

# **Mitigating Network Congestion: Analytical Models, Optimization Methods and their Applications**

THÈSE N° 4615 (2010)

PRÉSENTÉE LE 16 AVRIL 2010

À LA FACULTÉ SCIENCES DE BASE  
LABORATOIRE TRANSPORT ET MOBILITÉ  
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Carolina OSORIO PIZANO

acceptée sur proposition du jury:

Prof. F. Eisenbrand, président du jury  
Prof. M. Bierlaire, directeur de thèse  
Prof. J. Barceló, rapporteur  
Prof. A. Odoni, rapporteur  
Prof. P. Thiran, rapporteur



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

Suisse  
2010



*Para Bastien, Pa y Joe*



# Acknowledgments

My deepest gratitude goes to my thesis supervisor, Prof. Michel Bierlaire. His motivation, confidence, curiosity, rigor and intuition have inspired me throughout these years. He has been an excellent scientific mentor, who has made my experience as a research assistant a challenging, exciting and enriching one. I cannot thank him enough for encouraging and allowing me to constantly present my research to the scientific community. As a friend and as a colleague, I thank him for the confidence that he has shown in my work.

Prof. Thomas Liebling has encouraged and supported me throughout these years. I am thankful for the numerous teaching opportunities that he has provided me. I have fond memories that start with our very first encounter, which particularly touched me, and go until his precious advice minutes before defending my thesis.

Having joined his Operations Research group ROSO, I had the chance to be surrounded by motivating colleagues, that quickly became good friends. I thank them for their company. I would like to specially thank Lionel, Gautier, Michaël and Christophe.

I am very grateful to the members and former members of the Transport and Mobility laboratory. I have benefited from the friendly, motivating, relaxing and also challenging atmosphere of this lab. I thank all its members for their support and company. In particular, I would like to thank Gunnar, who read through a preliminary version of this manuscript. I appreciated his valuable feedback, as well as his support during the last months before defending.

I wish to thank Antoine, Emma and Mamy, with whom I shared offices. They supported, guided and shared with me on a daily basis.

I thank Prof. Moshe Ben-Akiva for giving me the opportunity to visit his Intelligent Transportation Systems laboratory at MIT. This research has benefited from the stimulating atmosphere that I experienced there, as well as from the constructive feedback that I received throughout my stay.

I would like to specially thank Maya and Tina for their warm welcoming, along with my friends in Boston for making the non-scientific moments of my stay memorable.

I am thankful to the members of my thesis committee: Professors Jaume Barceló, Amedeo Odoni and Patrick Thiran, for the valuable discussions and their constructive comments. I also thank Prof. Friedrich Eisenbrand for serving as the president of the committee.

This research was supported by the Swiss National Science Foundation grants 205321-107838 and 205320-117581.

The case study of Section 3.1 has been carried out for a project in collaboration with Dr. Philippe Garnerin and Pau Perez from the Division of Anesthesiology at the Geneva University Hospitals. I thank them both for their availability and willingness to introduce me to their field.

An ongoing project in collaboration with the Laboratory of Computational Systems Biotechnology (LCSB1) directed by Prof. Vassily Hatzimanikatis at EPFL triggered the formulation in Section 3.2. I thank both Dr. Luis Mier-y-Teran and Prof. Vassily Hatzimanikatis for the papers, references and numerical code of their model, that allowed me to understand the main features of this intricate biological problem.

I express my gratitude to the members of the Traffic Facilities Laboratory (LAVOC) at EPFL, who provided me with the Lausanne simulation model used in Chapters 4 and 5. Both Emmanuel Bert and Ashish Bhaskar have helped to resolve several technical issues. I thank them for their availability and efficiency.

My daily attempts to advance this research would not have gone far without the company, support and motivation from the members of what I call my *local families*.

Marianne, gracias pour ton soutien et ton écoute quotidiens. Tus consejos sencillos y sabios hacen que me sienta capaz de emprender grandes desafíos. Mi familia italiana: Marcolino, Céline, Gae, Tere, Teo y Leo el Cachetón, no tengo palabras para expresar lo invaluable que ha sido su compañía y su cariño. Rikkilina und Micha, sus sonrisas me acompañaron durante las numerosas noches de insomnio.

Ma famille *Montaignarde*, vous continuez à me soutenir comme si vous étiez tout près. Merci pour votre présence. KTP, your understanding of what this research means to me is the most valuable gift you have ever shared with me.

Las amigas rolas, las charlas que compartimos me llenan de fuerzas para intentar entender los desafíos que se nos presentan. ¡Gracias totales! Natis, tu cariño y tu presencia incondicionales me han llenado de paciencia y de coraje. Mari y Mafe, saben descifrar el significado de mis palabras de una forma sorprendente. Gracias por ser un espejo constante, ayudándome a crecer, a identificar y a enfrentar mis retos.

Clarita, siempre me haz acompañado como una hermana, cuidado como una madre y querido como la gran amiga que eres.

And last but not least, Bastien, Pa y Joe, su apoyo y su cariño me son vitales. La distancia no ha hecho sino recalcar cuánto los quiero.

# Abstract

Congestion is a phenomenon that arises in a variety of contexts. The most familiar representation is urban traffic congestion. Nonetheless, phenomena such as prison cell congestion, hospital bed blocking or, at a cellular scale, ribosome congestion, also arise and affect the performance of the underlying networks. The study of network congestion is therefore of interest in numerous application fields.

Analytical mathematical models enable the identification and the quantification of network congestion. Furthermore, these methods can be used to identify strategies that mitigate network congestion, by integrating them within optimization frameworks.

Deriving such models is an intricate task. Congested networks involve complex traffic interactions. Providing an analytical description of these intricate interactions is challenging. Furthermore, to identify traffic management strategies that indeed mitigate congestion, these models need to be realistic representations of the underlying process, while remaining computationally tractable such that efficient and operational optimization methods can be derived.

This thesis presents an analytical network model based on finite capacity queueing theory. Through a novel state space formulation and the use of structural parameters, the model provides a detailed decomposition of congestion. It describes congestion in terms of its sources, its propagation and dissipation rates as well as its frequency. The model is validated versus existing methods, exact results and simulation results.

Particularly tractable formulations are derived for single server bufferless queues in a tandem topology and for single server queues with finite buffers in an arbitrary topology network. Unlike existing models, the proposed model maintains the network topology and the queue capacities exogenous.

An urban vehicle traffic model is formulated based on this network model. A detailed formulation, based on national transportation standards, is provided. This model is then used to perform optimization for congested road networks. A traffic signal control problem is formulated and solved for the Lausanne city road network. The signal plans derived are evaluated at the microscopic scale with a calibrated simulation model, and compared to both an existing signal plan for the city of Lausanne and to signal plans derived by other methods. The proposed plans delay the propagation of congestion, and lead to improved

performance measures.

The contributions in the urban transportation field are two-fold. Firstly, the proposed model considers a set of intersections and analytically captures the interactions between queues, contrarily to existing analytic queueing models for urban networks which are formulated for a single intersection, and thus do not take such interactions into account.

Secondly, although there is a great variety of signal control methodologies in the literature, there is still a need for solutions that are appropriate and efficient under saturated conditions, where the performance of signal control strategies and the formation and propagation of queues are strongly related. To the best of our knowledge, the existing strategies have not taken urban spillbacks analytically into account.

A framework to perform simulation-based optimization, which combines structural information from the analytical queueing model and microscopic information from an urban traffic simulation model for the city of Lausanne, is presented. The framework resorts to a derivative-free trust region algorithm. It is used to solve a traffic signal control problem.

With this method well-performing signal plans can be identified given a tight computational budget. By combining a traditionally used functional metamodel with an application-specific analytical structural model, this algorithm overcomes the need for a substantial initial sample, and provides meaningful trial points since the very first iterations.

A network model is also formulated and used to evaluate congestion for two other applications. Firstly, the phenomenon of bed blocking in a network of operative and post-operative units of the Geneva University Hospitals is investigated. Three main sources of bed blocking are identified, and their impact upon the different hospital units is quantified.

We go beyond existing analytical queueing methods that have been used in the health care sector by allowing for networks with an arbitrary topology and with an arbitrary number of queues with finite capacity. Furthermore, the detailed performance measures provided by this approach respond to a recently stated need for methods that quantify in-patient bed blocking.

Secondly, the model is formulated for a protein synthesis network, where the traffic of ribosomes along mRNA (messenger ribonucleic acid) strands is of interest. This protein synthesis model consists of a system of linear and quadratic equations, which is a particularly simple and tractable formulation. Unlike other protein synthesis models, this formulation is numerically well-conditioned for highly congested scenarios, suitable for large-scale instances, and can be evaluated using simple numerical techniques.

Keywords: finite capacity queues, queueing networks, analytical network models, optimization, simulation-based optimization, congestion mitigation.

# Résumé

La congestion est un phénomène qui se produit dans de nombreux et divers contextes, tels que dans un réseaux routier, un réseau de cellules de prison, le blocage de lits entre différentes unités hospitalières, ou même à l'échelle cellulaire lors de la formation de protéines où la congestion se produit entre les ribosomes qui parcourent un même ARN messager.

Les modèles analytiques mathématiques nous permettent d'identifier et de quantifier cette congestion. Ces modèles peuvent être intégrés dans des cadres d'optimisation afin d'identifier des stratégies qui permettent de réduire cette congestion, ainsi que ses impacts.

Cependant, la formulation de tels modèles s'avère complexe pour deux raisons principales. Premièrement, dans des réseaux congestionnés, les flux interagissent de façon complexe. La description analytique de ces interactions est un défi. De plus, ces modèles doivent être suffisamment réalistes afin d'identifier des stratégies qui permettent en effet de réduire la congestion, tout en restant suffisamment simples pour que les méthodes d'optimisation restent efficaces et opérationnelles.

Cette thèse présente un modèle analytique de réseau basé sur la théorie de files d'attente à capacité finie. Le modèle décrit la congestion en termes de ses sources, ses taux de propagation et de dissipation, ainsi que de sa fréquence. Cette description détaillée de la congestion est possible grâce à une nouvelle définition de l'espace des états, ainsi qu'à l'utilisation de paramètres structuraux. Ce modèle est validé en le comparant avec d'autres modèles, avec des résultats exacts et avec des résultats de simulation.

Deux formulations qui sont particulièrement simples sont présentées: l'une concerne des réseaux de files à un serveur et sans buffer dans une topologie en série, l'autre concerne des files à un serveur dans une topologie quelconque et avec un buffer fini. Contrairement aux méthodes existantes, ce modèle préserve la topologie du réseau ainsi que la capacité des files comme des paramètres exogènes.

Un modèle de trafic de véhicules dans un réseau routier urbain est formulé, en se basant sur des normes routières nationales. Ce modèle est ensuite intégré dans un cadre d'optimisation. Un problème d'optimisation de feux est formulé et résolu pour le réseau routier de la ville de Lausanne. La performance des plans de feux proposés est évaluée avec un simulateur de trafic microscopique calibré pour la ville de Lausanne. Elle est comparée avec la performance d'un plan de feux existant pour la ville de Lausanne, et avec celle

de plans de feux proposés par d’autres méthodes. Les plans proposés par cette nouvelle méthode retardent la propagation de la congestion, et améliorent les principales mesures de performance du réseau.

Ce modèle contribue au secteur de transport urbain de deux manières. Premièrement, ce modèle considère un ensemble d’intersections et décrit analytiquement les interactions entre les files de véhicules, alors que les modèles analytiques de files d’attente proposés ont jusqu’à présent considéré une seule intersection, et n’ont pas pris en compte les interactions entre intersections voisines.

De plus, même si de nombreuses méthodes d’optimisation de feux existent, il y a encore un manque de méthodes appropriées et efficaces pour des scénarios congestionnés où la performance des plans de feux et la formation et propagation de files sont fortement liés. Il n’a y pas, à notre connaissance, de méthode prenant en compte la propagation de files analytiquement.

Un cadre d’optimisation basée sur la simulation est présenté. Il combine l’information structurelle du modèle analytique de files d’attente et l’information microscopique du simulateur de trafic de la ville de Lausanne. Ce cadre est basé sur un algorithme de région de confiance sans dérivées. Il est utilisé pour résoudre le problème de plans de feux. Cette méthode identifie des plans de feux performants, tout en limitant le nombre de simulations requises.

Cette thèse présente un modèle de réseau pour deux autres applications. Pour chacune d’entre elles, une formulation détaillée est donnée. Premièrement, le phénomène de “blocage de lits” au sein d’un réseau d’unités opératoires et post-opératoires des Hôpitaux Universitaires de Genève est étudié. Trois sources principales de blocage sont identifiées, leur impact sur les unités est quantifié.

Cette méthodologie va au-delà des modèles analytiques de files d’attente proposés pour le secteur hospitalier, en permettant l’analyse de réseaux avec une topologie quelconque, et sans limiter le nombre de files à capacité finie. De plus, les mesures de performance détaillées que cette méthode propose, répondent à un besoin récemment reconnu pour des méthodes permettant de quantifier le phénomène de “blocage de lits”.

Deuxièmement, le modèle est formulé pour un réseau d’ARN messagers, qui permettent la synthèse de protéines. Dans ce cas, nous étudions le trafic des ribosomes le long d’un ARN messager. Le modèle pour ce réseau consiste en un ensemble d’équations linéaires et quadratiques. C’est une formulation particulièrement simple. Contrairement à d’autres modèles de synthèse de protéines, ce modèle est numériquement bien conditionné pour des scénarios à haute congestion, il est approprié pour des réseaux à grande échelle, et peut être évalué avec des méthodes numériques simples.

Mots-clés: files d’attente à capacité finie, réseaux de files d’attente, modèles analytiques de réseaux, optimisation, simulation, congestion

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Thesis motivation and objectives . . . . .	2
1.2	Thesis contributions . . . . .	4
1.3	Thesis structure . . . . .	6
<b>2</b>	<b>Finite capacity queueing network model</b>	<b>9</b>
2.1	Introduction . . . . .	10
2.2	General framework . . . . .	11
2.3	Literature review . . . . .	12
2.3.1	Exact methods . . . . .	12
2.3.2	Approximation methods . . . . .	13
2.4	Model . . . . .	15
2.4.1	Global balance equations . . . . .	15
2.4.2	Transition rates . . . . .	17
2.4.3	System of equations . . . . .	24
2.5	Validation . . . . .	26
2.5.1	Validation versus existing methods . . . . .	26
2.5.2	Validation versus exact results . . . . .	28
2.5.3	Validation versus simulation results . . . . .	30
2.5.4	Convergence of the validation runs . . . . .	32
2.5.5	Tests on larger networks . . . . .	33
2.6	Conclusions and future work . . . . .	34
<b>3</b>	<b>Congestion evaluation: applications in health care and biology</b>	<b>35</b>
3.1	Network of hospital units . . . . .	36
3.1.1	Context . . . . .	36
3.1.2	Geneva University Hospitals network . . . . .	36
3.1.3	Comparison with simulation results . . . . .	38
3.1.4	Congestion analysis . . . . .	40
3.1.5	Conclusions . . . . .	41

3.2	Protein synthesis network . . . . .	42
3.2.1	Context . . . . .	42
3.2.2	Motivation . . . . .	43
3.2.3	Model . . . . .	43
3.2.4	System of equations . . . . .	50
3.2.5	Mapping of parameters and variables . . . . .	52
3.2.6	Validation . . . . .	53
3.2.7	Conclusions and future work . . . . .	54
<b>4</b>	<b>A surrogate for the optimization of congested urban road networks</b>	<b>57</b>
4.1	Motivation . . . . .	58
4.2	Literature review . . . . .	60
4.2.1	Analytic queueing models . . . . .	60
4.2.2	Traffic signal control . . . . .	61
4.3	Surrogate network model . . . . .	63
4.3.1	Queueing model . . . . .	63
4.3.2	Network topology . . . . .	63
4.3.3	Bounded queues . . . . .	64
4.3.4	Arrival and service rates . . . . .	65
4.3.5	System of equations . . . . .	66
4.4	Optimization problem . . . . .	68
4.5	Empirical analysis . . . . .	70
4.5.1	Microscopic traffic simulation model of the city of Lausanne . . . . .	70
4.5.2	Between-queue interactions . . . . .	72
4.5.3	Comparison with existing methods . . . . .	75
4.6	Large-scale networks . . . . .	80
4.6.1	Formulation . . . . .	80
4.6.2	Lausanne city network . . . . .	81
4.7	Conclusions and future work . . . . .	85
<b>5</b>	<b>A simulation-based optimization approach for the management of congested urban road networks</b>	<b>87</b>
5.1	Introduction . . . . .	88
5.2	Literature review . . . . .	90
5.3	Method . . . . .	93
5.3.1	Metamodel . . . . .	93
5.3.2	Algorithmic framework . . . . .	96
5.3.3	Algorithm . . . . .	96
5.3.4	Algorithmic details . . . . .	98

5.4	Optimization problem . . . . .	99
5.4.1	Traffic signal control . . . . .	99
5.4.2	Trust region subproblem . . . . .	101
5.4.3	Signal plan features . . . . .	101
5.5	Empirical analysis . . . . .	102
5.5.1	Lausanne subnetwork with simplified demand distribution . . . . .	102
5.5.2	Lausanne subnetwork with evening peak hour demand . . . . .	105
5.6	Conclusions and future work . . . . .	112
<b>6</b>	<b>Conclusions</b>	<b>115</b>
<b>A</b>	<b>Review of traffic signal control methodologies</b>	<b>119</b>
A.1	Fixed-time isolated strategies . . . . .	119
A.2	Fixed-time coordinated strategies . . . . .	119
A.3	Traffic-responsive methods . . . . .	120
<b>B</b>	<b>Webster and Highway Capacity Manual methods</b>	<b>123</b>
B.1	Webster's method . . . . .	123
B.2	Highway Capacity Manual method . . . . .	124
B.3	Equivalence between both methods . . . . .	124
<b>C</b>	<b>Trust region algorithms</b>	<b>127</b>
C.1	Basic trust region algorithm . . . . .	127
C.2	Derivative-free trust region algorithm . . . . .	128



# Chapter 1

## Introduction

### Contents

---

1.1	Thesis motivation and objectives . . . . .	2
1.2	Thesis contributions . . . . .	4
1.3	Thesis structure . . . . .	6

---

## 1.1 Thesis motivation and objectives

The study of network congestion is of interest in various fields, ranging from the analysis of spillbacks (i.e. the backwards propagation of congestion) in urban traffic or pedestrian traffic (Cheah and Smith, 1994) to that of hospital bed blocking (Koizumi et al., 2005), prison cell blocking (Korporaal et al., 2000) or even the congestion of ribosomes in the context of protein synthesis (Mehra and Hatzimanikatis, 2006).

Congestion is also a costly phenomenon. It impacts our environment and economy at a global scale, affects us at an individual scale, e.g. by hindering our mobility, our access to health care, and may also perturb fundamental cellular processes of our body at a nanoscopic scale.<sup>1</sup> It is therefore of wide interest to identify strategies that can mitigate congestion, e.g. urban traffic management schemes or hospital resource allocation strategies.

Mathematical methods that can evaluate network congestion are key tools to estimate its impact and cost, and ultimately to systematically identify strategies that enable its mitigation. Deriving such methods is an intricate task since congested networks involve complex traffic interactions that are typically difficult to both measure and model. In particular, as congestion increases so do the interactions between the different network components, as well as the correlation in their performance.

Consider, for instance, a given road of an urban transportation network. As the length of the queue of vehicles on that road increases and exceeds the length of the road, it will impact the queue of vehicles at upstream roads. Modeling how these queues spread, dissipate and, in particular, interact is challenging.

The complex interactions between the units that flow through the network and the network components have therefore lead to the development of detailed microscopic simulation models. Such models describe the interaction of each unit in the network. For instance, in the context of urban traffic simulators, disaggregate models describe the behavior of each driver within the network with regards to decisions such as whether to change lanes, or which route to choose. Microscopic simulation models are thus a popular approach to evaluate network performance measures in the context of scenario-based analysis (i.e. what-if analysis) or sensitivity analysis.

In practice, the use of simulators is limited by two main issues. Firstly, detailed and realistic performance estimates can be derived as long as these models are appropriately validated and calibrated. This requires substantial amounts of equally detailed data. Data collection is limited by both its cost, and for some applications (e.g. ribosome congestion)

---

<sup>1</sup>The case studies covered in this thesis will evaluate congestion at an urban, institutional, and cellular scale.

it is merely not possible to measure in such detail the interactions of interest.

Secondly, to systematically identify congestion mitigation strategies these simulators need to be embedded within an optimization framework. This is an intricate task for several reasons: the detailed underlying models lead to noisy nonlinear performance measures with no closed form available, their evaluation is also computationally expensive, not to mention the cost of evaluating derivative information. Thus when performing optimization the use of simulation models is limited across all application fields.

Given the complexity of performing simulation-based optimization, a common approach is to construct a simplified model of the simulation model. This lower fidelity model is referred to as a *surrogate* or a *metamodel*. It is less realistic but is also typically less expensive to evaluate. This less refined model may be a simplified physical or structural model of the simulator (e.g. of lower resolution); or an analytical model chosen for its tractability, which is fitted based on simulated observations.

The family of macroscopic models overcomes the two main limitations of simulation models by compromising on realism. Macroscopic models are flow-based methods that describe the traffic interactions at an aggregate scale. They attempt to capture only the main (e.g. first-order) features of these interactions. By focusing on a macroscopic scale, these models require less data, are more robust to data inaccuracies, more flexible and computationally more efficient. Furthermore, when formulated as analytical models, they fit naturally within classical optimization frameworks.

The main motivation of this work is to contribute in bridging the gap between these two research tracks of macroscopic and microscopic methods. Within this context, the objectives of this dissertation are three-fold.

1. To provide an analytical macroscopic yet detailed description of network congestion, e.g. how and where it arises, how it spreads, dissipates and impacts the networks performance. This involves identifying which traffic interactions are of first-order importance and which can be ignored at a macroscopic scale. Furthermore, providing an analytical description of the intricate behavior of congested networks is challenging.
2. To propose analytical surrogate models to perform optimization of congested networks, and to use them to derive congestion mitigation strategies. This requires models that are sufficiently detailed, such that the derived strategies do indeed mitigate congestion when evaluated at a finer scale (e.g. microscopic scale); while preserving tractability, such that the method remains computationally efficient.
3. To combine the advantages of both microscopic and macroscopic models, i.e. their respective realism and tractability, by proposing an optimization framework that combines both types of models.

## 1.2 Thesis contributions

A summary of the main contributions of this thesis follows.

### Analytical models

- This thesis formulates an analytical queueing network model based on finite capacity queueing theory (Chapter 2). Through a novel state space formulation and the use of structural parameters, the model provides a detailed decomposition of congestion. The model allows for networks with an arbitrary topology. Unlike existing models it maintains the network topology and the queue capacities exogenous, thus no constraints need to be checked a posteriori to ensure the validity of the estimates.
- To address large-scale networks, this thesis derives two formulations of the model that are particularly simple and tractable and therefore appealing for large-scale instances. These formulations are provided for single server bufferless queues in a tandem topology (Section 3.2), and for single server queues with finite buffers in an arbitrary topology network (Section 4.6).

### Optimization methods

- In this thesis two surrogates to perform optimization for congested urban road networks are presented (Chapters 4 and 5). As described in Section 1.1, the main motivation and challenge when deriving a surrogate is to achieve an appropriate trade-off between realism and tractability.

For each surrogate we provide a detailed formulation, we then integrate it within an optimization framework to perform traffic signal control. The results of these two chapters show that the signal plans derived by these methods lead to improved performance when evaluated at the microscopic scale and compared to signal plans derived by other methods.

- A framework to perform simulation-based optimization, which uses a metamodel that combines information from the queueing model and a microscopic urban traffic simulator, is presented (Chapter 5). The proposed metamodel innovates by combining two families of metamodels known as physical and functional metamodels.

The framework resorts to a derivative-free trust region algorithm. It is used to solve a traffic signal control problem. The results indicate that this framework allows the identification of well performing trial points (i.e. signal plans) given a tight computational budget, which is the main motivation of derivative-free methods.

By combining a traditionally used functional metamodel with an application-specific analytical structural model, this algorithm overcomes the need for a substantial initial sample, and provides meaningful trial points since the very first iterations.

## Applications

A general formulation has been preserved for the analytical model, such that it can be used to evaluate network congestion for various applications. This thesis applies this model to evaluate network congestion for three different application fields. For each application, a detailed formulation is presented.

- **Health care.** We study the phenomenon of *bed blocking* by modeling patient flow within a network of operative and post-operative units of the Geneva University Hospitals (Section 3.1). The detailed decomposition of congestion provided by the model allows us to identify three main sources of bed blocking and to quantify their impact upon the different hospital units. The performance measures of the model also reveal that although bed blocking may be a rare event, it may have a strong impact upon the performance of a given unit.

We go beyond existing analytical queueing methods that have been used in the health care sector by allowing for networks with an arbitrary topology and with an arbitrary number of queues with finite capacity. Furthermore, the detailed performance measures provided by this approach respond to a recently stated need for methods that quantify in-patient bed blocking.

- **Biology.** We propose a method to evaluate the congestion of ribosomes in protein synthesis networks (Section 3.2). The novel state space formulation of our model allows us to derive explicit information regarding ribosome blocking, providing a more detailed description of congestion compared to other protein synthesis models. Furthermore, the model is formulated as a system of linear and quadratic equations, which is a particularly simple and tractable formulation. Unlike other protein synthesis models, this formulation is numerically well-conditioned for highly congested scenarios, suitable for large-scale instances, and can be evaluated using simple numerical techniques.

- **Transportation.** We formulate both the modeling and the optimization framework for urban vehicle traffic networks (Chapters 4 and 5). The contributions in this field are three-fold. Firstly, the proposed model considers a set of intersections and analytically captures the interactions between queues, contrarily to existing analytic queueing models for urban networks which are formulated for a single intersection, and thus do not take such interactions into account.

Secondly, although there is a great variety of signal control methodologies in the literature, there is still a need for solutions that are appropriate and efficient under saturated conditions, where the performance of signal control strategies and the formation and propagation of queues are strongly related. To the best of our knowledge, the existing strategies have not taken urban spillbacks analytically into account. We use this

traffic model to solve a fixed-time signal control problem for a subnetwork of the Lausanne city center. We compare the performance of the derived signal plans with that of several other methods, illustrating the importance of capturing the between-queue interactions.

Thirdly, we derive a formulation of the traffic model suitable for large-scale networks. We use it to solve a signal control problem considering the entire city of Lausanne. The results indicate that the proposed method leads to improved average travel times when compared to an existing signal plan for the city of Lausanne.

## 1.3 Thesis structure

**Chapter 2** reviews finite capacity queueing network (FCQN) approaches and formulates an analytical FCQN model. In this chapter, the model is validated versus existing methods, exact results and simulation results. The methodology presented in this chapter has been published as:

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

**Chapter 3** uses the model formulated in Chapter 2 to evaluate network congestion considering two applications. Firstly, in Section 3.1 we study patient flow in a network of hospital operative and post-operative units. In particular, we study the phenomenon of *bed blocking*. The results presented have been carried out for a project in collaboration with Dr. Philippe Garnerin and Pau Perez from the Division of Anesthesiology at the Geneva University Hospitals. The results of this collaboration have been published as part of the previously mentioned paper and as:

Osorio, C., Weibel, C., Perez, P., Bierlaire, M. and Garnerin, P. (2006). *Patient flow simulation as a tool for estimating policy impact*, Swiss Medical Informatics 58:3336.

Secondly, in Section 3.2 we formulate this model to study congestion within a protein synthesis network and validate it versus an existing protein synthesis network model (Mehra and Hatzimanikatis, 2006). We derive a simple and particularly tractable model formulation for single server bufferless tandem networks. The method and results presented in this section have been carried out for an ongoing project in collaboration with the Laboratory of Computational Systems Biotechnology (LCSB1) directed by Prof. Hatzimanikatis at EPFL.

**Chapter 4** proposes the use of the model presented in Chapter 2 as a surrogate model to perform optimization. This chapter considers urban vehicle traffic and in particular

traffic signal control. It reviews both existing analytical queueing models for urban traffic, and traffic signal control methodologies. It proposes an urban traffic network model, which is then used as a surrogate to solve a fixed-time signal control problem. This method is applied to a subnetwork of the Lausanne city center, and its performance is compared to that of several other methods.

We then derive a formulation of the traffic model that is suitable for large-scale networks (Section 4.6). We use it to solve a signal control problem for the entire city of Lausanne, and compare the performance of the derived signal plan with that of an existing signal plan for the city of Lausanne.

The results of Sections 4.1-4.5 have been presented and published as:

Osorio, C. and Bierlaire, M. (2008). *Network performance optimization using a queueing model*, Proceedings of the European Transport Conference (ETC), Noordwijkerhout, The Netherlands.

Osorio, C. and Bierlaire, M. (2008). *A multiple model approach for traffic signal optimization in the city of Lausanne*, Proceedings of the Swiss Transport Research Conference (STRC), Ascona, Switzerland.

Osorio, C. and Bierlaire, M. (2008). *Signal control optimization with a queueing network model capturing congestion*, Proceedings of the Pan-American Conference on Traffic and Transportation Engineering (PANAM), Cartagena, Colombia.

Osorio, C. and Bierlaire, M. (2008). *A queueing network approach to the traffic signal optimization of the Lausanne city center*, Proceedings of the Latin-Ibero-American Conference on Operations Research (CLAIO), Cartagena, Colombia.

**Chapter 5** presents a framework to perform simulation-based optimization for the management of congested urban road networks. The metamodel proposed combines information from the analytical model presented in Chapter 4 and a calibrated microscopic traffic simulation model of the city of Lausanne. This metamodel is integrated within a derivative-free trust region algorithm. The framework is used to solve a fixed-time signal control problem for a subnetwork of the Lausanne city center. The performance of the signal plans derived by this approach is compared to that of signal plans derived by other approaches and to the performance of an existing plan for the city of Lausanne. Preliminary results of this framework have been presented and published as:

Osorio, C. and Bierlaire, M. (2009). *A multi-model algorithm for the optimization of congested networks*, Proceedings of the European Transport Conference (ETC), Noordwijkerhout, The Netherlands.

Osorio, C. and Bierlaire, M. (2009). *A simulation optimization framework for the management of congested urban road networks*, Proceedings of the Swiss Transport Research Conference (STRC), Ascona, Switzerland.

Osorio, C. and Bierlaire, M. (2009). *A metamodel approach for simulation optimization of congested urban road networks*. Computational Management Science Conference (CMS), Geneva, Switzerland.

**Chapter 6** presents conclusions and future lines of research that have arisen from the results of this thesis.

**Appendices.** The appendices provide: A) a review of traffic signal control methodologies; B) details regarding the traffic signal control methods proposed by Webster (1958) and by the Highway Capacity Manual (TRB, 2000); C) basic and derivative-free trust region algorithms.

