Urban Transportation Networks: Analytical Modeling of Spatial Dependencies and Calibration Techniques for Stochastic Traffic Simulators

by

Carter Wang

B.A., Rice University (2011)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Author .....
Department of Civil and Environmental Engineering
May 24, 2013

Certified by .....
Carolina Osorio
Assistant Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Heidi M. Nepf
Chair, Departmental Committee for Graduate Students

Urban Transportation Networks: Analytical Modeling of Spatial Dependencies and Calibration Techniques for Stochastic Traffic Simulators

by

Carter Wang

Submitted to the Department of Civil and Environmental Engineering
on May 24, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation

Abstract

Exact numerical evaluation of the stationary joint queue-length distribution of a Markovian finite capacity network with arbitrary size and topology can be obtained numerically. Nonetheless, the main challenge to such an approach remains the dimensionality of the joint distribution, which is exponential in the number of queues. This thesis proposes an analytical approximation of the joint distribution with a dimension that is linear in the number of queues. The method decomposes the network into overlapping subnetworks. The state of each subnetwork is described aggregately, i.e. in terms of a reduced state space, while ensuring consistency with the disaggregate, i.e., full state space, distribution. This aggregation-disaggregation technique is proposed for the analysis of Markovian tandem finite capacity queueing networks. The model is validated. We present its use to address an urban traffic control problem, and show the added value of accounting for higher-order spatial between-queue dependency information in the control of congested networks.

A second, distinct goal of this thesis is to examine the calibration of route choice parameters in microscopic traffic simulators. Automatically calibrating simulators using traffic counts requires describing the relationship between route choice and traffic flows. This thesis proposes an analytical finite capacity queueing model that accounts for the relationship between route choice and traffic flows. The method is embedded in a simulation-based optimization framework and applied to a calibration problem.

Thesis Supervisor: Carolina Osorio
Title: Assistant Professor of Civil and Environmental Engineering

Acknowledgments

To begin, I would like to thank Professor Carolina Osorio for providing countless ideas and encouragement in the pursuit of this research. This work would not have been possible otherwise, and I would certainly not be where I am today if not for her.

Thanks to Linsen, Franco, and Krishna for their input on various research related problems, their code, and occasionally, their computers.

I would also like to thank my friends and colleagues at MIT that I have made over the past two years. To you, I credit my sanity. The members of 1-151, Kanchana, Jameson, Ryan, Joel, Naomi, Andrés, Serdar, and Laura, have made the second half of my time especially enjoyable. I will certainly miss the lengthy discussions that served as a respite from the rigors of research and have certainly benefited from working with such a motivated, intelligent group of people.

Lastly, many thanks to the Center for Complex Engineering Systems, a collaboration between MIT and KACST that has provided funding for this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Global urbanization coupled with rising vehicle ownership has plagued many cities with congestion. The total vehicle stock is projected to increase from 800 million vehicles in 2002 to over 2 billion vehicles by 2030 (Dargay *et al.*, 2007), increasing faster in non-OECD countries. Additionally, studies have shown that vehicle-miles traveled increase proportionally to the available miles of roadway, suggesting that the construction of new roads does not effectively reduce congestion (Duranton and Turner, 2011). Congestion is thus an important issue in transportation policy, as it harms our environment and economy due to emissions and increased travel times.

Microscopic traffic simulators and macroscopic traffic models are important tools used by urban planners and transportation engineers to understand congestion. Traditionally, microscopic simulators have been used to estimate system performance and conduct scenario-based analyses of urban networks. These simulators model the complex decisions of individual drivers, such as the selection of routes and lane changing behavior, as well as the interactions between drivers. Thus, the evaluation of these models are computationally expensive. Macroscopic models have been used to determine strategies to reduce congestion, such as optimizing signal plans for a set of intersections, but rely on a less realistic set of assumptions. Improving the quality and use of existing microscopic and macroscopic models help address urban

congestion due to the low cost of implementation and the short-term solutions it can provide to existing transportation infrastructure.

1.2 Objectives

1.2.1 Analytical queueing model

Finite capacity queuing models are macroscopic models that have been used to address several different transportation optimization problems (Osorio and Chong, 2012b, Osorio and Nanduri, 2012, Chen *et al.*, 2012). The model, described in Osorio and Bierlaire (2009b), considers each lane within a road segment as a finite capacity queue. Interactions between queues, i.e. when a vehicle is blocked by a full downstream road segment, are accounted for indirectly through first order-moments, parameters that represent the arrival rate or service rate to a queue (road segment). This assumption improves the tractability of the model but lacks a detailed modeling of between-queue interactions.

The first objective is to build an analytical finite capacity queueing model that describes between-queue interactions in more detail by accounting for joint distributional information. Thus, a more accurate representation of congestion can be attained. This methodology is described in Chapter 2 and applied to a signal control problem in Section 2.4. The model is considered in a transportation context, yet is suitable for the analysis of other types of congested networks.

This work has been presented at the 15^{th} meeting of the EURO Working Group on Transportation and published in the conference proceedings Osorio and Wang (2012), as well as submitted for journal publication.

1.2.2 Simulation-based optimization

A second objective of this thesis is to develop efficient simulation-based optimization algorithms to address calibration problems for stochastic microscopic traffic simulators.

The applications of simulation-based optimization with microscopic traffic simulations extend beyond congestion mitigation techniques. Automatic calibration of the route choice parameters of microscopic traffic simulators is a difficult and practically relevant problem. A methodological challenge in this context is the formulation of tractable measurement equations that link available data to the model parameters. Current approaches require significant computational effort as they do not take advantage of the underlying problem structure. By combining microscopic traffic simulations with an analytical model that relates route choice parameters to traffic flow, gradient-based optimization routines can be used, assuming the analytical model is differentiable. Recent work by Flötteröd *et al.* (2011, 2012) and Flötteröd *et al.* (2012) demonstrates the use of an analytical approximation of the gradient in the calibration of route choice parameters from traffic counts.

This work has been accepted for presentation at the 8$^{\text{th}}$ Triennial Symposium on Transportation Analysis (TRISTAN VII) and submitted to the 2$^{\text{nd}}$ Symposium of the European Association for Research in Transportation (hEART 2013).

Chapter 2

Approximating the stationary joint aggregate queue-length distribution of tandem Markovian networks

2.1 Introduction

Probabilistic models of urban traffic can provide a detailed description of traffic, and can allow to solve robust and reliable formulations of traditional transportation optimization problems. There have been numerous attempts to develop probabilistic link models (i.e., models of vehicular traffic along a homogenous road segment) based on queueing theory. For a recent review, see Osorio (2010). More recently, a stochastic formulation of the link-transmission model (Yperman, 2007) was proposed (Osorio *et al.*, 2011, Osorio and Flötteröd, 2012). The link transmission model is an operational instance of Newell's simplified theory of kinematic waves (Newell, 1993).

As analytical and differentiable models, probabilistic link models are of wide interest to address a variety of network design and traffic management problems. Recently, such analytical models have been used to enhance the computational efficiency

of simulation-based optimization algorithms that embed detailed yet computationally inefficient microscopic traffic simulators (Osorio and Chong, 2012a, Osorio and Nanduri, 2012, Chen *et al.*, 2012, Osorio and Bierlaire, 2010a).

Existing probabilistic traffic models approximate either first-order moments of the main link performance metrics or the marginal probability distributions of these link metrics. This chapter proposes an approach to efficiently approximate the joint queue-length distribution of the main performance metrics of adjacent links. The purpose is to account for between-link dependencies beyond first-order moments, yet to do so in a tractable manner such that these techniques can be used for optimization purposes. The use of such higher-order distributional information may lead to more accurate path and network-wide performance estimates, and ultimately enhance network design and traffic management strategies. The main challenges that arise in such an approach are the dimensionality of the state space and the complexity of modeling network-wide dependency both analytically and tractably.

This chapter focuses on an application of finite capacity queueing network (FCQN) theory. Each link of an urban network is modeled as one (or potentially multiple) finite capacity queue(s). The term capacity refers to the space capacity of the queue. Such models differ from traditional queueing models in that they assume there is an upper bound on the number of vehicles that can fit within a queue.

The analytical stationary analysis of FCQNs is complex for various reasons. Firstly, unlike Jackson networks or BCMP networks (Jackson, 1957, 1963, Baskett *et al.*, 1975), such models do not have product-form joint queue-length distributions, i.e. their joint distribution cannot be decomposed as a product of its marginals (note that a product-form joint distribution does not imply that the queues are independent). Secondly, finite capacity leads to potential spillbacks (referred to in queueing theory as blocking). That is, the queue of vehicles along a road may extend beyond the road to upstream roads. Analyzing the blocking phenomenon analytically is challenging, as illustrated in Osorio and Bierlaire (2009a). Blocking is not captured with infinite capacity queues, but is prevalent in congested networks.

Exact numerical evaluation of the stationary joint distribution of a Markovian

network with arbitrary size and topology can be obtained by solving the global balance equations (presented in Section 2.2.1). A detailed description of these numerical methods can be found in Stewart (2000). Nonetheless, the main challenge to such an approach remains the dimensionality of the joint distribution. For instance, for a network with m queues each with space capacity k (hereafter referred to as capacity), the dimension of the state space of the joint distribution is $(k+1)^m$. Thus, the dimension is exponential in the number of queues. For realistically-sized networks exact numerical techniques lack computational tractability.

Thus, most analytical analysis of FCQNs consist of approximation methods. The most popular approximate approach are decomposition methods. The latter reduces the dimensionality of the system under study by decomposing the network into smaller subnetworks. Each subnetwork is then modeled independently. Some dependency is captured by approximating the structural parameters of a subnetwork (e.g. arrival rates, service rates) as a function of the performance of other subnetworks (e.g., flow conservation equations). Most decomposition methods decompose the network into single queues and approximate the marginal distributions of each queue. Decomposition methods that consider subnetworks with two or three queues have also been proposed. For a review of decomposition methods see Osorio and Bierlaire (2009a).

This chapter proposes to address the dimensionality issue by describing the state of the network aggregately. We propose a decomposition approach that decomposes the network into overlapping subnetworks. We approximate the stationary joint aggregate queue-length distribution of a subnetwork.

The main challenge is deriving an aggregate distribution that is consistent with the underlying disaggregate distribution. Aggregation-disaggregation techniques for queueing networks have addressed this issue in the past. To the best of our knowledge, the first such approach is that of Takahashi (1975). It considers an arbitrarily high-dimensional Markov chain. It clusters the states of the chain into a set of aggregate states. It proposes an exact numerical technique to efficiently derive the joint aggregate distribution (i.e., the probabilities of all feasible combinations of aggregate states). An approximate aggregation-disaggregation method for a Markovian FCQN

network is proposed by Takahashi (1985). Numerical results are presented for a network of 5 queues in tandem. Both the marginals of each queue and the 2-dimensional joints of pairs of adjacent queues are approximated. The work of Takahashi (1985) was extended to consider another blocking mechanism in Song and Takahashi (1991). Schweitzer (1984) formulates the Takahashi (1975) approach for arbitrary topology and size Markovian FCQNs, and for both the stationary and transient distributions. Schweitzer (1991) presents a survey of aggregation-disaggregation techniques.

This chapter proposes a methodology that approximates the joint aggregate distribution of a Markovian FCQN (Section 2.2). The approach considers a stationary regime and combines ideas from the methods of Takahashi (1975, 1985) and Schweitzer (1984) along with ideas from other decomposition techniques for FCQNs (Osorio and Bierlaire, 2009a) and probabilistic road traffic models (Osorio and Bierlaire, 2009b). The proposed approach is validated versus simulation results (Section 2.3). It is then used to investigate the added value of accounting for full distributional dependency in the context of urban traffic signal control (Section 2.4). Section 2.5 presents the main conclusions and discusses ongoing and future work.

2.2 Model formulation

2.2.1 Aggregation-disaggregation framework

In order to address the dimensionality issues mentioned in the previous section, we use the aggregation technique described by Schweitzer (1984). This section presents its main ideas. The technique considers a continuous or discrete time Markov chain with a finite and large state space. The Markov chain is assumed aperiodic and communicative. Let Ω denote the state space with $\text{card}(\Omega) = M$. The probability of being in an individual state $i \in \Omega$ at steady state is denoted by π_i. The rate at which a transition from state i to $j, i \neq j$, $(i, j) \in \Omega^2$, can take place is given by q_{ij}. The

steady state probabilities satisfy:

$$\begin{cases} \pi_i \sum_{j \in \Omega \setminus i} q_{ij} = \sum_{j \in \Omega \setminus i} \pi_j q_{ji}, \quad \forall i \in \Omega & \text{(2.1a)} \\ \sum_{i \in \Omega} \pi_i = 1. & \text{(2.1b)} \end{cases}$$

The above system of equations is referred to as the global balance equations. For a detailed derivation, see for instance Chapter 4.5 in Larson and Odoni (1981).

The global balance equations can be rewritten in matrix format as:

$$\begin{cases} \pi Q = 0 & \text{(2.2a)} \\ \sum_{i \in \Omega} \pi_i = 1, & \text{(2.2b)} \end{cases}$$

where Q is known as the transition rate matrix, and is defined as:

$$Q_{ij} = \begin{cases} q_{ij}, & \text{if} \quad i \neq j, \\ -\sum_{j \in \Omega \setminus i} q_{ij} & \text{if} \quad i = j \end{cases} \tag{2.3}$$

For Markov chains with a large number of states, Schweitzer (1984) proposes to partition the M states into $\overline{M}$ aggregate disjoint states, such that $\overline{M} \ll M$. Let $\bar{\Omega}$ denote the set of aggregate states.

Let Ω_a denote the set of disaggregate states within aggregate state a. The probability of being in aggregate state a, $\bar{\pi}_a$, is defined as:

$$\bar{\pi}_a = \sum_{i \in \Omega_a} \pi_i. \tag{2.4}$$

The global balance equations for the aggregate distribution are given by:

$$\begin{cases} \bar{\pi}_a \sum_{b \in \bar{\Omega} \setminus a} \bar{q}_{ab} = \sum_{b \in \bar{\Omega} \setminus a} \bar{\pi}_b \bar{q}_{ba}, \quad \forall a \in \bar{\Omega} & \text{(2.5a)} \\ \sum_{a \in \bar{\Omega}} \bar{\pi}_a = 1. & \text{(2.5b)} \end{cases}$$

where, $\bar{q}_{ab}$ is the transition rate from aggregate state a to aggregate state b, and is

referred to as an aggregate transition rate.

Schweitzer (1984) relates the aggregate transition rates to the disaggregate transition rates made from all disaggregate states into aggregate state a through the following equation (which corresponds to Equation (2.9) of Schweitzer (1984)):

$$\bar{q}_{ab} = \frac{\sum_{j\in\Omega_a}\sum_{i\in\Omega_b}\pi_j q_{ji}}{\sum_{j\in\Omega_a}\pi_j}, \quad (b,a)\in\bar{\Omega}^2, b\neq a. \tag{2.6}$$

2.2.2 Aggregate description of a single queue

In this section, we consider a single M/M/1/k queue and derive expressions for its aggregate transition rates. These expressions are then used in Sections 2.2.3 and 2.2.4 to derive the aggregate model for a tandem queueing network. The state of a queue is described by the number of jobs (e.g., vehicles), N, in the queueing system. The state space is given by $\Omega = \{0, 1, .., k\}$, where $k \in \mathbb{Z}^+$ is the queue capacity. The corresponding state transition diagram is displayed in Figure 2-1. Each circle denotes a state. The arrows denote possible transitions between the states, with their corresponding rates. In this case, arrivals are determined by the arrival rate, $\lambda \geq 0$, and departures are determined by the service rate, $\mu > 0$.

