.0062	NaN	0.5105	0.4189	0.0003
	80	0.0054	NaN	0.5181	0.4476	0.0001
	100	0.0045	NaN	0.5223	0.4713	0.0001
0.3	10	0.0081	0.0000	0.4654	0.1615	0.0036
	20	0.0206	0.0000	0.5071	0.2387	0.0045
	30	0.0237	0.0000	0.5214	0.2863	0.0041
	40	0.0223	NaN	0.5294	0.3215	0.0032
	60	0.0182	NaN	0.5384	0.3772	0.0016
	80	0.0145	NaN	0.5434	0.4215	0.0008
	100	0.0115	NaN	0.5458	0.4602	0.0004

Table 5: Time-average JSD metric of the UQ distribution. The value NaN denotes cases where the evaluation of the multivariate model exceeded the limit of 40 hours.

Experiment		Time-average JSD of the DQ distribution				
$\lambda(k)$	ℓ	Mixture	Multivariate	DetDet	DetExp	ExpDet
0.1	10	0.0007	0.0000	0.1550	0.0486	0.0028
	20	0.0002	0.0000	0.1530	0.0476	0.0028
	30	0.0000	0.0000	0.1486	0.0468	0.0027
	40	0.0000	NaN	0.1466	0.0460	0.0026
	60	0.0000	NaN	0.1401	0.0441	0.0025
	80	0.0000	NaN	0.1335	0.0422	0.0024
	100	0.0000	NaN	0.1270	0.0402	0.0022
0.2	10	0.0030	0.0000	0.2679	0.0527	0.0115
	20	0.0008	0.0000	0.2640	0.0538	0.0112
	30	0.0002	0.0000	0.2584	0.0538	0.0107
	40	0.0000	NaN	0.2528	0.0518	0.0105
	60	0.0000	NaN	0.2414	0.0497	0.0099
	80	0.0000	NaN	0.2301	0.0476	0.0095
	100	0.0000	NaN	0.2188	0.0456	0.0089
0.3	10	0.0077	0.0000	0.3677	0.0347	0.0262
	20	0.0033	0.0000	0.3811	0.0507	0.0217
	30	0.0007	0.0000	0.3760	0.0544	0.0202
	40	0.0002	NaN	0.3680	0.0539	0.0196
	60	0.0000	NaN	0.3512	0.0516	0.0184
	80	0.0000	NaN	0.3343	0.0492	0.0174
	100	0.0000	NaN	0.3173	0.0467	0.0164

Table 6: Time-average JSD metric of the DQ distribution. 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.

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

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.

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

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*

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions, *IEEE Transactions on Information Theory* **49**(7): 1858– 1860.

Flötteröd, G. and Osorio, C. (2014). Stochastic analytic dynamic qeueing network model with spillback, *Proceedings of the International Symposium of Dynamic Traffic Assignment (DTA)*.
Available at: http://web.mit.edu/osorioc/www/papers/floOso13Nwks.pdf .

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

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. and Liu, H. X. (2013). A stochastic model of traffic flow: Gaussian approximation and estimation, *Transportation Research Part B* **47**: 15–41.

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

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The annals of mathematical statistics* **22**(1): 79–86.

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

Larson, R. C. and Odoni, A. R. (1981). *Urban Operations Research*, Prentice-Hall, Inc., Englewood Cilffs, New Jersey, USA.

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

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

Lighthill, M. and Witham, J. (1955). On kinematic waves II. a theory of traffic flow on long crowded roads, *Proceedings of the Royal Society A* **229**: 317–345.

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

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

Osorio, C., Chen, X., Gao, J., Talas, M. and 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 (MIT). Available at: http://web.mit.edu/osorioc/www/papers/osoChenNYCDOTOfflineSO.pdf .

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. (Forthcoming). Analytical and scalable analysis of transient tandem Markovian finite capacity queueing networks, *Transportation Science* .
Available at: http://web.mit.edu/osorioc/www/papers/osoYamDynAggDisagg.pdf .

Reibman, A. (1991). A splitting technique for Markov chain transient solution, *in* W. J. Stewart (ed.), *Numerical solution of Markov chains*, Marcel Dekker, Inc, New York, USA, chapter 19, pp. 373–400.

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

Stafford, R. (2006). The theory behind the 'randfixedsum' function.
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.

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.

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