# Chapter 2

## Finite capacity queueing network model

### Contents

---

<b>2.1</b>	<b>Introduction . . . . .</b>	<b>10</b>
<b>2.2</b>	<b>General framework . . . . .</b>	<b>11</b>
<b>2.3</b>	<b>Literature review . . . . .</b>	<b>12</b>
2.3.1	Exact methods . . . . .	12
2.3.2	Approximation methods . . . . .	13
<b>2.4</b>	<b>Model . . . . .</b>	<b>15</b>
2.4.1	Global balance equations . . . . .	15
2.4.2	Transition rates . . . . .	17
2.4.3	System of equations . . . . .	24
<b>2.5</b>	<b>Validation . . . . .</b>	<b>26</b>
2.5.1	Validation versus existing methods . . . . .	26
2.5.2	Validation versus exact results . . . . .	28
2.5.3	Validation versus simulation results . . . . .	30
2.5.4	Convergence of the validation runs . . . . .	32
2.5.5	Tests on larger networks . . . . .	33
<b>2.6</b>	<b>Conclusions and future work . . . . .</b>	<b>34</b>

---

## 2.1 Introduction

Detecting the sources and effects of congestion within a network allows us to better understand its behavior and to improve its performance. As described in the introduction, the most common approach to analyze network congestion is the development of simulation models that capture the details of the underlying system, but are cumbersome to use within an optimization framework. On the other hand, analytic models naturally fit within such a framework but are rarely developed due to the complexity of modeling the propagation of congestion while preserving a flexible model. In this chapter, we focus on analytic network models and more specifically on analytic queueing network models.

When modeling a network using a queueing theory framework, it is important to capture the interactions between the queues. Consider a network of hospital units (e.g. operative and post-operative units) where each unit is modeled as a specific queue and where it is the patient flow that is of main interest. For such a network, understanding the correlation between the occupation of the different units can help to avoid bed blocking and to improve a patients recovery procedure. More generally, the between-queue correlation helps to explain the propagation of congestion as well as its effects (such as spillbacks). Moreover, in networks containing loops spillbacks are of special interest because they may lead to deadlocks (also known as gridlocks) (Daganzo, 1996).

The most researched queueing network model is the Jackson network model (Jackson, 1963; Jackson, 1957) which assumes infinite capacity (i.e. infinite buffer size) for all queues. Infinite capacity is a strong assumption that is often maintained due to the difficulty of grasping the between-queue correlation of finite capacity networks. In order to capture these between-queue interactions we resort to models with finite capacity queues.

A finite capacity queueing network (FCQN) consists of a network of queues with finite buffers. FCQN models are of interest for a variety of applications such as the study of hospital patient flow (Cochran and Bharti, 2006; Koizumi et al., 2005), manufacturing networks (Papadopoulos and Heavey, 1996), software architecture networks (Balsamo et al., 2003), circulation systems (e.g. corridors) (Cheah and Smith, 1994) and prison networks (Korporaal et al., 2000).

The spread of congestion is modeled in finite capacity queues by what is known as *blocking*. As is detailed in Section 2.2, a job (e.g. a patient, a prisoner) is said to be blocked if upon service completion it cannot proceed to the next queue on its path because that queue is full. The job must then remain at its current queue, i.e. it is blocked at its current location. Describing this blocking phenomenon (i.e. where and how often it occurs, as well as its duration) analytically is challenging; not to mention the added complexity of deriving a computationally efficient model.

The model developed in this chapter will be embedded within several optimization frameworks, in order to derive congestion mitigation strategies. Thus, it is important to propose

a tractable model. Given the complexity of deriving tractable analytical expressions for transient distributions, we focus on stationary distributions.

Exact methods to evaluate the stationary distribution of an FCQN exist only for networks with two or three queues with specific topologies. For more general networks, approximation methods are used to evaluate the stationary distributions. Existing analytic FCQN models based on approximation methods either revise queue capacities or vary the network topologies. If queue capacities are revised, then they become endogenous parameters. Moreover, approximations need to be used to ensure their integrality, and their positivity is only checked a posteriori. We propose an FCQN model which preserves these parameters as exogenous. Furthermore, we allow for networks with an arbitrary topology.

Additionally, in this model congestion is not regarded as an underlying phenomenon but is directly modeled. More specifically, we propose a novel formulation of the state space of the queues that explicitly models the blocking phase. Few analytic models incorporating blocking have been developed, and there is a recently recognized need for them (Cochran and Bharti, 2006). Our formulation yields performance measures that describe both the sources and the effects of congestion.

This chapter is structured as follows. We describe the FCQN framework and then review the existing models. The proposed model is then described, followed by its validation versus existing methods, exact results and simulation results.

## 2.2 General framework

We are interested in evaluating the performance of a network of queues. A job is the generic name for the units of interest that flow through the network, e.g. a vehicle, a prisoner, a patient. We consider open queueing networks where jobs are allowed to leave the network and where the external arrivals arise from an infinite population of jobs. We now describe the general process that a job goes through upon arrival to a queue.

Jobs arriving to a queue are either served immediately or wait until a server becomes available. Once a job is served, it is routed to its next queue according to a probabilistic routing model. We call this queue the target queue. If this target queue has finite capacity then it may be full. If it is full then the job is **blocked** at its current location. Once there is a place at the target queue, the job is unblocked and proceeds to the target queue. The jobs are unblocked with a first-in first-out (FIFO) mechanism.

Various blocking mechanisms have been defined in the literature (Balsamo et al., 2001). They differ either in the moment the job is considered to be blocked (e.g. before or after service) or in the routing mechanism of blocked jobs. The blocking mechanism that we have just described is known as blocking-after-service.

We introduce the main queueing theory terms. A given queue  $i$  has  $c_i$  parallel servers,

each one serving with rate  $\mu_i$ , and has an arrival rate of  $\lambda_i$ . The total number of jobs allowed in the queue is called the capacity of the queue,  $k_i$ , the buffer size is  $k_i - c_i$ . The possible routings among queues are given by the transition probability matrix  $(p_{ij})$ , where  $p_{ij}$  denotes the probability that a job at queue  $i$  is routed to queue  $j$ .

## 2.3 Literature review

A first survey of FCQN models was made by Perros (1984), who later on also wrote a historical overview of the research motivations and advances in networks with blocking (Perros, 2003). A detailed introductory book was written by Balsamo et al. (2001). Surveys focusing on specific application fields exist for the software architecture sector (Balsamo et al., 2003), the production and manufacturing sector (Papadopoulos and Heavey, 1996) and on retrial queues for the telecommunications sector (Artalejo, 1999).

The joint stationary distribution of the network, which contains the probability of each possible state of the network, allows us to derive the main network performance measures. We distinguish between models that allow the exact evaluation of this joint stationary distribution and those based on approximation methods.

### 2.3.1 Exact methods

Exact methods consist of either closed form expressions or numerical evaluation of the joint stationary distribution. For an FCQN, the between-queue correlation suggests a non-product form joint stationary distribution. Thus, closed form expressions are difficult to obtain.

Exact numerical methods exist for networks with two queues in tandem topologies (Grassman and Derkic, 2000; Langaris and Conolly, 1984), tandem topologies allowing for feedback (Akyildiz and von Brand, 1994; Latouche and Neuts, 1980; Konheim and Reiser, 1978; Konheim and Reiser, 1976) or two queues in a closed network (Balsamo and Donatiello, 1989).

Exact numerical evaluation of the joint stationary distribution can also be obtained by solving the global balance equations (these are detailed in Section 2.4.1). A detailed description of these numerical methods can be found in Stewart (2000). These equations require the construction of the transition rate matrix, i.e. the description of the transition rates between all feasible states of the network. This time consuming task is therefore only conceivable for small networks (i.e. small in the number of queues and their capacity). This approach also lacks flexibility because changes in the network topology require redefining the transition rate matrix. If the networks of interest have a more general topology or an arbitrary size, then their analysis is done by models based on approximation methods. The proposed model is based on an approximation method.

### 2.3.2 Approximation methods

Models based on approximation methods can be classified into either simulation-based or analytic models. The use of simulation models is the most popular approach to evaluate the performance of finite capacity queueing networks. Surveys of simulation models exist for sectors such as transportation (Nagel, 2002), healthcare (Fone et al., 2003; Jun et al., 1999), computer science (Sadoun, 2000; Obaidat, 1990) and the analysis of call centers (Koole and Mandelbaum, 2002; Mandelbaum, 2001). This approach, although more realistic and detailed, is cumbersome to optimize, and its accuracy is strongly dependent on the quality of the calibration data (Korporaal et al., 2000). Analytic models are simpler, less data expensive and more flexible.

The main motivation of analytic models based on approximation methods is to reduce the dimensionality of the system under study. Decomposition methods achieve this by decomposing the network into subnetworks and modeling each subnetwork independently. The structural parameters of each subnetwork (e.g. arrival and service rates) depend on the state of other subnetworks and thus capture the correlation with other subnetworks.

The main difficulty lies in obtaining good approximations for these parameters so that the stationary distribution of the subnetwork is a good estimate of its marginal stationary distribution. Given a subnetwork, its stationary distribution can be obtained by either establishing a behavioral analogy with a network whose distribution has a closed (and often product) form, or by exact numerical evaluation of the global balance equations which now have a smaller dimension but are often nonlinear.

Existing models based on decomposition methods have defined subnetworks consisting of single queues, pairs of queues or triplets. We call these methods single, two queue and three queue decomposition methods, respectively. If not stated otherwise the models concern open finite capacity networks with exponentially distributed service times.

The most commonly used decomposition method is single queue decomposition. The first model based on this method dates back to the work of Hillier and Boling (1967) who considered tandem single server networks. One of the most used models based on single queue decomposition concerns single server feed-forward networks where each finite capacity queue is transformed into an M/M/1 queue, and the blocking is taken into account by revising the arrival and service rates of the queues (Takahashi et al., 1980).

An extension of this model to queues with multiple servers is given by Koizumi et al. (2005). Each queue is treated as an M/M/c queue for which closed form expressions of the performance measures are used. The buffers are considered infinite for each isolated queue. This approximation holds for a given queue if either the expected number of jobs does not exceed its capacity, or if it does but the difference can be accommodated by predecessor queues. In this method this constraint is checked only a posteriori.

A model applicable to networks with an arbitrary topology is given by Korporaal et al. (2000). The individual queues are modeled as  $M/M/c/K$  queues for which closed form performance measures are used. As for the method of Koizumi et al. (2005) the capacity of the queues are revised. Here the average queue length updates the capacity of predecessor queues. They use linear interpolation in order to ensure the integrality of the capacities, and their positivity is verified a posteriori.

The Expansion Method (Kerbach and Smith, 2000, 1988, 1987) was developed for networks of  $M/M/1/K$  queues. Here a network reconfiguration expands all finite capacity queues to artificial infinite capacity holding queues, which register the blocked jobs. This model was later extended to multiple servers and applied to pedestrian traffic flows by Cheah and Smith (1994). Gupta and Kavusturucu (2000) applied this model to production feed-forward systems, where service interruptions are allowed. Singh and Smith (1997) used it to evaluate network performance measures within a buffer allocation problem. A similar transformation where all  $GE/GE/c/K$  queues are transformed into  $GE/GE/c$  queues, and thus the joint distribution is approximated by a product form joint distribution, was proposed by Tahilramani et al. (1999).

Models based on single queue decomposition have also been proposed for single server networks with phase-type service distributions for both tandem (Altioek, 1982) and feed-forward topologies (Altioek and Perros, 1987). Jun and Perros (1989) have extended this work to an arbitrary topology and have also considered general service times for an open tandem network in Jun and Perros (1990).

The use of a phase-type service distribution accounts for all possible blockings but, as stated in Altioek and Perros (1987), it requires the construction of very detailed phase-type service mechanisms, which is a cumbersome and CPU intensive task for large networks. In these models, the capacity of a given queue is also augmented. It is increased by the sum of all predecessor queue capacities, in order to account for blocked jobs at predecessor queues.

Few authors have considered subnetworks larger than single queues. Models based on two queue decomposition methods have been proposed for open tandem networks (Alfa and Liu, 2004; Brandwajn and Jow, 1988; Brandwajn and Jow, 1985) and for an arbitrary topology (Lee et al., 1998). Two queue decomposition was used by van Vuuren et al. (2005) to study multiple server tandem queues with generally distributed service times. As an extension of the work by Brandwajn and Jow (1988), Schmidt and Jackman (2000) proposed a model based on a three queue decomposition method for a single server arbitrary topology network. Subnetworks consisting of more than one queue can theoretically provide more accurate results than single queue decomposition, but are computationally more intensive (Perros, 1994).

In order to acknowledge the finite capacity property of the networks, the existing mod-

els modify either the network topologies or the queue capacities. In both cases, a posteriori validations are used. Additionally, if queue capacities are revised then approximations are needed in order to guarantee their integrality.

We believe that a flexible and optimization-friendly model is one that maintains the network topology and its configuration (number of queues and their capacities) as exogenous parameters. We propose such a method. We are also interested in explicitly modeling the blocking phase within our analytical approach. The outputs of this model therefore provide a description of both the causes and the effects of congestion.

## 2.4 Model

In this section, we describe a model that allows the analysis of a network of finite capacity queues. The model accounts for multiple server queues with an arbitrary topology and blocking-after-service. The model is an approximation method based on a decomposition of the network into single queues. Let  $\pi(i)$  denote the stationary distribution of the isolated queue  $i$ . The main aim of our method is to make  $\pi(i)$  a good estimate of the marginal stationary distribution of queue  $i$ .

To ensure tractability, we use classical distributional assumptions and approximations. For a given queue the inter-arrival times, the service times and the times between successive unblockings are assumed to be independent and identically distributed exponential variables. We model each queue as an M/M/c/K queue. This is discussed in detail in Section 2.4.2.5.

We recall the notation introduced so far:

- $\pi(i)$  stationary distribution of queue  $i$ ;
- $\lambda_i$  arrival rate;
- $\mu_i$  service rate of a server;
- $p_{ij}$  transition probability from queue  $i$  to queue  $j$ ;
- $c_i$  number of parallel servers;
- $k_i$  queue capacity (buffer size + number of servers).

### 2.4.1 Global balance equations

The distribution  $\pi(i)$  can be obtained via the global balance equations along with the use of a normalizing constraint:

$$\begin{cases} \pi(i)Q(i) & = 0 \\ \sum_{s \in \mathcal{S}(i)} \pi(i)_s & = 1, \end{cases} \quad \begin{matrix} (2.1a) \\ (2.1b) \end{matrix}$$

where  $\pi(i)_s$  denotes element number  $s$  of  $\pi(i)$ . The global balance equations involve the state space of queue  $i$ ,  $\mathcal{S}(i)$ , as well as the transition rate matrix,  $Q(i)$ , which is a square matrix. We now define these two elements.

#### *State space, $\mathcal{S}(i)$*

Since we are interested in explicitly modeling the blocking phase that a job may go through, we define the processing of a job as follows. A job:

1. arrives to a queue,
2. waits if all the servers are occupied,
3. is served (this is called the active phase),
4. is blocked if the next queue on its path is full (this is called the blocking phase),
5. leaves the queue.

The state of queue  $i$  at any point in time is thus described by the number of active jobs  $A_i$ , blocked jobs  $B_i$  and waiting jobs  $W_i$ . The sample space of this triplet of random variables  $(A_i, B_i, W_i)$  is called the state space and is defined as:  $\mathcal{S}(i) = \{(a, b, w) \in \mathbb{N}^3, a + b \leq c_i, a + b + w \leq k_i\}$ , where  $c_i$  is the number of servers and  $k_i$  is the capacity.

#### *Transition rate matrix, $Q(i)$*

The matrix  $Q(i)$  contains the transition rates between all pairs of states in  $\mathcal{S}(i)$ . The non-diagonal elements,  $Q(i)_{sj}$   $s \neq j$ , represent the rate at which the transition from state  $s$  to state  $j$  takes place. The diagonal elements are defined as:  $Q(i)_{ss} = -\sum_{j \neq s} Q(i)_{sj}$ . Thus  $-Q(i)_{ss}$  represents the rate of departure from state  $s$ . Each equation of the system of global balance equations can be written as:

$$\sum_{j \neq s} \pi(i)_j Q(i)_{js} = -\pi(i)_s Q(i)_{ss}, \quad (2.2)$$

it therefore balances the inflow and the outflow for a given state  $s$ .

We define  $Q(i)$  as a function of the following structural parameters:

- $\lambda_i$ : the arrival rate to queue  $i$ ;
- $\mu_i$ : the service rate of a server at queue  $i$ ;
- $\mathcal{P}_i$ : the probability of being blocked at queue  $i$ ;
- $\tilde{\mu}_{ib}$ : the unblocking rate at queue  $i$  given that there are  $b$  blocked jobs. The vector that considers all possible values of  $b$  is denoted  $\tilde{\mu}_{i\bullet}$ .

These four parameters allow us to describe the transition rates between the different states of queue  $i$ . We write  $Q(i) = f(\lambda_i, \mu_i, \tilde{\mu}_{i\bullet}, \mathcal{P}_i)$ .

As emphasized by Korporaal et al. (2000), the main challenge of models based on decomposition methods is to appropriately approximate these structural parameters so that

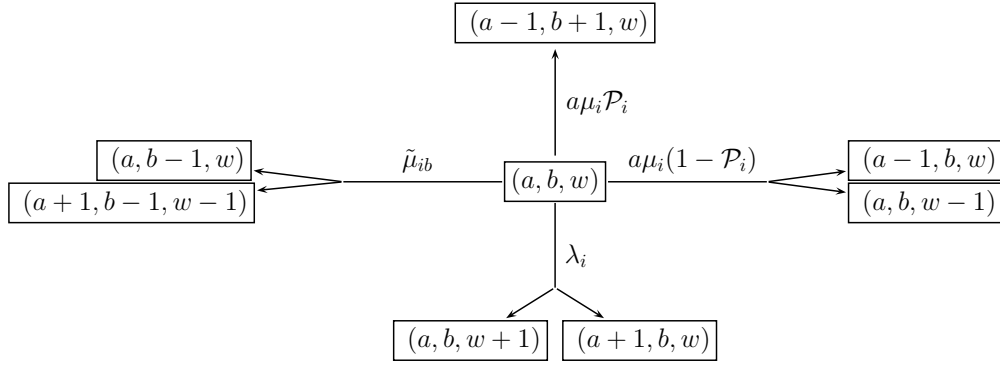


Figure 2.1: Possible transitions from state  $(a, b, w)$ .

initial state $s$	new state $j$	rate $Q(i)_{sj}$	condition
$(a, b, w)$	$(a+1, b, w)$	$\lambda_i$	$a+b+1 \leq c_i$
$(a, b, w)$	$(a, b, w+1)$	$\lambda_i$	$(a+b == c_i) \ \& \ (w+1 \leq k_i - c_i)$
$(a, b, w)$	$(a-1, b, w)$	$a\mu_i(1 - \mathcal{P}_i)$	$w == 0$
$(a, b, w)$	$(a, b, w-1)$	$a\mu_i(1 - \mathcal{P}_i)$	$w \geq 1$
$(a, b, w)$	$(a-1, b+1, w)$	$a\mu_i \mathcal{P}_i$	always possible
$(a, b, w)$	$(a, b-1, w)$	$\tilde{\mu}_{ib}$	$w == 0$
$(a, b, w)$	$(a+1, b-1, w-1)$	$\tilde{\mu}_{ib}$	$w \geq 1$

Table 2.1: Transition rates of queue  $i$ .

$\pi(i)$  is a good estimate of the marginal stationary distribution of queue  $i$ . We now describe how this is done.

### 2.4.2 Transition rates

Assume that queue  $i$  is in a given feasible state  $(a, b, w)$ . The possible transitions with their corresponding rates are displayed both in Figure 2.1 and in Table 2.1. Figure 2.1 displays the four different types of events that can arise: arrivals (characterized by the rate  $\lambda_i$ ), service completion followed by either a departure (with rate  $a\mu_i(1 - \mathcal{P}_i)$ ) or a blocking (with rate  $a\mu_i \mathcal{P}_i$ ), unblockings (with rate  $\tilde{\mu}_{ib}$ ).

Let us describe in more detail the set of states to where transitions can take place, by considering Table 2.1. Queue  $i$  is in a given feasible state  $s$ , such that  $s = (a, b, w)$ . This initial state is tabulated in column 1. The possible states to where a transition can take place are tabulated in the second column, the corresponding transition rate is in the third column, and the conditions under which such a transition can take place are in the last column.

The first two lines of the table distinguish between an arrival that can be served immediately and an arrival that must queue before being served. The next two lines concern the completion of an active phase that is not followed by a blocking phase. In the first case, the freed server remains available. In the second case, the freed server immediately starts serving a job that was waiting. The fifth line concerns jobs that have completed their

service and become blocked. The last two lines relate to the completion of the blocking phase. They differ in whether the server that was blocked stays available or immediately starts serving a job that was waiting. This table describes how we approximate the transition rates using structural parameters. We now describe the approximations used for these structural parameters.

#### 2.4.2.1 Arrival rate, $\lambda_i$

We model each queue as an M/M/c/K queue. For these models, known as *loss models*, all the arrivals that arise while the queue is full are considered to be lost. To characterize the arrival process we introduce the following notation:

- $\lambda_i$ : the total arrival rate to queue  $i$  (includes potentially lost arrivals);
- $\lambda_i^{\text{eff}}$ : the effective arrival rate to queue  $i$  (accounts only for the arrivals that are actually processed, i.e. excludes all lost arrivals);
- $\gamma_i$ : the external arrival rate to queue  $i$ ;
- $p_{ij}$ : transition probability from queue  $i$  to queue  $j$ .

For a given queue  $i$ , we denote by  $N_i$  the total number of jobs at queue  $i$  and by  $P(N_i = k_i)$  the probability that the queue is full, also known as the *blocking probability*. Since for these loss models the arrivals that arise while the queue is full are considered to be lost, the total and the effective arrival rates satisfy the following equation:

$$\lambda_i^{\text{eff}} = \lambda_i(1 - P(N_i = k_i)). \quad (2.3)$$

The arrivals to a given queue are composed of external arrivals (that arise from outside of the network) and internal arrivals (that arise from upstream queues). The total arrival rates are given by:

$$\lambda_i = \gamma_i + \sum_j p_{ji} \lambda_j^{\text{eff}}. \quad (2.4)$$

Note that if no arrivals are lost (e.g. for infinite capacity queues), the total arrival rates satisfy the classical *flow conservation* equations, also known as the *traffic equations*:

$$\lambda_i = \gamma_i + \sum_j p_{ji} \lambda_j. \quad (2.5)$$

By multiplying Equation (2.4) by  $(1 - P(N_i = k_i))$ , the effective arrival rates satisfy:

$$\lambda_i^{\text{eff}} = \gamma_i(1 - P(N_i = k_i)) + \sum_j p_{ji} \lambda_j^{\text{eff}}(1 - P(N_i = k_i)). \quad (2.6)$$

This equation states that the effective arrival rate is composed of the proportion of non-lost external arrivals and non-lost internal arrivals. In order to model congestion, i.e. the blocking of jobs within the network, we consider that only external arrivals are lost, whereas internal arrivals are blocked at their current location if the target queue is full. Thus, we approximate the effective arrival rates by:

$$\lambda_i^{\text{eff}} \approx \gamma_i(1 - P(N_i = k_i)) + \sum_j p_{ji}\lambda_j^{\text{eff}}. \quad (2.7)$$

#### 2.4.2.2 Probability of being blocked, $\mathcal{P}_i$

The probability of being blocked at queue  $i$ ,  $\mathcal{P}_i$ , helps us to describe the rate at which a job gets blocked after service. It is approximated by the weighted average of the blocking probabilities of all target queues:

$$\mathcal{P}_i \approx \sum_j p_{ij}P(N_j = k_j). \quad (2.8)$$

#### 2.4.2.3 Service and effective service rates, $\mu_i$ and $\mu_i^{\text{eff}}$

The total time spent by a job in front of a server, called the effective service time, is composed of the service time (active phase) and for some jobs of the blocked time (blocking phase). The expected effective service time is denoted  $1/\mu_i^{\text{eff}}$ .

Let  $T_{i,j}^B$  denote the blocked time of a job at queue  $i$  conditional on it being blocked by queue  $j$ , then the effective service rate is given by:

$$\frac{1}{\mu_i^{\text{eff}}} = \frac{1}{\mu_i} + \sum_j p_{ij}P(N_j = k_j)E[T_{i,j}^B]. \quad (2.9)$$

This equation states that a given job at queue  $i$  has an expected service time of  $1/\mu_i$ , and if its target queue  $j$  is full, then it is blocked with probability  $P(N_j = k_j)$  and has an expected blocked time of  $E[T_{i,j}^B]$ .

We approximate the expected blocked time due to all target queues by a common value, denoted  $E[T_i^B]$ , this leads to:

$$\frac{1}{\mu_i^{\text{eff}}} \approx \frac{1}{\mu_i} + \sum_j p_{ij}P(N_j = k_j)E[T_i^B] = \frac{1}{\mu_i} + \mathcal{P}_i E[T_i^B]. \quad (2.10)$$

This approximation is appropriate if the blocked times due to the different target queues have the same magnitude, otherwise it is not appropriate. Thus, the approximation for the

effective service rate is given by:

$$\frac{1}{\mu_i^{\text{eff}}} \approx \frac{1}{\mu_i} + \mathcal{P}_i E[T_i^B]. \quad (2.11)$$

In this equation,  $\mu_i$  is an exogenous parameter, the approximation of  $\mathcal{P}_i$  is given in Equation (2.8), and that of  $E[T_i^B]$  is detailed in Section 2.4.2.6 (Equation (2.25)).

#### 2.4.2.4 Unblocking rate, $\tilde{\mu}_{ib}$

We now describe how we approximate  $\tilde{\mu}_{ib}$ , the unblocking rate at queue  $i$  given that there are  $b$  blocked jobs. Suppose that queue  $i$  is in the state  $(a, b, w)$ . Then the service rate of the queue is  $a\mu_i$ , i.e. the active jobs are being processed by  $a$  **parallel** servers. In the state  $(a, b, w)$ , there are  $b$  blocked servers, but they do not all work in parallel, as we now describe.

Let  $D(i, b)$  denote the number of distinct target queues that are blocking the  $b$  jobs at queue  $i$ . Each target queue unblocks jobs at queue  $i$  at its own rate, which we call the acceptance rate of blocked jobs. We approximate the acceptance rate of all target queues by the average acceptance rate (the average is taken across the different target queues), denoted  $\tilde{\mu}_i^a$ .

Thus, if all  $b$  jobs are blocked by the same target queue, then they can be seen as forming a virtual queue in front of the blocking queue with a FIFO unblocking mechanism. The unblocking rate at queue  $i$  is then  $\tilde{\mu}_i^a$ . If the jobs are blocked by  $D(i, b)$  distinct target queues then they can be seen as forming  $D(i, b)$  virtual **parallel** queues, each with a FIFO unblocking mechanism. The unblocking rate at queue  $i$  is then  $D(i, b)\tilde{\mu}_i^a$ . More specifically, we have:

$$\frac{1}{\tilde{\mu}_{ib}} \approx \sum_{d=1}^{\min(b, \text{card}(\mathcal{I}^+))} P(D(i, b) = d) \frac{1}{d \tilde{\mu}_i^a}, \quad (2.12)$$

where  $\mathcal{I}^+$  represents the set of target queues of queue  $i$ , and  $\text{card}(\mathcal{I}^+)$  is its cardinality. Equation (2.12) is an approximation because we approximate the acceptance rate of the different target queues by a common acceptance rate,  $\tilde{\mu}_i^a$ . The approximation for  $P(D(i, b) = d)$  is described in Section 2.4.2.6 and involves only exogenous parameters. Thus we write  $\tilde{\mu}_{ib}$  in the form:

$$\tilde{\mu}_{ib} \approx \tilde{\mu}_i^a \phi(i, b), \quad (2.13)$$

where  $\phi(i, b)$  is exogenous and can be interpreted as the expected number of distinct target queues that are blocking the  $b$  jobs at queue  $i$  ( $\phi(i, b)$  is defined in Section 2.4.2.6 by Equation (2.20)). We now describe how we derive  $\tilde{\mu}_i^a$ .

The acceptance rate of blocked jobs,  $\tilde{\mu}_i^a$

The scalar  $\tilde{\mu}_i^a$  denotes the rate at which a target queue of queue  $i$  accepts (i.e. unblocks) jobs that are blocked at queue  $i$ . We denote by:

- $\tilde{p}_{ij}$ : the transition probabilities conditional on a job being blocked at queue  $i$ , i.e.  
 $\tilde{p}_{ij} = p_{ij}P(N_j = k_j)/\mathcal{P}_i$ .
- $r_{ij}$ : the proportion of arrivals to queue  $j$  that arise from blocked jobs at queue  $i$ , i.e.  
 $r_{ij} = \tilde{p}_{ij}\lambda_i^{\text{eff}}/\lambda_j^{\text{eff}}$ .

Suppose queue  $j$  is blocking jobs at predecessor queues. It is therefore full and is serving at rate  $\mu_j^{\text{eff}}c_j$ . It accepts jobs that are blocked at queue  $i$  at the rate  $r_{ij}\mu_j^{\text{eff}}c_j$ . By averaging over the possible target queues of queue  $i$  we obtain:

$$\frac{1}{\tilde{\mu}_i^a} = \sum_{j \in \mathcal{I}^+} \tilde{p}_{ij} \frac{1}{r_{ij}\mu_j^{\text{eff}}c_j} = \sum_{j \in \mathcal{I}^+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}}\mu_j^{\text{eff}}c_j}. \quad (2.14)$$

#### 2.4.2.5 Distributional approximations

**Inter-arrival times.** The arrivals to a given queue are composed of external and internal arrivals. The arrival process is therefore determined by the external arrival process and the departure process of upstream queues. For a given queue  $i$  with finite capacity where blocking can arise, the departure process of its upstream queues is a function of its blocking probability,  $P(N_i = k_i)$ . Thus, determining the exact distribution of the inter-arrival times is an intricate task.

For the classical infinite capacity Jackson networks, if the network topology contains cycles (i.e. it is not a feed-forward network), then the arrival process to any queue that belongs to a cycle is not Poisson (Melamed, 1979; Burke, 1976). Furthermore, for finite capacity queues with blocking the inter-arrival times are no longer independent. In order to derive a tractable formulation, we approximate the inter-arrival times of a given queue as exponentially distributed and independent random variables with expectation  $1/\lambda_i$ .

**Service and unblocking times.** Service times are assumed to be independent and identically distributed exponential variables with expectation  $1/\mu_i$ . If service times are exponential, then the times between successive unblockings consist of a sum of exponential variables with different scale parameters, i.e. they follow phase-type distributions.

The use of phase-type distributions provides a detailed description of the unblocking mechanisms. This approach has been used in some of the existing methods (Jun and Perros, 1989; Altioek, 1989). Nonetheless, constructing these phase-type service mechanisms requires enumerating all possible blocking configurations, which is a CPU intensive task. This approach is therefore limited to small networks (small in the

number of queues and in the number of servers per queue). We approximate the times between successive unblockings as independent and identically distributed exponential variables with expectation  $1/\tilde{\mu}_{ib}$ .

By explicitly modeling both the service and the blocking phase, the number of jobs in front of the servers becomes a two dimensional system  $(a, b)$  composed of the active and the blocked jobs. We are thus in the presence of an M/M/c/K model with a three-dimensional state space  $(a, b, w)$ .

#### 2.4.2.6 Derivation of $P(D(i, b) = d)$ and of $E[T_i^B]$

##### Approximation of $P(D(i, b) = d)$

$P(D(i, b) = d)$  represents the probability that  $d$  distinct queues are blocking the  $b$  blocked jobs at queue  $i$ . Consider  $R(i, b, d)$  the random vector containing the  $b$  target queues of the blocked jobs,  $d$  of which are distinct, and let  $\mathcal{R}(i, b, d)$  be its sample space.