We aggregate the $k+1$ states into the following three states: the queue is empty, the queue is full, the queue is neither empty nor full. The choice of these three states is based on insights from urban traffic intersection models, where between-link interactions (i.e. interactions of links that are connected via an intersection) are mainly determined based on whether a vehicle is ready to be sent from the upstream link to the downstream link (i.e. non-empty upstream queue) and whether there is space downstream to receive this vehicle (i.e. non-full downstream queue). There are now three aggregate states: state 0, state k, and the state defined by the dashed lines in Figure 2-1.

Figure 2-2 represents the state transition diagram of the aggregate queueing system. The states 0, 1 and 2 denote, respectively, the disaggregate states 0, $\{1, ..., k-1\}$ and k. As represented in Figure 2-2, the aggregate system is now fully described by a set of four rates: λ, μ, $\bar{\lambda}$, and $\bar{\mu}$, where $\bar{\mu}$ and $\bar{\lambda}$ describe the transition rates from

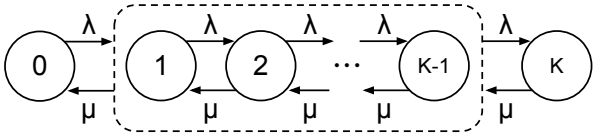

Figure 2-1: Single disaggregate queue.

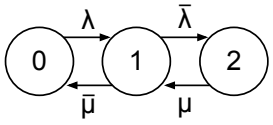

Figure 2-2: Single aggregate queue.

the new aggregate state to one of the other states (they are the aggregate transition rates).

The aggregate states are described by the random variable N_A:

- Empty queue: $N_A = 0, \quad \Omega_0 = \{N = 0\}$,
- Non-empty and non-full queue: $N_A = 1, \quad \Omega_1 = \{N \in [1, k-1]\}$,
- Full queue: $N_A = 2, \quad \Omega_2 = \{N = k\}$.

Hereafter the disaggregate (resp. aggregate) state probability π_i (resp. $\bar{\pi}_i$) is denoted $P(N = i)$ (resp. $P(N_A = i)$). The global balance equations satisfied by the aggregate state probabilities are:

$$\begin{cases} \lambda P(N_A = 0) = \bar{\mu} P(N_A = 1) & \text{(2.7a)} \\ \mu P(N_A = 2) = \bar{\lambda} P(N_A = 1) & \text{(2.7b)} \\ \sum_{i=0}^{2} P(N_A = i) = 1 & \text{(2.7c)} \end{cases}$$

Following Equation (2.6), the aggregate transition rates are given by:

$$\begin{cases} \bar{\lambda} = \dfrac{\sum_{j \in \Omega_1} \sum_{i \in \Omega_2} P(N = j) q_{ji}}{\sum_{j \in \Omega_1} P(N = j)} & \text{(2.8a)} \\ \bar{\mu} = \dfrac{\sum_{j \in \Omega_1} \sum_{i \in \Omega_0} P(N = j) q_{ji}}{\sum_{j \in \Omega_1} P(N = j)}. & \text{(2.8b)} \end{cases}$$

In an M/M/1/k queue we have (see Figure 2-1):

$$q_{jk} = \begin{cases} \lambda & \text{if } j = k-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

$$q_{j0} = \begin{cases} \mu & \text{if } j = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Inserting Equations (2.9) and (2.10) into (2.8), and noting that $\Omega_0 = \{0\}$ and $\Omega_2 = \{k\}$, we obtain:

$$\begin{cases} \bar{\lambda} = \dfrac{P(N = k-1)\lambda}{\sum_{j \in \Omega_1} P(N = j)} = \lambda \dfrac{P(N = k-1)}{P(N_A = 1)} & \text{(2.11a)} \\ \bar{\mu} = \dfrac{P(N = 1)\mu}{\sum_{j \in \Omega_1} P(N = j)} = \mu \dfrac{P(N = 1)}{P(N_A = 1)} & \text{(2.11b)} \end{cases}$$

The above equations can be rewritten as:

$$\begin{cases} \bar{\lambda} = \lambda\, P(N = k-1 \mid N_A = 1) & \text{(2.12a)} \\ \bar{\mu} = \mu\, P(N = 1 \mid N_A = 1). & \text{(2.12b)} \end{cases}$$

In summary, in order to determine the aggregate transition rates $\bar{\lambda}$ and $\bar{\mu}$, we need to evaluate the probabilities $P(N = k-1 \mid N_A = 1)$ and $P(N = 1 \mid N_A = 1)$. We call these probabilities disaggregation probabilities, since they represent the probabilities of being in a disaggregate state of a given aggregate state.

For a single M/M/1/k queue, there is a closed-form expression for the stationary queue-length distribution (Bocharov *et al.*, 2004):

$$P(N = n) = \frac{(1-\rho)\rho^n}{1-\rho^{k+1}} \quad \forall\, n \in [0, k], \tag{2.13}$$

where ρ is known as the traffic intensity and is defined as λ/μ. This expression holds for $\rho \neq 1$. We assume hereafter that $\rho \neq 1$. Thus, we can obtain an exact closed-form

expression for the disaggregation probabilities.

$$
\begin{aligned}
P(N=1 \mid N_A=1) = \frac{P(N=1)}{P(N_A=1)} &= \left(\frac{(1-\rho)\rho}{1-\rho^{k+1}}\right)\left(\sum_{j=1}^{k-1}\frac{(1-\rho)\rho^j}{1-\rho^{k+1}}\right)^{-1} \\
&= \rho\left(\sum_{j=1}^{k-1}\rho^j\right)^{-1} = \rho\left(\rho\frac{1-\rho^{k-1}}{1-\rho}\right)^{-1} \\
&= \frac{1-\rho}{1-\rho^{k-1}}. \qquad (2.14)
\end{aligned}
$$

We can proceed similary to obtain:

$$
P(N=k-1 \mid N_A=1) = \frac{(1-\rho)\rho^{k-2}}{1-\rho^{k-1}}. \qquad (2.15)
$$

2.2.3 Aggregate description of a three queue tandem network

We consider a network of three queues in a tandem (i.e. series) topology. This is the simplest topology in which a queue is affected by both upstream and downstream traffic conditions. Hereafter index i refers to a queue index.

Queue i ($i \in \{1,2,3\}$) has finite space capacity $k_i \in \mathbb{Z}^+$ and independent exponentially distributed service times with parameter μ_i. For each queue, external arrivals (i.e. arrivals that come from outside the network) follow a Poisson process with rate parameter γ_i. The joint aggregate probabilities are denoted by $P(N_A = s)$, where an aggregate state s is defined as the triplet: $s = (j_i, j_{i+1}, j_{i+2})$, and $j_i \in \{0,1,2\}$. For a three queue network with three aggregate states, the state space is aggregated into $3^3 = 27$ distinct states. The dimension of the state space is now independent of the individual queue capacities.

The main challenges in extending the approach from a single queue to network of queues arise as a result of the possibility of blocking. Blocking occurs when a job completes service yet cannot proceed because the downstream queue is full. This is known as spillback in urban traffic. We assume here blocking-after-service (Balsamo *et al.*, 2001), where a blocked job continues to occupy the underlying server until it is unblocked. Thus, a job that cannot proceed downstream due to lack of space is

blocked, while also blocking the use of the underlying server. A job becomes unblocked when its downstream destination queue can accommodate it.

Two main challenges that arise due to blocking are:

1. the rate of job departures at a blocked queue depends on both the state and the service rates of downstream queues,

2. a service completion at a blocking queue (i.e. a queue that is blocking jobs at upstream queues) triggers state changes at upstream blocked queues.

Methods to approximate stationary marginal queue-length distributions that explicitly describe blocking states have been proposed (e.g., Osorio and Bierlaire (2009a)). They illustrate the complexity of approximating the blocking probabilities, and the effective service rates (service rates that account for the occurrence of blocking). In this work, we propose simple approximations to account for both challenges, described respectively in Sections 2.2.3 and 2.2.3.

Aggregate transition rates

Within a three queue network, we assume that the aggregate transition rates for a given queue i are given by (2.12), and that the disaggregation probabilities have the functional form given by Equations (2.14) and (2.15). The disaggregation probabilities are each a function of the traffic intensity ρ of the underlying queue. We now describe how we approximate ρ.

The traffic intensity is defined as the ratio between an arrival rate and a service rate. In a finite capacity queuing network, the prevailing arrival and service rates of a given queue may be state-dependent. This is due to the occurrence of blocking. To illustrate this, consider a given queue i that is the most upstream queue of a three-queue tandem network. Consider a job that is occupying a server at queue i. It can either be:

1. undergoing service, which will be completed with rate μ_i,

2. blocked by its directly downstream queue which is itself not blocked, the job will therefore be unblocked with rate μ_{i+1},

3. blocked by its directly downstream queue which is itself blocked, the job will therefore be unblocked with rate μ_{i+2}.

Thus, depending on the state of that job (under service, blocked and if so blocked by which queue), the queue will have a different prevailing service rate.

In order to approximate ρ, we use a simple and exogenous approximation for the state-dependent service rates. For state $s = (1, j_{i+1}, j_{i+2})$, the state-dependent (prevailing) service rate, $\mu_{i,s}$, is given by:

$$\mu_{i,s} = \begin{cases} \mu_i & \text{if} \quad j_{i+1} < 2 \\ \mu_{i+1} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} < 2 \\ \mu_{i+2} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} = 2 \end{cases} \tag{2.16}$$

This approximation states that if queue i has one or several consecutive downstream queues that are full, then its prevailing service rate is that of the most downstream queue that is full. Recall that for queue i, the only aggregate state with more than one disaggregate state (i.e. the only state where disaggregation probabilities are needed) is aggregate state 1. This is why, when approximating the disaggregation probabilities of queue i, we need only consider states s with $j_i = 1$.

The arrival rate of queue i is obtained by solving the flow conservation equation (which is derived and further detailed in Osorio and Bierlaire (2009a)):

$$\lambda_i = \gamma_i + \frac{\lambda_{i-1} P(N_{i-1} < k_{i-1})}{P(N_i < k_i)}. \tag{2.17}$$

where:

$$P(N_i < k_i) = P(N_{A,i} < 2) \tag{2.18}$$

and N_i (resp. $N_{A,i}$) represents the disaggregate (resp. aggregate) number of jobs in queue i.

To summarize, the state-dependent approximation for the disaggregation probabilities (Equations (2.14) and (2.15)) are denoted $\alpha^{\mathrm{f}}_{i,s}$ and $\alpha^{\mathrm{e}}_{i,s}$, and are given by:

$$\begin{cases} \alpha^{\mathrm{f}}_{i,s} = P(N_i = k_i - 1 \mid N_{A,i} = 1, N_A = s) = \dfrac{(1 - \lambda_i/\mu_{i,s})}{1 - (\lambda_i/\mu_{i,s})^{k_i - 1}}, & \text{(2.19a)} \\ \alpha^{\mathrm{e}}_{i,s} = P(N_i = 1 \mid N_{A,i} = 1, N_A = s) = \dfrac{(1 - \lambda_i/\mu_{i,s})(\lambda_i/\mu_{i,s})^{k_i - 2}}{1 - (\lambda_i/\mu_{i,s})^{k_i - 1}}. & \text{(2.19b)} \end{cases}$$

where N_A is the state vector $(N_{A,i}, N_{A,i+1}, N_{A,i+2})$, and the superscripts e and f refer to empty and full, respectively (since these expressions are used to approximate the transition rates towards empty and full states, respectively).

Blocking probabilities

Recall that we describe the state of a queue as either empty, full or 'non-empty and non-full'. Given a job occupying a server, this state description does not distinguish between a job undergoing service or one that has completed service and is blocked. This section presents a simple approximation of the probability of a job being blocked, i.e., the blocking probability.

The following example allows us to introduce the notion of blocking probability and the complex between-queue dependencies that arise due to blocking. Consider a state $s = (1, 2, 2)$ where queue i (i.e., the most upstream queue) is in (aggregate) state 1, and queues $i+1$ and $i+2$ are in aggregate state 2, i.e. they are full. Assume there is a service completion at queue $i + 2$. This service completion can trigger a transition to one of the following states:

- if queue $i + 2$ is not blocking queue $i + 1$, then the new state is $(1, 2, 1)$;
- if queue $i + 2$ is blocking queue $i + 1$ and is not blocking queue i, then the new state is $(1, 1, 2)$;
- if queue $i+2$ is blocking queue $i+1$ and is blocking queue i, then the new state can be either $(1, 2, 2)$ or $(0, 2, 2)$.

In order to determine the new state to which a transition can take place, we use

state-dependent, yet simple and exogenous, approximations for the blocking probabilities. The approximation assumes that service completions follow an exponential distribution. We use the following property of exponential random variables. For a set of n independent service durations $\{X_\ell\}_{\ell=1:n}$, which are exponentially distributed random variables with rate parameter μ_ℓ, the probability that the first service completion is of type ℓ is given by (e.g., Larson and Odoni (1981), Chapter 2.12.4, Equation (2.62)):

$$P(X_\ell < X_i \quad \forall i \neq \ell) = \frac{\mu_\ell}{\sum_{j=1}^{n} \mu_j}. \tag{2.20}$$

We use this property to approximate the blocking probabilities. The states that are affected by the blocking probabilities are listed in Table 2.1. This table lists the queues that are blocked (column 1), the queue that is at the source of (i.e. causes) the blocking (column 2), the feasible joint states where such blocking can occur (column 3) and the corresponding probability with which this blocking occurs (column 4). We assume that no queue is initially blocked in any of the feasible joint states and approximate the probability that blocking occurs before the joint aggregate state transitions. For brevity, multiple states for the initial joint states are listed in braces. The approximations of the blocking probabilities are given by:

$$\beta_{i,1} = \frac{\mu_i}{\mu_i + \mu_{i+1}} \tag{2.21}$$

$$\beta_{i,2} = \frac{\mu_i}{\mu_i + \mu_{i+1} + \mu_{i+2}} \frac{\mu_{i+1}}{\mu_{i+1} + \mu_{i+2}} + \frac{\mu_{i+1}}{\mu_i + \mu_{i+1} + \mu_{i+2}} \frac{\mu_i}{\mu_i + \mu_{i+2}} \tag{2.22}$$

$$\beta_{i,3} = \frac{\mu_{i+1}}{\mu_{i+1} + \mu_{i+2}} \tag{2.23}$$

$$\beta_{i,4} = \frac{\mu_{i+1}}{\mu_i + \mu_{i+1} + \mu_{i+2}} \frac{\mu_{i+2}}{\mu_i + \mu_{i+2}} \tag{2.24}$$

Equations (2.21) and (2.23) (rows 1 and 3 of Table 2.1) are derived directly from Equation (2.20), and represent the probability that the first (resp. second) queue finishes service before the second (resp. third) queue. That is, blocking occurs due to the queue immediately downstream. Equation (2.22) (row 2 of Table 2.1) considers

Blocked queues	Source queue	Initial joint states	Blocking probability
i	$i+1$	$(\{1,2\},2,\{0,1\})$	$\beta_{i,1}$
$i,\ i+1$	$i+2$	$(\{1,2\},2,2)$	$\beta_{i,2}$
$i+1$	$i+2$	$(\{0,1,2\},1,2),\ (0,2,2)$	$\beta_{i,3}$
$i+1$	$i+2$	$(\{1,2\},2,2)$	$\beta_{i,4}$

Table 2.1: Blocking probabilities.

the scenario where both the first and the second queue are blocked by the third queue. This occurs when both the first and second queues complete their service before the third queue. The equation sums the probabilities of the independent events in which either the first or the second queue finish first out of the three and then the remaining queue of the first two finishes service before the third. Equation (2.24) (row 4 of Table 2.1) considers the case where the second queue is blocked by the third queue, but the first queue is not blocked. This is the probability that the second queue finishes service before the first and third queues and that the next queue to finish service is the third queue.