In order to derive an expression for  $P(D(i, b) = d)$ , we sum over all possible realizations of  $R(i, b, d)$ .

$$P(D(i, b) = d) = \sum_{r \in \mathcal{R}(i, b, d)} P(R(i, b, d) = r) = \sum_{r \in \mathcal{R}(i, b, d)} \tilde{p}_{ir_1} \tilde{p}_{ir_2} \dots \tilde{p}_{ir_b} = \sum_{r \in \mathcal{R}(i, b, d)} \prod_{j \in \mathcal{I}^+} \tilde{p}_{ij}^{\ell(i, b, d)_j}, \quad (2.15)$$

where  $\ell(i, b, d)_j$  is the number of jobs blocked by queue  $j$  at queue  $i$  (given that there are a total of  $b$  blocked jobs that are blocked by  $d$  distinct target queues). This last equation shows that for a given realization of  $R(i, b, d)$ , what is of interest in determining  $P(D(i, b) = d)$  is the occurrence of each target queue (i.e. the vector  $\ell(i, b, d)$ ), the ordering of the target queues is not important. Thus, instead of summing over  $\mathcal{R}(i, b, d)$ , we sum over the set of  $\ell(i, b, d)$  vectors. This reduces the size of the space over which we sum. The set of such vectors is noted  $\mathcal{L}(i, b, d)$  and is defined by:

$$\ell(i, b, d) \in \mathcal{L}(i, b, d) \Leftrightarrow \begin{cases} \sum_{j \in \mathcal{I}^+} \ell(i, b, d)_j = b \\ \sum_{j \in \mathcal{I}^+} \mathbf{1}(\ell(i, b, d)_j > 0) = d \\ \ell(i, b, d)_j \geq 0 \quad \forall j \in \mathcal{I}^+, \end{cases} \quad (2.16)$$

where  $\mathbf{1}(x)$  is the indicator function. The first equation of the System of Equations (2.16) means that there are a total of  $b$  jobs blocked at queue  $i$ . The second means that these jobs are blocked by  $d$  different target queues.

For a given vector  $\ell(i, b, d)$  that satisfies the System of Equations (2.16), there are  $b! / (\prod_{j \in \mathcal{I}^+} \ell(i, b, d)_j!)$  different realizations of  $R(i, b, d)$  that are associated with it. This

corresponds to the number of permutations of a vector of  $b$  elements where element  $j$  is repeated  $\ell(i, b, d)_j$  times. Therefore we obtain:

$$P(D(i, b) = d) = \sum_{\ell(i, b, d) \in \mathcal{L}(i, b, d)} \frac{b!}{\prod_{j \in \mathcal{I}^+} \ell(i, b, d)_j!} \prod_{j \in \mathcal{I}^+} \tilde{p}_{ij}^{\ell(i, b, d)_j}. \quad (2.17)$$

Coming back to Equation (2.12) and replacing  $P(D(i, b) = d)$  by the expression that we have just derived, we obtain:

$$\frac{1}{\tilde{\mu}_{ib}} \approx \frac{1}{\tilde{\mu}_i^a} \sum_{d=1}^{\min(b, \text{card}(\mathcal{I}^+))} \frac{1}{d} \sum_{\ell(i, b, d) \in \mathcal{L}(i, b, d)} \frac{b!}{\prod_{j \in \mathcal{I}^+} \ell(i, b, d)_j!} \prod_{j \in \mathcal{I}^+} \tilde{p}_{ij}^{\ell(i, b, d)_j}. \quad (2.18)$$

The size of the space  $\mathcal{L}(i, b, d)$  is still considerably large. Therefore, when approximating  $\tilde{\mu}_{ib}$  we use an exogenous approximation of  $\tilde{p}_{ij}$ :

$$\tilde{p}_{ij} = \frac{p_{ij} P(N_j = k_j)}{\mathcal{P}_i} = \frac{p_{ij} P(N_j = k_j)}{\sum_l p_{il} P(N_l = k_l)} \approx \frac{p_{ij}}{\sum_l p_{il}}. \quad (2.19)$$

This approximation makes both summations of Equation (2.18) exogenous. These two summations are therefore evaluated only once when solving the entire system of equations. This approximation is appropriate if the blocking probabilities of the target queues have the same magnitude, otherwise it is inadequate. The only endogenous parameter remaining in Equation (2.18) is  $\tilde{\mu}_i^a$ . Thus, we have written  $\tilde{\mu}_{ib}$  in the form:  $\tilde{\mu}_{ib} \approx \tilde{\mu}_i^a \phi(i, b)$ , where

$$\frac{1}{\phi(i, b)} = \sum_{d=1}^{\min(b, \text{card}(\mathcal{I}^+))} \frac{1}{d} \sum_{\ell(i, b, d) \in \mathcal{L}(i, b, d)} \frac{b!}{\prod_{j \in \mathcal{I}^+} \ell(i, b, d)_j!} \prod_{j \in \mathcal{I}^+} \left( \frac{p_{ij}}{\sum_k p_{ik}} \right)^{\ell(i, b, d)_j}. \quad (2.20)$$

### Approximation of $E[T_i^B]$

Given a blocked job at queue  $i$ ,  $E[T_i^B]$  represents its expected blocked time. Recall that  $B_i$  denotes the number of blocked jobs at queue  $i$ . We derive an expression for  $E[T_i^B]$  by conditioning on the length of the blocked queue:

$$\begin{aligned} E[T_i^B] &= E[E[T_i^B \mid B_i]] = \sum_{b \geq 0} P(B_i = b \mid B_i > 0) E[T_i^B \mid B_i = b] \\ &= \sum_{b \geq 1} \frac{P(B_i = b)}{P(B_i > 0)} E[T_i^B \mid B_i = b]. \end{aligned} \quad (2.21)$$

Let  $T(i, b)_j$  denote the blocked time of the job that was unblocked in  $j^{\text{th}}$  position given

that there were  $b$  blocked jobs. We have:

$$E[T_i^B \mid B_i = b] = \frac{1}{b} \sum_{j=1}^b E[T(i, b)_j]. \quad (2.22)$$

We know that the expected time between successive departures given that there are  $b$  blocked jobs at queue  $i$  is represented by  $1/\tilde{\mu}_{ib}$ , thus the expected blocked time of the first job to be unblocked is given by  $1/\tilde{\mu}_{ib}$ , that of the second job to be unblocked by  $1/\tilde{\mu}_{ib} + 1/\tilde{\mu}_{i(b-1)}$  and that of the  $j^{th}$  by:

$$E[T(i, b)_j] = \sum_{k=b-j+1}^b \frac{1}{\tilde{\mu}_{ik}}. \quad (2.23)$$

Putting the last two equations together and then interchanging the summations, we obtain:

$$E[T_i^B \mid B_i = b] = \frac{1}{b} \sum_{j=1}^b \sum_{k=b-j+1}^b \frac{1}{\tilde{\mu}_{ik}} = \frac{1}{b} \sum_{k=1}^b \frac{1}{\tilde{\mu}_{ik}} \sum_{j=b-k+1}^b 1 = \frac{1}{b} \sum_{k=1}^b \frac{k}{\tilde{\mu}_{ik}}. \quad (2.24)$$

Therefore our expression for  $E[T_i^B]$  is given by:

$$E[T_i^B] = \sum_{b \geq 1} \frac{P(B_i = b)}{P(B_i > 0)} \sum_{k=1}^b \frac{k}{b} \frac{1}{\tilde{\mu}_{ik}}. \quad (2.25)$$

### 2.4.3 System of equations

The main aim is to obtain the stationary distributions of each queue,  $\pi(i)$ . The main equations consist of the global balance equations (Equations (2.1)), which require the definition of the transition rate matrix (Table 2.1). We have directly implemented these equations as a single set:

$$\pi(i)g(\lambda_i, \mu_i, \tilde{\mu}_{i\bullet}, \mathcal{P}_i) = 0. \quad (2.26)$$

The system of nonlinear equations is given (on the following page) by:

$$\begin{aligned}
& \left\{ \begin{aligned} \pi(i)g(\lambda_i, \mu_i, \tilde{\mu}_{i\bullet}, \mathcal{P}_i) &= 0 & (2.27a) \\ \lambda_i &= \lambda_i^{\text{eff}} / (1 - P(N_i = k_i)) & (2.27b) \\ \lambda_i^{\text{eff}} &= \gamma_i(1 - P(N_i = k_i)) + \sum_j p_{ji} \lambda_j^{\text{eff}} & (2.27c) \\ \frac{1}{\mu_i^{\text{eff}}} &= \frac{1}{\mu_i} + \mathcal{P}_i E[T_i^B] & (2.27d) \\ \tilde{\mu}_{ib} &= \tilde{\mu}_i^a \phi(i, b) & (2.27e) \\ \frac{1}{\tilde{\mu}_i^a} &= \sum_{j \in \mathcal{I}^+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}} \mu_j^{\text{eff}} c_j} & (2.27f) \\ \mathcal{P}_i &= \sum_j p_{ij} P(N_j = k_j) & (2.27g) \\ E[T_i^B] &= \sum_{b \geq 1} \frac{P(B_i = b)}{P(B_i > 0)} \sum_{k=1}^b \frac{k}{b} \frac{1}{\tilde{\mu}_{ik}}. & (2.27h) \end{aligned} \right.
\end{aligned}$$

It is solved simultaneously for all queues. For each queue, the exogenous parameters are  $c_i, k_i, p_{ij}, \mu_i, \gamma_i, \phi(i, b)$ . The system of equations has been implemented in terms of six endogenous parameters:  $\lambda_i, \tilde{\mu}_i^a, \mu_i^{\text{eff}}, \mathcal{P}_i, P(N_i = k_i), P(B_i > 0)$ . For a given queue, the dimension of its distribution is equal to  $\text{card}(\mathcal{S}_i) = (c_i + 1)(k_i + 1 - \frac{c_i}{2})$ . Thus the total size of the system of equations is:  $\sum_i ((c_i + 1)(k_i + 1 - \frac{c_i}{2}) + 6)$ .

Existing methods that require a posteriori validations, (e.g. to ensure the integrality of endogenous queue capacities) resort to iterative methods. For a given iteration the system of equations for each queue is solved sequentially. Since our method requires no a posteriori validations we are able to solve the set of equations associated to all queues simultaneously.

The system is solved by using the Matlab routine *fsolve*, which implements a trust-region dogleg algorithm based on the method described by Powell (1970). The Jacobian of the system has been calculated analytically and implemented. In order to ensure the positivity of the distributions, the system of equations has been implemented in terms of an auxiliary variable  $y(i)$  such that  $y(i)^2 = \pi(i)$ .

For a given tolerance,  $tol$ , convergence is attained when either the absolute values of all equations are smaller than  $tol$  or when both the sum of squares of the system of equations is smaller than  $\sqrt{tol}$  and the change of its relative value is smaller than  $\max(tol^2, eps)$ , where  $eps$  is the machine precision which is of magnitude  $10^{-16}$ . This choice is based on the criteria given in Dennis and Schnabel (1996).

The endogenous parameters are initialized as follows. The arrival rates,  $\lambda$ , are initialized using the arrival rates that satisfy the classical flow conservation laws. The distributions,  $\pi$ , are initialized using uniform distributions, thus no a priori information concerning the stationary behavior of the queues is required, but such information could be used if available.

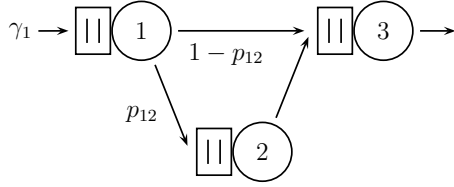


Figure 2.2: Triangular topology.

scenario	1	2	3	4	5	6	7	8	9	10
$\mu_1$	1	1	1	1	1	1	1	1	1	1
$\mu_2$	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
$\mu_3$	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3

Table 2.2: Increasing service rate scenarios.

The other endogenous parameters are deduced from these initializations.

## 2.5 Validation

### 2.5.1 Validation versus existing methods

#### Triangular topology

We first compare our method with that of Altıok and Perros (1987) and that of Takahashi et al. (1980). The latter considered a single server network with triangular topology (depicted in Figure 2.2) and the following configuration:  $p_{12} = \frac{1}{2}$ ,  $\gamma_1 = 1$ . They considered two cases according to the buffer size of the queues: a null buffer and a buffer of size two. For each case, they considered a set of scenarios with increasing service rates for queues two and three. These scenarios are displayed in Table 2.2. The chosen performance measure was the blocking probability of queue one,  $P(N_1 = k_1)$ .

They then compared their estimates with either simulation results or with exact results derived by using the global balance equations of the entire network. The relative error of the estimates of the different methods are displayed in Figure 2.3. The left plot considers a null buffer, and the right plot considers a buffer of size two. For both cases, all methods yield accurate estimates, the relative error remaining under 7% for the first case and 4% for the second case. We yield similar estimates to those of Takahashi et al. (1980). For the first case, Altıok and Perros (1987) yields the most accurate estimates.

#### Two queues in a tandem topology

Bell (1982) derived a theoretical upper bound on the mean throughput rate of M/M/c/K networks. By considering two queues in a tandem topology under a set of scenarios with varying queue capacities, he showed that several models based on decomposition methods violate this upper bound. We compare the mean throughput estimates of our method with the methods of Singh and Smith (1997), Kerbache and Smith (1988), Boxma and Konheim (1981), Takahashi et al. (1980) and Hillier and Boling (1967).

The configuration of the network is:  $\mu_1 = 3, \mu_2 = 1, c_1 = c_2 = 1$ , and  $\gamma_1 = 1$ . The different scenarios are given in Table 2.3 and the mean throughput estimates of the various methods are depicted in Figure 2.4. Our mean throughput is estimated by using the effective

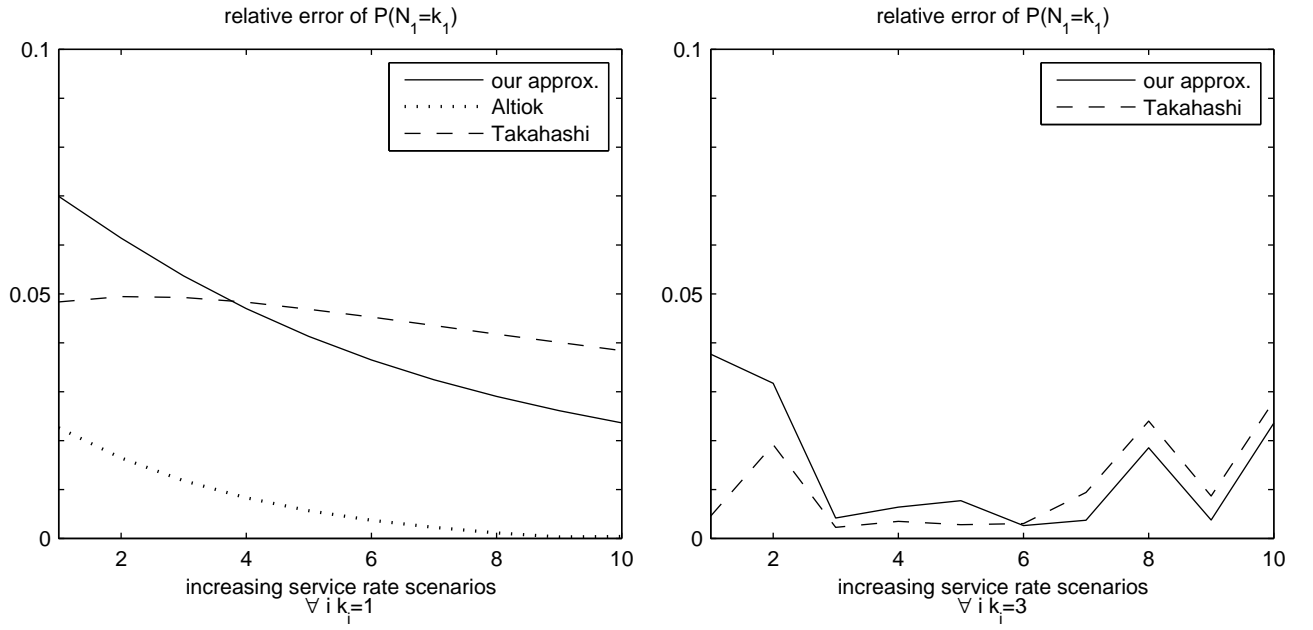


Figure 2.3: Comparison with the methods of Altiok and Perros (1987) and of Takahashi et al. (1980) under two capacity configurations.

scenario	1	2	3	4	5	6	7	8	9
$k_1 - c_1$	1	1	2	2	2	3	4	5	10
$k_2 - c_2$	1	2	1	2	3	3	4	5	10

Table 2.3: Increasing buffer size scenarios.

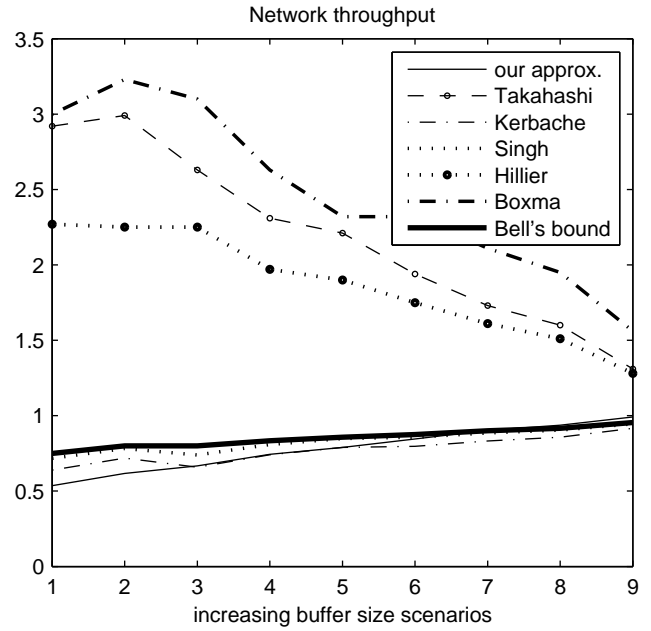


Figure 2.4: Comparison of the mean throughput estimate of various decomposition methods with the theoretical upper bound derived by Bell (1982).

scenario	1	2	3	4
$\mu_1$	1	1.2	1	1
$\mu_2$	1	1	1.2	1
$\mu_3$	1	1	1	1.2

Table 2.4: Service rate scenarios

scenario	1	2	3	4
$\gamma_1$	0.5	1	1.5	2

Table 2.5: External arrival rate scenarios with increasing congestion

departure rate at queue two,  $\lambda_2^{\text{eff}}$ .

Figure 2.4 shows that our mean throughput estimate remains near the upper bound, and is similar to that of the Expansion Method of Singh and Smith (1997) and Kerbache and Smith (1988). For the last three scenarios, it violates the bound by 0.3%, 2.2% and 3.8%, respectively. Our method therefore yields consistent throughputs unlike the methods of Takahashi et al. (1980), Hillier and Boling (1967) and Boxma and Konheim (1981).

## 2.5.2 Validation versus exact results

In this section, we compare the queue length distributional estimates derived by the proposed method with the exact distributions. The exact distributions are obtained by numerically solving the global balance equations of the full networks. This requires enumerating all possible network states, as well as all possible transitions with their corresponding rates.

We consider networks with three bufferless queues, where each queue has two servers. We consider two topologies: a tandem topology and a triangular topology (depicted in Figure 2.2, with  $p_{12} = \frac{1}{2}$ ).

### Scenarios with varying levels of congestion

For each topology, we consider a set of four service rate scenarios, which are displayed in Table 2.4. For each service rate scenario, we consider four external arrival rate scenarios, which are given in Table 2.5. That is, for each topology we consider a total of 16 scenarios with varying congestion levels and varying bottleneck locations.

For each one of these 16 scenarios, we compare the queue length distributions of each queue. For both networks and all 16 scenarios, we evaluate the difference between the proposed queue length probabilities and their exact values. These are referred to as the errors of the distributional estimates.

Figure 2.5 displays a histogram of these errors. There are a total of 288 error values, 50% of the absolute errors are smaller than 0.009, 70% smaller than 0.02 and 90% smaller than 0.042. This figure confirms that for scenarios with different levels and locations of congestion, the distributional estimates provided by the proposed approximation method remain accurate when compared to the exact distributions.

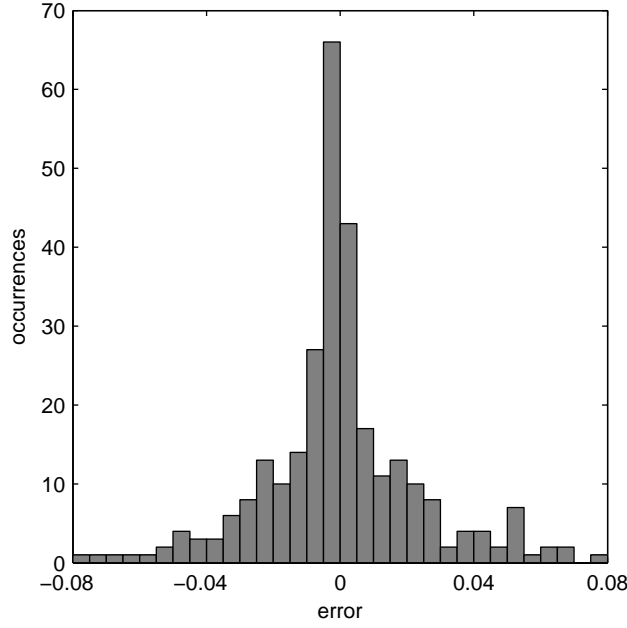


Figure 2.5: Histogram of the errors of the queue length probabilities for scenarios with varying levels of congestion.

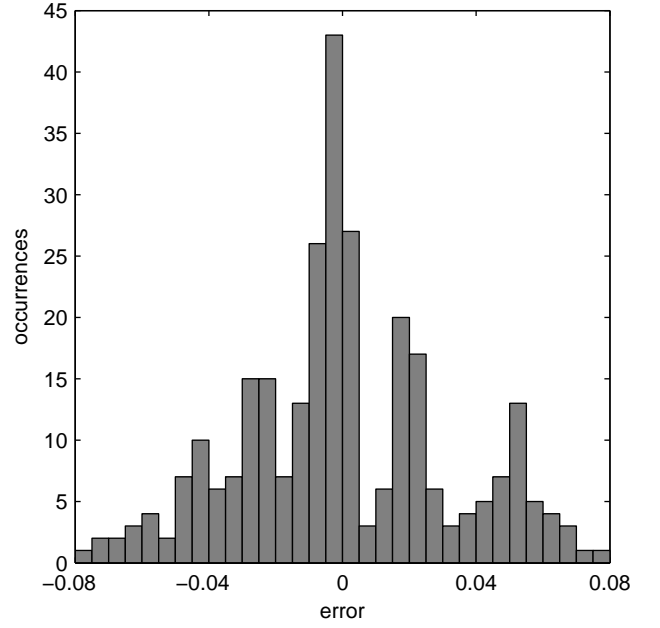


Figure 2.6: Histogram of the errors of the queue length probabilities for highly congested scenarios

scenario	1	2	3	4
$\gamma_1$	1.5	1.7	1.9	2

Table 2.6: Highly congested external arrival rate scenarios

### Highly congested scenarios

To investigate further the performance of the proposed method, we consider highly congested scenarios. We consider the same service rate scenarios (Table 2.4). The demand scenarios are given in Table 2.6.

Once again, for both networks and all 16 scenarios, we evaluate the difference between the proposed queue length probabilities and their exact values. The histogram of these errors is presented in Figure 2.6. Of the 288 values 50% are smaller in absolute value than 0.019, 70% smaller than 0.032 and 90% smaller than 0.055.

We now consider in detail the scenarios with the highest congestion. These correspond to service rate scenario 1 (Table 2.4) and all four arrival rate scenarios (Table 2.6). Figure 2.7 displays the queue length distributions obtained by exact evaluation and by the proposed method. This figure consists of nine plots.

Each row of plots corresponds to a given queue. The upper row plots display the distributions of the first queue (the most upstream). The middle and the lower row plots display the distributions of queues two and three, respectively.

Each column of plots corresponds to a given state. The first column of plots considers the probability that the queue is empty,  $P(N_i = 0)$ . Columns two and three display the

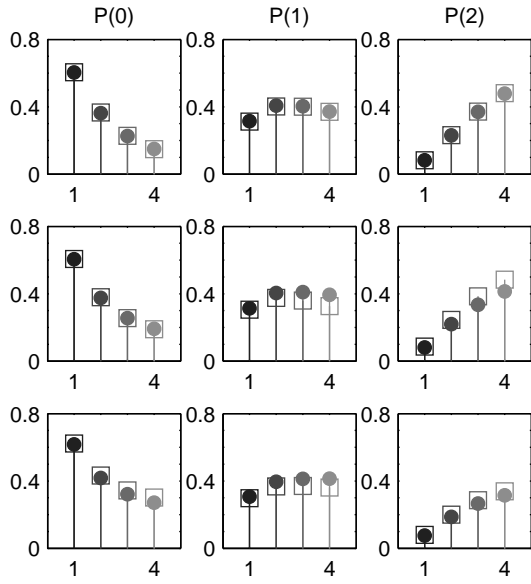


Figure 2.7: Tandem topology. Queue length probabilities for highly congested scenarios.

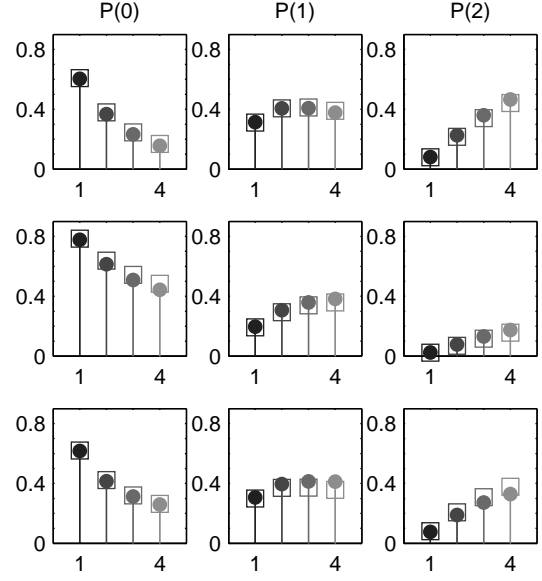


Figure 2.8: Triangular topology. Queue length probabilities for highly congested scenarios.

probabilities  $P(N_i = 1)$  and  $P(N_i = 2)$ , respectively.

Each plot displays four probabilities, which correspond to the four arrival rate scenarios (Table 2.6). The probabilities have a lighter color as congestion increases.

To summarize, each plot displays the probability that a given queue is in a given state under increasing arrival rate scenarios. The estimates of the proposed method are denoted by filled circles, whereas the exact probabilities are represented by empty squares.

Considering the same scenarios, the queue length probabilities of the triangular network are presented in Figure 2.8. Figures 2.7 and 2.8 both illustrate that for highly congested scenarios, as congestion increases the proposed queue length distributions accurately follow the trend of the exact distributions. Overall, the proposed method leads to accurate estimates for the queue length distributions, under a variety of supply and demand scenarios.

### 2.5.3 Validation versus simulation results

Of main interest in our method are the marginal distributional estimates, which allow us to derive the performance measures that describe congestion. These could not be compared with existing methods because we know of no method that defines the state space in such a way. We resort to simulation results in order to validate our method on a larger set of scenarios and topologies.

We consider three different topologies. Each network consists of nine queues, all of which are bufferless with three servers. For each network, we consider a set of scenarios with increasing external arrival rates. The network configurations and scenario definitions

	$i:$	1	2	3	4	5	6	7	8	9
Network A	$\gamma_i$	-	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0
	$\mu_i$	0.3	0.3	0.3	0.1	0.01	0.014	0.1	0.4	0.5
	scenario	1	2	3	4					
	$\gamma_1$	0.1	0.2	0.3	0.4					
	$\gamma_7$	0.1	0.3	0.5	0.7	0.9				
	$i:$	1	2	3	4	5	6	7	8	9
Network B	$\gamma_i$	-	0	0	0	0	0	-	0	0
	$\mu_i$	0.3	0.3	0.3	0.6	0.6	0.6	0.3	0.3	0.3
	scenario	1	2	3	4	5				
	$\gamma_1$	0.1	0.3	0.5	0.7	0.9				
	$\gamma_7$	0.1	0.3	0.5	0.7	0.9				
	$i:$	1	2	3	4	5	6	7	8	9
Network C	$\gamma_i$	-	0	0	0	0	0	0	0	0
	$\mu_i$	0.3	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.3
	scenario	1	2	3	4	5				
	$\gamma_1$	0.1	0.3	0.5	0.7	0.9				
	$\gamma_7$	0.1	0.3	0.5	0.7	0.9				

Table 2.7: Configuration and scenario definitions for networks A, B and C.

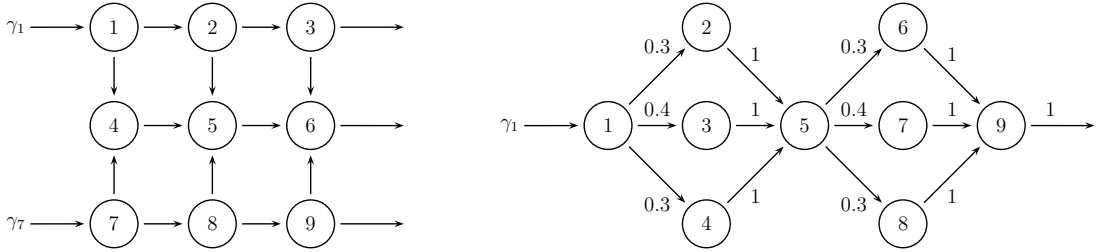


Figure 2.9: Topologies of networks B and C (left and right hand side, respectively).

of networks A, B and C are displayed in Table 2.7.

Network A is a simplified version of the case study network presented in Section 3.1. Its topology and transition probabilities are the same as that of the case study. They are displayed in Figure 3.1 and Table 3.2, respectively. The simplifications with regards to the case study concern the number of servers per queue and the external arrival rates. The topologies of networks B and C are displayed in Figure 2.9. For a given queue of network B, the transition probabilities are uniformly distributed among the possible target queues. For network C, the transition probabilities are indicated in Figure 2.9.

In order to validate our results, we developed the corresponding simulation models using a discrete event simulator, ProModel version 4.1 (ProModel, 1997). Let  $t_o$  denote the temporal unit of the transition rates (e.g. minutes, hours). The simulation runs consist of 20 replications with a warm-up time of 10000  $t_o$  and further run time of 40000  $t_o$ .

For all three networks, all scenarios, queues and states, we consider the errors of the distributional estimates:  $\pi(i)_{(a,b)} - \pi^*(i)_{(a,b)}$ , where  $\pi(i)_{(a,b)}$  denotes our estimate of the probability that queue  $i$  is in state  $(a,b)$  and  $\pi^*$  is the simulation estimate. Figure 2.10 displays a histogram of the errors of the distributional estimates. There are a total of 1200 estimates. 70% of the absolute errors are smaller than 0.0065, 80% smaller than 0.0129 and

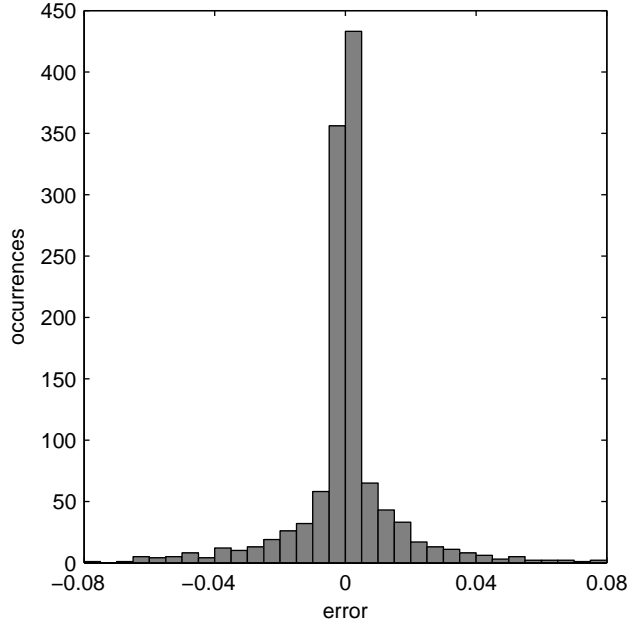


Figure 2.10: Histogram of the errors of the distributional estimates for all scenarios of networks A, B and C.

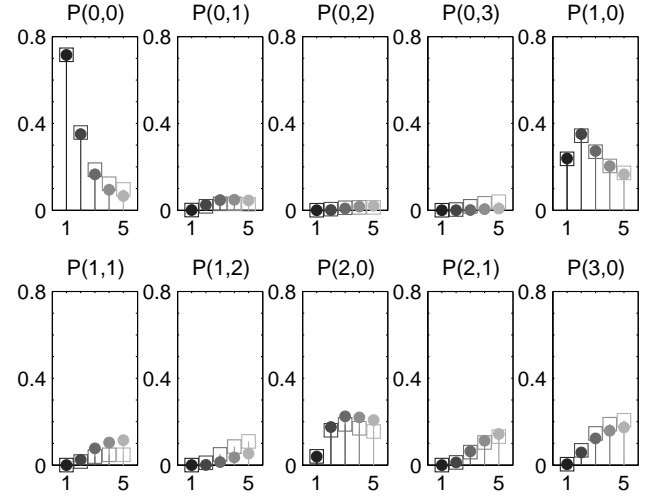


Figure 2.11: Distribution of queue 5 for network C across all scenarios.

90% smaller than 0.0245. Our method therefore yields accurate distributional estimates.

In order to illustrate the blocking information derived by our method we consider the scenarios of network C (Table 2.7). Figure 2.11 displays the estimates of the distribution of queue five given by our method and those obtained via simulation. Each plot considers a given state  $(a, b)$  and plots  $\pi(5)_{(a,b)}$  and  $\pi^*(5)_{(a,b)}$  for all scenarios. The simulated distribution is depicted as empty squares, whereas our estimates are represented by filled circles. The scenarios are in a lighter color as the external arrival rate of queue one increases.

The figure shows that as the external arrival rate increases the states with blocked jobs become more likely, e.g. states  $(a, b)$  in  $\{(1, 1), (1, 2), (2, 1)\}$ . Take for example state  $(2, 1)$  where there are two active jobs and one blocked job. The probability  $P(2, 1)$  gradually increases from zero at scenario one to 0.14 at scenario five. For all states, our estimates follow the trend of the simulated probabilities. Overall the estimates are very accurate.

## 2.5.4 Convergence of the validation runs

The stopping criterion is detailed in Section 2.4.3. The tolerance is chosen as  $tol = 10^{-6}$ . This choice is based on the criteria given in Dennis and Schnabel (1996). If after 150 iterations there is no convergence the run is stopped and initialized again with a new starting point.

A description of the convergence of the algorithm under the different validation runs is tabulated in Table 2.8. Columns two and three contain the average number of initializations

Case		number of				time [sec]	total number of scenarios	
		initializations		iterations				
Triangular	bufferless	1	(0)	7	(1)	0.08	(0.02)	10
	buffer of size 2	7	(4)	65	(13)	0.47	(0.1)	10
Two queues in tandem		3	(7)	37	(51)	0.2	(0.2)	9
Networks A, B and C		10	(11)	57	(46)	1.53	(1.1)	14

Table 2.8: Convergence of validation runs.

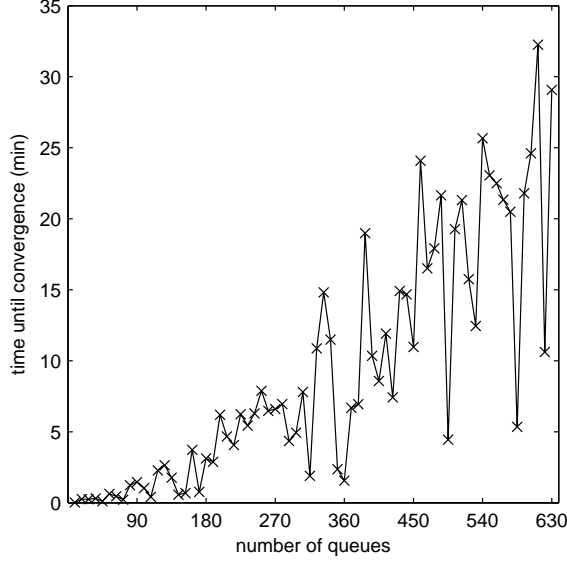


Figure 2.12: Time until convergence.

required until convergence and their standard errors, respectively. Columns four to six concern the converged run. They give the average number of iterations, their standard errors, the average execution time and their standard errors, respectively. Column seven gives the total number of considered scenarios.

### 2.5.5 Tests on larger networks

In order to further evaluate the speed of our method, we have applied it to a set of larger networks. We use network C as a building block. We construct the full networks by putting a set of C networks in a tandem configuration. We evaluated 70 networks, where the  $n^{th}$  network has  $n$  instances of network C in tandem. This corresponds to networks with 9 to 630 queues. Only the first queue has external arrivals,  $\gamma_1 = 0.3$ . Recall that the distributions  $\pi$  are initialized with the uniform distribution (Section 2.4.3). Note that in practice a priori information would be used to initialize  $\pi$ . The average number of iterations required until convergence was 275 with a standard deviation of 125. Figure 2.12 displays the time until convergence across the networks in minutes.

## 2.6 Conclusions and future work

We have presented an analytic queueing network model that preserves the finite capacity property of the real system. The model is formulated for multiple server finite capacity queueing networks with an arbitrary topology and blocking-after-service. The model is based on a decomposition of the network into single queues. The structural parameters of the queues are approximated so that they can account for the between-queue interactions. Unlike existing methods, the network topology and its configuration (number of queues and their capacity) are preserved throughout the analysis thus no constraints need to be checked a posteriori.

The originality of this method also lies in its ability to explicitly model the blocking phase that jobs may go through under congested traffic conditions. This is done via a novel state space formulation, and provides detailed stationary distributions. Furthermore, the endogenous parameters of the model quantify the occurrence, duration and effect of blocking. This leads to detailed performance measures that describe congestion in terms of its sources, its frequency and its impact. The case studies presented in the next chapter will illustrate how these endogenous parameters can be used to identify and describe congestion.

Performance measures have been validated by comparison with existing methods on networks with varying buffer sizes or service rates. The distributional approximations have been compared versus both exact results and simulation results, on a set of networks under a set of scenarios with varying arrival rates, namely under high intensity traffic. In all three types of validations, the results illustrate the good accuracy of the model.

This model assumes an infinite population of jobs. This assumption can be relaxed by making the external arrival rate an endogenous parameter. This is carried out for the protein synthesis application in Section 3.2.

Additional validation runs to test the sensitivity of the approximations would be desirable. In particular, tests to examine the robustness of this methodology to the distributional assumptions are of interest. For applications where these assumptions do not hold, the methods with phase-type distributions are adequate since phase-type distributions are dense within the class of continuous distributions (Inman, 1999; Altıok, 1989).

# Chapter 3

## Congestion evaluation: applications in health care and biology

### Contents

---

<b>3.1</b>	<b>Network of hospital units . . . . .</b>	<b>36</b>
3.1.1	Context . . . . .	36
3.1.2	Geneva University Hospitals network . . . . .	36
3.1.3	Comparison with simulation results . . . . .	38
3.1.4	Congestion analysis . . . . .	40
3.1.5	Conclusions . . . . .	41
<b>3.2</b>	<b>Protein synthesis network . . . . .</b>	<b>42</b>
3.2.1	Context . . . . .	42
3.2.2	Motivation . . . . .	43
3.2.3	Model . . . . .	43
3.2.4	System of equations . . . . .	50
3.2.5	Mapping of parameters and variables . . . . .	52
3.2.6	Validation . . . . .	53
3.2.7	Conclusions and future work . . . . .	54

---

In this chapter, we formulate the model presented in Chapter 2 for two applications and use it to evaluate network congestion. Section 3.1 considers a network of operative and post-operative hospital units, and studies what is known as *bed blocking*. This case study has been carried out for a project in collaboration with Dr. Philippe Garnerin and Pau Perez from the Division of Anesthesiology at the Geneva University Hospitals.

We then consider a biological application in Section 3.2, where ribosome congestion within a protein synthesis network is of interest. These results have been carried out for an ongoing project in collaboration with the Laboratory of Computational Systems Biotechnology (LCSB1) directed by Prof. Hatzimanikatis at EPFL.

## 3.1 Network of hospital units

### 3.1.1 Context

We apply our model to the study of patient flow in a network of hospital operative and post-operative units. Clinically, bed blocking may occur for example when a recovered intensive care patient cannot proceed to the intermediate care facility due to unavailable beds. The patient is said to be blocked until his (or her) placement is possible. Studies have acknowledged that bed unavailability renders the emergency and surgical admissions procedure less flexible and less responsive (Mackay, 2001).

Modeling bed blocking and estimating its effects would bring both patient care and budgetary improvements (Cochran and Bharti, 2006; Koizumi et al., 2005). This shows the importance of modeling the bed blocking phase within a patients recovery procedure. Although few analytic models incorporating blocking have been developed, there is a recently recognized need for them (Cochran and Bharti, 2006). The existing analytic models that account for blocking in the healthcare sector have limited their study to feed-forward networks with at most three finite capacity queues (Koizumi et al., 2005; Weiss and McClain, 1987; Hershey et al., 1981).

### 3.1.2 Geneva University Hospitals network

The hospital of interest is the Hôpitaux Universitaires de Genève (HUG, Geneva University Hospitals). The considered units with their corresponding queue index in parenthesis are the emergency operating suite (indexed as queue 1), elective operating suite (2), otorhinolaryngology operating suite (3), surgical intensive care (4), medical intensive care (5), medical intermediate care (6), neuro-surgical intermediate care (7), elective recovery (8) and otorhinolaryngology recovery (9). Hereafter, we refer to the units by using either their full name or their queue index.

$i$	1	2	3	4	5	6	7	8	9
$c_i$	4	8	5	18	18	4	4	10	6
$\gamma_i$	0.39	0.5	0.25	0.06	0.18	0.03	0.01	0.16	0
$\mu_i$	0.32	0.26	0.34	0.01	0.02	0.01	0.02	0.22	0.52
$card(\mathcal{S}_i)$	15	45	21	190	190	15	15	66	28

Table 3.1: Configuration of the HUG network.

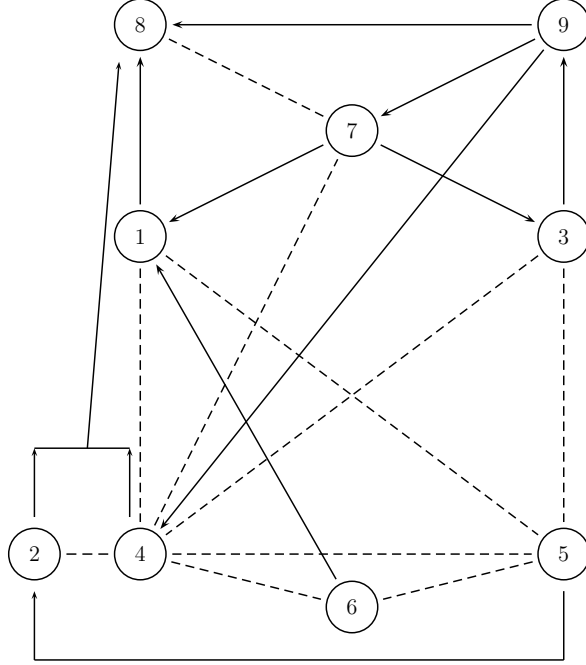


Figure 3.1: HUG network topology.

The patients are modeled as jobs and the beds as servers. Since there is no waiting space, each unit is modeled as a bufferless queue. The blocking-after-service mechanism of our model accurately mimics in-patient bed blocking.

The capacities of the different units were estimated according to the evaluations of HUG members. HUG members also extracted patient flow data, which we used to estimate the exogenous parameters  $\gamma, \mu$  and  $p_{ij}$ . Maximum likelihood estimates were used for  $\gamma$  and  $\mu$ , the transition probabilities were estimated by the transition frequencies. The data consisted of 25336 patient records ranging over a year.

The configuration of the network is presented in Table 3.1 and its topology is given in Figure 3.1. In this figure, the dashed lines correspond to bi-directional arrows. The network consists of 9 operative and post-operative units, with 31 possible transitions, containing numerous cycles. This makes the network prone to blocking.

Table 3.2 contains the transition probability matrix. In this table, the null probabilities are denoted by dashes. Note that the sum of the transition probabilities for a given unit (i.e. a given line) may not sum to 1, in this case  $1 - \sum_j p_{ij}$  represents the probability of exiting the network given that the job is at queue  $i$ .

	1	2	3	4	5	6	7	8	9
1	-	-	-	.16	.02	-	-	.71	-
2	-	-	-	.07	-	-	-	.84	-
3	-	-	-	.03	.01	-	-	-	.95
4	.18	.01	.03	-	.03	.01	.11	.03	-
5	.05	.01	.01	.01	-	.07	-	-	-
6	.02	-	-	.01	.1	-	-	-	-
7	.05	-	.05	.04	-	-	-	.01	-
8	-	-	-	-	-	-	.01	-	-
9	-	-	-	.05	-	-	.05	.02	-

Table 3.2: Transition probability matrix of the HUG network,  $p_{ij}$ .

### 3.1.3 Comparison with simulation results

We have also carried out this case study using the discrete event simulation model described in Section 2.5.3. This allowed us to compare our distributional estimates with those obtained via simulation. The simulation setup is the same as that of Section 2.5.3. The threshold for the stopping criteria of the algorithm is chosen as  $10^{-6}$ . Convergence was attained after 325 iterations and 84 seconds whereas the time required to complete the simulation was 25 minutes.

We consider once again the absolute errors of the marginal distributional estimates; their 90<sup>th</sup>, 95<sup>th</sup> and 99<sup>th</sup> percentiles are 0.008, 0.02 and 0.07, respectively. We have four estimates that have an absolute error larger than 0.1. Overall the distributional estimates are very good. The cumulative distribution function for the total number of jobs at each queue are depicted in Figure 3.2. The estimates of our method are represented by filled circles, whereas the simulation estimates are denoted by empty squares. All queues except queues seven and nine have very accurate estimates.

Three of the four previously mentioned estimates with large errors concern queue seven, the fourth error concerns queue nine. Explaining the cause of these large errors is not a straightforward task given the correlation between the endogenous parameters of our system of equations. The detailed distributions of queues seven and nine are displayed in Figure 3.3. The estimates of our method are represented by filled circles, whereas the simulation estimates are denoted by empty squares. The states  $(a, b)$  are ordered by increasing number of active jobs and then increasing number of blocked jobs.

This figure shows that for queue seven the state (4,0) is underestimated and for queue nine it is the blocked states (0,1) and (0,2) that are underestimated. These misestimations may be correlated since  $\tilde{p}_{97} = 0.82$  (displayed in Table 3.3 and discussed later on), i.e. given that a job is blocked at queue nine the probability that it has been blocked by queue seven is 0.82. Thus, the underestimation of the occupation of queue seven may lead to an underestimation of the blocking at queue nine.

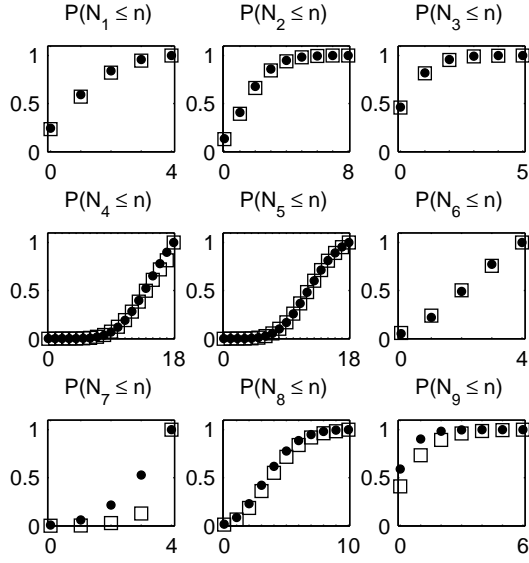


Figure 3.2: Comparison of the cumulative distribution function,  $P(N_i \leq n)$  for all queues.

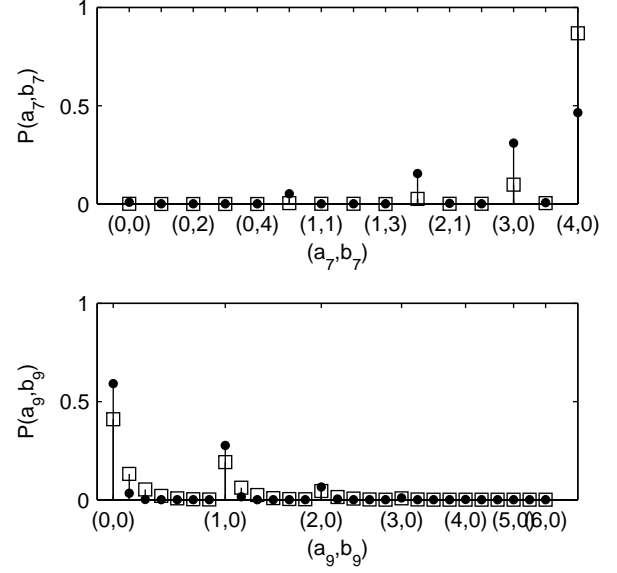


Figure 3.3: Distributions of queues seven and nine.

		1	2	3	4	5	6	7	8	9
Emergency OS	1	-	-	-	.76	.04	-	-	.19	-
Elective OS	2	-	-	-	.59	-	-	-	.41	-
ORL OS	3	-	-	-	.87	.13	-	-	-	.01
Surgical IC	4	.12	-	-	-	.02	.04	.82	-	-
Medical IC	5	.11	-	-	.05	-	.83	-	-	-
Medical IMC	6	.13	-	-	.16	.71	-	-	-	-
Neuro-surgical IMC	7	.34	-	.01	.65	-	-	-	.01	-
Elective REC	8	-	-	-	-	-	-	1	-	-
ORL REC	9	-	-	-	.18	-	-	.82	-	-

Table 3.3: Transition probabilities conditional on a patient being blocked,  $\tilde{p}_{ij}$ .

### 3.1.4 Congestion analysis

#### The sources of congestion

The outputs of our model help us to quantify the blocking and also investigate its causes. The transition probabilities conditional on a patient being blocked,  $\tilde{p}_{ij}$ , are displayed in Table 3.3. These probabilities can help to determine the source of blocking. The probabilities have been rounded to  $10^{-2}$ , those smaller than 0.005 are denoted by dashes. This table uses the following abbreviations: OS for operating suite, IC for intensive care, IMC for intermediate care, REC for recovery and ORL for otorhinolaryngology. For a given unit (i.e. a given line in the table) we can identify the target units that are more likely to block patients.

This table helps us to detect three main sources of blocking. The medical intensive care and the medical intermediate care units mutually block each others patients ( $\tilde{p}_{56} = 0.83, \tilde{p}_{65} = 0.71$ ). The same holds for the surgical intensive care and the neuro-surgical intermediate care units ( $\tilde{p}_{47} = 0.82, \tilde{p}_{74} = 0.65$ ). This first type of blocking (mutual blocking) may be irrelevant in practice given that the swapping of patients can be identified and carried out easily.

The second source of blocking, which may be more difficult to solve, is the blocking at the operating suites due to the surgical intensive care unit ( $\tilde{p}_{14} = 0.76, \tilde{p}_{24} = 0.59, \tilde{p}_{34} = 0.87$ ). Moreover, the performance of the emergency operating suite is strongly linked to its responsiveness, which is deteriorated by blocking. The third source of blocking occurs at the recovery units and is due to the neuro-surgical intermediate care unit ( $\tilde{p}_{87} = 1, \tilde{p}_{97} = 0.82$ ).

#### The frequency and effects of congestion

By explicitly modeling the blocking phase, our model yields novel performance measures that quantify the occurrence as well as the impact of congestion. Table 3.4 displays several performance measures of the different units. It recalls the capacity,  $k_i$ , which is equal to the number of servers  $c_i$ , and the estimate of the expected service time,  $1/\mu_i$ , of the units, which are exogenous parameters.  $1/\mu_i$  is given in hours. This table presents for each unit the probability of being blocked,  $\mathcal{P}_i$ , the expected number of blocked patients,  $E[B_i]$ , and the expected total number of patients,  $E[N_i]$ .

The parameter  $\mathcal{P}_i$  quantifies the occurrence of blocking at a given unit, but it does not capture the impact that a given blocking event may have on that unit. This impact can be described with  $E[B_i]/E[N_i]$ , which approximates the proportion of patients that are blocked. Consider the otorhinolaryngology recovery (ORL REC) unit, where  $\mathcal{P}_9$  equals 0.03, that is the probability of a patient getting blocked at that unit is 0.03. Blocking may then be considered as rare. For this unit  $E[B_9]/E[N_9]$  equals 0.11, i.e. approximately 11% of the occupied beds are blocked, thus blocking has a strong impact on the performance of the ORL REC unit. Although blocking may be rare, the impact that it may have on the unit

$i$	1	2	3	4	5	6	7	8	9
$k_i, c_i$	4	8	5	18	18	4	4	10	6
$\frac{1}{\mu_i}$	3.1	3.9	3.0	76.9	66.7	71.4	66.7	4.6	1.9
$\mathcal{P}_i$	0.02	0.01	0.00	0.06	0.02	0.01	0.01	0.00	0.03
$E[B_i]$	0.04	0.01	0.01	0.22	0.04	0.01	0.01	0.00	0.06
$E[N_i]$	1.37	2.00	0.77	14.03	12.56	2.46	3.19	4.04	0.53

Table 3.4: Performance measures for the HUG network.

or on the patient is not to be ignored.

### 3.1.5 Conclusions

We have formulated a queueing network model to study bed blocking in a network of operative and post-operative units of the Geneva University Hospitals. We have validated the distributional estimates with those obtained via simulation. These comparisons highlight the important gain in computation time since the time to estimate the parameters of our model is negligible compared to that required via simulation.

We have provided a detailed analysis of congestion, which illustrates how the performance measures derived by the model can be used to provide a fine description of congestion. In particular, we have identified three main sources of bed blocking and have quantified their impact upon the different hospital units.

We go beyond existing analytical queueing methods that have been used in the health care sector by allowing for networks with an arbitrary topology and with an arbitrary number of queues with finite capacity. Furthermore, the detailed performance measures provided by this approach respond to a recently stated need for methods that quantify in-patient bed blocking. Additionally, the blocking-after-service mechanism of the queueing model accurately mimics the bed blocking phenomenon.

This methodology has allowed us to investigate and quantify bed blocking. Such an analysis is useful to evaluate the quality of the health care service provided, to improve the allocation of resources across the different units, and also to identify and quantify the need for increased capacity.

## 3.2 Protein synthesis network

### 3.2.1 Context

In this section, we evaluate congestion in the context of protein synthesis. We first describe the main aspects of protein synthesis that we are interested in modeling. For a more detailed description of protein synthesis see Mehra and Hatzimanikatis (2006).

To synthesize proteins, the information of an mRNA (messenger RiboNucleic Acid) is translated. An mRNA consists of a strand of codons, i.e. a set of codons in a tandem topology. The information of an mRNA is encoded in these codons (i.e. each codon codes for an amino acid) and is translated to form proteins using ribosomes as catalysts.

Protein synthesis involves three main phases: initiation, elongation and termination. These are depicted in Figure 3.4. This figure presents an mRNA strand that consists of  $N$  codons. Each codon is depicted by a vertical line on the mRNA. There are four ribosomes on the mRNA. Each ribosome is  $L$  codons long.

During the initiation phase, the ribosome binds to the mRNA at the first codon (which is known as the start codon). Then the ribosome advances along the mRNA, one codon at a time. At each codon, elongation takes place. During elongation the corresponding codon (i.e. the underlying amino acid) is added to the growing protein chain. Termination occurs when the ribosome encounters the last codon (which is known as the termination codon). Both the ribosome and the newly formed protein are released, i.e. they unbind from the mRNA, and the ribosome is once again available for other translations.

For a given mRNA, the binded ribosomes advance along its codons, and may therefore be blocked by downstream ribosomes. Within a cell there are numerous mRNA's competing for resources and, in particular, for ribosomes. The blocking of ribosomes on an mRNA strand therefore reduces the protein synthesis rate for that mRNA, and may also affect that of all other mRNA's by reducing the probability that a ribosome is available for translation.

The frequency and effect of ribosome blocking is determined by the codon-specific initiation, elongation and termination rates, which therefore play an important role in the protein

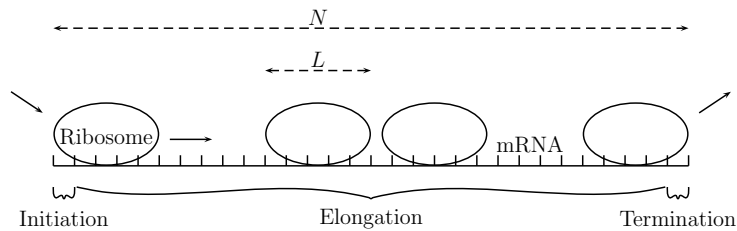


Figure 3.4: Ribosomes on an mRNA strand. Adapted from Mehra and Hatzimanikatis (2006).

synthesis rate. To study how these translation rates affect ribosome congestion and protein synthesis, we formulate the finite capacity queueing network (FCQN) model presented in Chapter 2 for a protein synthesis network. This is an ongoing project, where the ultimate goal is to understand how the synthesis of individual proteins varies with different levels of mRNA and of ribosomes.

### 3.2.2 Motivation

The basis of this work is the paper by Mehra and Hatzimanikatis (2006) entitled *An algorithmic framework for genome-wide modeling and analysis of translation networks*. This paper provides an analytical model, denoted hereafter as the *initial model*, that describes a codon-scale model of the translation of mRNAs into proteins. This model explicitly describes the phases of initiation, elongation and termination, and yields stationary distributions for each codon. A slightly modified version of this model is also given in Mier-y-Teran-Romero et al. (2009).

There are two main motivations for this work. Firstly, for highly congested scenarios, the initial model is numerically ill-conditioned, and therefore fails to provide reliable estimates. Secondly, there is a need for tractable models that can address large-scale problems. For instance, in a small-genome organism the number of codons is of the order of 400,000 (Mehra and Hatzimanikatis, 2006). We address these issues by deriving a queueing approach to this problem.

### 3.2.3 Model

We introduce the following notation:

- $N$  total number of codons on an mRNA strand;
- $L$  number of codons covered by a single ribosome;
- $t_i$  probability that codon  $i$  is not occupied by the head of a ribosome;
- $y_i$  probability that codon  $i$  is occupied by the head of a *blocked* ribosome;
- $z_i$  probability that codon  $i$  is occupied by the head of an *active* ribosome.

#### 3.2.3.1 Steady state distributions

The main purpose of the initial model is to determine the steady-state location of the different ribosomes on an mRNA. The location of the ribosomes is described based on the location of their *heads*. Recall that a ribosome occupies  $L$  consecutive codons. The head of a ribosome refers to the part of the ribosome that occupies the most downstream of these  $L$  codons.

The initial model considers each codon, and derives the steady-state probability that

there is a head of a ribosome at the  $i^{th}$  codon of an mRNA. This probability is the main variable of the model, and is denoted  $x_i$ . These variables satisfy a set of differential equations (Equations (1), (2) and (3) in Mehra and Hatzimanikatis (2006)). At steady state, these equations equal zero, which means that the probability of observing a transition out of a given state  $n$  in the next  $\Delta t$  time interval must be equal to the probability of observing a transition into state  $n$ . It is exactly this type of reasoning that leads in queuing theory to the *global balance equations*, which were detailed in Section 2.4 (Equation (2.1)).

We therefore follow the same reasoning as in Mehra and Hatzimanikatis (2006). We model each codon, and determine the distribution of the location of the head of the ribosomes. Each codon is modeled as a queue, and we investigate how the ribosomes advance along the network of codons.

Each codon is modeled as a bufferless queue with one server. Thus, an mRNA consists of a network of single server bufferless queues in tandem. This topology simplifies the global balance equations, as is detailed in the next section. We provide a more detailed distribution, than that of the initial model, by considering that each codon can be in one of three states:

- the codon is occupied by the head of an *active* ribosome
- the codon is occupied by the head of a *blocked* ribosome
- the codon is not occupied by the head of a ribosome, i.e. it is either not covered by a ribosome at all, or it is covered by a part of the ribosome that is not the head.

In other words, given that a codon is occupied by the head of a ribosome, we distinguish between whether the ribosome is active or blocked. It is this more detailed state space formulation that will allow us to quantify the impact of congestion along an mRNA strand.

As in Mehra and Hatzimanikatis (2006), we consider that a ribosome is  $L$  codons long. This has two main implications. Firstly, the last  $L$  codons of an mRNA cannot be blocked by a downstream ribosome. They are denoted *terminal* codons. All other codons can be blocked and are denoted *non-terminal* codons. Terminal and non-terminal codons are depicted in Figure 3.5. Secondly, if there is a head of a ribosome at a given codon  $i$ , blocking can occur if there is a head of a ribosome  $L$  codons downstream.

### 3.2.3.2 Single server networks

We apply the equations of Section 2.4 (System (2.27)) to single server networks. The specific case of single server networks leads to the following simplifications. The vector denoted by  $\tilde{\mu}_{ib}$  reduces to a single value that is now denoted  $\tilde{\mu}_i$ . Furthermore,  $E[T_i^B] = \frac{1}{\tilde{\mu}_i}$ ,  $\phi(i, b) = 1$  and  $c_i = 1$ . The parameter previously denoted by  $P(N_i = k_i)$  is now denoted  $(1 - t_i)$ . Thus,

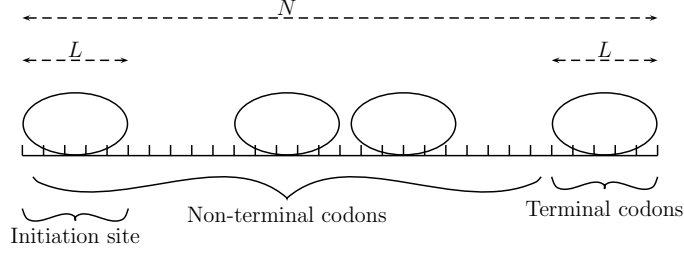


Figure 3.5: Initiation site of an mRNA strand, terminal and non-terminal codons.

the system of equations for single server networks is given by:

$$\begin{cases} \pi(i)g(\lambda_i, \mu_i, \tilde{\mu}_i, \mathcal{P}_i) = 0 & (3.1a) \end{cases}$$

$$\begin{cases} \lambda_i = \lambda_i^{\text{eff}}/t_i & (3.1b) \end{cases}$$

$$\begin{cases} \lambda_i^{\text{eff}} = \gamma_i t_i + \sum_j p_{ji} \lambda_j^{\text{eff}} & (3.1c) \end{cases}$$

$$\begin{cases} \frac{1}{\mu_i^{\text{eff}}} = \frac{1}{\mu_i} + \mathcal{P}_i / \tilde{\mu}_i & (3.1d) \end{cases}$$

$$\begin{cases} \frac{1}{\tilde{\mu}_i} = \sum_{j \in \mathcal{I}^+} \lambda_j^{\text{eff}} / (\lambda_i^{\text{eff}} \mu_j^{\text{eff}}) & (3.1e) \end{cases}$$

$$\begin{cases} \mathcal{P}_i = \sum_j p_{ij} (1 - t_j). & (3.1f) \end{cases}$$

We now show that for a tandem topology of bufferless queues this system of equations simplifies further. We first detail the simplifications regarding the equations of the structural parameters (i.e. Equations (3.1b) to (3.1f)), we then derive the simplifications of the global balance equations (Equation (3.1a)).