Let us summarize the procedure to derive the joint aggregate distribution of a three-queue tandem network. The joint distribution $P(N_A)$ is obtained by solving the corresponding (aggregate) global balance equations (2.5), which depend on the transition rate matrix, $\bar{Q}$. The latter is defined as:

$$\bar{Q} = f(\gamma, \mu, k, \alpha^{\mathrm{f}}, \alpha^{\mathrm{e}}, \beta), \tag{2.25}$$

where γ, μ, and k are exogenous parameters, α^{f} and α^{e} are the state-dependent disaggregation probabilities (defined by Equations (2.17)-(2.19)), and β denotes the blocking probabilities (Table 2.1). For a three queue system, this results in $27+12=39$ variables, representing the 27 joint states of the system as well as the 12 state-dependent disaggregation probabilities (following (2.16) there are 3 disaggregation probabilities for the first queue, 2 for second queue 3 for the third queue). The full transition rate matrix, $\bar{Q}$, of a 3-queue tandem network is given in Appendix A.

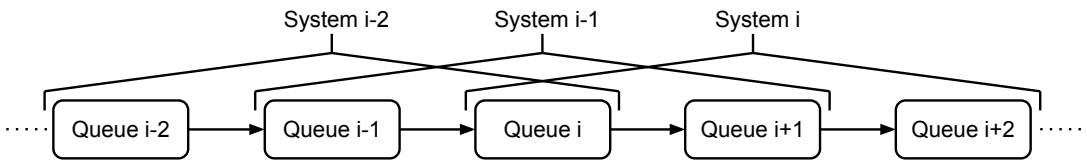

Figure 2-3: Overlapping systems of three tandem queues.

2.2.4 Aggregate description of a tandem network with I queues

This section generalizes the above approach to a tandem network with I queues. If the above approach is directly applied to an I-queue network, the state space would be of dimension 3^I, and would thus increase exponentially with the number of queues in the system. Instead, we view the network as a set of overlapping systems of three tandem queues. The set of I queues is decomposed into $I-2$ overlapping systems each with three queues in tandem. This decomposition is illustrated in Figure 2-3. Thus, the number of states is linearized with the number of queues.

Each of the $I-2$ systems can be viewed as a single 3-queue system. For each system, the approach of Section 2.2.3 is applied, i.e. the joint aggregate distribution of each system satisfies the system of equations described in Section 2.2.3.

Consider a 3-queue system within a larger network (e.g., system $i-1$ of Figure 2-3). In order to account for its dependencies with adjacent queues, we need to approximate the arrival rate to the most upstream queue, and the effective service rate of its most downstream queue. In the 3-queue network of Section 2.2.3, these two rates were exogenous, however, in a larger tandem network, these rates are now endogenous.

For a given system with queues indexed by $(i, i+1, i+2)$, the arrival rate to the most upstream queue (queue i) is given by the flow conservation equations (2.17). The service rate of the most downstream queue (queue $i+2$) is obtained by following the ideas in Osorio and Bierlaire (2009a), where the effective service rate of a queue (which accounts for service and for potential blocking from downstream queues) is

given by:

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + p_i^b \frac{1}{\tilde{\mu}_i}, \tag{2.26}$$

where $\hat{\mu}_i$ denotes the effective service rate, μ_i represents the exogenous service rate, p_i^b denotes the blocking probability, and $\tilde{\mu}_i$ is the unblocking rate.

The approximation for the unblocking rate is derived in Osorio and Bierlaire (2009a). Its simplified expression for a single server queue is detailed in Osorio and Bierlaire (2009b), and is given by:

$$\frac{1}{\tilde{\mu}_i} = \frac{\lambda_{i+1} P(N_{i+1} < k_{i+1})}{\lambda_i P(N_i < k_i)} \frac{1}{\hat{\mu}_i}, \tag{2.27}$$

which is equivalent to:

$$\frac{1}{\tilde{\mu}_i} = \frac{\lambda_{i+1} P(N_{A,i+1} < 2)}{\lambda_i P(N_{A,i} < 2)} \frac{1}{\hat{\mu}_i}. \tag{2.28}$$

The probability that queue i is blocked, p_i^b, is approximated by:

$$p_i^b = P(N_{A,i+1} = 2) \frac{\mu_i}{\mu_i + \mu_{i+1}}, \tag{2.29}$$

which considers the probability that the immediately downstream queue is full and the probability that the upstream queue serves before the downstream queue.

Thus, when analyzing a 3-queue system (with queues indexed by $(i, i+1, i+2)$) that is embedded within a larger network of queues, we apply the procedure of Section 2.2.3, with the arrival to queue i given by Equation (2.17) and the effective service rate of queue $i+2$, $\hat{\mu}_{i+2}$, given by Equations (2.26)-(2.29).

Since the systems overlap (see Figure 2-3), there are individual and pairs of queues that are common to multiple systems. We include the following system of linear constraints in order to ensure that the one or two-dimensional marginal distributions obtained are equivalent regardless of the system from which they are obtained.

For an I-queue network, there are $I-3$ pairs of queues that are contained in two

different systems. This leads to 3^2 equations (a pair of queues has 3^2 joint states) for each overlapping pair of queues:

$$\sum_{a=0}^{2} P_{i-1}(i-1=a, i=b, i+1=c) = \sum_{d=0}^{2} P_i(i=b, i+1=c, i+2=d),$$
$$(b,c) \in \{0,1,2\}^2, i \in \{2,\ldots,I-2\} \quad (2.30)$$

where P_i denotes the distribution obtained from analyzing the i^{th} 3-queue system, for $i \in [2, I-2]$. Enforcing consistency between systems with 2 overlapping queues also ensures consistency between systems with 1 overlapping queue.

For a network of I queues, the joint states and state-dependent disaggregation probabilities are represented by $(27+12)(I-2)$ variables and $2(I-2)$ variables represent the arrival and service rates for the first and (resp.) last queues of systems. Each of these variables has an associated equation, and ensuring consistency adds $9(I-3)$ equations without adding any variables. Thus, there are $41(I-2)+9(I-3)$ equations and $41(I-2)$ variables for a tandem network of I queues.

2.3 Validation

We compare the joint aggregate distributions obtained by the proposed model with those estimated from a discrete event simulation model of a Markovian FCQN. For all experiments in this section, 10,000 simulation replications are run, each with a duration of 1,000 time units in order to ensure stationarity. For each replication, the disaggregate states are evaluated at time 1,000, from which we derive the aggregate states.

Let p_s denote the probability of being in a given aggregate state s. A 95% confidence interval for p_s is given by: $\hat{p}_s \pm 1.96\sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{10{,}000-1}}$ (see, for instance, Section 7.3.3 of Rice (1994)), where $\hat{p}_s$ is the simulated estimate of p_s. This confidence interval is displayed as error bars in the figures of this section.

The aggregate joint stationary distribution is solved using the nonlinear systems of equations solver of Matlab ("Levenberg-Marquardt" algorithm of the *fsolve* function)

Scenario	1	2	3	4	5	6	7	8	9
μ_1	2	2	2	2	2	2	6	6	6
μ_2	2	2	2	4	4	4	4	4	4
μ_3	2	2	2	6	6	6	2	2	2
k_1	2	5	10	2	5	10	2	5	10
k_2	2	5	10	2	5	10	2	5	10
k_3	2	5	10	2	5	10	2	5	10

Table 2.2: Three-queue network scenarios.

(Mathworks, Inc., 2012), with a tolerance of 10^{-6}. Initial values are obtained by approximating the marginal distribution of each queue using the FCQN model given in Appendix C, and then approximating the joint distribution as the product of the marginals.

Three queue network

We assume that external arrivals only occur at the first queue, with $\gamma_1 = 1.8$. We consider nine scenarios with differing service rates and queue capacities, displayed in Table 2.2. In all scenarios, the minimum service rate is 2, implying an approximate traffic intensity of 0.9, a high level of congestion.

Each plot of Figure 2-4 displays the aggregate joint stationary distribution for each scenario. For scenarios 1, 4, and 7, with queue capacities of $(2, 2, 2)$, the aggregate state 1 maps directly onto one disaggregate state, and thus the expressions for the state-dependent disaggregation probabilities, α^{e} and α^{f}, are exact and equal to 1. In these cases, the errors between the proposed model and the simulation estimates are due to errors in blocking probabilities. There is an excellent match between the two distributions in scenarios 1, 4 and 7. The approximations in scenarios 2, 3, 5 and 6 are also excellent.

In scenarios 7−9, blocking is most likely to occur as a result of the third queue, which leads to the most complex blocking configurations. For these three scenarios, the accuracy of the proposed approximation decreases as the space capacity increases. This can be partly explained by the increasing difficulty to approximate the disag-

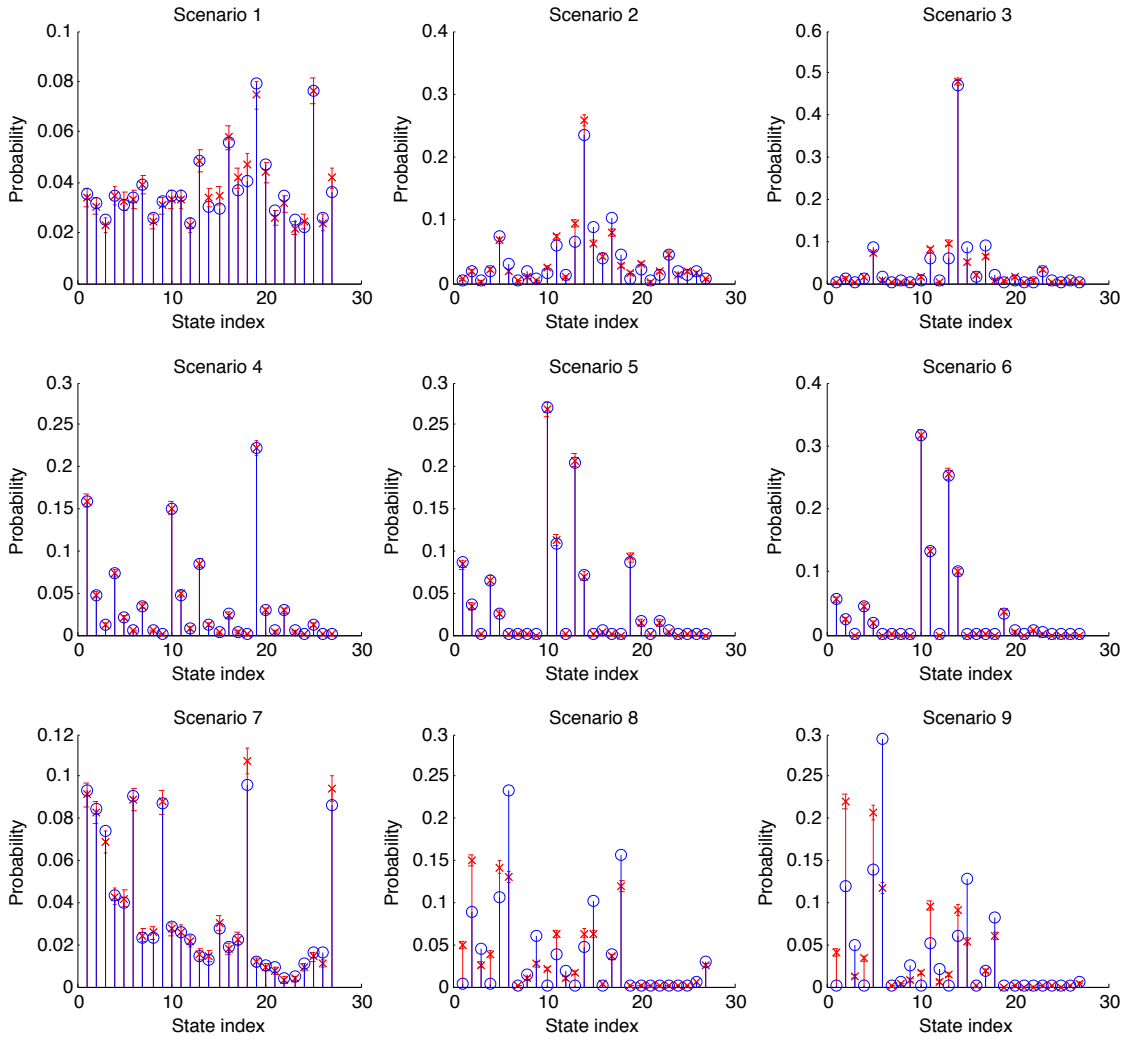

Figure 2-4: Stationary joint aggregate distribution for all scenarios. The blue circles represent the proposed model predictions, and the red crosses represent the simulation estimates with error bars for the probability of being in each of the 27 states.

Queue i	1	2	3	4	5
γ_i	3	0	3	3	0
μ_i	10	10	10	10	10
k_i	25	10	25	10	25

Table 2.3: Five-queue network scenario.

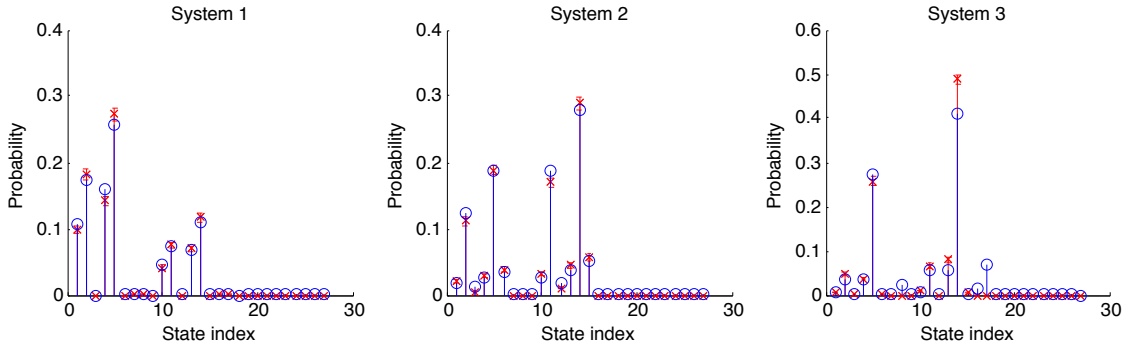

Figure 2-5: Stationary distributions for the three systems in a 5-queue network. The blue circles represent the proposed model predictions, and the red crosses represent the simulation results.

gregation probabilities as the space capacity increases. Additionally, scenarios 8 and 9 illustrate the complexity of accurately approximating both the intricate blocking effects between queues, as well as their interaction with the within-queue disaggregate states (i.e., the disaggregation probabilities). Thus, the resulting validation scenarios will consider this complex setting where both the space capacity is large, and blocking is triggered by queues downstream of the network.

Five queue network

We consider a 5-queue network with exogenous queue parameters defined in Table 2.3. This configuration also leads to a traffic intensity of approximately 0.9 for queues 4 and 5, which is a high level of congestion. Thus, blocking is likely to occur. For a 5-queue system, there are three 3-queues systems. The three corresponding joint distributions are displayed in Figure 2-5. The distributions obtained analytically approximate well those obtained through simulation.

Queue i	1	2	3	4	5	6	7	8
γ_i	4	0	1	1	0	2	0	1
μ_i	10	10	10	10	10	10	10	10
k_i	25	10	25	10	25	10	25	10

Table 2.4: Arrival rates, service rates, and queue capacities for the 8-queue scenario.