### 3.2.3.3 Structural parameters

In this protein synthesis context, there are three differences with the equations of Section 2.4.

1. There is a fixed and limited number of ribosomes that can bind to the mRNAs. The external arrival rate,  $\gamma$ , is therefore a function of the number of available (i.e. non-binding) ribosomes. It is no longer exogenous. Its expression will be detailed further on. This model therefore assumes a finite population of jobs (i.e. ribosomes).
2. To start the transcription process, a ribosome binds to the first codon of the mRNA. If this first codon is full, the ribosome cannot bind. Since a ribosome covers  $L$  consecutive codons, the first codon is full if there is a head of a ribosome on any of the *first  $L$  codons*. These first  $L$  codons are called the *initiation site*, and are depicted in Figure 3.5. The

probability that the first codon is free, i.e. that the initiation site does not contain a head of a ribosome, is denoted  $w_1$ .

3. Once a job has completed its service at a given queue, it may get blocked in a tandem network if its downstream queue is full. In this context, a ribosome may get blocked if there is a head of a ribosome  $L$  codons downstream.

Each mRNA is modeled as a tandem network of queues. In a tandem network, the effective arrival rate,  $\lambda_i^{\text{eff}}$ , is equal for all queues. In particular, Equation (3.1b) gives:

$$\forall i \quad \lambda_i^{\text{eff}} = \lambda_i t_i = \text{constant}. \quad (3.2)$$

The constant is given by the external arrival rate  $\gamma$ , and the probability that the first codon is free,  $w_1$ . Thus the effective arrival rate for all queues is given by:

$$\forall i \quad \lambda_i^{\text{eff}} = \lambda_i t_i = \gamma w_1. \quad (3.3)$$

We first present the system of equations for the structural parameters, we then detail its derivation. The system is given by:

### Terminal queues

$$\forall i \in [N - L + 1, N], \begin{cases} \lambda_i^{\text{eff}} = \gamma w_1 & (3.4a) \\ \mu_i^{\text{eff}} = \mu_i & (3.4b) \\ \frac{1}{\tilde{\mu}_i} = 0 & (3.4c) \\ \mathcal{P}_i = 0 & (3.4d) \end{cases}$$

### Non-terminal queues

$$\forall i \in [1, N - L], \begin{cases} \lambda_i^{\text{eff}} = \gamma w_1 & (3.5a) \\ \frac{1}{\mu_i^{\text{eff}}} = \frac{1}{\mu_i} + (1 - t_{i+L}) \frac{1}{\mu_{i+L}^{\text{eff}}} & (3.5b) \\ \tilde{\mu}_i = \mu_{i+L}^{\text{eff}} & (3.5c) \\ \mathcal{P}_i = 1 - t_{i+L}, & (3.5d) \end{cases}$$

where:

$$w_1 = 1 - \sum_{i=1}^L (y_i + z_i) \quad (3.6)$$

$$\gamma = a_0 + a_1 \sum_{i=1}^N (y_i + z_i). \quad (3.7)$$

Equations (3.4a) and (3.5a) result from Equation (3.3). The System (3.4) concerns terminal queues. These queues cannot be blocked, thus their expected blocked time (Equation (3.4c)) and their probability of being blocked (Equation (3.4d)) are null. Inserting Equations (3.4c) and (3.4d) in (3.1d) yields Equation (3.4b).

For non-terminal queues, blocking at queue  $i$  occurs if there is a head of a ribosome  $L$  codons downstream. Thus, the probability of being blocked at queue  $i$ ,  $\mathcal{P}_i$ , is given by the probability that there is a head of a ribosome at queue  $i + L$ ,  $1 - t_{i+L}$ . This is described by Equation (3.5d). Furthermore, the unblocking rate at queue  $i$ ,  $\tilde{\mu}_i$ , is determined by the effective service rate of queue  $i + L$ ,  $\mu_{i+L}^{\text{eff}}$  (Equation (3.5c)). Inserting Equations (3.5c) and (3.5d) in (3.1d) yields Equation (3.5b).

Equation (3.6) gives the probability that the first codon is free, this approximation is based on that given in Mehra and Hatzimanikatis (2006) (Equation (5)). Equation (3.7) concerns the external arrival rate, and gives its expression as a function of the number of available ribosomes (i.e. ribosomes not yet binded to an mRNA) and of two exogenous parameters  $a_0$  and  $a_1$ . The expression is taken from Equations (5) and (6) of Mehra and Hatzimanikatis (2006), or equivalently from their scaled versions which appear as Equation (11a) in Mier-y-Teran-Romero et al. (2009). The expressions for  $a_0$  and  $a_1$  are given in Table 3.5.

### 3.2.3.4 Global balance equations

In this section, we show that for a network of single server bufferless queues in a tandem topology the global balance equations given by Equation (2.1) lead to the following systems of equations:

#### Terminal queues

$$\forall i \in [N - L + 1, N], \begin{cases} t_i + z_i = 1 & (3.8a) \\ y_i = 0 & (3.8b) \\ \mu_i z_i = \gamma w_1. & (3.8c) \end{cases}$$

#### Non-terminal queues

$$\forall i \in [1, N - L], \begin{cases} t_i + y_i + z_i = 1 & (3.9a) \\ y_i = (y_{i+L} + z_{i+L})^2 & (3.9b) \\ \mu_i z_i = \gamma w_1. & (3.9c) \end{cases}$$

We now detail how these equations are derived. In the case of terminal queues the global balance equations are given by:

$$\begin{cases} t_i + z_i = 1 & (3.10a) \\ y_i = 0 & (3.10b) \\ \lambda_i t_i - \mu_i z_i = 0. & (3.10c) \end{cases}$$

Since  $\forall i \lambda_i t_i = \gamma w_1$  (Equation (3.3)), then the Systems of Equations (3.10) and (3.8) are equivalent. In the case of non-terminal queues, the global balance equations are given by:

$$\begin{cases} t_i + y_i + z_i = 1 & (3.11a) \\ -\tilde{\mu}_i y_i + \mathcal{P}_i \mu_i z_i = 0 & (3.11b) \\ \lambda_i t_i - \mu_i z_i = 0. & (3.11c) \end{cases}$$

Note that Equation (3.11b) balances blocking and unblocking events, while Equation (3.11c) balances arrival and service.

As for terminal queues, we can use Equation (3.3) to obtain the equivalence between Equations (3.9c) and (3.11c). Thus, to show that the Systems (3.9) and (3.11) are equivalent we need to show the equivalence between Equations (3.9b) and (3.11b). To do so, we will proceed by recursion and prove the following lemma.

**Lemma 1** *Let  $H(i)$  denote the hypothesis that Equation (3.9b) holds for queue  $i$ . Then  $H(N - L)$  holds, and if  $H(k)$  holds  $\forall k \in [i + 1, N - L]$ , then  $H(i)$  holds.*

**Proof.** We first show that  $H(N - L)$  holds. Combining Equations (3.11c) and (3.3) yields:

$$\mu_i z_i = \gamma w_1. \quad (3.12)$$

Inserting this into (3.11b) gives:

$$y_i = (\mathcal{P}_i \gamma w_1) / \tilde{\mu}_i. \quad (3.13)$$

Hereafter, we denote in brackets the equations used at each step. Since queue  $N - L$  is non-terminal, System (3.5) applies:

$$\begin{aligned} y_i &= \frac{\mathcal{P}_i \gamma w_1}{\mu_{i+L}^{\text{eff}}} & [3.5c] \\ &= \frac{(1 - t_{i+L}) \gamma w_1}{\mu_{i+L}^{\text{eff}}}. & [3.5d] \end{aligned} \quad (3.14)$$

Since queue  $N$  is terminal, Systems (3.4) and (3.8) apply:

$$\begin{aligned}
y_i &= \frac{(1 - t_{i+L})\gamma w_1}{\mu_{i+L}} & [3.4b] \\
&= \frac{z_{i+L}\gamma w_1}{\mu_{i+L}} & [3.8a] \\
&= z_{i+L}^2 & [3.8c] \\
&= (y_{i+L} + z_{i+L})^2. & [3.8b]
\end{aligned} \tag{3.15}$$

This gives  $H(N - L)$ .

We assume that  $H(k)$  holds  $\forall k \in [i + 1, N - L]$ . Since queue  $i$  is non-terminal, we can proceed as for  $H(N - L)$ :

$$\begin{aligned}
y_i &= \frac{\mathcal{P}_i \mu_i z_i}{\tilde{\mu}_i} & [3.11b] \\
&= \frac{\mathcal{P}_i \lambda_i t_i}{\tilde{\mu}_i} & [3.11c] \\
&= \frac{P_i \gamma w_1}{\tilde{\mu}_i} & [3.3] \\
&= \frac{P_i \gamma w_1}{\mu_{i+L}^{\text{eff}}} & [3.5c] \\
&= \frac{(1 - t_{i+L})\gamma w_1}{\mu_{i+L}^{\text{eff}}}. & [3.5d]
\end{aligned} \tag{3.16}$$

We distinguish between two cases. Firstly, if queue  $i + L$  is terminal we have:

$$\begin{aligned}
y_i &= \frac{(1 - t_{i+L})\gamma w_1}{\mu_{i+L}} & [3.4b] \\
&= \frac{z_{i+L}\gamma w_1}{\mu_{i+L}} & [3.8a] \\
&= (z_{i+L})^2 & [3.8c] \\
&= (y_{i+L} + z_{i+L})^2. & [3.8b]
\end{aligned} \tag{3.17}$$

Secondly, if queue  $i + L$  is non-terminal, then:

$$\begin{aligned}
y_i &= (1 - t_{i+L})\gamma w_1 \left( \frac{1}{\mu_{i+L}} + (1 - t_{i+2L}) \frac{1}{\mu_{i+2L}^{\text{eff}}} \right) & [3.5b] \\
&= (1 - t_{i+L})\gamma w_1 \left( \frac{1}{\mu_{i+L}} + (1 - t_{i+2L}) \frac{1}{\tilde{\mu}_{i+L}} \right) & [3.5c] \\
&= (1 - t_{i+L})\gamma w_1 \left( \frac{1}{\mu_{i+L}} + \mathcal{P}_{i+L} \frac{1}{\tilde{\mu}_{i+L}} \right) & [3.5d] \\
&= (1 - t_{i+L})\lambda_{i+L} t_{i+L} \left( \frac{1}{\mu_{i+L}} + \mathcal{P}_{i+L} \frac{1}{\tilde{\mu}_{i+L}} \right) & [3.3]
\end{aligned} \tag{3.18}$$

$$\begin{aligned}
y_i &= (1 - t_{i+L})\mu_{i+L} z_{i+L} \left( \frac{1}{\mu_{i+L}} + \mathcal{P}_{i+L} \frac{1}{\tilde{\mu}_{i+L}} \right) & [3.11c] \\
&= (y_{i+L} + z_{i+L}) \left( z_{i+L} + \mathcal{P}_{i+L} \mu_{i+L} z_{i+L} \frac{1}{\tilde{\mu}_{i+L}} \right) & [3.11a] \\
&= (y_{i+L} + z_{i+L})(z_{i+L} + y_{i+L}) & [3.11b] \\
&= (y_{i+L} + z_{i+L})^2.
\end{aligned} \tag{3.19}$$

This concludes the recurrence.  $\square$

Thus the global balance equations for single server bufferless queues in a tandem topology are given by the Systems (3.8) and (3.9).

### 3.2.4 System of equations

The system of nonlinear equations for single server bufferless queues in a tandem topology is described by the Systems of Equations (3.4)-(3.9). The only exogenous parameter is  $\mu_i$ , all other variables are endogenous.

These systems can be decoupled. In particular, we have implemented the following system:

$$\forall i \in [N - L + 1, N], \begin{cases} t_i + z_i = 1 & (3.20a) \\ y_i = 0 & (3.20b) \\ \mu_i z_i = \gamma w_1 & (3.20c) \end{cases}$$

$$\forall i \in [1, N - L], \begin{cases} t_i + y_i + z_i = 1 & (3.21a) \\ y_i = (y_{i+L} + z_{i+L})^2 & (3.21b) \\ \mu_i z_i = \gamma w_1 & (3.21c) \end{cases}$$

$$\begin{cases} w_1 = 1 - \sum_{i=1}^L (1 - t_i) & (3.22a) \\ \gamma = a_0 + a_1 \sum_{i=1}^N (1 - t_i). & (3.22b) \end{cases}$$

Once this system is evaluated, all the other variables of the full system can be deduced.

For a set of  $N$  codons, the system of equations consists of  $3N + 2$  equations. There are  $N + L + 2$  linear equations and  $2N - L$  quadratic equations. We compare this formulation to that of Mehra and Hatzimanikatis (2006) (referred to as the initial model). We then detail the contributions of the proposed formulation.

The initial model derives the stationary distributions of each codon. At steady state the initiation, elongation and termination rates are equal. Each one of these rates is given, in an aggregate form, by the following expressions. The initiation rate is expressed as:

$$v_I = b_0 \left( 1 - \sum_{s=1}^L x_s \right) \left( b_1 + b_2 \sum_{s=1}^N x_s \right), \quad (3.23)$$

where  $b_0, b_1$  and  $b_2$  are exogenous parameters, and  $x_s$  represents the probability that codon  $s$  is occupied by the head of a ribosome. As was described in Section 3.2.3.3, we have derived the expressions for  $w_1$  and for  $\gamma$  from Equation (3.23). The latter is equivalent to the product of Equations (3.22a) and (3.22b).

The elongation rates for non-terminal codons are given by:

$$v_j = c_j x_j \frac{1 - \sum_{s=1}^L x_{j+s}}{1 - \sum_{s=1}^{L-1} x_{j+s}}, \quad (3.24)$$

where  $(c_j)$  are exogenous parameters. The fraction approximates the conditional probability that codon  $j + 1$  is free given that codon  $j$  is occupied.

The elongation rates for terminal codons are given by:

$$v_j = d_j x_j, \quad (3.25)$$

where  $(d_j)$  are exogenous parameters.

**Congestion decomposition** One of the contributions of the paper of Mehra and Hatzimanikatis (2006) is to acknowledge the interactions between the initiation, elongation, termination and protein synthesis rates. In particular, given a set of ribosomes on an mRNA, the model acknowledges that their translation rate may be deteriorated by the presence of downstream ribosomes, that prevent the ribosome from advancing. This is captured by the fraction of Equation (3.24).

We go beyond this by describing these ribosome congestion effects in more detail. By using the *blocking* phenomenon of finite capacity queueing theory, and the detailed state space formulation of the FCQN model, we disaggregate the state “a codon is occupied” into two states “occupied and blocked” and “occupied and active”. By distinguishing between active and blocked codons, we provide more detailed distributional estimates. Additionally, as was illustrated in Section 3.1, the endogenous parameters of the proposed model provide a fine decomposition of congestion (e.g. in terms of its sources, frequency, impact).

**Conditioning** Equation (3.24) describes how blocking decreases the elongation rates. For highly congested scenarios, the denominator of the fraction in that equation tends to

zero. This leads to a numerically ill-conditioned model and to unreliable estimates.

This fraction approximates the probability that a codon is “occupied and active”. The queueing model explicitly considers that state, and does therefore not need to approximate this conditional probability. It is this simplification that leads to a well-conditioned specification for congested scenarios.

**Computational efficiency** To evaluate the stationary distributions of each codon, two procedures have been used in Mehra and Hatzimanikatis (2006) and in Mier-y-Teran-Romero et al. (2009). The first solves a bilevel nonlinear optimization problem, the second solves a system of ordinary differential equations. The procedure proposed in this section consists of a system of linear and quadratic equations, its implementation is straightforward, and it can be solved with less complex numerical methods.

Nonetheless, if transient distribution are of interest these can only be derived by the method of Mier-y-Teran-Romero et al. (2009).

**Scalability** In a small-genome organism the number of codons is of the order of 400,000 (Mehra and Hatzimanikatis, 2006). Such problems are considered of large-scale. The simplicity and tractability of the formulation presented in this section ensures its applicability for large-scale instances.

Furthermore, the number of equations that need to be implemented can be substantially reduced by identifying the queues that have equal service rates,  $\mu_i$ . In this case, Equations (3.20c) and (3.21c) indicate that these queues also have a common value for  $z_i$ , since

$$z_i = \gamma w_1 / \mu_i = \text{constant}. \quad (3.26)$$

If among the  $N$  queues there are  $D$  distinct service rates, then the number of equations reduces to  $2N - L + D + 2$ . In the case of protein synthesis, this can occur if the codons have common elongation rates. There are then three distinct service rates: initiation rate, termination rate and elongation rate, and the number of equations becomes  $2N - L + 5$ .

### 3.2.5 Mapping of parameters and variables

The parameters and variables used by Mehra and Hatzimanikatis (2006) and by Mier-y-Teran-Romero et al. (2009) have a straightforward mapping with the queueing theory parameters and variables. These are given in Table 3.5.

The rates in the second column (i.e. column of Mier-y-Teran-Romero et al., 2009) are scaled versions of those of the first column. The first line considers the main variable, which

	Mehra and Hatzimanikatis (2006)	Mier-y-Teran-Romero (2009)	FCQN
Codon distribution	$x_i^l$	$x_i$	$1 - t_i$
Initiation rate	$k_{if}^l M_l (R_T - \sum_l M_l \sum_i x_i^l)$	$\alpha \mu (r_T - \mu \sum_i x_i)$	$\gamma$
Elongation rate	$k_{E,j}^l M_l$	$\mu \beta_i$	$\mu_i$
Termination rate	$k_T^l M_l$	$\mu \gamma$	$\mu_N$

Table 3.5: Mapping of parameters and variables.

is the stationary probability that a codon is occupied by the head of a ribosome. The formulation in Mehra and Hatzimanikatis (2006) accounts for different mRNA species,  $x_i^l$  denotes the stationary probability for the  $l^{th}$  mRNA species.

For a given species, the number of mRNA strands is denoted  $M_l$ . The queueing formulation considers a single species and a single strand ( $l = 1, M_l = 1$ ). The parameters  $k_{if}, k_{E,j}$  and  $k_T$  are initiation, elongation and termination rate constants. They are exogenous parameters, and are represented by the service rates in the queueing model  $\mu_i$ . The initiation rate (second line of the table) is a function of the initiation constant  $k_{if}$  and of the expected number of free (non-binding) ribosomes. This expectation is approximated by  $(R_T - \sum_l M_l \sum_i x_i^l)$ , where  $R_T$  denotes the total number of ribosomes, and  $M_l \sum_i x_i^l$  approximates the expected number of ribosomes on the mRNA's of species  $l$ .

### 3.2.6 Validation

We now validate this model by comparing its distributional estimates with those of the system of equations given in both Mehra and Hatzimanikatis (2006) and Mier-y-Teran-Romero et al. (2009). The code to solve the system of differential equations has been provided to us by members of the LCSB1 laboratory. It implements the system of differential equations of Mehra and Hatzimanikatis (2006) with two minor simplifications: 1) it assumes a single mRNA strand and a single mRNA species, 2) it assumes no reversible binding (i.e. the last term of Equation (4) in Mehra and Hatzimanikatis (2006) is not accounted for). Given these simplifications, the systems of equations in Mehra and Hatzimanikatis (2006) and Mier-y-Teran-Romero et al. (2009) are equivalent. The validation scenarios have also been defined by the LCSB1 members.

We consider an mRNA with 144 codons, and assume that each ribosome covers 12 codons, i.e.  $N = 144, L = 12$ . The elongation rates of all codons are assumed equal. We compare the distributional estimates derived by both methods.

Since the distribution defined by the system of differential equations is less detailed than that provided by the queueing model (because the state space does not explicitly account for blocking), we compare the distribution of  $x_i$  with that of the aggregated states  $y_i + z_i$ , both representing the probability that a codon is covered by a head of a ribosome.

We consider a set of 27 scenarios with varying initiation, elongation and termination

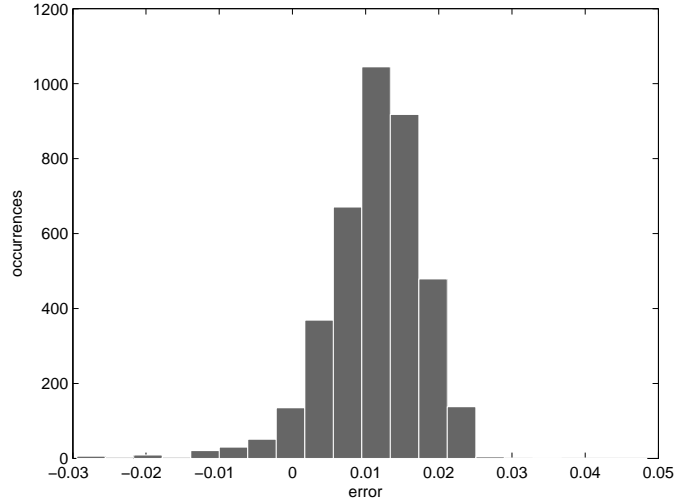


Figure 3.6: Histogram of the errors of the distributional estimates for all 144 codons and 27 scenarios.

rates. For each scenario we compare the distribution obtained for each one of the 144 codons.

Figure 3.6 displays a histogram of the difference between  $y_i + z_i$  and  $x_i$ . The distributional differences are of the order of  $10^{-2}$ , which shows that for these scenarios both methods yield similar values for the probabilities that the head of a ribosome is occupying a given codon. The queueing approach provides larger estimates.

Figure 3.7 groups the 144 codons into three segments: codons 1-48 (i.e. the most upstream codons), codons 49-96 (i.e. the middle codons), codons 97-144 (i.e. the most downstream codons). This figure illustrates that as the codon index increases, so does the difference in the estimates provided by the queueing model and the differential equations.

Experimental data is not available. Validation therefore consists of comparing the estimates provided by various models. Given the lack of data, it is intricate to draw conclusions from the differences in these estimates. That is, we cannot state whether these larger estimates reveal improved or decreased accuracy. Nonetheless, it is of interest to investigate further the main differences between these two approaches.

### 3.2.7 Conclusions and future work

We have presented a queueing network formulation to study protein synthesis. Each codon of an mRNA is modeled as a queue. Each mRNA strand is modeled as a tandem network of single server bufferless queues. The methodology derives a distribution for each codon, that evaluates whether or not there is a head of a ribosome on that codon, and in particular identifies whether ribosomes are blocked by downstream ribosomes. The main modeling assumptions are those of Mehra and Hatzimanikatis (2006).

We have shown that for a tandem network of single server bufferless queues, the sys-

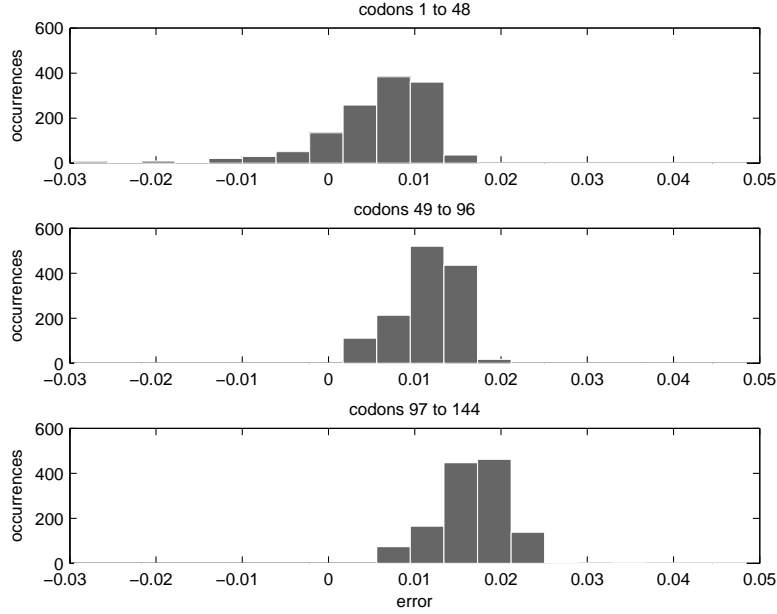


Figure 3.7: Histogram of the errors of the distributional estimates. Each plot considers a different group of 48 codons and all 27 scenarios.

tem of nonlinear equations simplifies, and leads to a particularly tractable set of linear and quadratic equations. The distributional estimates provided by this method have been validated versus those in Mehra and Hatzimanikatis (2006).

This simple formulation overcomes the main limitations of the Mehra and Hatzimanikatis (2006) model. It yields a numerically well-conditioned and computationally efficient formulation, that can address large-scale problems.

More research is needed. First, we need to compare the distributional estimates with those of other methods. Once the validation phase has been completed, this model will be applied to a more general context considering 1) multiple mRNA strands, 2) multiple mRNA species, 3) codon-specific elongation rates. These three factors are captured by the exogenous parameters of our model; taking them into account is therefore straightforward. Then we will investigate the performance of the proposed approach on large-scale instances.

We will also use the detailed blocking information provided by this queueing model to study how ribosome congestion affects protein synthesis. In particular, ribosome blocking may reduce protein synthesis by reducing the probability that the initiation sites are free, or also by reducing the number of ribosomes available. It is also of interest to investigate which rates are limiting protein synthesis; this can be done by studying the sensitivity of the protein synthesis rate to the initiation, elongation and termination rates.



# Chapter 4

## A surrogate for the optimization of congested urban road networks

### Contents

---

<b>4.1</b>	<b>Motivation . . . . .</b>	<b>58</b>
<b>4.2</b>	<b>Literature review . . . . .</b>	<b>60</b>
4.2.1	Analytic queueing models . . . . .	60
4.2.2	Traffic signal control . . . . .	61
<b>4.3</b>	<b>Surrogate network model . . . . .</b>	<b>63</b>
4.3.1	Queueing model . . . . .	63
4.3.2	Network topology . . . . .	63
4.3.3	Bounded queues . . . . .	64
4.3.4	Arrival and service rates . . . . .	65
4.3.5	System of equations . . . . .	66
<b>4.4</b>	<b>Optimization problem . . . . .</b>	<b>68</b>
<b>4.5</b>	<b>Empirical analysis . . . . .</b>	<b>70</b>
4.5.1	Microscopic traffic simulation model of the city of Lausanne . . . .	70
4.5.2	Between-queue interactions . . . . .	72
4.5.3	Comparison with existing methods . . . . .	75
<b>4.6</b>	<b>Large-scale networks . . . . .</b>	<b>80</b>
4.6.1	Formulation . . . . .	80
4.6.2	Lausanne city network . . . . .	81
<b>4.7</b>	<b>Conclusions and future work . . . . .</b>	<b>85</b>

---

## 4.1 Motivation

Road traffic congestion is a costly phenomenon that is common to the vast majority of urban road networks. A recent European Commission report emphasizes that to alleviate congestion “in certain cases new infrastructure might be needed, but the first step should be to explore how to make better use of existing infrastructure” (CEC, 2007). Thus, the importance of understanding the origins of congestion, of quantifying its effects, and of controlling traffic in order to optimize the use of existing infrastructure.

Vehicle traffic can be modeled at either a macroscopic or a microscopic scale. Microscopic models represent each vehicle in the network. Disaggregate traffic models are then used to describe how each vehicle interacts with the network supply and with other vehicles: e.g. car following, lane changing or route choice models.

By relying on disaggregate behavioral models, these simulators yield detailed performance measures, such as vehicle specific trajectories. A variety of microscopic traffic simulation models have been developed, and they remain popular tools to evaluate traffic management schemes.

Nevertheless, as was mentioned in the introduction of this thesis (Section 1.1), the use of microscopic models is limited due to two main issues. Firstly, their validation and calibration requires large amounts of detailed data, which is expensive and cumbersome to collect. Secondly, the use of such detailed behavioral models leads to performance measures that have no closed-form available, are stochastic and computationally expensive to estimate. Thus, their use to perform network optimization to derive traffic management schemes remains a difficult task.

Macroscopic models are flow-based models. For these models, vehicle traffic may be represented continuously (e.g. fluid approximation) or discretely (vehicles or groups of vehicles, e.g. queueing theory). These models may be analytical or simulation-based. They yield aggregate performance measures.

Given the increasing incidence and cost of urban congestion, it is of great importance to formulate models that can be embedded within optimization frameworks in order to identify suitable traffic management schemes. Analytical macroscopic models with sound mathematical properties naturally fit within such optimization frameworks. The main challenge when deriving such models is to ensure an appropriate trade-off between efficiency (i.e. computational tractability) and realism (i.e. capturing the complex behavior of congested traffic).

As described in the introduction of this thesis, models that are developed to perform optimization and that are often a simplified version of an underlying realistic but cumbersome to optimize model, are known as surrogates. Surrogate models are less realistic but are also typically cheaper to evaluate and more tractable. Within this context, the contributions of this chapter are two-fold.

Firstly, we present a surrogate analytical macroscopic model for congested network optimization. It is an extension of the stochastic network model derived in Chapter 2. Existing analytical queueing network models for vehicle traffic have focused on the study of uninterrupted traffic flow.<sup>1</sup> To the best of our knowledge, the few studies that consider interrupted traffic flow are formulated for a single intersection. They therefore do not take into account the interactions between upstream and downstream roads. The framework that we present models a set of urban intersections. It captures the interactions between consecutive roads by combining approximation methods with finite capacity queueing theory. It therefore provides a detailed description of congestion and, in particular, identifies both bottlenecks and spillbacks; and quantifies their impact upon the overall network performance.

The second contribution of this chapter concerns the improvement of the use of existing infrastructure. We formulate a fixed-time signal control problem where the network model is included as a set of constraints. To the best of our knowledge, the existing signal control strategies based on analytical network models have not taken urban spillbacks into account. More generally, most signal control strategies do not account for saturated or highly congested networks where spillbacks are likely to occur (Papageorgiou et al., 2003). We therefore believe that the considered network model is an appropriate tool to improve urban signal settings, namely during peak hours.

This chapter is structured as follows. We present in Section 4.2 a literature review of analytic queueing models for urban traffic and of signal control methodologies. We describe the surrogate network model (Section 4.3) and formulate the optimization problem (Section 4.4). We then present and discuss the role of a microscopic traffic simulator used in this framework (Section 4.5.1).

The methodology is applied to a subnetwork of the Lausanne city center. The optimized signal plan is then compared with plans generated by several other methods. Section 4.5.2 compares it with the plan derived by the queueing model assuming independent queues. This section analyzes the added value of modeling the between-queue interactions. In Section 4.5.3, the signal plan is compared to an existing plan for the city of Lausanne and to the plans derived by the methods proposed in Webster (1958) and in the Highway Capacity Manual (TRB, 2000).

Section 4.6 presents a formulation of the surrogate model that is suitable for large-scale networks. It is used to solve a signal control problem, considering the road network of the entire city of Lausanne. We consider three different initial signal plans, and compare the performance of the derived plans to the initial plans.

In this and the following chapter, the term *capacity* refers to what is known in traffic theory as flow capacity (VSS, 1998), that is the maximum flow rate expected to cross a

---

<sup>1</sup>Flow is denoted as uninterrupted flow when it is regulated by interactions between vehicles, e.g. on a highway or at unsignalized intersections. Interrupted flow is regulated by an external mean, e.g. at signalized intersections.

given roadway per unit of time. It is typically given in vehicles per hour and corresponds in queueing theory terms to the *service rate*. The term denoted in queueing theory as queue capacity,  $k_i$ , will be denoted hereafter as the upper bound of the queue length.