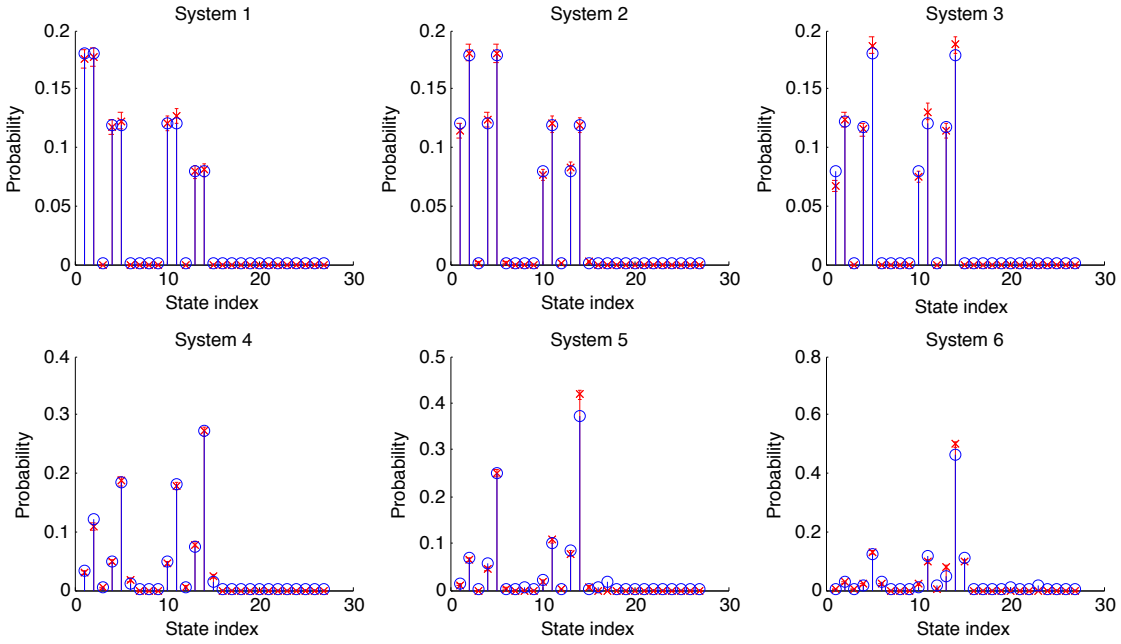

Figure 2-6: Stationary distributions for the six systems of an 8-queue network. The blue circles represent the model predictions, and the red crosses represent the simulation results.

Eight queue network

The 8-queue network considers a similar scenario as for the 5-queue network: external arrivals along the series of queues lead to downstream queues with high traffic intensity. The parameter values are given in Table 2.4. Figure 2-6 displays the joint stationary distributions of the six 3-queue tandem systems. The approximations provided by the analytical model accurately mimic those estimated via simulation. Thus, the proposed approach can capture the propagation of congestion (e.g., spillbacks) accurately.

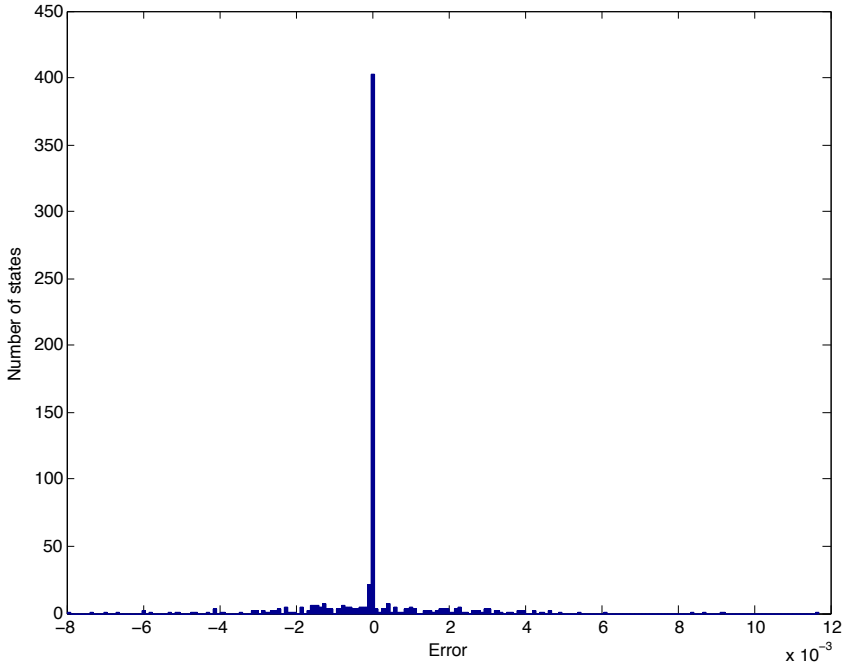

Figure 2-7: Histogram of the errors between the simulated and predicted state probabilities for each of the $23(27) = 621$ states.

Twenty-five queue network

The 25-queue network has a service rate of 10 for each queue and an alternating capacity of 25 and 10, as in the 8-queue scenario. External arrivals occur at queues 1, 11 and 21, with a rate of 2. A histogram of the error between the simulated and analytically approximated state probabilities is displayed in Figure 2-7. Again, the approximations are very accurate.

2.4 Case study

In this section, we use the proposed model to address a traditional urban traffic signal control problem. The purpose of this section is to investigate the added value of accounting for joint distributional information. We compare the performance of the proposed approach with that of an equivalent approach that yields only one-dimensional marginal (disaggregate) queue-length distributions. We call the latter approach the independent queueing model. It is derived in Osorio and Bierlaire

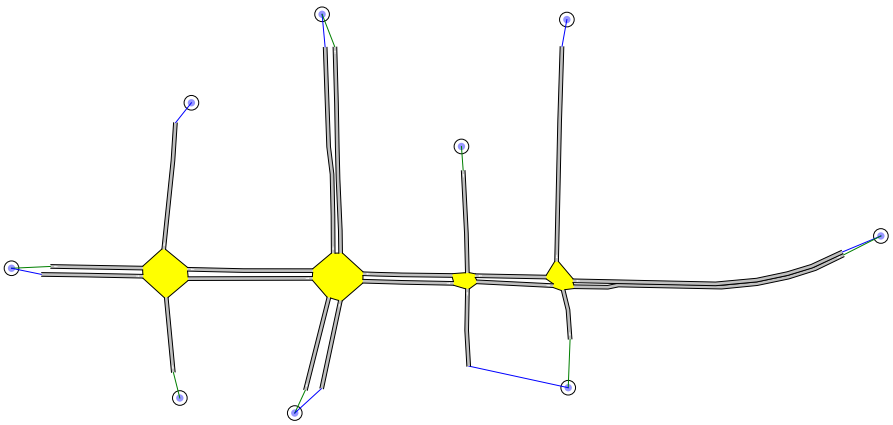

Figure 2-8: Network of single lane roads used for the case study.

(2009b) and described in Appendix C. In order to model an urban road network as an FCQN, we follow the approach in Osorio and Bierlaire (2009b). Each lane in the road network is modeled as a queue. The service rate of a queue corresponds to the flow capacity of the underlying lane.

2.4.1 Network

We consider the road network displayed in Figure 2-8. It considers 20 single-lane roads (i.e. 20 queues) and 4 intersections, each with 2 endogenous phases. All west-bound links of the main artery are modeled jointly, as are all east-bound links. The cross streets (north-bound and south-bound) are modeled independently (following the independent queueing formulation of Appendix C). Drivers travel along a single direction (i.e. they do not turn within the network). External arrivals and departures to the network occur at the boundaries of the network (represented by the circles in Figure 2-8).

We consider two different demand scenarios. For the medium demand scenario, the east-bound and west-bound demand (i.e. demand along the main arterial, noted as E−W and W−E) is 700 vehicles per hour in each direction. This increases to 900 vehicles per hour in the high demand scenario. Numbering each intersection from left to right, the north and southbound demand for the cross streets is listed in Table 2.5.

Demand	E−W	W−E	N−S (1)	N−S (2)	S−N (2)	S−N (3)	N−S (4)
Medium	700	700	100	600	600	100	100
High	900	900	100	600	600	200	200

Table 2.5: Demand in vehicles per hour for the medium and high demand scenarios.

The performance of the signal plans proposed by both the independent and joint formulations are evaluated by a microscopic traffic simulation model, implemented in AIMSUN, version 6.1 (TSS, 2011). Fifty simulation replications are run, each for one hour with a warm-up period of fifteen minutes. For each replication, we obtain a given simulated performance measure (e.g., average trip travel time), and compare the cumulative distribution functions (cdf's) obtained from the 50 observations.

2.4.2 Problem formulation

The signal control problem that we consider is formulated in detail in Osorio and Bierlaire (2009b) and presented briefly here. For a review of traffic signal control terminology and formulations, we refer the reader to Appendix A of Osorio (2010). We consider a fixed-time (also called time of day or pre-timed) control strategy. These are strategies that use historical traffic patterns to derive a fixed signal plan for a given time period. The signal control problem is solved offline. The signal plans of multiple intersections are determined jointly. The decision variables are the green splits (i.e., normalized green times) of phases of the different intersections. All other traditional control variables (e.g., cycle times, offsets, stage structure) are assumed fixed.

To formulate this problem we introduce the following notation:

b_i	available cycle ratio of intersection i;
$x(j)$	green split of phase j;
x_L	vector of minimal green splits;
$\mathcal{I}$	set of intersection indices;
$\mathcal{L}$	set of indices of the signalized lanes;
$\mathcal{P}_I(i)$	set of phase indices of intersection i;
$\mathcal{P}_L(\ell)$	set of phase indices of lane ℓ.

The problem is formulated as follows:

$$\min_x T(x, y; u) \tag{2.31}$$

subject to

$$\sum_{j \in \mathcal{P}_I(i)} x(j) = b_i, \ \forall i \in \mathcal{I} \tag{2.32}$$

$$\mu_\ell - \sum_{j \in \mathcal{P}_L(\ell)} x_j s = 0, \ \forall \ell \in \mathcal{L} \tag{2.33}$$

$$h(y; u) = 0 \tag{2.34}$$

$$y \geq 0 \tag{2.35}$$

$$x \geq x_L, \tag{2.36}$$

where the decision vector x consists of the green splits for each phase. Constraints (2.32) ensure that for a given intersection the available cycle time is distributed among all phases. Constraint (2.33) relates the service rate (i.e., link flow capacity) of a signalized queue to the saturation flow s (set to 1800 vehicles per hour) and to its green split x_j. Equation (2.34) represents the queuing model, i.e. the system of equations that is solved in order to yield the queue-length distributions, and the corresponding delays. The queuing model h depends on a vector of endogenous queuing variables y (e.g., disaggregation probabilities) and a set of exogenous parameters u (e.g., external arrival rates, space capacities). The endogenous queuing variables are subject to

positivity constraints (2.35). Green splits have lower bounds (Equation (2.36)), which are set to 4 seconds in this work (following the transportation norms VSS (1992)). The objective function $T(x, y; p)$ represents the expected trip travel time.

The expected time in the system is obtained by applying Little's law: (Little, 2011, 1961):

$$T(x, y; u) = \frac{\sum_i E[N_i]}{\sum_i \gamma_i P(N_i < k_i)}, \tag{2.37}$$

where $E[N_i]$ represents the expected number of vehicles in queue i and the summation considers all queues in the network (queues are indexed by i). The derivation of $E[N_i]$ is given in Appendix B.

2.4.3 Implementation notes

The initial green times are given by allocating the available green time at an intersection equally between its phases. The signal control problem is solved analytically with the independent queueing model. The independent queueing model uses the *fsolve* algorithm (Mathworks, Inc., 2012) to solve for an initial point by using the initial green times. The signal optimization problem for the independent queueing model allows the green times to vary, adding variables representing the green times and constraints representing the available green times and minimum green times. The problem is solved by the "active-set" algorithm within the *fmincon* function (Mathworks, Inc., 2012) with constraint and function tolerances of 10^{-6} and 10^{-3}, respectively. This yields an optimal set of green times that is used by this joint signal control formulation as a better initial point than the initial plan used by the independent model.

The green times output by the independent approach are used as initial green times for the joint model, and the initial values for the joint model are solved first before allowing the green times to vary.

When using the joint model to address the control problem, we also implement the expected number of vehicles in a queue, $E[N_i]$, and the probability that a queue is full as variables. The approximation of $E[N_i]$ is derived in Appendix B. The

probability that a queue is full is taken from summing the relevant states from the joint distributions. Thus, in an I-queue network we have $2I$ additional variables and constraints, resulting in $(41)(I-2)+2I$ variables and $(41)(I-2)+2I+9(I-3)$ constraints. This road network consists of two 5-queue systems modeled jointly, and the remaining 10 queues are modeled independently. Each 5-queue system leads to $(41)(5-2)+2(5)=133$ variables and $(41)(5-2)+2(5)+9(5-3)=151$ equations. Queues modeled independently each lead to 5 variables and 5 equations. The joint model therefore consists of 316 variables and 352 constraints. This initial problem is solved with the Sequential Quadratic Programming (SQP) algorithm within the *fmincon* function of Matlab, with a constraint tolerance of 10^{-6} and a function tolerance of 10^{-3}.

The signal control problem has the following additional endogenous variables and constraints: 8 phase variables (2 per intersection) with their corresponding lower bound constraints (Equation (2.36)), 4 green split allocation linear constraints (Equation (2.32)) (1 per intersection). Thus, the signal control system consists of 324 variables, 356 equality constraints, and 8 inequality constraints. The optimization problem is solved with the Sequential Quadratic Programming (SQP) algorithm within the fmincon function of Matlab, with a constraint tolerance of 10^{-6} and a function tolerance of 10^{-3}.

2.4.4 Results

Medium demand scenario

The cdf of the average (resp. total) trip travel times is presented in the left (resp. right) plot of Figure 2-9. The signal plan proposed by the joint model outperforms the signal plan proposed by the independent model.

We test the hypothesis that the average trip travel time derived from the joint model is equal to the time derived by the independent model by conducting a paired t-test. The sample mean of the average trip travel times for the 50 replications with the signal plan derived from the independent model is 1.0375 minutes with

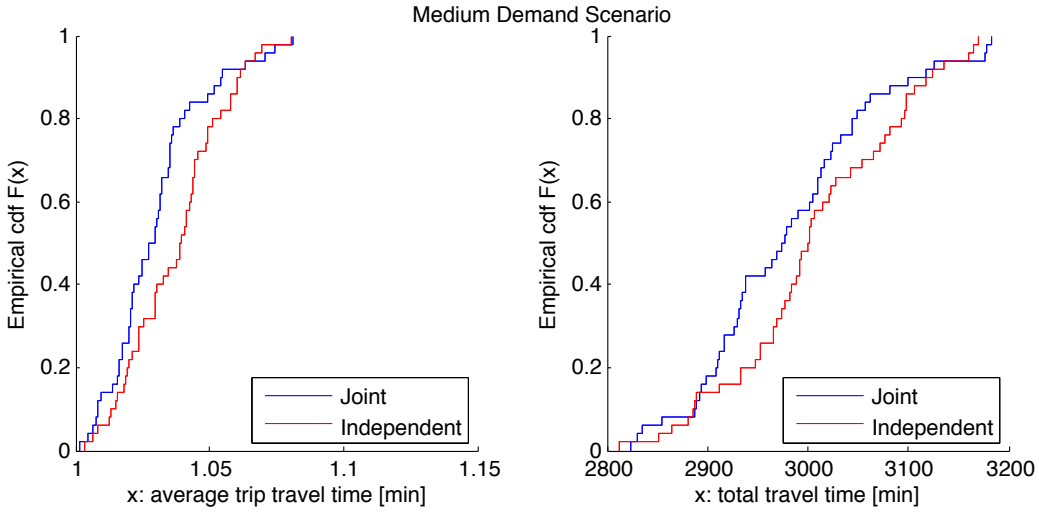

Figure 2-9: The left (resp. right) plot displays the cdf's of the average trip travel time (resp. total travel time) for the medium demand scenario.

a sample standard deviation of 0.0180. For the joint model, the resulting signal plan led to a sample mean of 1.0300 minutes and a sample standard deviation of 0.0179. Representing the sample mean by $\overline{Y}$, sample variance by s^2, and number of observations O, the test statistic is:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s_1^2/O + s_2^2/O}} \tag{2.38}$$

Thus, the test statistic for this experiment is 2.096 There are $2(O) - 2$ degrees of freedom, such that d.f. $= 98$ (Hogg and Tanis, 2006, p. 486). The null hypothesis is rejected with a significance level of 0.05, as the critical value, $t_{1-0.05,98} = 1.661$, is less than the absolute value of the test statistic for this experiment, 2.096.

High demand scenario

The high demand scenario increases the demand relative to the medium demand scenario along the main arterial and two of the cross streets. The cdf of the average (resp. total) trip travel time is presented in the left (resp. right) plot of Figure 2-10. The signal plan proposed by the joint model outperforms that proposed by the

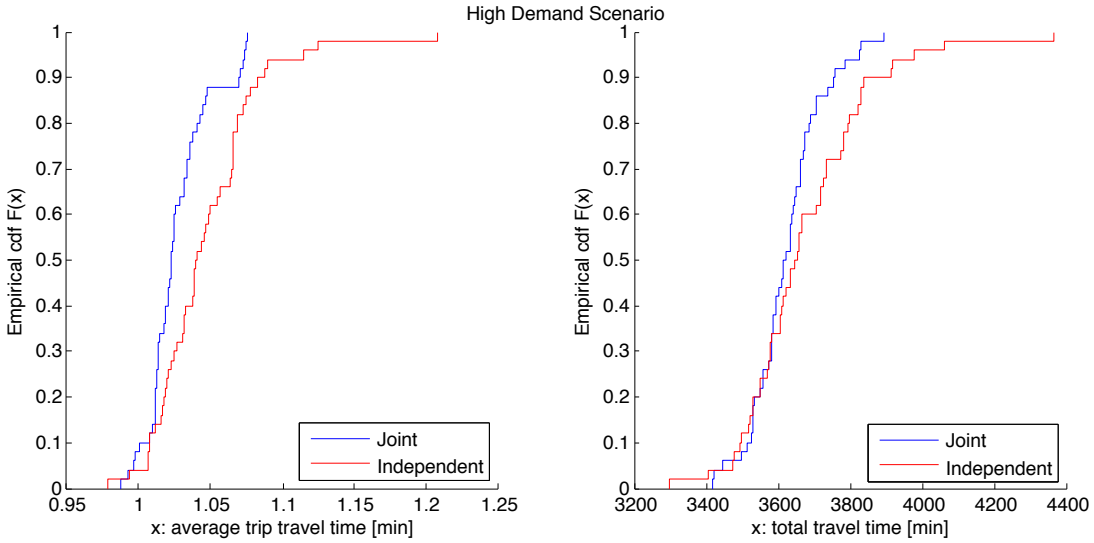

Figure 2-10: The left (resp. right) plot displays the cdf's of the average trip travel time (resp. total travel time) for the high demand scenario.

independent model. Additionally, it leads to signal plans with reduced variability in the average trip travel time.

The additional demand in the high demand scenario leads to an increase in the average trip travel times, as congestion increases and a greater number of vehicles take the longest route, the main arterial. Figure 2-11 presents the cdf's of the total number of trips for the high demand scenario. We test the hypothesis that the total number of trips completed derived from the joint model is the same as the number of trips completed in the independent model using a paired t-test (Equation (2.38)). The sample means for the joint and independent models are 3529 and 3510, respectively, with standard deviations of 49.2 and 59.5, respectively. The test statistic for this experiment is therefore 1.700. There are 98 degrees of freedom. The null hypothesis is rejected, as the test statistic is greater than the critical value, $t_{1-0.05,98} = 1.661$.

We test the hypothesis that the average trip travel derived from the joint model offers no improvement over the independent model by conducting a paired t-test (Equation (2.38)). The sample mean of the expected trip travel times for the 50 replications with the signal plan derived from the joint (resp.) independent model is 1.0804 (resp. 1.1973) minutes with a sample standard deviation of 0.0374 (resp.

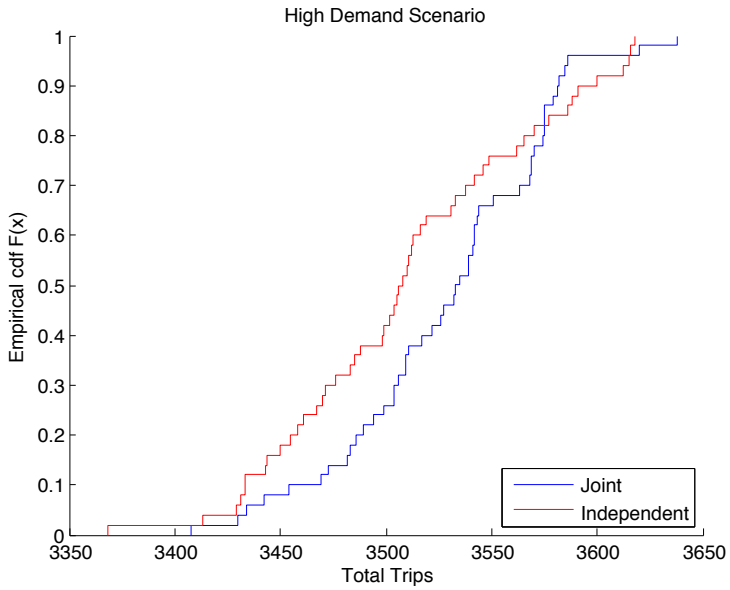

Figure 2-11: Cdf's of the total number of completed trips for the high demand scenario for each model.

0.0904). For 98 degrees of freedom, the null hypothesis is rejected, as the test statistic of 8.4539 is greater than the critical value, $t_{1-0.05,98} = 1.661$.

2.5 Conclusion

This chapter proposes an analytical approximation of the stationary joint aggregate queue-length distribution of a tandem Markovian network. The method combines ideas from decomposition methods, finite capacity queueing network models and aggregation-disaggregation techniques. The state space of the proposed distribution increases linearly with the number of queues, rather than exponentially, making it suitable for the analysis of large-scale networks. The analytical joint distribution is validated versus simulation estimates. This queueing method is then used to model a congested urban traffic network and to address a traditional signal control problem. The problem is solved with the proposed joint modeling approach and with an analytical model that only captures first-order between-queue dependency. The proposed model yields signal plans with significantly lower average trip travel times.

This case study illustrates the added value of using higher-order spatial dependency information for traffic control.

The extension of the proposed stationary model to account for arbitrary topology networks, as well as its use to enhance the computational efficiency of simulation-based optimization methods (Chen *et al.*, 2012, Osorio and Chong, 2012a, Osorio and Nanduri, 2012, Osorio and Bierlaire, 2010a) is of interest. Ongoing work focuses on the analytical approximation of the joint aggregate transient distribution.

Chapter 3

Efficient calibration techniques for stochastic traffic simulators

3.1 Introduction

Microscopic traffic simulators provide a detailed representation of urban transportation system dynamics. The greater resolution of microscopic models, as compared to macroscopic (flow-based) models, accounts for additional interactions and can address a variety of transportation problems (Osorio and Chong, 2012b, Osorio and Nanduri, 2012, Chen *et al.*, 2012, Nagel and Flötteröd, 2012). Due to their increased complexity, microscopic traffic simulators require more parameters to be calibrated, i.e. estimate the parameters used in the simulators from measured data. Relating available data to the model parameters in a computationally efficient manner is a largely unresolved challenge.

Most existing approaches to parameter calibration use black-box optimization routines, rather than exploiting the underlying structure of the problem by capturing the relationship between model parameters and the traffic data. Examples of computationally intensive strategies include the simultaneous perturbation stochastic approximation (SPSA; Spall (1992)), the Kalman Filter (Kalman, 1960), and derivative-free search techniques. See Balakrishna (2006) and Antoniou (2004) for examples, or Ben-Akiva *et al.* (2012) for a literature review. Recent contributions by

Flötteröd *et al.* (2011, 2012), Flötteröd *et al.* (2012) approach the calibration problem efficiently by analytically approximating the gradient of the measurement equation.

This chapter presents a methodology for the calibration of stochastic traffic simulation models from network flows. This approach distinguishes itself from existing literature by combining traffic simulation observations with a tractable analytical approximation of non-linear traffic flow dynamics, allowing for the application of gradient-based optimization routines to the calibration problem.

3.2 Methodology

3.2.1 Formal problem statement

Each origin-destination pair (OD) in an urban network is connected by a set of routes. For origin, O, and destination, D, the set of routes is denoted by R_{OD}. The total demand for an OD pair connected by a route r is given by $d_{OD(r)}$. The probability of that a vehicle selects route r is given by $P(r; \mathbf{x}, \theta)$, where $\mathbf{x}$ represents network attributes (in particular, travel times) that describe alternatives routes, and θ is a vector of parameters that governs the route choice selection. Denote the set of routes that contain a link segment i as R^i. Assuming no losses, the expected flows q on a link segment i for parameters θ are given by:

$$q_i(\theta) = \sum_{r \in R^i} d_{OD(r)} P(r; \mathbf{x}, \theta)) \tag{3.1}$$

Because the network travel times, $\mathbf{x}$, depend on the network flows, defined in Equation (3.1), the network travel times and flows depend on each other. These equations can be solved iteratively. In a traffic microsimulation, these iterations can be interpreted as a learning process, where users select routes based on the recent network conditions $\mathbf{x}$, which updates the future network conditions (route flows and travel times).

Denoting y_i as the given a traffic count on link segment i, a non-linear least square

formation of the calibration problem is given by:

$$\min_{\theta} \sum_{i} \left(y_i - q_i(\theta)\right)^2 \tag{3.2}$$

The calibration problem is difficult due to the complexity of the relationship between route flows and travel times, described Equation (3.1). This relationship is not available in closed form, but can be evaluated through microscopic traffic simulation. An efficient simulation-based optimization approach to the calibration problem can be formulated by embedding structural information that analytically approximates the relationship between route flows and travel times.

3.2.2 Simulation-based optimization formulation

Our simulation-based optimization (SO) approach combines information from the simulator with information from an analytical model that relates route flows and travel times, described in Section 3.2.4. This SO approach has been used to address large-scale urban traffic management problems with detailed, yet inefficient, microscopic traffic simulators (Osorio and Chong, 2012b, Osorio and Nanduri, 2012, Chen *et al.*, 2012). This formulation is based on the metamodel (or surrogate model) framework presented in Osorio and Bierlaire (2010b).

Each iteration of the SO algorithm uses a microscopic traffic simulator to determine equilibrium traffic flows for a given value of the behavioral parameter. For the calibration problem, estimates for the equilibrium flows are extracted from the simulator and compared with the given flows. These observations are then used to fit an analytical approximation of predicted flows. This approximation is referred to as a metamodel, and is used to derive a new trial point (i.e. value of the behavioral parameter) that minimizes the difference between the given flows and the predicted flows. The trial point is then evaluated by the simulator, and the process continues for a fixed computational budget.

Polynomials are often chosen for metamodels due to their analytic properties, yet may not provide a satisfactory global fit for the relationship between the behavioral

parameter and the predicted flows. By combining simulated information with the approximations derived from the analytical queueing model, physical characteristics of the network are captured and provide structure for the relationship between expected travel time and route choice. The incorporation of the analytical model to the metamodel is expected to improve the accuracy of the metamodel, leading to better trial points and faster convergence.

3.2.3 Metamodel formulation

For the calibration problem, the metamodel m_i fits the flow predicted by the analytical queueing model described in Section 3.2.4 on link segment i to the equilibrium flows determined by the microscopic simulator. The flows predicted on link segment i by the analytical queueing model for the behavioral parameter θ are given by the total arrival rate to the link segment, $\lambda_i(\theta)$ (Equation (C.1)). The equilibrium flows for a link segment i are evaluated by the simulator for a value of the behavioral parameter and are denoted by $\hat{f}_i(\theta)$. The metamodel used for each link segment i combines the flows predicted by the analytical queueing model with a linear term:

$$m_i(\theta; \alpha_i, \beta_i) = \alpha_i \lambda_i(\theta) + \beta_{i,0} + \beta_{i,1}\theta. \tag{3.3}$$

At each iteration c, the parameters of the metamodel, α^c and β^c, are fit by performing a least squares regression to the flow observations for each trial point, $\hat{f}(\theta_a)$, $a \in [1, c]$. The least squares problem is formulated for link segments $i \in \bar{L}$ as follows:

$$\min_{\alpha_i^c, \beta_i^c} \sum_{a=1}^{k} w_a \left(\hat{f}_i(\theta_a) - m_i^c(\theta_a; \alpha_i^c, \beta_i^c) \right)^2 + w_0 \left((\alpha_i^c - 1)^2 + (\beta_{i,0}^c)^2 + (\beta_{i,1}^c)^2 \right) \tag{3.4}$$

The first term in Equation (3.4) represents the weighted distance between the flows predicted by the metamodel and estimated by the simulator. As in Osorio and Bierlaire (2010a), the weights for each observation are proportional to the inverse of the distance from the current iterate. The weight parameters for each observation

$a \in [1, c]$, w_a, are given for current iterate θ_c by:

$$w_a = \frac{1}{1 + ||\theta_c - \theta_a||_2}. \tag{3.5}$$

The second term in Equation (3.4) ensures that the least square matrix is of full rank. The initial values that result from these terms, $\alpha = 1$ and $\beta = 0$, lead to an initial metamodel that is based solely on the queueing model.

3.2.4 Analytical model formulation

Urban traffic can be modeled as a finite capacity queueing network (FCQN). Each link in the urban network is modeled as one or more finite capacity queues. A link is therefore divided into a set of link segments, such that each segment is modeled by a single finite capacity queue with inter-arrival times and service times that follow negative exponential distributions (i.e. the M/M/1/k queues). This differs from traditional queueing models, as such models assume that there is no bound on the number of vehicles that can fit within a queue. The analytical queueing model formulated by Osorio and Bierlaire (2009b) and described by Equations (C.1) through (C.6) captures between-queue interactions through structural parameters, and assumes that turning probabilities are fixed. The variables used in the model for a given link segment (i.e. queue) i are listed in Table C.1.

The model proposed by Osorio and Bierlaire (2009b) and described in Section 3.2.4 does not use a behavioral parameter to relate route choice with expected route travel time. In order to calibrate this behavioral parameter, route choice must be endogenous.

Route choice

Denote, for a set of routes R_{OD} between an origin, O, and destination, D, the travel time of a route $r \in R_{OD}$ by t_r. For simplicity, it is assumed that route choice selection follows a multinomial logit model that varies across alternatives, yet is uniform for all users (see Ben-Akiva and Lerman (1985)); more general specifications such as

path-size logit or C-logit could be inserted into the framework as well (Ben-Akiva and Bierlaire, 1999, Cascetta *et al.*, 1996). Thus, the probability of taking a route r is given by:

$$P(r;\theta) = \frac{e^{\theta t_r}}{\sum_{s \in R_{OD}} e^{\theta t_s}}, \tag{3.6}$$

for a negative scale parameter, θ, representing the sensitivity that users have to differences in route travel times. We assume that users select from a fixed set of routes between any pair of origin and destinations and that the demand for a given OD pair is exogenous and fixed, independent of congestion levels. This route choice model will in the following be inserted into the queueing model of Osorio and Bierlaire (2009b), resulting in an endogenous representation of turning probabilities as a function of congestion-sensitive route choice probabilities.