## 4.2 Literature review

### 4.2.1 Analytic queueing models

Queueing models have been used in transportation mainly to model highway traffic (Garber and Hoel, 2002). Several simulation models have been developed, but few studies have explored the potential of the queueing theory framework to develop analytical urban traffic models. Furthermore, existing urban queueing models have mainly focused on unsignalized intersections.

Heidemann and Wegmann (1997) give a literature review for exact analytical queueing models of unsignalized intersections. They model the minor stream as an M/G2/1 queue. They emphasize the importance of the pioneer work of Tanner (1962). Heidemann also contributed to the study of signalized intersections (Heidemann, 1994), and presented a unifying approach to both signalized and unsignalized intersections (Heidemann, 1996).

These models combine a queueing theory approach with a realistic description of traffic processes for a given lane at a given intersection. They yield detailed performance measures such as queue length distributions or sojourn time distributions. Nevertheless, as exact analytical methods, they are difficult to generalize to consider multiple lanes, not to mention multiple intersections.

To the best of our knowledge, no method has been proposed to model the traffic process for a set of urban intersections using an analytic queueing network framework. Nevertheless the methods proposed by Jain and Smith (1997) and Van Woensel and Vandaele (2007), which are both based on the Expansion Method (Kerbache and Smith, 2000; the main ideas of this method were described in Section 2.3.2) and formulated for highway traffic, could be extended to consider an urban setting.

As described in detail in Section 2.3.1, methods that allow the exact evaluation of the joint stationary distribution of the network of queues are difficult to obtain, let alone transient distributions (Newell, 1979). The method proposed here is an extension of the model presented in Chapter 2. It provides estimates of the stationary marginal distributions of each queue.

The purpose of this model is to embed it within an optimization framework in order to derive traffic control schemes for congested networks. The main challenge is to achieve an adequate tradeoff between realism and tractability. Our aim is therefore to provide a detailed analytical description of how queues interact under congested conditions, and to ensure tractability by limiting the analysis to the stationary regime.

### 4.2.2 Traffic signal control

Traffic signal setting strategies can be either fixed-time or traffic-responsive strategies. *Fixed-time* (also called *pre-timed*) strategies use historical traffic data, and yield one traffic signal setting for the considered time of day. The traffic signal optimization problem is solved offline. On the other hand, *traffic-responsive* (also called *real-time*) methods use real-time data to define timings for immediate implementation that are used over a short time horizon.

Furthermore, signal timings can be derived by considering either a single or a set of intersections. These methods are called *isolated* methods and *coordinated* methods, respectively (Papageorgiou et al., 2003). Methods that handle individual intersections are based on models that capture the local behavior of the network. They describe in detail the traffic interactions at an intersection, but at the expense of capturing less well the interactions among intersections.

A *phase* is defined as a set of streams that are mutually compatible and that receive identical control. The cycle of a signal plan is divided into a sequence of periods called *stages*. Each stage consists of a set of mutually compatible phases that all have green. Methods where the stage structure (i.e. the sequence of stages) is given are known as *stage-based* approaches, whereas methods where the stage structure is endogenous are referred to as *phase-based* or *group-based* approaches.

Delay minimization and reserve capacity maximization are the most common objective functions used by existing methods. Delay may be directly measured, leading to a data-driven approach, or estimated (model-based approach). The first approximate expression for the delay at an intersection was given by Webster (1958), and is still widely used. Other expressions include those of McNeil (1968), Newell (1965) and Miller (1963). Viti (2006) provides a review of delay models; Dion et al. (2004) compare the performance of different delay models, and Chow and Lo (2007) derive approximate delay derivatives that can be integrated within a simulation-based signal setting optimization context in order to reduce the computation time required to obtain numerical derivatives. The notion of the *reserve capacity* of an intersection is defined by Wong and Yang (1997) as the greatest common multiplier of existing flows that can be accommodated subject to saturation and signal timing constraints. This notion has been extended to consider several intersections (Ziyoun and Yifan, 2002; Wong and Yang, 1997).

The works of Allsop (1992) and of Shepherd (1994) review signal control methods. Allsop (1992) describes in detail the corresponding terminology as well as the different formulations for isolated methods. More recently, reviews of traffic control methods are given by Papageorgiou et al. (2003) and by Cascetta et al. (2006). Papageorgiou et al. (2003) review urban traffic control methods, while highlighting their applications (either via simulation or field implementations). They also consider freeways and route guidance methodologies. A detailed review of signal control methodologies is provided in Appendix A of this thesis.

The method proposed in this chapter belongs to the category of fixed-time coordinated methods. Traditionally, fixed-time strategies have been considered suitable only for under-saturated traffic conditions (Chow and Lo, 2007; Papageorgiou et al., 2003; Abu-Lebdeh and Benekohal, 1997; Shepherd, 1994). Thus, methods for saturated conditions have focused on real-time strategies. Nevertheless, we believe that the development of optimal fixed-time methods is of primary importance. First, they can be used as benchmark solutions to evaluate traffic-responsive strategies. Second, they represent robust control solutions (Yin, 2008). Finally, they may be used as building blocks to derive real-time methods.

Although there is a vast range of signal control methodologies in the literature, there is still a need for solutions that are appropriate and efficient under saturated conditions (Dinopoulou et al., 2006). Under congested conditions the performance of signal control strategies and the formation and propagation of queues are strongly related.

Models that ignore the spatial extension of queues are known as vertical or point queueing models. Such models fail to capture congestion effects such as spillbacks and gridlocks. Adopting a vertical queueing model is therefore only reasonable when the degree of saturation is moderate. Both Chow and Lo (2007) and Abu-Lebdeh and Benekohal (1997), illustrate the effects of ignoring this spatial dimension. Therefore a signal control strategy suitable for congested conditions must take into account the interactions between queues.

Nevertheless, most existing strategies do not account for these interactions and are thus unsuitable for highly congested networks (Papageorgiou et al., 2003; Abu-Lebdeh and Benekohal, 2003). Furthermore Abu-Lebdeh and Benekohal (2003) emphasize the importance of accounting for the effects of queue propagation within a signal timing framework.

The recently proposed TUC (Traffic-responsive Urban Control) method (Dinopoulou et al., 2006) focuses on saturated traffic conditions. It overcomes the exponential complexity of the existing methods by avoiding the use of discrete variables, which “is of paramount importance because it opens the way to the application of a number of highly efficient optimization and control methods”. Following these ideas, we also model the flow as a continuous variable. The consequences of this assumption are detailed by Dinopoulou et al. (2006).

Additionally, the most recent real-time methods, such as TUC and RHODES (Mirchandani and Head, 2001), overcome the need for analytically grasping the interaction between queues by assuming that measurements are available either on every link or on every signalized link of the network. By capturing the between-queue interactions such assumptions could be relaxed, allowing these methods to be applicable on a wider range of networks. Therefore, the queueing model proposed in this chapter is an appropriate tool both to improve urban signal settings during peak hours and to emphasize the importance of accounting for the between-queue interactions.

## 4.3 Surrogate network model

This section formulates the surrogate model for urban vehicle traffic. We consider an urban transportation network composed of a set of both signalized and unsignalized intersections. The traffic model is an extension of the queueing network model presented in Chapter 2. In this section we detail its formulation and adaptation for urban traffic networks.

We study a fixed-time signal control problem where the offsets, the cycle times and the all-red durations are fixed. The stage structure is also given. In other words, the set of lanes associated with each stage as well as the sequence of stages are both known.

The objective is to minimize the average time  $T$  spent in the network by adjusting the green splits at each intersection (i.e. the proportion  $x(j)$  of cycle time that is allocated to each phase  $j$ ). The travel time is derived from a traffic model which combines both exogenous (fixed) parameters  $q$ , such as the total demand, the route choice decisions and the topological structure of the street network, with endogenous variables  $y$ , such as the capacities and the probability of spillbacks. The latter are directly linked with the decision variables  $x$ . We now present the traffic model that derives  $T$  from  $x$ ,  $y$  and  $q$ . We then formulate the signal control problem.

### 4.3.1 Queueing model

In the model presented in Chapter 2, we assume both the total demand and the capacities to be given, and derive a set of performance measures such as stationary distributions and congestion indicators. Each queue is defined according to a set of exogenous structural parameters. We extend here the formulation by considering the capacities endogenous, as they are determined by the decision variables (i.e. the green splits). All distributional assumptions and approximations of the model of Chapter 2 are preserved in this framework.

### 4.3.2 Network topology

Each road in the network is divided into segments such that the number of lanes is constant on each segment. Segment boundaries are therefore either intersections or locations where the number of lanes changes between intersections. They correspond to changes of capacity. Figure 4.1 illustrates how a road with two main lanes and a right turn lane is represented as two segments (depicted as rectangles delimited in bold).

A queue is associated with each lane of each segment in the network (similarly to the supply simulator in the DynaMIT system, see Ben-Akiva et al. (2001)). Each queue is connected to the downstream segments where a turning of the underlying lane is permitted. Note that connecting a queue to a segment means that it is connected to all of the queues in that segment.

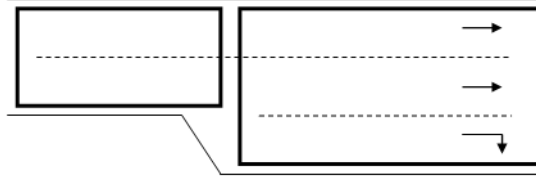


Figure 4.1: Example of how a road is mapped to a set of segments.

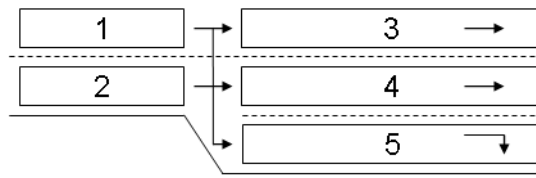


Figure 4.2: Example of how a road is mapped to a set of queues.

Figure 4.2 illustrates how a road composed of two main lanes and one right turn lane is modeled as two upstream queues (indexed 1 and 2) followed by three downstream queues (indexed 3-5). In particular, if queue 5 spills back it will block the through movements of the upstream queues. The interactions among the queues are explicitly captured by linking the parameters of the queues (such as the capacity and the arrival flow) with the state of other queues.

### 4.3.3 Bounded queues

In order to account for the limited physical space that a queue of vehicles may occupy, we resort to finite capacity queueing theory, where there is a finite upper bound on the length of each queue. The use of a finite bound allows us to capture the impact of queues on upstream segments (e.g. spillbacks), and to consider congested scenarios where traffic demand may exceed capacity. In queueing theory terms, this corresponds to a traffic intensity that may exceed one. This is the key distinction between classical queueing theory and finite capacity queueing theory.

The upper bound of the length of queue  $i$  is denoted  $k_i$  (known as the capacity of the queue in queueing theory). Heidemann (1996), as well as Van Woensel and Vandaele (2007), divide each road into parts of length  $1/k_{\text{jam}}$ , where  $k_{\text{jam}}$  is the jam density, and thus  $1/k_{\text{jam}}$  represents the minimal length that each vehicle needs. We also follow this type of reasoning and define  $k_i$  as:

$$k_i = \lfloor (\ell_i + d_2) / (d_1 + d_2) \rfloor,$$

where  $\ell_i$  denotes the length of lane  $i$ ,  $d_1$  is the average vehicle length (e.g. 4 meters), and  $d_2$  is the minimal inter-vehicle distance (e.g. 1 meter). The fraction is then rounded down to the nearest integer.

The physical space occupied by a queue is represented by a server followed by a buffer.

All queues have one server, which represents the service due to the change of capacity at the boundary of a segment.

#### 4.3.4 Arrival and service rates

The exogenous parameters used to describe the distribution of the demand throughout the network are the external arrival rates and the transition probabilities. The external arrival rate of a queue corresponds to vehicles reaching the queue coming from outside of the network and not from another queue. This typically applies to the boundaries of the network or parking lots inside the network. The transition probability between queue  $i$  and queue  $j$  is the proportion of flow from queue  $i$  that goes to queue  $j$ . These turning proportions may be obtained from a route choice model (Bierlaire and Frejinger, 2008).

The service rates of the queues are defined as the capacities of the underlying lanes. For segments that lead to intersections, the service rate of its queues is defined as the capacity of the intersection for that approach or lane. We derive formulations for the capacities of the different types of intersections based on the Swiss national transportation standards.

For unsignalized intersections (e.g. two-way stop controlled intersections, yield-controlled intersections) the standard VSS (1999a) is used. The turning movements are ranked. For each movement, the conflicting flow is calculated based on a set of equations that depend on the type of movement and its rank. Then its potential capacity and its movement capacity are calculated. Finally the capacity of the lanes with multiple turnings are adjusted to take into account the lack of turning lanes.

The capacity of the lanes leading to, on, or exiting roundabouts is derived based on the standard VSS (2006). This approach is similar to that for unsignalized intersections, it takes into account the same parameters but is based on a different set of equations. This standard accounts for roundabouts with either one lane or one large lane. For networks that contain roundabouts with two lanes, the capacity of these lanes is calculated based on the equations for roundabouts with one large lane.

For signalized intersections, we use the standard VSS (1999b), which defines the capacity of a lane as the product of the saturation flow and the proportion of green time allocated to that lane per cycle. This approach is also proposed in Chapter 16 of the Highway Capacity Manual (TRB, 2000).

When a segment does not lead to an intersection (e.g. segments where all of the vehicles leave the network or segments that lead directly to another segment), the service rate of its queues is set to the saturation flow of the corresponding lane.

### 4.3.5 System of equations

We adapt the equations proposed in Section 2.4 to this context. We recall the notation of Section 2.4 that will be used throughout this chapter. The index  $i$  refers to a given queue.

$\gamma_i$	external arrival rate;
$\lambda_i$	total arrival rate;
$\lambda_i^{\text{eff}}$	effective arrival rate (accounts only for the arrivals that are actually processed, i.e. excludes all lost arrivals);
$\mu_i$	service rate of a server;
$\tilde{\mu}_i$	unblocking rate;
$\mu_i^{\text{eff}}$	effective service rate (accounts for both service and eventual blocking);
$\mathcal{P}_i$	probability of being blocked at queue $i$ ;
$p_{ij}$	transition probability from queue $i$ to queue $j$ ;
$k_i$	upper bound of the queue length;
$N_i$	total number of vehicles in queue $i$ ;
$P(N_i = k_i)$	probability of queue $i$ being full, also known as the blocking probability;
$\mathcal{I}^+$	set of downstream queues of queue $i$ .

The main feature of the queueing model of Section 2 is the detailed state space, which distinguishes between active and blocked jobs. Having such a detailed state space is of minor interest in this context. Here the main motivation is to capture spillbacks. These can be described based on the probability that a queue is full,  $P(N_i = k_i)$ , which is known as the *blocking probability* in finite capacity queueing theory. This leads us to use the closed form expression available for the blocking probability of finite capacity queues (Bocharov et al., 2004), instead of resorting to the *global balance equations* as in Section 2.4.

Since we have modeled the road network as a network of single server queues, we can use the system of equations for single server networks that was derived in Section 3.2.3.2 (Equations (3.1)). Recall that  $P(N_i = k_i)$  is denoted  $(1 - t_i)$  in Section 3.2.3.2.

The system of equations is given (on the following page) by:

$$\begin{cases}
\lambda_i = \lambda_i^{\text{eff}} / (1 - P(N_i = k_i)) & (4.1a) \\
\lambda_i^{\text{eff}} = \gamma_i (1 - P(N_i = k_i)) + \sum_j p_{ji} \lambda_j^{\text{eff}} & (4.1b) \\
\frac{1}{\mu_i^{\text{eff}}} = \frac{1}{\mu_i} + \mathcal{P}_i \frac{1}{\tilde{\mu}_i} & (4.1c) \\
\frac{1}{\tilde{\mu}_i} = \sum_{j \in \mathcal{I}^+} \lambda_j^{\text{eff}} / (\lambda_i^{\text{eff}} \mu_j^{\text{eff}}) & (4.1d) \\
\mathcal{P}_i = \sum_j p_{ij} P(N_j = k_j) & (4.1e) \\
P(N_i = k_i) = \frac{1 - \rho_i}{1 - \rho_i^{k_i+1}} \rho_i^{k_i} & (4.1f) \\
\rho_i = \frac{\lambda_i}{\mu_i^{\text{eff}}}, & (4.1g)
\end{cases}$$

where  $\rho$  is known in queueing theory as the traffic intensity (or the utilization ratio) of a queue.

We recall the main features of these equations. Equations (4.1a) and (4.1b) define the total and effective arrival rates by combining flow conservation with loss model information. Equation (4.1c) defines the effective service rate as a function of its capacity, of the probability that a spillback occurs from its downstream lanes and of their corresponding spillback dissipation rates. Equation (4.1d) defines the unblocking rate  $\tilde{\mu}_i$ , which describes the rate at which downstream spillbacks dissipate.

Equation (4.1e) gives the probability with which spillbacks occur at lane  $i$  due to its downstream lanes. Equation (4.1f) is the closed form expression for the blocking probability, it gives the probability that traffic at lane  $i$  spills back to upstream lanes. It is a function of the traffic intensity,  $\rho_i$ , which is defined by Equation (4.1g).

Unlike the network model formulation of Chapter 2,  $\mu_i$  is now an endogenous variable. The exogenous parameters are  $\gamma_i$ ,  $p_{ij}$  and  $k_i$ . This system has been implemented in terms of 6 endogenous variables per queue ( $\lambda_i$ ,  $\mu_i$ ,  $\tilde{\mu}_i$ ,  $\mu_i^{\text{eff}}$ ,  $P(N_i = k_i)$ ,  $\mathcal{P}_i$ ). Thus, the system consists of  $6n$  equations, where  $n$  is the total number of queues.

## 4.4 Optimization problem

In order to formulate the signal control problem we introduce the following notation:

$b_i$	available cycle ratio of intersection $i$ (one minus the ratio of all-red time to cycle time);
$x(j)$	green split of phase $j$ (green time of phase $j$ divided by the cycle time of its corresponding intersection);
$x_L$	vector of minimal green splits for each phase (minimal green time allowed for each phase divided by the cycle time of its corresponding intersection);
$s$	saturation flow rate [veh/h];
$y$	endogenous queueing model variables, $y = (\lambda, \mu, \tilde{\mu}, \mu^{\text{eff}}, P(N = k), \mathcal{P})$ ;
$q$	exogenous queueing model parameters, $q = (\gamma, k, (p_{ij}))$ ;
$\mathcal{I}$	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 $\ell$ .

The problem is formulated as follows:

$$\min_{x,y} T(x, y; q) \quad (4.2)$$

subject to

$$\sum_{j \in \mathcal{P}_I(i)} x(j) = b_i, \quad \forall i \in \mathcal{I} \quad (4.3)$$

$$\mu_\ell - \sum_{j \in \mathcal{P}_L(\ell)} x(j)s = 0, \quad \forall \ell \in \mathcal{L} \quad (4.4)$$

$$h_1(y; q) = 0 \quad (4.5)$$

$$x \geq x_L \quad (4.6)$$

$$y \geq 0, \quad (4.7)$$

where  $h_1$  represents the traffic model presented in Section 4.3.5.

The objective is to reduce the average time that vehicles spend in the network, which is represented by  $T$  (Equation (4.2)). The expression for  $T$  is derived based on Little's law applied to the network and taking into account the finite capacity queueing framework. This leads to the following nonlinear objective function:

$$T(x, y; q) = \frac{\sum_i E[N_i]}{\sum_i \gamma_i (1 - P(N_i = k_i))}, \quad (4.8)$$

where  $E[N_i]$  is given by:

$$E[N_i] = \rho_i \left( \frac{1}{1 - \rho_i} - (k_i + 1) \frac{\rho_i^{k_i}}{1 - \rho_i^{k_i+1}} \right). \quad (4.9)$$

The details of how Equation (4.9) is derived are given at the end of this section.

The linear constraints (4.3) link the green times with the available cycle time for each intersection. Equation (4.4) defines the capacities of the signalized lanes as a function of the green times. Equation (4.5) consists of the system of nonlinear equations given in Section 4.3.5. The bounds (4.6) correspond to minimal green time values for each phase. These have been set to 4 seconds according to the Swiss standard VSS (1992).

The optimization problem is solved with the Matlab routine for constrained nonlinear problems, *fmincon*, which resorts to a sequential quadratic programming method (Coleman and Li, 1996, 1994). A feasible initial point is obtained by fixing a control plan and solving the network model (Equations (4.4), (4.5) and (4.7)). Details on the solution procedure of this system of equations as well as on its own initialization settings are given in Section 2.4.3.

At the solution, both the maximum constraint violation and the relative gradient of the Lagrangian are smaller than the threshold,  $10^{-6}$ . The use of relative gradient information to test for optimality is detailed in Dennis and Schnabel (1996), and further described in the context of constrained optimization in Conn et al. (2000). The choice of the threshold is based on the criteria given in Dennis and Schnabel (1996).

### Derivation of $E[N]$

$E[N]$  is defined as:

$$E[N] = \sum_{n=0}^k n P(N = n). \quad (4.10)$$

The stationary probabilities for each queue,  $P(N = n)$ , are given in Bocharov et al. (2004) by:

$$P(N = n) = \frac{1 - \rho}{1 - \rho^{k+1}} \rho^n. \quad (4.11)$$

Inserting Equation (4.11) into (4.10), and then rearranging the terms yields

$$\begin{aligned} E[N] &= \sum_{n=0}^k n \frac{1 - \rho}{1 - \rho^{k+1}} \rho^n \\ &= \sum_{n=1}^k n \frac{1 - \rho}{1 - \rho^{k+1}} \rho^n. \end{aligned} \quad (4.12)$$

$$\begin{aligned}
E[N] &= \frac{1-\rho}{1-\rho^{k+1}} \sum_{n=1}^k n\rho^n \\
&= \frac{1-\rho}{1-\rho^{k+1}} \rho \sum_{n=1}^k n\rho^{n-1}.
\end{aligned} \tag{4.13}$$

We then derive an expression for the last summation as follows. For a geometric series, such that  $\rho \neq 1$ , we have:

$$\sum_{n=0}^k \rho^n = \frac{\rho^{k+1} - 1}{\rho - 1}. \tag{4.14}$$

We differentiate this formula with respect to  $\rho$  and obtain:

$$\sum_{n=1}^k n\rho^{n-1} = \frac{1-\rho^{k+1}}{(1-\rho)^2} - \frac{(k+1)\rho^k}{1-\rho}. \tag{4.15}$$

Inserting this expression into the equation of  $E[N]$ , and rearranging the terms gives:

$$\begin{aligned}
E[N] &= \frac{1-\rho}{1-\rho^{k+1}} \rho \left( \frac{1-\rho^{k+1}}{(1-\rho)^2} - \frac{(k+1)\rho^k}{1-\rho} \right) \\
&= \rho \left( \frac{1}{1-\rho} - \frac{(k+1)\rho^k}{1-\rho^{k+1}} \right).
\end{aligned} \tag{4.16}$$

## 4.5 Empirical analysis

### 4.5.1 Microscopic traffic simulation model of the city of Lausanne

To perform the empirical analysis, we use a calibrated microscopic traffic simulation model of the Lausanne city center. This model (Dumont and Bert, 2006) is implemented with the AIMSUN simulator (TSS, 2008). Figure 4.3 presents a map of the city of Lausanne, along with the delimited area that is accounted for in the simulation model. The corresponding road network model is displayed in Figure 4.4. It contains 652 roads and 231 intersections, 49 of which are signalized. We use this model for two purposes.

Firstly, we use it to extract the network data (e.g. road characteristics, demand distribution) needed to estimate the exogenous parameters of the queueing model. The intersection characteristics include an existing fixed-time signal control plan of the city of Lausanne. For more information concerning this control plan we refer the reader to Dumont and Bert (2006). Based on this control plan, we give initial values to the capacities of the signalized lanes.

The demand distribution is described in terms of roads, whereas we require lane specific distributions. We describe how we convert the road-specific distribution to the lane-specific

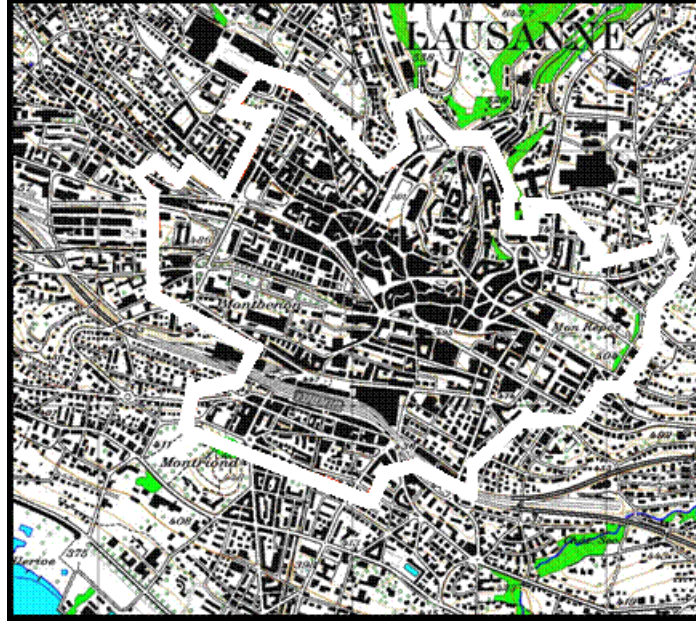


Figure 4.3: Map of the city of Lausanne. Adapted from Dumont and Bert (2006).

distribution. For each road, we have three types of flow data: external outflow (flow that leaves the network), road-to-road turning flow and external inflow (flow that arises from outside of the network). In order to obtain lane specific distributions, we disaggregate the flow data as follows.

**External outflow** We assume that this flow is distributed with equal probability across all the lanes of the road. If the road is modeled with several segments, the outflow is associated with the last (most downstream) segment. In other words, departures only occur at the end of the road.

**Turning flow** We consider that this flow is distributed with equal probability across all the lanes involved in the turning.

**External inflow** If the road is modeled with several segments, then the external inflow is distributed with equal probability across all the lanes of the first (most upstream) segment. In other words, arrivals only occur at the beginning of the road.

If the road is modeled as a single segment, then the external inflow that arrives to a given road  $r_1$  and then proceeds to road  $r_2$ , is distributed with equal probability across all the lanes of  $r_1$  that can turn to  $r_2$ .

Secondly, we use this simulation model to evaluate and compare the performance of different signal plans. Once a new plan is determined, it is integrated in the simulation model, its performance is evaluated and then compared with that of other plans.

We now compare the performance of several methodologies, by considering a subnetwork of the Lausanne city center. The subnetwork is located in the city center, and is delimited in



Figure 4.4: Lausanne city road network model.

Figure 4.4 by an oval. The subnetwork is also displayed in detail in Figure 4.5. It contains 48 roads and 15 intersections. Nine intersections are signalized and control the flow of 30 roads. There are a total of 51 phases that are considered variable. The intersections have a cycle time of either 90 or 100 seconds. For each methodology, we derive the optimal signal plan for the subnetwork, and then use the simulation model to evaluate its effect upon the entire Lausanne network.

The simulation setup consists of 100 replications of the evening peak period (17h-19h), preceded by a 15 minute warm-up time. Within this time period, congestion gradually increases. The average flow of the roads in the subnetwork steadily decreases from 339 to 25 (veh/h); and the average density increases from 10 to 57 (veh/km).

The queueing model of this subnetwork consists of 102 queues. The optimization problem consists of 621 endogenous variables with their corresponding lower bound constraints, 408 nonlinear equality constraints and 171 linear equality constraints.

### 4.5.2 Between-queue interactions

The queueing model proposed in this chapter describes congestion by taking into account the interactions between upstream and downstream roads. In this section, we illustrate the added value of accounting for these interactions. We compare this model with the same

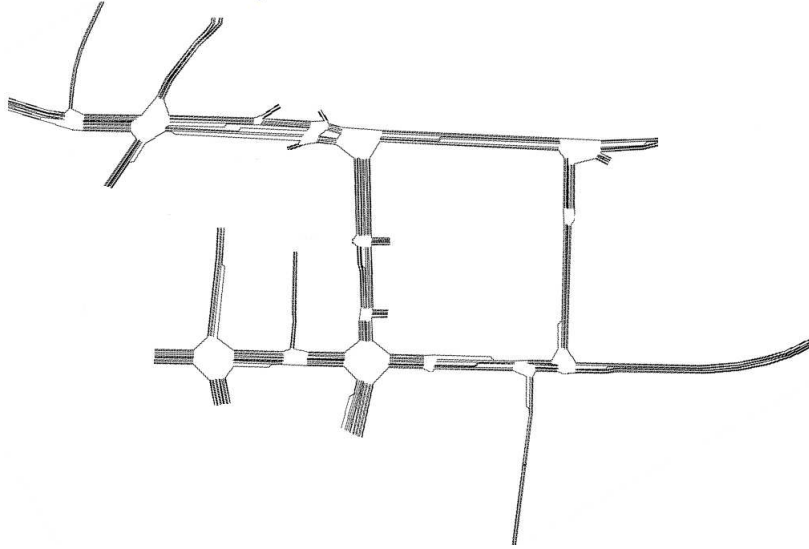


Figure 4.5: Subnetwork of the Lausanne city center.

model where independence of the queues is assumed.

The optimization problem is solved for both queueing models (correlated queues versus independent queues), and the performance of the corresponding signal plans are compared. We denote these as the *correlated* and the *independent* plans, respectively.

Assuming independent queues leads to the following simplifications:

- the arrival rates are now exogenous, they are determined by flow conservation equations;
- the effective service rates are no longer linked to the potential spillbacks of downstream roads, i.e. the total time spent on a road is entirely determined by its capacity.

We consider the network and simulation setup described in Section 4.5.1. Figure 4.6 displays the empirical cumulative distribution functions (cdf's) of the average travel times over the 100 replications for both methods. This figure shows that the correlated method improves the distribution of the average travel times.

To test whether the difference in the average travel times is significant we perform a t-test. We assume that the observed average travel times arise from a normal distribution with common but unknown variance. The null hypothesis assumes that the expected travel time is the same for both methods, whereas the alternative hypothesis assumes that they differ.

The travel time averages and standard deviations that the t-test is based on are given in Table 4.1. Table 4.2 displays the results of the t-test that examines whether the difference in the average travel time, between the correlated and the independent plan, is statistically significant. We set the confidence level of the test to 0.05. Each sample consists of 100 observations, which leads to a test with 198 degrees of freedom. The corresponding critical

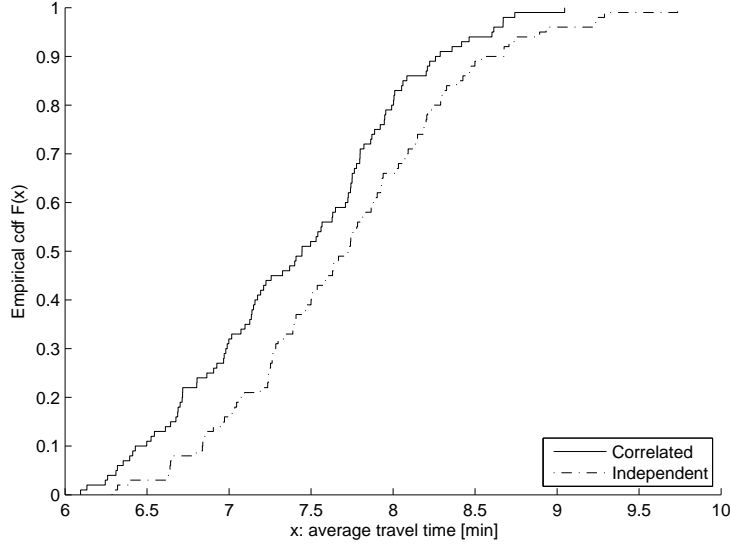


Figure 4.6: Empirical cumulative distribution functions of the average travel times for the independent and the correlated methods.