Link segment travel times

Denoting the set of link segments that comprise route r as Ω_r and the expected travel time on a link segment i as $\tilde{t}_i$, the expected travel time of a route is described by the sum of the expected travel time on each link segment within the route:

$$t_r = \sum_{i \in \Omega_r} \tilde{t}_i. \tag{3.7}$$

Expected link segment travel times are approximated by summing the free flow travel time with the expected delay. The free flow travel time is exogenous, given by length of the link segment, ℓ_i, divided by its speed limit, s_i. The expected delay in link segment i is denoted by $E[W_i]$, and thus the travel time along a link segment, $\tilde{t}_i$, is given by:

$$\tilde{t}_i = E[W_i] + \ell_i / s_i. \tag{3.8}$$

The expected delay is estimated by using the relationship between expected delay and queue-length described by Little's Law (Little, 1961) applied to each link segment

independently. The delay is a function of the total arrival rate to the queue λ_i and the expected queue-length, denoted by $E[N_i]$:

$$E[W_i] = \frac{E[N_i]}{\lambda_i}. \tag{3.9}$$

In order to calculate the delay, the expected number of vehicles in the queue must be determined. Using N as a random variable representing the number of vehicles in a single M/M/1/k queue, there is a closed-form expression for the stationary queue-length distribution (see Equation 4.54 in Chapter 4.6.3 of Larson and Odoni (1981)):

$$P(N = n) = \frac{(1-\rho)\rho^n}{1-\rho^{k+1}} \quad \forall\, n \in \{0, \ldots, k\}. \tag{3.10}$$

For $\rho_i = \lambda_i/\hat{\mu}_i$, the expected number of vehicles in queue i is given by:

$$\begin{aligned}
E[N_i] &= \sum_{n=0}^{k_i} n\frac{(1-\rho_i)\rho_i^n}{1-\rho_i^{k_i+1}} = \frac{(1-\rho_i)\rho_i}{1-\rho_i^{k_i+1}} \sum_{n=0}^{k_i} n\rho_i^{n-1} \\
&= \frac{(1-\rho_i)\rho_i}{1-\rho_i^{k_i+1}} \sum_{n=0}^{k_i} \frac{d}{d\rho_i}\rho_i^n = \frac{(1-\rho_i)\rho_i}{1-\rho_i^{k_i+1}} \frac{d}{d\rho_i}\left(\sum_{n=0}^{k_i} \rho_i^n\right) \\
&= \frac{(1-\rho_i)\rho_i}{1-\rho_i^{k_i+1}} \frac{d}{d\rho_i}\left(\frac{1-\rho_i^{k_i+1}}{1-\rho_i}\right) \\
&= \frac{(1-\rho_i)\rho_i}{1-\rho_i^{k_i+1}} \left(\frac{1-\rho_i^{k_i+1}}{(1-\rho_i)^2} - \frac{(k_i+1)\rho_i^{k_i}}{1-\rho_i}\right) \\
&= \frac{\rho_i}{1-\rho_i} - \frac{(k_i+1)\rho_i^{k_i+1}}{1-\rho_i^{k_i+1}}.
\end{aligned} \tag{3.11}$$

Thus, the link segment travel time is given by:

$$\tilde{t}_i = \frac{1}{\lambda_i}\left(\frac{\rho_i}{1-\rho_i} - \frac{(k_i+1)\rho_i^{k_i+1}}{1-\rho_i^{k_i+1}}\right) + \ell_i/s_i. \tag{3.12}$$

The link segment travel time is added as a variable to the queueing model designed by Osorio and Bierlaire (2009b), described in Section 3.2.4, increasing the number of variables per link segment from five (Equations (C.1) to (C.5)) to six.

External arrival rate

Routes for the same origin-destination pair may begin at multiple different links, as an origin may connect to many road segments. The set of routes beginning with queue i is denoted by $\{R(0) = i\}$. Denoting the demand for the OD pair connected by a route r as $d_{OD(r)}$, the external arrival rate to queue i is given by:

$$\gamma_i = \sum_{r \in \{R(0)=i\}} P(r;\theta) d_{OD(r)}. \tag{3.13}$$

Vehicles arriving to a full link are lost and do not enter the system. This external arrival rate includes vehicles that are lost.

Turning probabilities

In the queueing model presented in Appendix C, the turning probabilities from queue to queue are exogenous. As route choice becomes endogenous, so do the turning probabilities. By denoting the set of routes containing upstream link segment i and downstream link segment j as R^{ij}, and the routes containing link segment i as R^i, the turning probabilities are given by:

$$p_{ij} = \frac{\sum_{r \in R^{ij}} P(r;\theta) d_{OD(r)}}{\sum_{s \in R^i} P(s;\theta) d_{OD(s)}}. \tag{3.14}$$

The denominator is the total demand for routes that contain the link segment i, whereas the numerator is the portion of the total demand for routes that contain a turn from link segment i to j. This ratio gives the turning probabilities, assuming that all vehicles enter the system. This expression is used by the analytical queueing model as the probability that a vehicle in link segment i turns into link segment j.

3.3 Simulation-based optimization algorithm

The algorithm used to solve the calibration optimization problem is a trust region method adapted from Osorio and Bierlaire (2010a). For iteration c of the algorithm,

define:

- θ_c as the current iterate, the value of the behavioral parameter
- m_i^c as the metamodel for link i
- Δ_c as the current trust region radius
- α_i^c, $\beta_{i,0}^c$, and $\beta_{i,1}^c$ as the metamodel parameters for link segment i
- ν_c as a vector containing the metamodel parameters for each link segment i, α_i^c, $\beta_{i,0}^c$, and $\beta_{i,1}^c$ for $i \in L$
- n_c as the number of simulated points

0. Initialization

The algorithm is initialized with a given value of the behavioral parameter, θ_0. The analytical model is solved for the initial point and the simulated flows, $\hat{f}(\theta_0)$, are evaluated, such that the initial metamodel parameters, α and β, can be fit.

The initial value for the trust region radius, Δ_0 is selected in the range $(0, \Delta_{max}]$. The trust region radius limits the step size to the local region around the current value of the iterate.

The parameters that remain fixed for the entirety of the algorithm include:

- Δ_{max}, the upper bound on the trust region radius
- Δ_{min}, the lower bound on the trust region radius
- θ_{max}, the upper bound on the value of the behavioral parameter
- θ_{min}, the lower bound on the value of the behavioral parameter
- $\eta \in (0, 1)$, the threshold for accepting new trial points
- $\bar{\tau} \in (0, 1)$, the threshold for sampling new points
- γ_{inc} and γ_{dec}, the factors for increasing or decreasing the trust region radius
- $\bar{u}$, the threshold on the number of consecutive rejected trial points required to reduce the trust region radius
- n_{max}, the maximum number of simulation runs, i.e. computational budget

1. Step calculation

At each iteration c, a new trial point is determined by minimizing the sum of the squared distance between the flows predicted by the metamodel and the observed flows:

$$\min_{\theta} \sum_{i \in \bar{L}} (m_i^c(\theta) - y_i)^2 \tag{3.15}$$

subject to:

$$h(\theta) = 0 \tag{3.16}$$

$$||\theta - \theta_c||_2 \leq \Delta_c \tag{3.17}$$

$$\theta_{min} \leq \theta \leq \theta_{max}. \tag{3.18}$$

Equation (3.16) represents the modified queueing model, consisting of the system of nonlinear equations described by Equations (C.1) to (C.5) and (3.12). Constraint (3.17) is the trust region constraint, which bounds the step size that behavioral parameter θ can take from the current value of the behavioral parameter θ_c. Constraint (3.18) places an upper and lower bound on the value of the behavioral parameter. The solution to this constrained nonlinear optimization problem is the new trial point, denoted by θ^*.

2. Trial point test

The simulated flows for the trial point, $\hat{f}(\theta^*)$, are evaluated. Compute:

$$\sigma_c = \frac{\sum_{i \in \bar{L}} (\hat{f}_i(\theta_c) - y_i)^2 - \sum_{i \in \bar{L}} (\hat{f}_i(\theta^*) - y_i)^2}{\sum_{i \in \bar{L}} (m_i^c(\theta_c) - y_i)^2 - \sum_{i \in \bar{L}} (m_i^c(\theta^*) - y_i)^2}. \tag{3.19}$$

- If $\sigma_c \geq \eta$, then accept the trial point, $\theta_{c+1} = \theta^*$, and set $u_c = 0$
- Otherwise, reject the trial point, $\theta_{c+1} = \theta_c$, and and set $u_{c+1} = u_c + 1$

Regardless of whether the trial point is accepted or rejected, the simulated flow observations for the trial point are included in the set of simulation flow observations used to fit the parameters of the metamodel for all subsequent iterations, $n_c = n_c + 1$

3. Model improvement

The metamodel parameters are determined by solving the least squares problem defined by Equation (3.3)-(3.5). This yields a new set of metamodel parameters, denoted ν_{c+1}.

If the relative change in the model parameters as given by the Euclidean norm is less than the threshold:

$$\frac{||\nu_{c+1} - \nu_c||_2}{||\nu_c||_2} < \bar{\tau}, \tag{3.20}$$

a new point, θ^+, is sampled uniformly from the feasible region of the behavioral parameter, given by $[\theta_{min}, \theta_{max}]$. The simulated and predicted flows, $\hat{f}(\theta^+)$ and $\lambda(\theta^+)$, are derived for the sampled point, and the point is included in the set of simulated flow observations used to fit the parameters of the metamodel, $n_c = n_c + 1$. The parameters for the metamodel are then updated to reflect the new set of observations by solving the least squares problem defined by Equation (3.3)-(3.5).

4. Updating the trust region radius

If $\sigma_c > \eta$, then the trust region radius is increased, such that:

$$\Delta_{c+1} = \min\{\gamma_{inc}\Delta_c, \Delta_{max}\}. \tag{3.21}$$

If $\rho_c \leq \eta$ and the number of consecutive rejected points exceeds the threshold, i.e. $u_c > \bar{u}$, then the trust region radius is decreased, such that:

$$\Delta_{c+1} = \max\{\gamma_{dec}\Delta_c, \Delta_{min}\}. \tag{3.22}$$

In this case, u_c is reset to 0, such that the trust region radius can only be decreased every $(\bar{u}+1)$ iterations.

After the trust region radius is updated, the algorithm increases by a step, from c to $c+1$, and returns to the step calculation. It terminates when the number of simulation runs reaches the exceeds the maximum budgeted, $n_c > n_{max}$.

3.4 Implementation

3.4.1 Test network description

Consider the network in Figure 3-1. Each link in the network consists of a single lane and is modeled by a single queue. Table 3.1 details the properties of the links within the network.

The network has one origin-destination pair, nodes 1 and 5 respectively, with a demand of 1400 vehicles per hour. There are two routes that connect the pair, a route to the north (through nodes 1, 2, 3, 5) and a route to the south (through nodes 1, 2, 4, 5). The northern route has a shorter free flow travel time, yet faces delay at a signalized intersection at node 3, whereas the southern route is entirely unsignalized and free of bottlenecks. The signal at node 3 is green for 30 seconds out of the 90 second cycle time. The free flow travel time on the northern route is 18 minutes, as compared to the free flow travel time on the southern route, which is 20.14 minutes.

The queueing model parameters that represent rates (γ, λ, and μ) are given in terms of vehicles per lane per hour. Thus, for an unsignalized link i, the service rate, μ_i, is given by the saturation rate of 1800 vehicles per hour. The service rate of a signalized link is given by the green split multiplied by the saturation rate. This leads to a service rate of $\frac{30}{90}1800 = 600$ vehicles per hour for link 2. Lastly, the external arrival rate to link 1, described by Equation (3.13), is 1400 vehicles per hour, as both routes in the network begin on link 1[1].

The northern route is faster and hence preferred in the absence of congestion.

[1]For numerical reasons, the rates to the queueing model are scaled by the saturation flow, such that the service rate for an unsignalized link is 1.

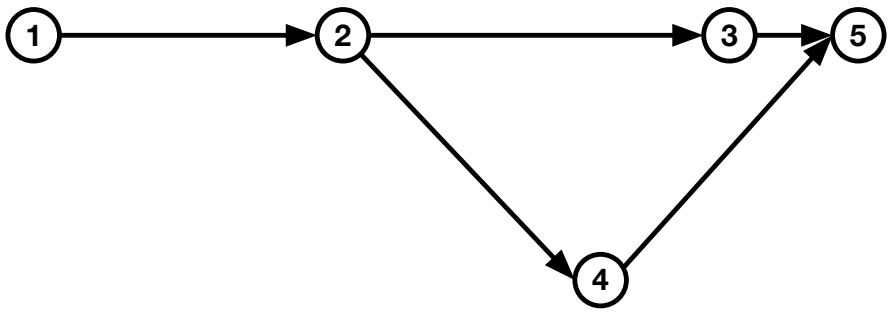

Figure 3-1: Test network diagram.

As more vehicles take this route, it becomes congested, its delay increases, and the southern route becomes increasingly attractive. The stochastic user equilibrium in the simulator produces mutually consistent route flows and travel times for a given route choice model.

Link	Nodes connected	Length [m]	Speed limit [km/hr]	Signalized
1	$1 \rightarrow 2$	5000	50	unsignalized
2	$2 \rightarrow 3$	7500	50	signalized
3	$2 \rightarrow 4$	7071	60	unsignalized
4	$3 \rightarrow 5$	7071	60	unsignalized
5	$4 \rightarrow 5$	2500	50	unsignalized

Table 3.1: Link properties of the test network.

3.4.2 Traffic simulation model

The network is implemented in AIMSUN, version 6.1 (TSS, 2011). Of interest is determining the stochastic user equilibrium that corresponds to the aforementioned consistency between route flows and route travel times for a given route choice model. We assume a multinomial logit model and evaluate the equilibrium based on a given value of the behavioral parameter. The computation of this equilibrium is detailed in the following paragraphs.

Initial route choice

For a given value of θ, the initial values for the route travel times, t_r^0 are based on the relationship between expected link segment travel time $\tilde{t}_i(\theta)$, and route travel time, described by Equations (3.7) and (3.12):

$$t_r^0 = \sum_{i \in \Omega_r} \tilde{t}_i(\theta) \tag{3.23}$$

Route choice probabilities are given by the multinomial logit model for a given θ (Equation (3.6)). These probabilities are determined before the traffic simulation is run and fixed for the entirety of a given simulation. A given run of the simulator contains a 90 minute warm-up period before collecting the route travel times and simulated traffic counts for 3 hours. The simulator is run with the initial route choice, with $\hat{t}_r^0$ representing the average travel time on route r and $\hat{f}_i^0(\theta)$ representing the average flow per hour on link segment i for the 3 hour simulation.

Stochasticity in the simulator lead to significant fluctuations in simulated flows and travel times for the same route choice. Filtering can reduce the oscillations in the system in order to better estimate the stochastic user equilibrium. The filtered travel times serve as estimators of the expected travel times, leading to less variable route choice. The initial use of a fixed filtering weight results in relatively fast convergence to a near-equilibrium state, which then is refined by the method of successive averages. The description of the filtering used is described in the following paragraphs.