	Average [min]	Standard deviation
Correlated	7.41	0.68
Independent	7.71	0.70

Table 4.1: Travel time statistics based on 100 replications.

value for is 1.97. The t-statistic is equal to -3.07 which indicates that there is a statistically significant difference in average travel time.

To further evaluate the performance of the plans, we consider the average number of vehicles that have exited each origin-destination (OD) pair at a given time. The simulation time is segmented into 40 3-minute intervals. Figure 4.7 displays for each time interval a boxplot of the difference between the average number of vehicles for the independent and the correlated plans. Each point within a boxplot represents this difference for a given OD pair. This figure illustrates how the number of OD pairs that have a higher flow under the correlated plan than under the independent one increases as congestion increases.

This figure also shows that there is no difference for the majority of the OD pairs. It makes sense since only 51% of the 2096 OD pairs have more than 2 trips assigned per hour,

T-statistic	-3.07
degrees of freedom:	198
confidence level:	0.05
hypothesized average difference	0

Table 4.2: T-statistics assuming equal variance. The statistic considers the difference between the correlated method and the independent method.

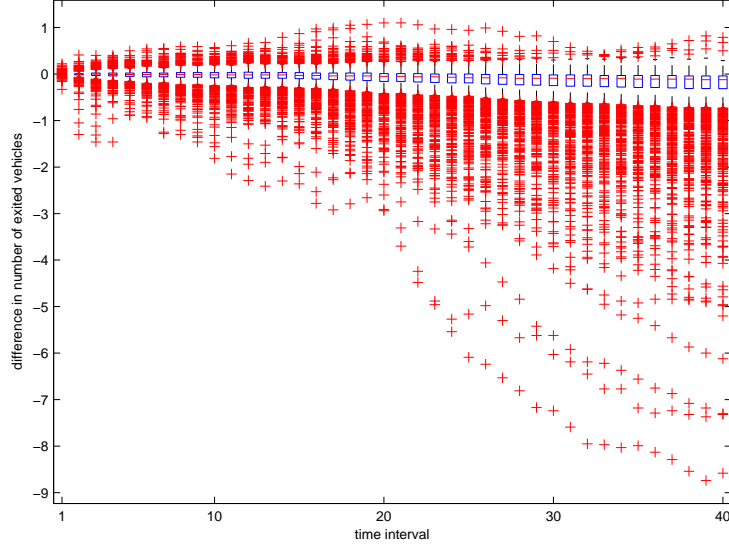


Figure 4.7: Difference in the average number of vehicles that have exited each OD pair versus time. Difference between the independent and the correlated plans. Negative values correspond to an increase in throughput.

14% have more than 10 trips and 6.6% have more than 20 trips. Thus, for the majority of the OD pairs we would not expect a difference larger than a couple of vehicles.

Figure 4.8 displays the empirical cdf of these differences for the intervals 10, 20, 30 and 40. It also shows that as congestion increases there is a higher proportion of OD pairs that perform better when the correlation is taken into account. The asymmetry of Figures 4.7 and 4.8 are evidence of the added value of accounting for the dependence of queues in signal optimization.

Figure 4.9 displays the density averaged across replications and across the roads of the network and of the subnetwork, respectively. These averages are plotted versus time. The crosses denote the independent plan, the circles represent the correlated plan. Recall that we consider a highly congested scenario, where congestion gradually increases. These figures illustrate how the proposed method delays the propagation of congestion, by delaying the increase in density at both the subnetwork and the network scale. These results illustrate well the added value of the method, not only on the global throughput but also locally.

### 4.5.3 Comparison with existing methods

We now compare the signal settings derived by the method proposed in this chapter (previously referred to as the correlated method) with an existing fixed-time signal setting for the city of Lausanne, with the method derived by Webster (1958) and with the method suggested in the Highway Capacity Manual (TRB, 2000).

**Base plan** The calibrated simulation model of the Lausanne city center is based on an existing fixed-time signal control plan. For more information concerning this control

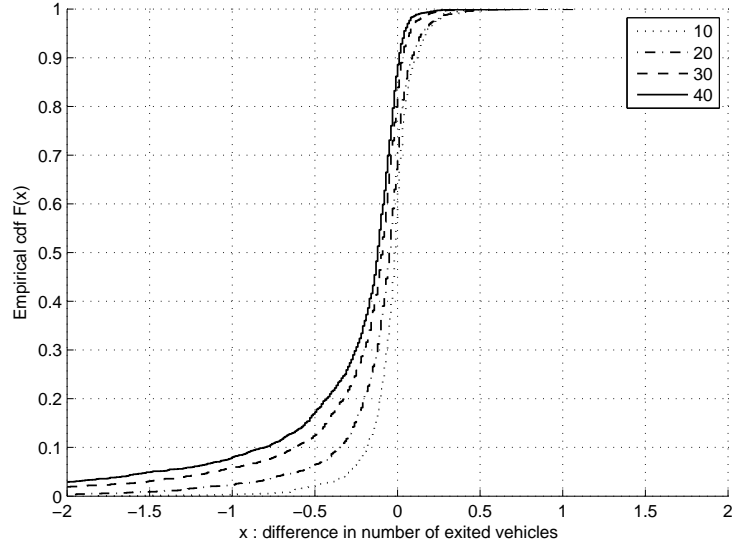


Figure 4.8: Empirical cumulative distribution function of the difference in the average number of vehicles that have exited the OD pairs for time intervals 10, 20, 30 and 40. Difference between the independent and the correlated plans. Negative values correspond to an increase in throughput.

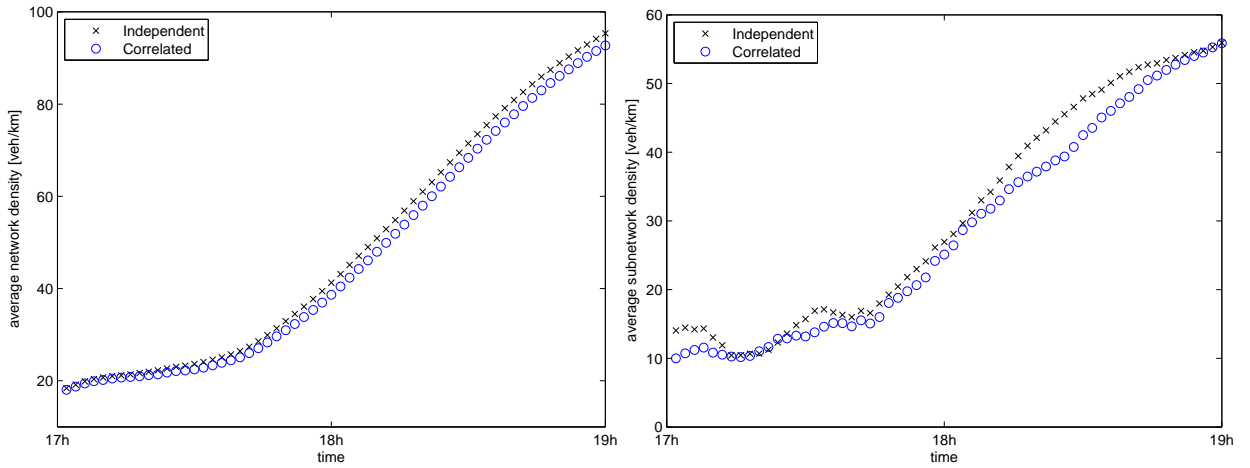


Figure 4.9: Average network and subnetwork density plotted versus time.

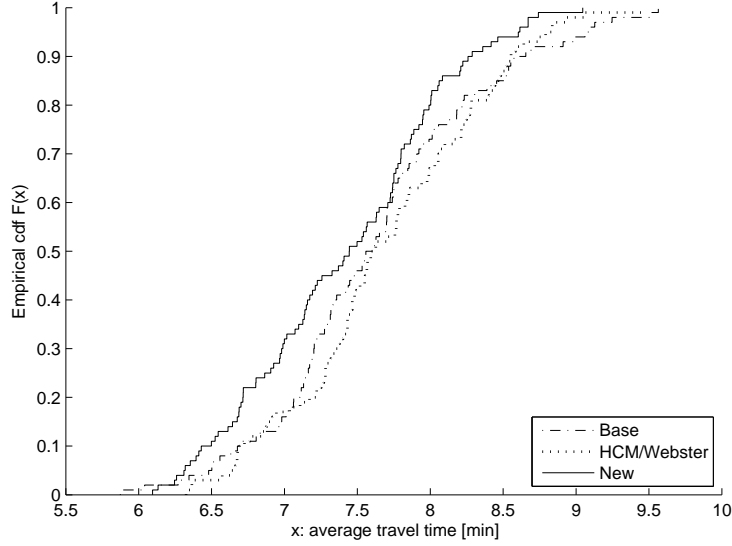


Figure 4.10: Empirical cumulative distribution functions of the average travel times for the base plan, HCM/Webster plan and the new plan.

plan, we refer the reader to Dumont and Bert (2006). This signal plan will be referred to as the *base* plan.

**HCM/Webster** Webster’s method is described in detail in Appendix B. By allocating the green times such that the flow to capacity ratios for the critical movements of each phase are equal, the method suggested in the Highway Capacity Manual (TRB, 2000) leads to the same green split equations as Webster’s method (1958). This equivalence is detailed in Appendix B.

Once again, we consider the network and simulation setup described in Section 4.5.1. Figure 4.10 displays the empirical cdf’s of the average travel times over the 100 replications for both methods. This figure shows that the plan derived by the proposed method improves the distribution of the average travel times, when compared to both the base plan and the HCM/Webster plan.

We perform t-test’s to determine whether the difference in the average travel times is statistically significant. Table 4.3 presents the average and standard deviations of each sample. Table 4.4 displays the results of the t-tests. Recall that the critical value is 1.97. Table 4.4 indicates that the difference in average travel times is significant. Thus, there is an improvement in average travel time when comparing the proposed method to the base plan and also to the HCM/Webster plan.

We compare the methods in terms of the average number of vehicles that have exited each OD pair across time. The description of how these comparisons are carried out is given in Section 4.5.2. Figure 4.11 displays the empirical cdf’s when comparing the new plan to the base plan (left plot) and to the HCM/Webster plan (right plot). This figure shows that there is a high proportion of OD pairs for which the new plan yields an increase in

	Average [min]	Standard deviation
New	7.41	0.68
HCM/Webster	7.69	0.67
Base	7.63	0.75

Table 4.3: Travel time statistics based on 100 replications.

	T-statistic
HCM/Webster	-2.95
Base	-2.17
degrees of freedom: 198	
confidence level: 0.05	
hypothesized mean difference: 0	

Table 4.4: T-statistics assuming equal variance. Each statistic considers the difference between the proposed (new) method and the method indicated in the corresponding row.

outflow. Furthermore, this proportion increases with congestion. The asymmetry of this figure illustrates the superiority of the proposed method.

Figure 4.12 displays the average density across the network and the subnetwork. The densities are plotted versus time. The crosses, squares and circles denote the base plan, the HCM/Webster plan and the new plan, respectively. As in the previous section, we observe how the proposed method delays the propagation of congestion, by delaying the increase in density at both the subnetwork and the network scale.

Figure 4.13 considers the flow and the travel time averaged across the roads of the subnetwork and across replications. These plots illustrate how the new plan leads to improved average travel times and a slight improvement in flow.

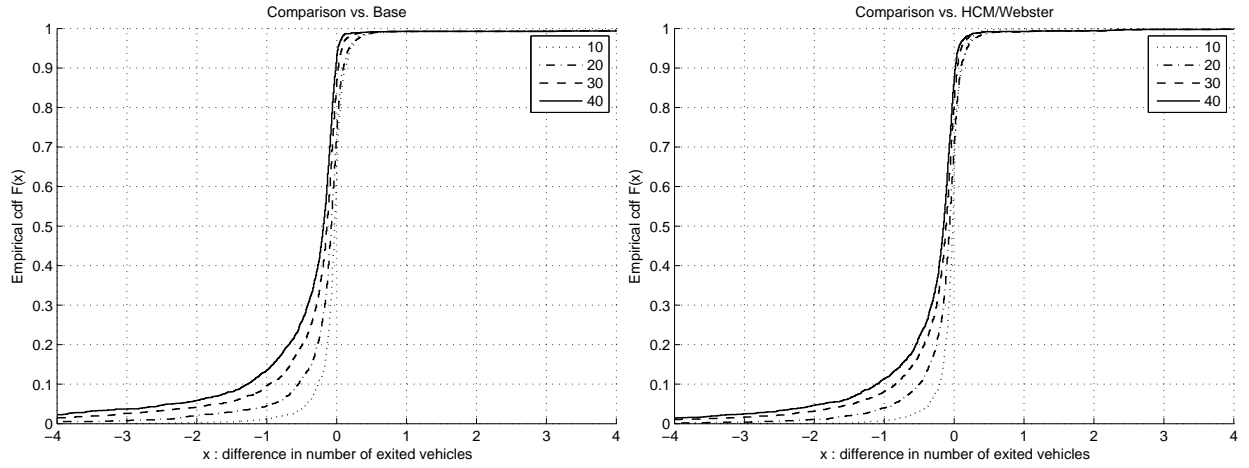


Figure 4.11: Empirical cumulative distribution function of the difference in the average number of vehicles that have exited the OD pairs for time intervals 10, 20, 30 and 40. The left (resp. the right) plot displays the difference between the base (resp. HCM/Webster) plan and the proposed method.

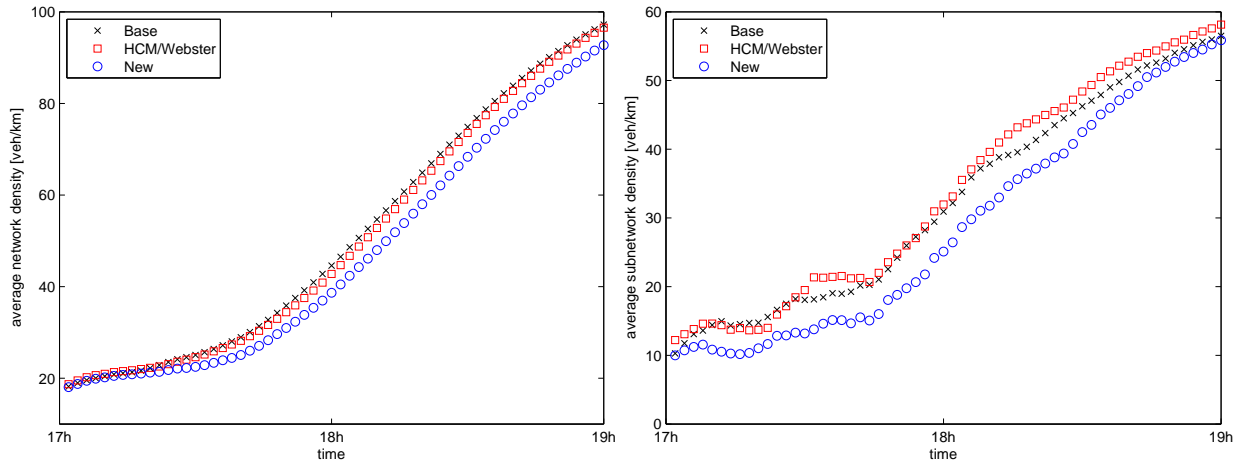


Figure 4.12: Average network and subnetwork density plotted versus time.

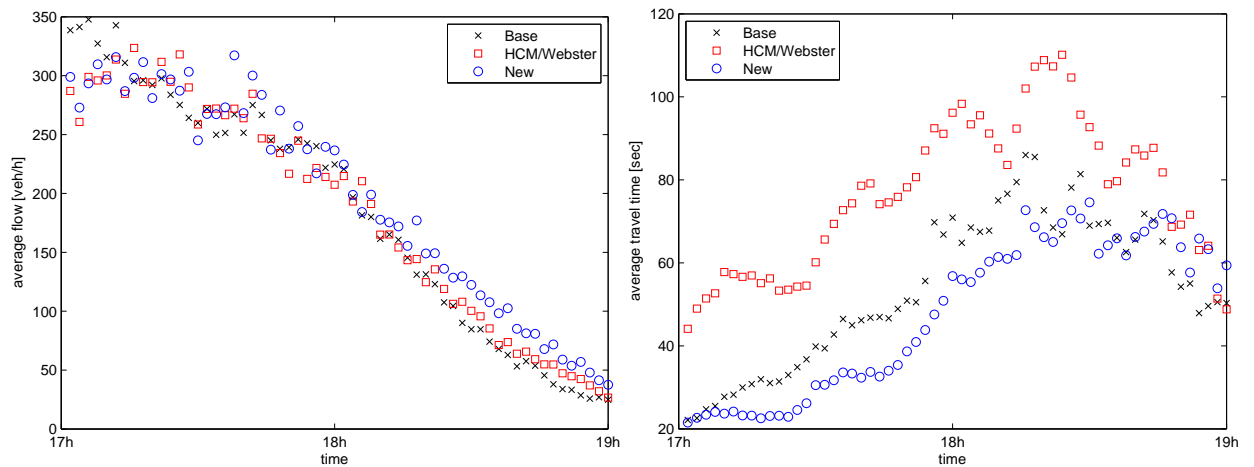


Figure 4.13: Average flow and travel time of the roads of the subnetwork, plotted versus time.

## 4.6 Large-scale networks

As detailed in Section 4.3.5, the System of Equations (4.1) consists of  $6n$  equations, where  $n$  is the number of queues. The queueing model for the entire Lausanne city road network consists of 922 queues, this leads to a set of 5532 equations. In this section, we present a formulation of the traffic model that reduces the number of equations to  $3n$ . In the case of the Lausanne city network this yields a system of 2766 equations. We then use this formulation to solve a signal control problem for the entire Lausanne city road network.

### 4.6.1 Formulation

First, we insert Equation (4.1a) in Equation (4.1g) to obtain:

$$\rho_i = \frac{\lambda_i^{\text{eff}}}{\mu_i^{\text{eff}}(1 - P(N_i = k_i))}. \quad (4.17)$$

Second, we insert Equation (4.1d) in (4.1c), which yields:

$$\frac{1}{\mu_i^{\text{eff}}} = \frac{1}{\mu_i} + \mathcal{P}_i \sum_{j \in \mathcal{I}^+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}} \mu_j^{\text{eff}}}. \quad (4.18)$$

This last equation is equivalent to:

$$\frac{\lambda_i^{\text{eff}}}{\mu_i^{\text{eff}}} = \frac{\lambda_i^{\text{eff}}}{\mu_i} + \mathcal{P}_i \sum_{j \in \mathcal{I}^+} \frac{\lambda_j^{\text{eff}}}{\mu_j^{\text{eff}}}. \quad (4.19)$$

Inserting Equation (4.17) in (4.19) yields:

$$\rho_i(1 - P(N_i = k_i)) = \frac{\lambda_i^{\text{eff}}}{\mu_i} + \mathcal{P}_i \sum_{j \in \mathcal{I}^+} \rho_j(1 - P(N_j = k_j)). \quad (4.20)$$

Finally, we insert (4.1e) in this last equation and obtain:

$$\rho_i(1 - P(N_i = k_i)) = \frac{\lambda_i^{\text{eff}}}{\mu_i} + \left( \sum_j p_{ij} P(N_j = k_j) \right) \left( \sum_{j \in \mathcal{I}^+} \rho_j(1 - P(N_j = k_j)) \right). \quad (4.21)$$

Thus, the traffic model is also defined by the System of Equations (4.1a), (4.1b), (4.1d)-(4.1f), (4.17) and (4.21). The main feature is that these equations can now be decoupled; the system can be implemented in terms of only three endogenous variables:  $P(N_i = k_i)$ ,  $\lambda_i^{\text{eff}}$ ,  $\rho_i$ , which are given by the three equations: (4.1b), (4.1f) and (4.21). The system corresponds

to the following equations:

$$\lambda_i^{\text{eff}} = \gamma_i(1 - P(N_i = k_i)) + \sum_j p_{ji}\lambda_j^{\text{eff}} \quad (4.22a)$$

$$P(N_i = k_i) = \frac{1 - \rho_i}{1 - \rho_i^{k_i+1}} \rho_i^{k_i} \quad (4.22b)$$

$$\rho_i(1 - P(N_i = k_i)) = \frac{\lambda_i^{\text{eff}}}{\mu_i} + \left( \sum_j p_{ij}P(N_j = k_j) \right) \left( \sum_{j \in \mathcal{I}^+} \rho_j(1 - P(N_j = k_j)) \right) \quad (4.22c)$$

The system of equations consists of one linear and two nonlinear equations, and a total of  $3n$  equations, where  $n$  is the total number of queues. This formulation allows us to tackle larger networks.

Unless scalability is an issue, we have so far no reason to prefer one formulation over the other. It would be of interest to compare the performance, and in particular the numerical conditioning, of these two equivalent formulations on a variety of networks and scenarios.

#### 4.6.2 Lausanne city network

We use the traffic model presented in the previous section to solve the traffic control problem presented in Section 4.4. We consider the full road network of the city of Lausanne. The model considers 603 roads and 231 intersections. We determine the signals of 17 intersections, which have a cycle time of either 80, 90 or 100 seconds. There are a total of 99 phases that are considered variable. The road network is displayed in Figure 4.14, where the 17 intersections are depicted as filled squares.

The queueing model consists of 922 queues. The optimization problem consists of 2986 endogenous variables with their corresponding lower bound constraints, 1844 nonlinear equality constraints and 1060 linear equality constraints.

The simulation setup consists of 50 replications of the evening peak period (17h-19h), preceded by a 15 minute warm-up time. As described in Section 4.5.1, within this time period congestion gradually increases.

Firstly, we consider the base plan as the initial plan, and solve the optimization problem. Figure 4.15 displays the empirical cdf's of the average travel times over the 50 replications for both the proposed plan and the initial (base) plan. This figure shows that the proposed method improves the distribution of the average travel times.

Figure 4.16 displays the network density averaged across the 50 replications. These averages are plotted versus time. Figure 4.17 presents the network flow and travel time averaged across replications, and plotted versus time. Figures 4.16 and 4.17 illustrate how the proposed plan delays the propagation of congestion, by delaying the increase in density, the increase in travel time, as well as the decrease in flow.

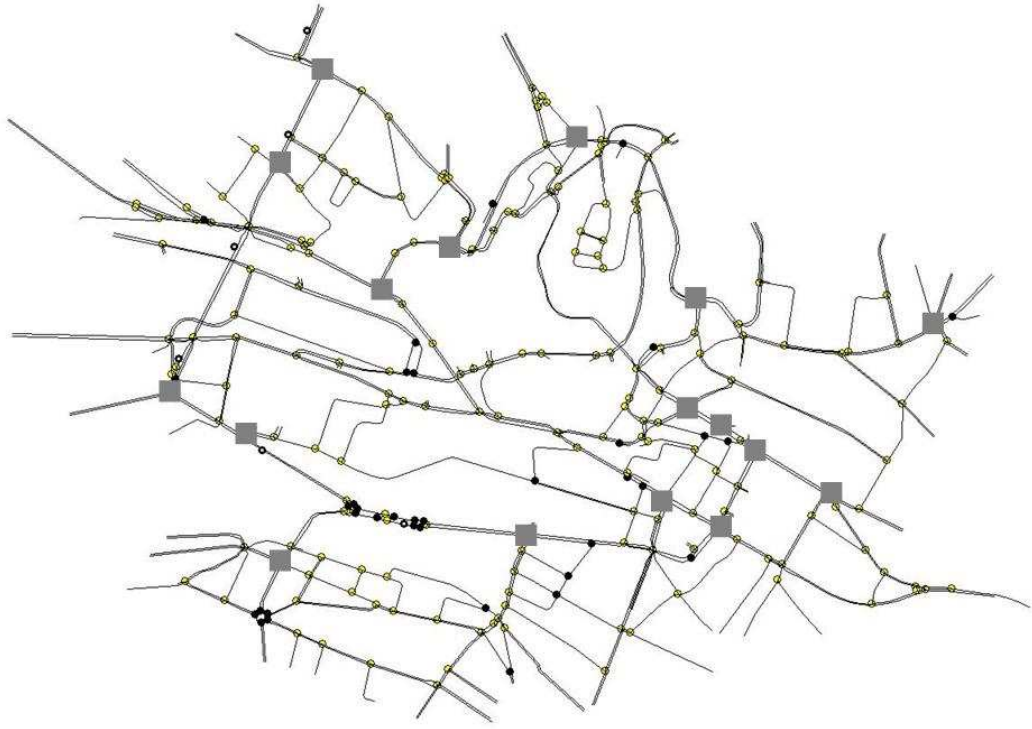


Figure 4.14: Set of endogenous signal plans in the Lausanne city network.

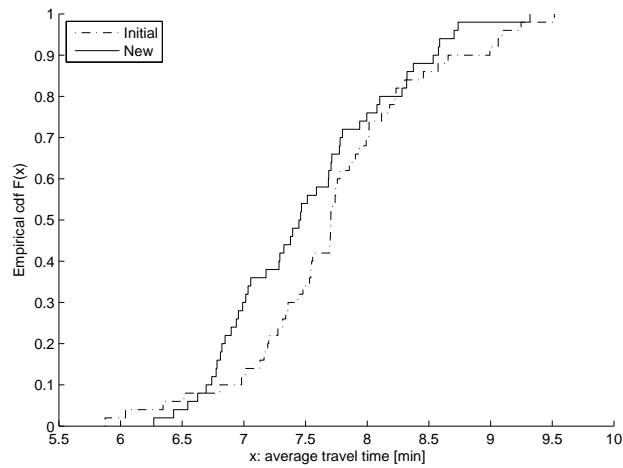


Figure 4.15: Empirical cumulative distribution functions of the average travel times for the initial and the proposed (new) plans, considering the base plan as the initial plan.

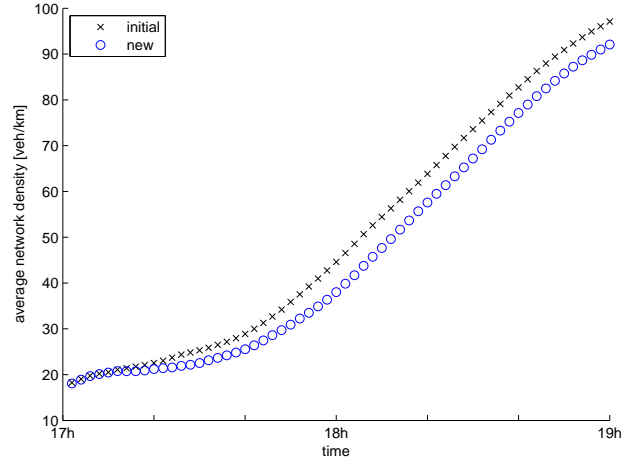


Figure 4.16: Average network density plotted versus time. The initial plan corresponds to the base plan.

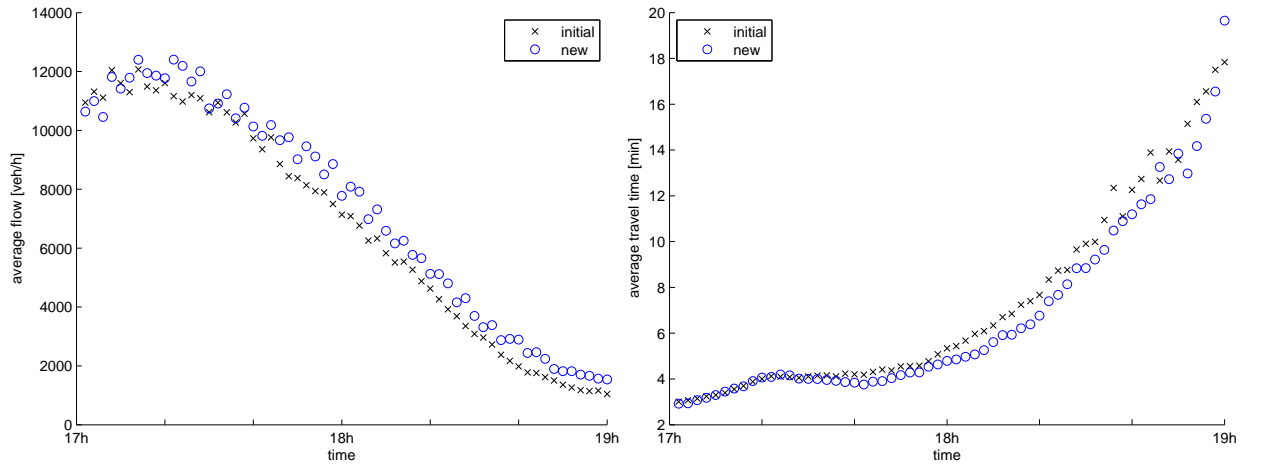


Figure 4.17: Average network flow and travel time plotted versus time. The initial plan corresponds to the base plan.

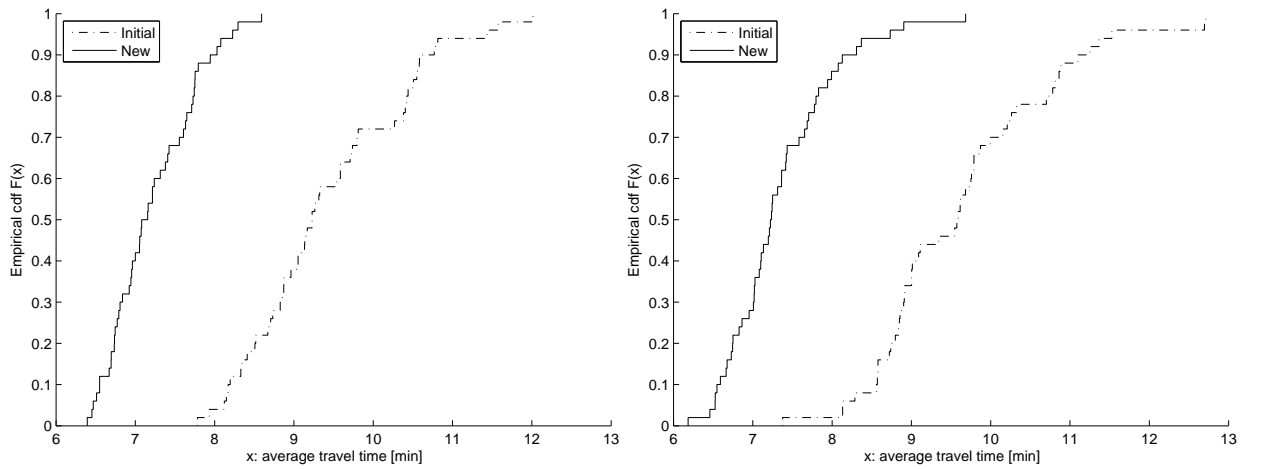


Figure 4.18: Empirical cumulative distribution functions of the average travel times for the initial and the optimal plans. Each plot considers a random uniformly drawn initial plan.