Filtering travel times

Multiple simulation runs are used to filter the values for the route travel times and flows. The value of the travel time on a route r at simulation run v is denoted by t_r^v. The average travel time on a route r for the simulation at simulation run v is given by $\hat{t}_r^v$, and the average flow per hour on a link segment i for the simulation at simulation run v is given by $\hat{f}_i^v(\theta)$. The initial values for the travel time at $v = 0$ are

given in Equation (3.23). For $0 < v < 20$ and $w = 1/10$:

$$t_r^{v+1} = w\hat{t}_r^v + (1-w)t_r^v. \tag{3.24}$$

The route travel time estimates switch from averaging simulated travel times, $\hat{t}_r^v$, with a fixed weight w to averaging the travel times with a weight $w(v)$ that decreases as v increases. This relationship is traditionally given by:

$$w(v) = \frac{1}{v+1}. \tag{3.25}$$

However, when switching to a method of successive averages, the initial weight after the switch is set to equal the fixed weight. Switching at $c = 20$ implies that:

$$w(20) = \frac{1}{20+m+1} \tag{3.26}$$

for some value of m. For $w = 1/10$, $m = -11$. Thus for $v > 19$:

$$t_r^{v+1} = \frac{1}{v-10}\hat{t}_r^v + \frac{v-11}{v-10}t_r^v. \tag{3.27}$$

The algorithm terminates at $v = 60$, once 60 simulations have been run. The stochastic user equilibrium route flows for a given value of θ are taken to be the average of the last 20 route flows output by the simulator:

$$\hat{f}_i(\theta) = \sum_{v=41}^{60} \frac{1}{20}\hat{f}_i^v(\theta). \tag{3.28}$$

3.4.3 Algorithm details

The parameters that remain fixed for the entirety of the algorithm include:

- $\Delta_{max} = 10^{10}$
- $\Delta_{min} = 10^{-2}$
- $\Delta_0 = 10^{-1}$
- $\eta = 10^{-3}$

- $\bar{\tau} = 0.1$
- $\gamma_{inc} = 1.2$
- $\gamma_{dec} = 0.9$
- $\bar{u} = 10$
- $n_{max} = 50$
- $w_0 = 10^{-3}$

Fitting the metamodel

The *lsqlin* routine in Matlab (MATLAB, 2012) was used to solve the least squares problem for fitting the parameters of the metamodel.

Step calculation: trust region subproblem

New trial points are determined by minimizing the sum of the squared difference between the flows predicted by the metamodel and the observed flows. This objective function for this problem is described by Equations (3.15), and the constraints for this problem are the equations used by the modified queueing model (Equations (C.1) to (C.5) and (3.12)), as well as the trust region inequalities (Equation (3.17)), and the bounds on the behavioral parameter (Equation (3.18)). This nonlinear optimization problem is solved in Matlab using the "active-set" algorithm within the *fmincon* routine (Mathworks, Inc., 2012), with function and constraint tolerances set at the default of 10^{-6}.

3.4.4 Test network results

Given flows for the a value of the behavioral parameter, -0.2, were calculated by averaging 10 simulated equilibrium flows evaluated using the methodology in Section 3.4.2. The simulation-based optimization algorithm detailed in Section 3.3 was then used to calibrate the behavioral parameter, given an initial value of -0.4.

The results given by this full metamodel formulation were compared with results given by a metamodel implemented with the polynomial term only, i.e. $\alpha_i = 0, \forall i \in \bar{L}$

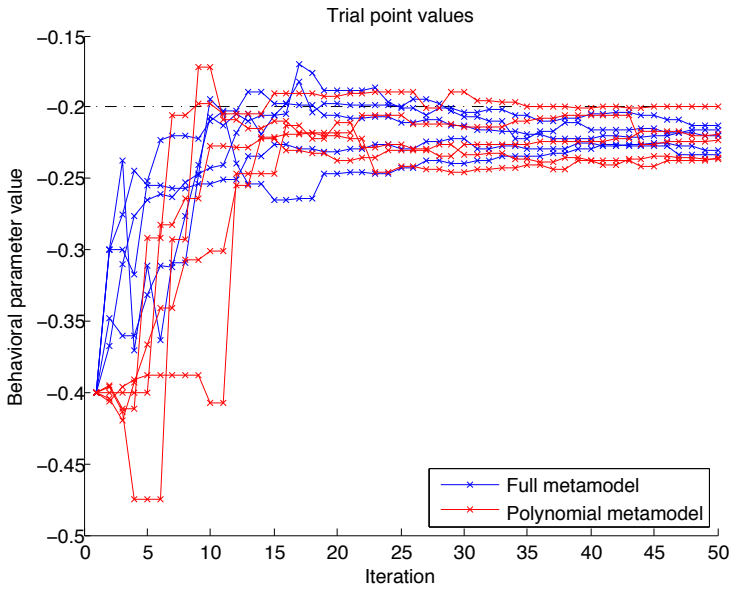

Figure 3-2: The figure plots the value of the trial points for each iteration of the SO algorithm for each metamodel with the initial point of -0.4.

in Equation (3.3). Each metamodel variation was run through the SO algorithm five times. The trial points for each iteration of the algorithm are shown in Figure 3-2. The mean values of the trial points at each iteration are displayed in Figure 3-3. For the simple topology of the network, the metamodel that incorporates the analytical queueing model converges faster than the linear metamodel.

3.5 Conclusion

This chapter proposes an analytical finite capacity queueing model that accounts for the relationship between route choice and traffic flow. The model is adapted from Osorio and Bierlaire (2009b) and used in the metamodel within a simulation-based optimization framework to calibrate behavioral parameters from measured flows. A sample network is considered and calibrated based on simulated flows. The added value of using the analytical model in the metamodel is determined by comparing the convergence of the SO algorithm with a polynomial metamodel. Ongoing work

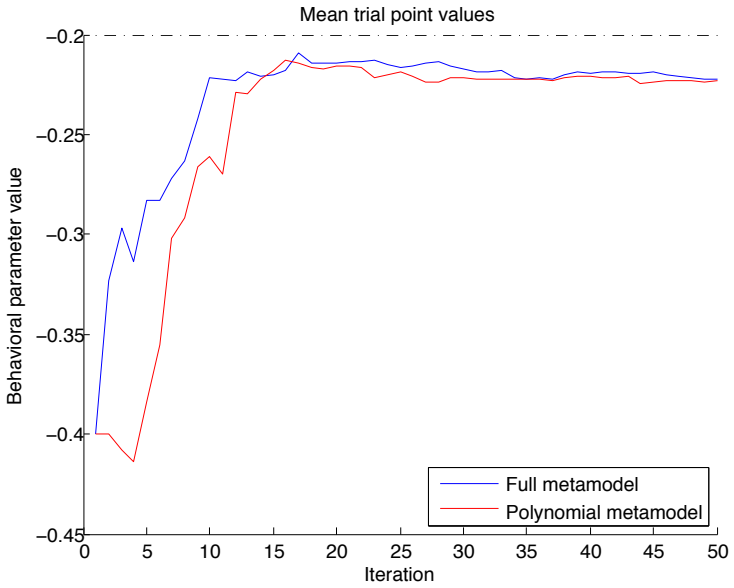

Figure 3-3: The figure plots the mean values of the trial points for each iteration of the SO algorithm for each metamodel with the initial point of -0.4.

focuses on increasing the tractability of the analytical queueing model for large-scale networks.

Chapter 4

Conclusion

Chapter 2 describes the formulation of an analytical queueing model that approximates the stationary joint aggregate distribution of a tandem network. The model was validated with a discrete event simulator and used to solve a signal optimization problem. The model, accounting for joint information, produced a signal plan that reduced average travel time, as compared to the model presented by Osorio and Bierlaire (2009b). Further research is required to apply the model to more realistic road networks, incorporating the additional complexity that results from dedicated turning lanes and various types of intersections. The joint model can also be embedded in a simulation-based optimization framework, as in Osorio and Bierlaire (2010b), to produce signal plans that outperform the analytical solution. Improving the computational efficiency of the joint approach may be required in order to make these problems tractable for large-scale systems.

Chapter 3 formulates an analytical queueing model that accounts for route choice (Section 3.2.4) and presents a methodology to calibrate route choice parameters of a microscopic traffic simulator using simulation-based optimization techniques (Section 3.3). Future work can add flexibility to the route choice model, considering additional variables that affect route choice. Extending the model to large-scale systems may require a more tractable formulation of the analytical queueing model in order to address more complex calibration problems.

Appendix A

Transition rate matrix derivation

Table A.1 details the full transition rate matrix of a three queue system with queues indexed by $(i, i+1, i+2)$, within a general I-queue network. The parameters μ, γ and k are exogenous. The variables λ_i and $\hat{\mu}_{i+2}$ are endogenous. Recall that λ_i approximates the arrivals to queue i, which consist of either external arrivals, γ_i, or arrivals from queues upstream of i which are not in the $(i, i+1, i+2)$ 3-queue system. Similarly, the service rate of queue $i+2$ (the most downstream of the 3-queue system), $\hat{\mu}_{i+2}$, is also endogenous, since it is determined by the traffic at further downstream queues.

External arrivals are allowed to all queues. The first three sets of rates define the transitions that may occur when an external arrival occurs at queue i, $i+1$ and $i+2$, respectively. An external arrival to a given queue can cause the queue to transition from aggregate state 0 (resp. 1) to aggregate state 1 (resp. 2). Upon an external arrival to, for instance, queue i, the transition from aggregate state 0 to 1 occurs with probability 1, and the transition from aggregate state 1 to 2 occurs if queue i is in the disaggregate state $k_i - 1$, occurring with probability $\alpha^{\mathrm{f}}_{i,s}$ (defined in Section 2.2.3).

The fourth set of rates considers a service completion at queue i. Such an event can cause queue i to transition from aggregate state 1 (resp. 2) to 0 (resp. 1). Upon service completion, the transition from state 2 to 1 occurs with probability 1, and the transition from 1 to 0 occurs if queue i is in the disaggregate state 1, occurring with probability $\alpha^{\mathrm{e}}_{i,s}$ (defined in Section 2.2.3). Additionally, a service completion at

queue i can cause queue $i+1$ to transition from aggregate state 0 (resp. 1) to 1 (resp. 2), which occurs with probability 1 (resp. $\alpha^{\mathrm{f}}_{i+1,s}$).

The fifth set of rates considers a service completion at queue $i+1$. The rates are obtained through similar reasoning as for service completion at queue i. Additionally, if a job at queue i is being blocked by queue $i+1$, then a service completion at queue $i+1$ may trigger a change in the state of queue i. This is described via the blocking probability $\beta_{i,1}$ (defined in Section 2.2.3). More specifically, if queue i is blocked by queue $i+1$, then a service completion at queue $i+1$ will:

1. send the job that has completed service at queue $i+1$ to queue $i+2$, which may lead queue $i+2$ to transition from aggregate state 0 (resp. 1) to 1 (resp. 2);

2. unblock a job at queue i, which may lead queue i to transition from aggregate state 1 (resp. 2) to 0 (resp. 1);

3. have no impact on the state of queue $i+1$ (since an arrival and a departure occur simultaneously).

The final set of rates considers a service completion at queue $i+2$. The rates are obtained through similar reasoning as for service completion at queue $i+1$. Since queue $i+2$ may block both queue $i+1$ and queue i, then a service completion at queue $i+2$ may trigger changes in the states of both queues i and $i+1$. This unblocking is described via the blocking probabilities $\beta_{i,2}, \beta_{i,3}$ and $\beta_{i,4}$ (defined in Section 2.2.3).

Table A.1: Transition rate matrix for a 3-queue system with the 3 queues indexed by $(i, i+1, i+2)$. Enumeration of all possible transitions assuming an initial state $s = (j_i, j_{i+1}, j_{i+2})$ and a new state t.

External arrival to queue i		
New state t	Initial conditions	Rate
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 0$	λ_i
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = \{0,1\}$	$\lambda_i \alpha^{\mathrm{f}}_{i,s}$
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = \{0,1\}$	$\lambda_i \alpha^{\mathrm{f}}_{i,s}$
$(j_i+1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 2,\ j_{i+2} = 2$	$\lambda_i \alpha^{\mathrm{f}}_{i,s}$

External arrival to queue $i+1$		
New state t	Initial conditions	Rate
$(j_i, j_{i+1}+1, j_{i+2})$	$j_{i+1} = 0$	γ_{i+1}
$(j_i, j_{i+1}+1, j_{i+2})$	$j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\gamma_{i+1} \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i, j_{i+1}+1, j_{i+2})$	$j_{i+1} = 1,\ j_{i+2} = 2$	$\gamma_{i+1} \alpha^{\mathrm{f}}_{i+1,s}$

External arrival to queue $i+2$		
New state t	Initial conditions	Rate
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+2} = 0$	γ_{i+2}
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+2} = 1$	$\gamma_{i+2} \alpha^{\mathrm{f}}_{i+2,s}$

Service completion at queue i		
New state t	Initial conditions	Rate
$(j_i, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 0$	$\mu_i(1 - \alpha^{\mathrm{e}}_{i,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 0$	$\mu_i \alpha^{\mathrm{e}}_{i,s}$
$(j_i, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i(1 - \alpha^{\mathrm{e}}_{i,s})\alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i \alpha^{e}_{i,s}(1 - \alpha^{\mathrm{f}}_{i+1,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i \alpha^{\mathrm{e}}_{i,s} \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i(1 - \alpha^{\mathrm{e}}_{i,s})\alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i \alpha^{\mathrm{e}}_{i,s}(1 - \alpha^{\mathrm{f}}_{i+1,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 1,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i \alpha^{\mathrm{e}}_{i,s} \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 2,\ j_{i+1} = 0$	μ_i
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = \{0,1\}$	$\mu_i(1 - \alpha^{\mathrm{f}}_{i+1,s})$
$(j_i-1, j_{i+1}+1, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i \alpha^{\mathrm{f}}_{i+1,s}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i = 2,\ j_{i+1} = 1,\ j_{i+2} = 2$	$\mu_i(1 - \alpha^{\mathrm{f}}_{i+1,s})$

Service completion at queue $i+1$		
New state t	Initial conditions	Rate
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+1} = 1,\ j_{i+2} = 0$	$\mu_{i+1}(1 - \alpha^{\mathrm{e}}_{i+1,s})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_{i+1} = 1,\ j_{i+2} = 0$	$\mu_{i+1} \alpha^{\mathrm{e}}_{i+1,s}$

$(j_i, j_{i+1}, j_{i+2}+1)$	$j_{i+1}=1,\ j_{i+2}=1$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i+1,s})\alpha^{\mathrm{f}}_{i+2,s}$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_{i+1}=1,\ j_{i+2}=1$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i+1,s}(1-\alpha^{\mathrm{f}}_{i+2,s})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_{i+1}=1,\ j_{i+2}=1$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i+1,s}\alpha^{\mathrm{f}}_{i+2,s}$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i=0,\ j_{i+1}=2,\ j_{i+2}=0$	μ_{i+1}
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i=0,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i=0,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i=\{1,2\},\ j_{i+1}=2,\ j_{i+2}=0$	$\mu_{i+1}(1-\beta_{i,1})$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i=\{1,2\},\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})(1-\beta_{i,1})$
$(j_i, j_{i+1}-1, j_{i+2}+1)$	$j_i=\{1,2\},\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}(1-\beta_{i,1})$
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_i=1,\ j_{i+1}=2,\ j_{i+2}=0$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i,s})\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i=1,\ j_{i+1}=2,\ j_{i+2}=0$	$\mu_{i+1}\alpha^{\mathrm{e}}_{i,s}\beta_{i,1}$
$(j_i, j_{i+1}, j_{i+2}+1)$	$j_i=1,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}(1-\alpha^{\mathrm{e}}_{i,s})\alpha^{\mathrm{f}}_{i+2,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i=1,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})\alpha^{\mathrm{e}}_{i,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i=1,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}\alpha^{\mathrm{e}}_{i,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i=2,\ j_{i+1}=2,\ j_{i+2}=0$	$\mu_{i+1}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2}+1)$	$j_i=2,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}\alpha^{\mathrm{f}}_{i+2,s}\beta_{i,1}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i=2,\ j_{i+1}=2,\ j_{i+2}=1$	$\mu_{i+1}(1-\alpha^{\mathrm{f}}_{i+2,s})\beta_{i,1}$

Service completion at queue $i+2$

New state t	Initial conditions	Rate
$(j_i, j_{i+1}, j_{i+2}-1)$	$j_{i+2}=1$	$\hat{\mu}_{i+2}\alpha^{\mathrm{e}}_{i+2,s}$
$(j_i, j_{i+1}, j_{i+2}-1)$	$j_{i+1}=0,\ j_{i+2}=2$	$\hat{\mu}_{i+2}$
$(j_i, j_{i+1}, j_{i+2}-1)$	$j_{i+1}=\{1,2\},\ j_{i+2}=2$	$\hat{\mu}_{i+2}(1-\beta_{i,3})$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_{i+1}=1,\ j_{i+2}=2$	$\hat{\mu}_{i+2}\alpha^{\mathrm{e}}_{i+1,s}\beta_{i,3}$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i=0,\ j_{i+1}=2,\ j_{i+2}=2$	$\hat{\mu}_{i+2}\beta_{i,3}$
$(j_i, j_{i+1}-1, j_{i+2})$	$j_i=\{1,2\},\ j_{i+1}=2,\ j_{i+2}=2$	$\hat{\mu}_{i+2}\beta_{i,4}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i=1,\ j_{i+1}=2,\ j_{i+2}=2$	$\hat{\mu}_{i+2}\alpha^{\mathrm{e}}_{i,s}\beta_{i,2}$
$(j_i-1, j_{i+1}, j_{i+2})$	$j_i=2,\ j_{i+1}=2,\ j_{i+2}=2$	$\hat{\mu}_{i+2}\beta_{i,2}$

Appendix B

Expected number of vehicles

This section derives the analytical expression for the expected number of vehicles in queue i, $E[N_i]$. We have:

$$\begin{aligned} E[N_i] &= E[E[N_i \mid N_{A,i}]] \\ &= 0P(N_{A,i} = 0) + E[N_i \mid N_{A,i} = 1]P(N_{A,i} = 1) + k_i P(N_{A,i} = 2) \end{aligned} \tag{B.1}$$

We derive an analytical approximation for $E[N_i \mid N_{A,i} = 1]$. By definition:

$$E[N_i \mid N_{A,i} = 1] = \sum_{n=1}^{k_i - 1} nP(N_i = n \mid N_{A,i} = 1) \tag{B.2}$$

We approximate $P(N_i \mid N_{A,i} = 1)$ by using the functional form for the stationary distribution of a single M/M/1/k queue (Equation (2.13)), and following a similar derivation to that of (2.14), for a given value of the traffic intensity, ρ, we obtain:

$$P(N_i = n \mid N_{A,i} = 1) = \frac{(1 - \rho)\rho^{n-1}}{1 - \rho^{k-1}}. \tag{B.3}$$

Inserting (B.3) into (B.2):

$$
\begin{aligned}
E[N_i \mid N_{A,i} = 1] &= \sum_{n=1}^{k_i-1} n \frac{(1-\rho)\rho^{n-1}}{1-\rho^{k_i-1}} = \frac{1-\rho}{1-\rho^{k_i-1}} \sum_{n=1}^{k_i-1} n\rho^{n-1} \\
&= \frac{1-\rho}{1-\rho^{k_i-1}} \sum_{n=1}^{k_i-1} \frac{d}{d\rho}\rho^n = \frac{1-\rho}{1-\rho^{k_i-1}} \frac{d}{d\rho} \left(\sum_{n=1}^{k_i-1} \rho^n \right) \\
&= \frac{1-\rho}{1-\rho^{k_i-1}} \frac{d}{d\rho} \left(\rho \sum_{n=0}^{k_i-2} \rho^n \right) \\
&= \frac{1-\rho}{1-\rho^{k_i-1}} \frac{d}{d\rho} \left(\rho \frac{1-\rho^{k_i-1}}{1-\rho} \right) \\
&= \frac{1-\rho}{1-\rho^{k_i-1}} \left(\frac{1-\rho^{k_i-1}}{1-\rho} + \rho \left[\frac{1-\rho^{k_i-1}}{(1-\rho)^2} - \frac{(k_i-1)\rho^{k_i-2}}{1-\rho} \right] \right) \\
&= 1 + \frac{\rho}{1-\rho} - \frac{(k_i-1)\rho^{k_i-1}}{1-\rho^{k_i-1}}
\end{aligned} \tag{B.4}
$$

The above expression depends on the traffic intensity ρ. We proceed as in Section 2.2.3 and provide the following state-dependent approximation of ρ, for a state $s = (1, j_{i+1}, j_{i+2})$:

$$
\rho_{i,s} = \begin{cases}
\lambda_i/\mu_i & \text{if} \quad j_{i+1} < 2 \\
\lambda_i/\mu_{i+1} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} < 2 \\
\lambda_i/\hat{\mu}_{i+2} & \text{if} \quad j_{i+1} = 2 \quad \text{and} \quad j_{i+2} = 2
\end{cases} \tag{B.5}
$$

where λ_i is obtained through (2.17), μ is exogenous and $\hat{\mu}$ is obtained through (2.26). To summarize, for a given queue i, we consider the 3-queue system where i is the most upstream of the 3 queues. This allows us to derive the value of the disaggregate state N_i by conditioning on the states of the 2-downstream queues. In this way, we account for blocking that arises from either of these 2 downstream queues. Letting P_i denote the joint aggregate distribution of the system $(N_{A,i}, N_{A,i+1}, N_{A,i+2})$:

$$
\begin{aligned}
E[N_i \mid N_{A,i} = 1] &= E[E[N_i \mid N_{A,i} = 1, N_A = s]] \\
&= E[N_i \mid N_{A,i} = 1, N_{A,i+1} < 2] P_i(N_{A,i} = 1, N_{A,i+1} < 2) \\
&+ E[N_i \mid N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} < 2] P_i(N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} < 2) \\
&+ E[N_i \mid N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} = 2] P_i(N_{A,i} = 1, N_{A,i+1} = 2, N_{A,i+2} = 2),
\end{aligned}
\tag{B.6}
$$

where the three expectations on the right-hand side of the above equation are given by inserting the expression in (B.4) and their corresponding ρ values defined by (B.5).

Appendix C

Marginal finite capacity queueing model

The finite capacity queuing model from (Osorio and Bierlaire, 2009b) assumes independence between queues and captures interactions through structural parameters. The variables used in this model for a given queue i are listed in Table C.1.

γ_i	external arrival rate;
λ_i	total arrival rate;
μ_i	service rate of a server;
$\tilde{\mu}_i$	unblocking rate;
$\hat{\mu}_i$	effective service rate;
P_i^f	probability of being blocked at queue i;
p_{ij}	transition probability from queue i to queue j;
k_i	upper bound on queue length;
N_i	number of vehicles in queue i;
$P(N_i = k_i)$	probability that queue i is full;
$\mathcal{I}^+$	set of downstream queues to queue i.

Table C.1: List of variables used in independent finite capacity queuing model.

The equation for blocking probability is based on the closed-form expression available for single finite capacity queues, the same as in Equation (2.13). The system of equations is given by:

$$\lambda_i = \gamma_i + \frac{\sum_j p_{ij}\lambda_j(1 - P(N_j = k_j))}{(1 - P(N_i = k_i))}, \tag{C.1}$$

$$\frac{1}{\tilde{\mu}_i} = \sum_{j \in \mathcal{I}^+} \frac{\lambda_j(1 - P(N_j = k_j))}{\lambda_i(1 - P(N_i = k_i))\hat{\mu}_j} \tag{C.2}$$

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i}, \tag{C.3}$$

$$P_i^f = \sum_j p_{ij} P(N_j = k_j), \tag{C.4}$$

$$P(N_i = k_i) = \frac{1 - \rho_i}{1 - \rho_i^{k_i+1}} \rho^{k_i}, \tag{C.5}$$

$$\rho_i = \frac{\lambda_i}{\hat{\mu}_i}. \tag{C.6}$$

The exogenous parameters used in the system of equations are γ_i, p_{ij}, k_i, and μ_i. All other variables are endogenous. Between queue interactions are captured by determining the probability a queue is blocked by a downstream queue that is full, P_i^f, and the rate at which the queue becomes unblocked, $\tilde{\mu}_i$, related by Equations (C.2), (C.3), and (C.4).

Bibliography

Antoniou, C. (2004). *On-line calibration of dynamic traffic assignment models*. Ph.D. thesis, Massachusetts Institute of Technology.

Balakrishna, R. (2006). *Off-line calibration of dynamic traffic assignment models*. Ph.D. thesis, Massachusetts Institute of Technology.

Balsamo, S., De Nitto Persone, V., and Onvural, R. (2001). *Analysis of Queueing Networks with Blocking*, volume 31 of *International Series in Operations Research and Management Science*. Kluwer Academic Publishers, Boston.

Baskett, F., Chandy, K. M., Muntz, R., and Palacios, F. (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery*, **22**(2), 248–260.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions. In R. Hall, editor, *Handbook of Transportation Science*, pages 5–34. Kluwer.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Applications to Travel Demand*, chapter 5, page 103. MIT Press, USA.

Ben-Akiva, M., Gao, S., Lu, L., and Wen, Y. (2012). Combining disaggregate route choice estimation with aggregate calibration of a dynamic traffic assignment model. In *Proceedings of the Fourth International Symposium on Dynamic Traffic Assignment*, Marthas Vineyard, Massachusetts, USA.

Bocharov, P. P., D'Apice, C., Pechinkin, A. V., and Salerno, S. (2004). *Queueing theory*, chapter 3, pages 96–98. Modern Probability and Statistics. Brill Academic Publishers, Zeist, The Netherlands.

Cascetta, E., Nuzzolo, A., Russo, F., and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks. In J.-B. Lesort, editor, *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, pages 697–711, Lyon, France. Pergamon.

Chen, X., Osorio, C., and Santos, B. F. (2012). A simulation-based approach to reliable signal control. In *Proceedings of the International Symposium on Transportation Network Reliability (INSTR)*.

Dargay, J., Gately, D., and Sommer, M. (2007). Vehicle ownership and income growth, worldwide: 1960-2030. *The Energy Journal*, **28**(4), 143–170.

Duranton, G. and Turner, M. (2011). The fundamental law of road congestion: Evidence from us cities. *American Economic Review*, **101**(6), 2616–2652.

Flötteröd, G., Bierlaire, M., and Nagel, K. (2011). Bayesian demand calibration for dynamic traffic simulations. *Transportation Science*, **45**(4), 541–561.

Flötteröd, G., Chen, Y., and Nagel, K. (2012). Behavioral calibration and analysis of a large-scale travel microsimulation. *Networks and Spatial Economics*, **12**, 481–502.

Flötteröd, G., Chen, Y., and Nagel, K. (2012). Choice model refinement from network data. In *Proceedings of IATBR 2012, The 13th International Conference on Travel Behavior Research*, Toronto, Canada.

Hogg, R. and Tanis, E. (2006). *Probability and Statistical Inference*, volume 7e. Pearson Ediucation, Inc., Upper Saddle River.

Jackson, J. R. (1957). Networks of waiting lines. *Oper. Res.*, **5**(4), 518–521.

Jackson, J. R. (1963). Jobshop-like queuing systems. *Management Sci.*, **10**(1), 131–142.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, **82D**, 35–45.

Larson, R. and Odoni, A. (1981). *Urban operations research*. Prentice-Hall.

Little, J. D. C. (1961). A proof for the queuing formula: L= λ W. *Operations Research*, **9**(3), 383–387.

Little, J. D. C. (2011). Little's law as viewed on its 50th anniversary. *Operations Research*, **59**(3), 536–549.

Mathworks, Inc. (2012). *Optimization Toolbox Version 6.2. User's Guide Matlab*. Natick, MA, USA.

MATLAB (2012). *version 7.14.0/739 (R2012a)*. The MathWorks Inc., Natick, Massachusetts.

Nagel, K. and Flötteröd, G. (2012). Agent-based traffic assignment: going from trips to behavioral travelers. In R. Pendyala and C. Bhat, editors, *Travel Behaviour Research in an Evolving World*, chapter 12, pages 261–293. Emerald Group Publishing, Bingley, United Kingdom.

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*. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Osorio, C. and Bierlaire, M. (2009a). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research*, **196**(3), 996–1007.

Osorio, C. and Bierlaire, M. (2009b). A surrogate model for traffic optimization of congested networks: an analytic queueing network approach. Technical Report 090825, Transport and Mobility Laboratory, ENAC, Ecole Polytechnique Fédérale de Lausanne.

Osorio, C. and Bierlaire, M. (2010a). A simulation-based optimization approach to perform urban traffic control. In *Proceedings of the Triennial Symposium on Transportation Analysis (TRISTAN)*, Tromsø, Norway.

Osorio, C. and Bierlaire, M. (2010b). A simulation-based optimization framework for urban traffic congestion management. In *Proceedings of the World Conference on Transport Research (WCTR)*, Lisbon, Portugal.

Osorio, C. and Chong, L. (2012a). An efficient simulation-based optimization algorithm for large-scale transportation problems. In *Winter Simulation Conference (WSC)*, Berlin, Germany.

Osorio, C. and Chong, L. (2012b). Large-scale simulation-based traffic signal control. In *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA.

Osorio, C. and Flötteröd, G. (2012). Capturing dependency among link boundaries in a stochastic network loading model. In *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA.

Osorio, C. and Nanduri, K. (2012). Energy-efficient traffic management: a microscopic simulation-based approach. In *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA.

Osorio, C. and Wang, C. (2012). An analytical approximation of the joint distribution of queue-lengths in an urban network. In *Procedia Social and Behavioral Sciences. Papers selected for the 15th meeting of the EURO Working Group on Transportation*.

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.

Rice, J. A. (1994). *Mathematical statistics and data analysis*. Duxbury Press, Belmont CA USA.

Schweitzer, P. J. (1984). Aggregation methods for large Markov chains. In G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, pages 275–286, North-Holland, Amsterdam.

Schweitzer, P. J. (1991). A survey of aggregation-disaggregation in large Markov chains. In W. J. Stewart and M. Dekker, editors, *Numerical Solution of Markov Chains*, pages 63–88, New York.

Song, Y. and Takahashi, Y. (1991). Aggregate approximation for tandem queueing systems with production blocking. *Journal of the Operations Research Society of Japan*, **34**(3), 329–353.

Spall, J. (1992). Multivariate stochastic approximation using a simultaneous pertubation gradient approximation. *IEEE Transactions on Automatic Control*, **37**(3), 332–341.

Stewart, W. J. (2000). Numerical methods for computing stationary distributions of finite irreducible Markov chains. In W. Grassmann, editor, *Computational Probability*, chapter 4. Kluwer Academic Publishers, Boston, USa.

Takahashi, Y. (1975). A lumping method for numerical calculations of stationary distributions of Markov chains. *Research Reports on Information Sciences, Series B: Operations Research*.

Takahashi, Y. (1985). A new type aggregation method for large Markov chains and its application to queueing networks. In *Proceedings of the International Teletraffic Congress 11*, Kyoto, Japan.

TSS (2011). *AIMSUN 6.1 Microsimulator Users Manual*. Transport Simulation Systems.

VSS (1992). *Norme Suisse SN 640837 Installations de feux de circulation; temps transitoires et temps minimaux*. Union des professionnels suisses de la route, VSS, Zurich.

Yperman, I. (2007). *The link transmission model for dynamic network loading*. Ph.D. thesis, Katolieke Universiteit Leuven.