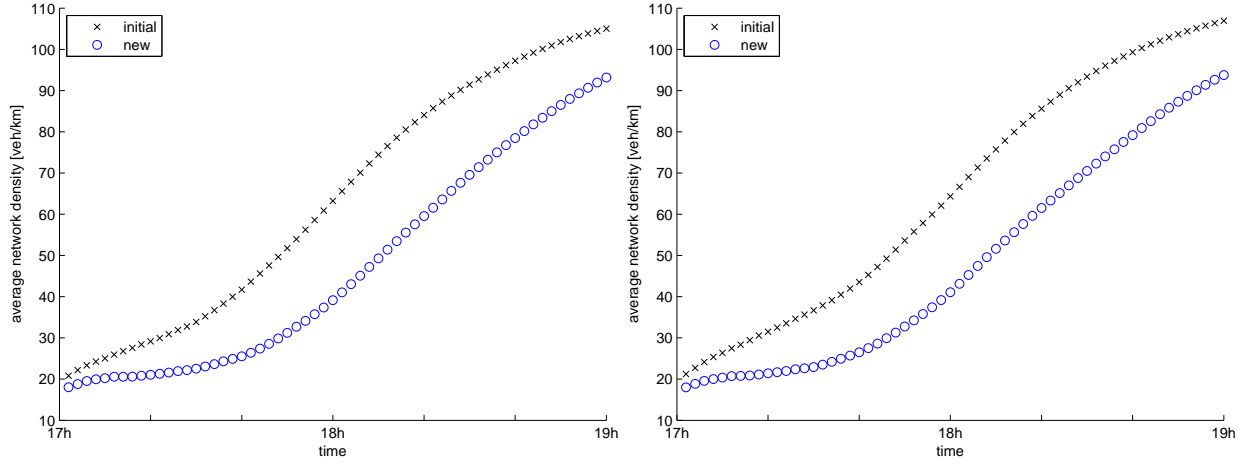


Figure 4.19: Average density plotted versus time. Each plot considers a random uniformly drawn initial plan.

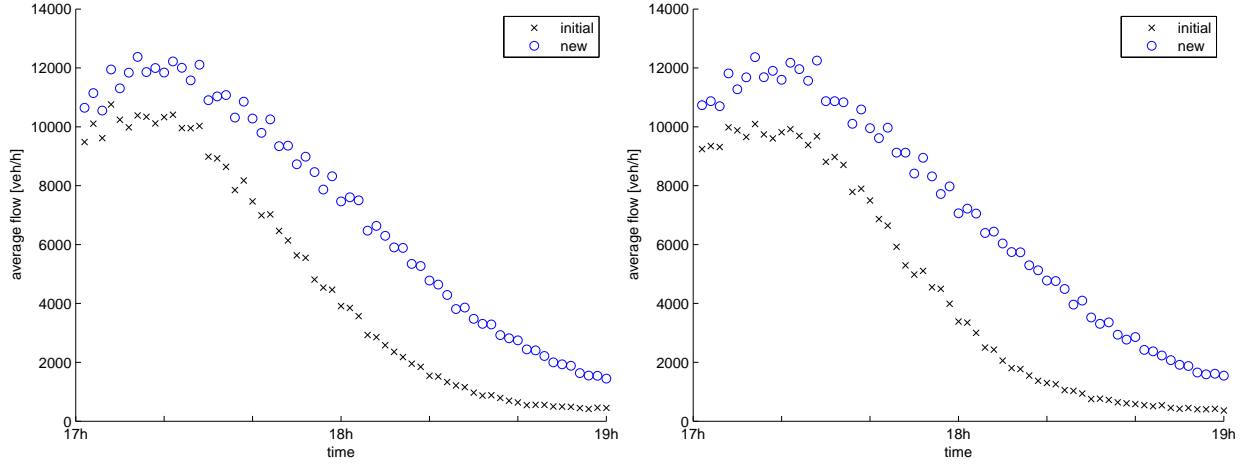


Figure 4.20: Average flow plotted versus time. Each plot considers a random uniformly drawn initial plan.

Secondly, we evaluate the performance of this method given uniformly drawn random initial signal plans. To generate these plans we use the method of Stafford (2006). We consider two different random plans.

Each plot of Figure 4.18 displays the empirical cdf's of the average travel times over the 50 replications for both the proposed plan and the random initial plan. The proposed plans improve the distribution of the average travel times.

Figure 4.19 displays the network density averaged across replications. These averages are plotted versus time. Each plot of the figure considers a random initial plan. Figures 4.20 and 4.21 present the average flow and travel times for each initial plan. These performance measures are also averaged across replications, and plotted versus time. Figures 4.19, 4.20 and 4.21, illustrate how the proposed method identifies well-performing signal plans when initialized with a random plan. For both initial plans and for all three performance measures, the proposed plans delay the propagation of congestion.

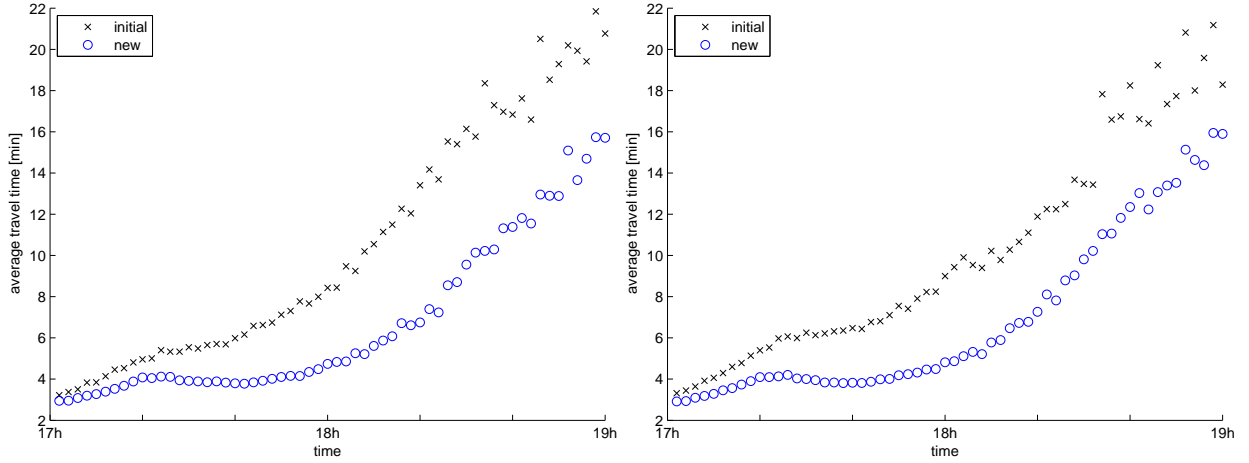


Figure 4.21: Average travel time plotted versus time. Each plot considers a random uniformly drawn initial plan.

## 4.7 Conclusions and future work

We have proposed a model based on a queueing network framework, which is to be used as a surrogate of traffic simulators to perform optimization for congested urban networks. As a specific example of such an approach, we have formulated a fixed-time traffic signal optimization problem, and have used the surrogate as the network model.

We have solved the signal control problem for a subnetwork of the city of Lausanne. The new signal plan has been evaluated with a microscopic traffic simulation tool. Its performance has been compared with the same model assuming independent queues, with a fixed-time plan that exists for the city of Lausanne, with Webster's method and with the method proposed by the Highway Capacity Manual. As congestion increases, the new method leads to improved performance measures.

We have also derived a formulation of the traffic model that is suitable for large-scale networks, and have used it to solve the signal control problem considering the entire Lausanne road network. We have considered three different initial plans: the existing Lausanne city signal plan and two uniformly drawn random plans. We have compared the performance of the proposed plans with that of the initial plans. For all three cases, the proposed plans improve the distribution of the average travel times, and delay the propagation of congestion.

The formulation of an urban road network using finite capacity queueing theory and accounting for multiple intersections is novel. Additionally, by using a set of structural parameters that capture the between-queue interactions, this queueing model approximates how congestion arises and how it spreads sufficiently well to be used as an appropriate surrogate. It characterizes congestion in terms of its sources, its frequency, its propagation and its impact. This approach, based on a fine decomposition of the phenomenon of congestion, is of general interest for traffic control, and particularly appropriate for the study and

management of congested urban networks.

There is a need for operational signal control methodologies that are suitable for congested conditions and that capture the complex features of congested traffic flows such as spillbacks. This model contributes to the development of these methodologies thanks to an analytical framework that makes an attractive trade-off between capturing the complexity of congested traffic flows and analytical tractability.

Clearly, the proposed model is not a fully realistic representation of spillbacks in signalized arterials. Therefore, there is a potential to investigate if more sophistication would preserve the tractability of the model, while enhancing the optimization. The realism of the model can be increased by accounting for the dynamic nature of traffic. This would allow us to describe in more detail the build-up and dissipation of congestion, and to evaluate the impact that this transient behavior has on congestion and on network performance.

# Chapter 5

## A simulation-based optimization approach for the management of congested urban road networks

### Contents

---

<b>5.1</b>	<b>Introduction . . . . .</b>	<b>88</b>
<b>5.2</b>	<b>Literature review . . . . .</b>	<b>90</b>
<b>5.3</b>	<b>Method . . . . .</b>	<b>93</b>
5.3.1	Metamodel . . . . .	93
5.3.2	Algorithmic framework . . . . .	96
5.3.3	Algorithm . . . . .	96
5.3.4	Algorithmic details . . . . .	98
<b>5.4</b>	<b>Optimization problem . . . . .</b>	<b>99</b>
5.4.1	Traffic signal control . . . . .	99
5.4.2	Trust region subproblem . . . . .	101
5.4.3	Signal plan features . . . . .	101
<b>5.5</b>	<b>Empirical analysis . . . . .</b>	<b>102</b>
5.5.1	Lausanne subnetwork with simplified demand distribution . . . . .	102
5.5.2	Lausanne subnetwork with evening peak hour demand . . . . .	105
<b>5.6</b>	<b>Conclusions and future work . . . . .</b>	<b>112</b>

---

## 5.1 Introduction

In the previous chapter, we derived an analytical urban traffic model. We then used it as a surrogate model to perform optimization, and in particular to solve a traffic signal control problem. In this chapter, we address the same problem from a simulation-based perspective. That is, we use observations from an urban traffic simulation model throughout the optimization process. The main motivation is to benefit from the more realistic and detailed performance measure estimates that can be obtained with microscopic simulation tools.

Microscopic urban simulators capture in detail the behavior of drivers as well as their interaction with the network infrastructure. They can provide accurate network performance estimates in the context of scenario-based analysis or sensitivity analysis. They are therefore often used to evaluate traffic management schemes. Nevertheless, using them to derive appropriate management schemes (i.e. to perform optimization) is a complex task.

An optimal traffic management scheme can be formulated as:

$$\min_{x, z \in \Omega} E[f(x, z; p)], \quad (5.1)$$

where the objective is to minimize the expected value of a suitable network performance measure,  $f$ . This performance measure is a function of a decision or control vector  $x$ , endogenous variables  $z$  and exogenous parameters  $p$ . The feasible space  $\Omega$  consists of a set of constraints that link  $x$  to  $z$ ,  $p$  and  $f$ .

For instance, a traffic signal control problem can take  $f$  as the average vehicle travel time and  $x$  as the green splits for the signalized lanes. Elements such as the total demand or the network topology will be captured by  $p$ , while  $z$  will account, for instance, for the capacities of the signalized lanes.

The various traffic models embedded within the simulator make it a detailed and realistic model, but lead to noisy nonlinear objective functions containing potentially several local minima. These objective functions have no available closed form; we can only derive estimates for  $E[f(x, z; p)]$ . Additionally, evaluating these estimates is computationally expensive because they involve running numerous replications. As a nonlinear, stochastic and computationally-expensive problem, it is complex to address.

### Metamodel methods

As is detailed in Section 5.2, one approach to perform simulation-based optimization (SO) is to build an analytical model of the objective function based on a sample of simulated

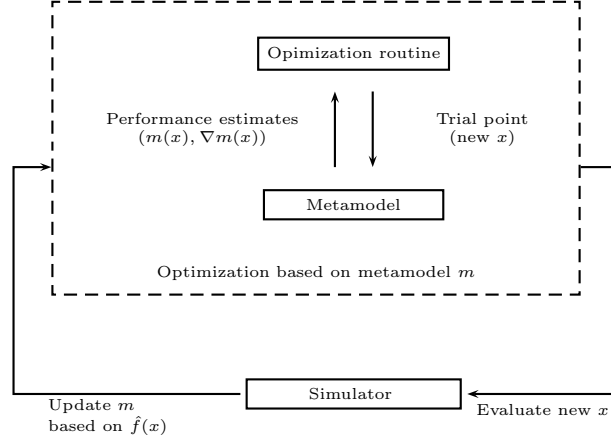


Figure 5.1: Metamodel simulation-based optimization methods. Adapted from Alexandrov et al. (1999).

observations. This analytical model is referred to as a metamodel or a surrogate model. This family of approaches is known as metamodel methods. Once the metamodel is constructed (e.g. fitted) it is used to perform optimization.

This approach is depicted in Figure 5.1. The metamodel is denoted as  $m$ , whereas the simulation response is denoted as  $\hat{f}$ . This figure illustrates the two main steps of metamodel methods. Firstly, the metamodel is constructed based on simulated observations. Secondly, once the metamodel  $m$  has been fitted, the optimization process derives a trial point based on the predictions and properties of  $m$ .

These steps can then be iterated as follows. For a given sample the metamodel is fitted, the optimization problem is solved, deriving a trial point. Then the performance of this trial point can be evaluated by the simulator, which leads to a new observation. As new observations become available the accuracy of the metamodel can be improved, leading ultimately to more reliable trial points.

Metamodels are typically deterministic functions that are cheaper to evaluate. By replacing the stochastic response of the simulation by a deterministic function, deterministic optimization techniques can be used. Furthermore, by using metamodels that are cheap to evaluate, the number of objective function evaluations is no longer a limitation. The main limitation remains the number of simulation runs needed such that an accurate metamodel can be built and well-performing trial points can be derived.

The most common metamodels (also called surrogates) used to perform simulation-based optimization are general-purpose models (e.g. polynomials, splines) that can be used to approximate any objective function, but capture little information about the structure of the underlying problem. Furthermore, they require a large initial sample to be fitted, and are thus inappropriate for applications with a tight computational budget.

We believe that in order to perform SO for congested urban networks given a limited

computational budget, these generic metamodels should be combined with surrogate network models that analytically capture the structure of the underlying problem, and potentially improve the short-term behavior of SO algorithms.

## Derivative-free optimization

Both the noise inherent in simulation outputs along with their high computational cost, makes the accurate estimation of derivatives an expensive and difficult task. When derivative information is either unavailable, available at a high cost or unreliable, then derivative-free (DF) optimization methods are an appropriate and common approach to tackle such problems.

Given the lack of derivative information, DF methods typically perform less well than their derivative-based counterparts. In particular, the scale of the problems that can be efficiently tackled is substantially reduced. Currently, the most recent DF methods are limited to around 20 variables for unconstrained problems and their convergence is typically rather slow (Conn et al., 2009b), not to mention their limitations in the presence of constraints. By using a surrogate that integrates structural information, we expect to be able to address both larger and constrained problems more efficiently.

Furthermore, DF methods are often used when function evaluations are computationally expensive. They therefore attempt to identify trial points with improved performance, given a fixed, and typically tight, computational budget. We expect the added structural information of the metamodel to allow the identification of good trial points even for tight computational budgets. In this chapter, we will evaluate the performance of the proposed metamodel considering scenarios with tight computational budgets and assuming that there are initially no observations available.

This chapter is structured as follows. Firstly, we present a literature review of the surrogate models used to perform SO and of the existing optimization algorithms that allow for arbitrary surrogates (Section 5.2). In Section 5.3, we present both the surrogate model and the optimization algorithm that will be used. We then show how this methodology applies to a fixed-time traffic signal control problem (Section 5.4), and present empirical results evaluating its performance (Section 5.5).

## 5.2 Literature review

There are two types of approaches to address SO problems. Firstly, there is the family of direct-search methods, which rely only on objective function evaluations and do not resort to any direct or indirect (i.e. explicit or implicit) derivative approximation or model building. In these methods, the search is based on a pre-determined geometric pattern or grid.

Secondly, there are the methods that do use gradient information. Barton and Meckesheimer (2006) review such methods and classify them into direct gradient and metamodel methods. Direct gradient methods estimate the gradient of the simulation response, whereas metamodel methods use an indirect-gradient approach by computing the gradient of the metamodel, which is a deterministic function.

In this chapter, we focus on the second family of methods, and in particular on metamodel methods. Although there have been significant advances and novel approaches for gradient estimation (Fu, 2006; Fu et al., 2005), methods that rely on direct derivative information often require more function evaluations, and are therefore inappropriate for applications with a limited computational budget. Additionally, by resorting to a metamodeling approach the stochastic response of the simulation is replaced by a deterministic metamodel response function, such that deterministic optimization techniques can be used.

Recent reviews of metamodels are given by Conn et al. (2009b), by Barton and Meckesheimer (2006) and by Søndergaard (2003). Metamodels are classified in the literature as either physical or functional metamodels (Søndergaard, 2003; Serafini, 1998). Physical metamodels consist of application-specific metamodels, whose functional form and parameters have a physical or structural interpretation.

Functional metamodels are generic (i.e. general-purpose) functions, that are chosen based on their analytical tractability, but do not take into account any information with regards to the specific objective function, let alone the structure of the underlying problem. They are often a linear combination of basis functions from a parametric family.

The most common approach is the use of low-order polynomials (e.g. linear or quadratic). Quadratic polynomials are used as surrogates in most trust region methods (Conn et al., 2000). Spline models have also been used, although their use within an SO framework has focused on univariate or bivariate functions, and as Barton and Meckesheimer (2006) mention: “unfortunately, the most popular and effective multivariate spline methods are based on interpolating splines, which have little applicability for SO”. Radial basis functions (Ouvray and Bierlaire, 2009; Wild et al., 2008) and Kriging surrogates (Booker et al., 1999) have also been proposed.

The existing metamodels consist of either physical or functional components. The metamodel proposed here goes beyond existing approaches by combining both a physical and a functional component. It combines an analytical network traffic model with a quadratic polynomial. The physical component is captured by the traffic model, whose parameters have a structural interpretation. For a given problem, the traffic model will yield a different functional form for the objective function. The functional component is captured by the general purpose quadratic polynomial.

In order to integrate the proposed metamodel within an existing optimization method,

we review the algorithms that allow for an arbitrary metamodel. These methods are called multi-model or hybrid methods. They share a common motivation, which is to combine the use of models with varying evaluation costs (low versus high-fidelity models, or coarse versus fine models).

A trust region optimization framework for unconstrained problems allowing for multiple models was proposed by Carter (1986) (see references herein for previous multi-model frameworks). His work analyses the theoretical properties and derives a global convergence theory for several types of multi-model algorithms. It allows for nonquadratic models as long as at least one model is a standard quadratic with uniformly bounded curvature.

The Approximation and Model Management Optimization/Framework (AMMO or AMMF) is a trust region framework for generating and managing a sequence of metamodels. There are several versions of the algorithm: for unconstrained problems (Alexandrov et al., 1998), bound constrained (Alexandrov et al., 2000), inequality constrained (Alexandrov et al., 1999) and generally constrained problems (Alexandrov et al., 2001). Although no restrictions are imposed on the type of surrogates allowed, it is a first-order method that requires that the model and the objective function, as well as their first-order derivatives, coincide at each major (or accepted) iterate. Thus the metamodel must always behave as a first-order Taylor series approximation. This is a strong restriction if the function is noisy and expensive to evaluate.

The Surrogate-Management Framework (SMF) proposed by Booker et al. (1999) is a derivative-free method for bound constrained problems. It is based on a direct search technique called pattern search. Since direct search techniques typically require many function evaluations, they use a surrogate model of the objective function to improve the performance of the algorithm. The surrogate model used is an interpolated Kriging model.

The Space Mapping (SM) technique and its many versions (Bandler et al., 2006; Bandler et al., 2004) is a simulation-based optimization technique that uses two metamodels: a fine and a coarse model. Both models are often simulation-based. The coarse model is constructed based on a transformation of the endogenous variables (“space mapping”) that minimizes the error for a sampled set of high-fidelity response values. Nevertheless, SM relies on the assumption that via a transformation of the endogenous variables the coarse model will exhibit the physical/mathematical properties of the fine model (Alexandrov and Lewis, 2001) and as Bandler et al. (2004) mention “the required interaction between coarse model, fine model, and optimization tools makes SM difficult to automate within existing simulators”. Alexandrov and Lewis (2001) give a comparison of the AMMO, the SMF and the SM methods.

Conn et al. (2009a) recently proposed a trust region derivative-free framework for unconstrained problems. This framework allows for arbitrary metamodels and makes no assumption on how these metamodels are fitted (interpolation or regression). To ensure global

convergence, a model improvement algorithm guarantees that the models achieve a uniform local behavior (i.e. satisfy Taylor-type bounds) within a finite number of steps.

Derivative-free (DF) methods do not require nor do they explicitly approximate derivatives. Resorting to a DF algorithm, rather than to first or second order algorithms, is appropriate when the derivatives are difficult to obtain, unreliable or computationally expensive to evaluate. This is the case for noisy problems, for problems where the evaluation of the objective function is computationally expensive, or for problems where the simulation source code is unavailable (Moré and Wild, 2009). In the field of transportation, most simulators fall into all three of these categories. Thus we will opt for a DF approach.

Among the two main strategies used to ensure global convergence, line search and trust region methods, the latter are more appropriate for our context since they “extend more naturally than line search methods to models that are not quadratics with positive Hessians” (Carter, 1986). The most common approach for fitting metamodels within a trust region (TR) framework is interpolation. Nevertheless, for noisy functions we believe that regression is more appropriate since it is less sensitive to the inaccuracy of the observations.

The framework proposed by Conn et al. (2009a), as a derivative-free TR method that allows for arbitrary models and does not impose interpolation, is therefore particularly appealing. We will therefore integrate the proposed metamodel within this framework.

## 5.3 Method

### 5.3.1 Metamodel

The metamodel combines information from two models: a simulation model and an analytical network model. We first present these two models, we then describe how they are combined.

**Simulation model.** We use the calibrated microscopic traffic simulation model of the Lausanne city center, which was presented in Section 4.5.1. For a given decision vector  $x$ , the simulator provides a realization  $\hat{f}(x, z; p)$  of the performance measure  $f(x, z; p)$  (presented in Equation (5.1)).

**Analytical queueing model.** The model used in this framework is the analytical urban traffic model formulated in Section 4.3. Alternatively, for large-scale networks the model of Section 4.6 can also be used.

In the previous chapter, this queueing model was used as a surrogate to perform optimization, and in particular to solve a fixed-time traffic control problem. In this chapter, it will be used to formulate a metamodel to perform simulation-based optimization.

We briefly, recall the main components of this model. By resorting to finite capacity queueing theory, it captures the key traffic interactions and the underlying network structure, e.g. how upstream and downstream queues interact, and how this interaction is linked to network congestion. The model consists of a system of nonlinear equations. It is formulated based on a set of exogenous parameters  $q$  that capture the network topology, the total demand, as well as the turning probabilities. A set of endogenous variables  $y$  describe the traffic interactions, e.g. spillback probabilities, average rates at which a spillback diffuses. For a given decision vector  $x$ , the network model yields the objective function  $T(x, y; q)$ , which is a deterministic approximation of  $E[f(x, z; p)]$ .

We recall here the notation that we have introduced so far:

- $x$  decision vector;
- $T$  approximation of the objective function derived by the queueing model;
- $\hat{f}$  performance measure observation derived by the simulation model;
- $y$  endogenous queueing model variables;
- $z$  endogenous simulation variables;
- $q$  exogenous queueing model parameters;
- $p$  exogenous simulation parameters.

We now describe how  $\hat{f}$  and  $T$  are combined to derive the metamodel  $m$ . The main idea of trust region methods is to build, at each iteration, a model of the objective function which one “trusts” in a neighborhood of the current iterate, the *trust region*. The most common approach is to use a quadratic polynomial. The proposed metamodel combines a quadratic polynomial with a deterministic approximation of the objective function, provided by the analytical network model. The functional form of  $m$  is:

$$m(x, y; \alpha, \beta, q) = \alpha T(x, y; q) + \phi(x; \beta), \quad (5.2)$$

where  $\phi$  is a quadratic polynomial in  $x$ ,  $\alpha$  and  $\beta$  are parameters of the metamodel.

The polynomial  $\phi$  is quadratic in  $x$  with a diagonal second derivative matrix. This choice is based on existing numerical experiments for derivative-free TR methods, which show that these types of quadratic polynomials are often more efficient than full quadratic polynomials (Powell, 2003).

$$\phi(x; \beta) = \beta_1 + \sum_{j=1}^d \beta_{j+1} x_j + \sum_{j=1}^d \beta_{j+d+1} x_j^2, \quad (5.3)$$

where  $d$  is the dimension of  $x$ ,  $x_j$  and  $\beta_j$  are the  $j^{th}$  components of  $x$  and  $\beta$ , respectively.

At each iteration of a trust region algorithm the objective function is evaluated at a set of points. The model is then constructed based on objective function observations. Traditionally, trust region methods fit the polynomial via interpolation. In this framework, we fit the metamodel via regression. At each iteration, the simulator and the queueing model are evaluated at one (in some cases two) point(s). The metamodel is fitted using the observations obtained at the current iteration, as well as all observations collected at previous iterations.

The parameters  $\beta$  and  $\alpha$  of the metamodel are fitted by solving a least squares problem. At a given iteration, the model approximates the objective function in a neighborhood of the current iterate. In order to give more importance to observations that correspond to points that are near the current iterate, we associate weights to each observation. The least squares problem is formulated as follows.

$$\min_{\alpha, \beta} \sum_{i=1}^{n_k} \left\{ w_{ki} \left( \hat{f}(x^i, z^i; p) - m(x^i, y^i; \alpha, \beta, q) \right) \right\}^2 + (w_0(\alpha - 1))^2 + \sum_{i=1}^{2d+1} (w_0 \beta_i)^2, \quad (5.4)$$

where  $x^i$  represents the  $i^{th}$  point in the sample, with corresponding endogenous simulation variables  $z^i$ , endogenous queueing model variables  $y^i$  and observation  $\hat{f}(x^i, z^i; p)$ . The sample size at iteration  $k$  is  $n_k$ . The weight associated at iteration  $k$  to the  $i^{th}$  observation is denoted  $w_{ki}$ . The parameter  $w_0$  represents a fixed weight, its role will be discussed further on.

The first squared term of Equation (5.4) represents the weighted distance between the simulated observations and the metamodel predictions. The next two squared terms measure the distance between the parameters and their initial values. These terms ensure that the least squares matrix is of full rank. The initial values used here (one for  $\alpha$  and zero for  $\beta$ ) lead to an initial metamodel that is based only on the queueing model. This is of interest when starting off the algorithm with few or even no observations.

The weights  $w_{ki}$  capture the importance of each point with regards to the current iterate. The work of Atkeson et al. (1997) gives a survey of weight functions and analyzes their theoretical properties. We use what is known as the *inverse distance* weight function along with the Euclidean distance. This leads to the following weight parameters:

$$w_{ki} = \frac{1}{1 + \|x_k - x^i\|_2}, \quad (5.5)$$

where  $x_k$  is the current iterate, and  $x^i$  is the  $i^{th}$  sample point.

The weight of a given point is therefore inversely proportional to its distance from the current iterate. This allows us to approximately have a Taylor-type behavior, where observations corresponding to local points have more weight. The least squares problem is solved

using the Matlab routine *lsqlin* (The Mathworks, 2008).

### 5.3.2 Algorithmic framework

For an introduction to trust region (TR) methods, we refer the reader to Conn et al. (2000). They summarize the main steps of a TR method in the *Basic trust region algorithm*. This algorithm is presented in Appendix C.1. The method proposed by Conn et al. (2009a) builds upon the *Basic TR algorithm* by adding two additional steps: a model improvement step and a criticality step. This algorithm is given in Appendix C.2. For a detailed description, see Conn et al. (2009a).

A given iteration  $k$  of the algorithm considers a metamodel  $m_k$ , an iterate  $x_k$  and a TR radius  $\Delta_k$ . Hereafter, the subscript  $k$  refers to the iteration. Each iteration consists of 5 steps:

- **Criticality step.** This step may modify  $m_k$  and  $\Delta_k$  if the measure of stationarity is close to zero.
- **Step calculation.** Approximately solve the TR subproblem to yield a trial point.
- **Acceptance of the trial point.** The actual reduction of the objective function is compared to the reduction predicted by the model, this determines whether the trial point is accepted or rejected.
- **Model improvement.** Either certify that  $m_k$  is *fully linear* (i.e. satisfies Taylor-type bounds) in the TR or attempt to improve the accuracy of the metamodel.
- **TR radius update.**

### 5.3.3 Algorithm

The algorithm used in this framework follows. We then provide details regarding the implementation of each step of the algorithm.

#### 0. Initialization. Set

- an initial point  $x_0$ ,
- an upper bound for the trust region radius  $\Delta_{max} > 0$ ,
- an initial trust region radius  $\Delta_0 \in (0, \Delta_{max}]$ ,
- the parameters  $\eta_1, \gamma, \gamma_{inc}, \epsilon_c, \bar{\tau}, \bar{d}, \bar{u}$  such that
  - $0 < \eta_1 < 1$ ,
  - $0 < \gamma < 1 < \gamma_{inc}$ ,

- $\epsilon_c > 0$ ,
  - $0 < \bar{\tau} < 1$ ,
  - $0 < \bar{d} < \Delta_{max}$ ,
  - $\bar{u} \in \mathbb{N}^*$ ,
- the maximum number of function evaluations (i.e. simulation runs) permitted  $n_{max}$ .
  - Define
    - $\alpha_k$  and  $\beta_k$  as the metamodel parameters at iteration  $k$ ,
    - $\nu_k$  as the vector of parameters of  $m_k$ ,  $\nu_k = (\alpha_k, \beta_k)$ ,
    - $n_k$  as the sample size,
    - $u_k$  as the number of successive trial points rejected,
    - $g_k$  the gradient of the Lagrangian evaluated at  $x_k$ .
  - Compute  $T$  and  $\hat{f}$  at  $x_0$ , fit an initial model  $m_0$ , and compute  $\nu_0$ .
  - Set  $k = 0, n_0 = 1, u_0 = 0$ .
1. **Criticality step.** If  $\|g_k\| \leq \epsilon_c$ , then switch to *conservative mode* (detailed in Section 5.3.4).
  2. **Step calculation.** Compute a step  $s_k$  that “sufficiently reduces the model”  $m_k$  and such that  $x_k + s_k \in B(x_k; \Delta_k)$  (i.e. approximately solve the TR subproblem).
  3. **Acceptance of the trial point.** Compute  $\hat{f}(x_k + s_k)$  and

$$\rho_k = \frac{\hat{f}(x_k) - \hat{f}(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

- If  $\rho_k \geq \eta_1$ , then accept the trial point:  $x_{k+1} = x_k + s_k$ ,  $u_k = 0$ .
- Otherwise, reject the trial point:  $x_{k+1} = x_k$ ,  $u_k = u_k + 1$ .

Include the new observation in the sample set ( $n_k = n_k + 1$ ), and fit the new model  $m_{k+1}$ .

4. **Model improvement.** Compute

$$\tau_{k+1} = \frac{\|\nu_{k+1} - \nu_k\|}{\|\nu_k\|}. \quad (5.6)$$

If  $\tau_{k+1} < \bar{\tau}$ , then improve the model by sampling a new point  $x$ , evaluate  $T$  and  $\hat{f}$  at  $x$ . Include this point in the sample set ( $n_k = n_k + 1$ ). Update  $m_{k+1}$ .

## 5. Trust region radius update.

- If  $\rho_k > \eta_1$ , then increase the trust region radius:  

$$\Delta_{k+1} = \min\{\gamma_{inc}\Delta_k, \Delta_{max}\}.$$
- Otherwise,
  - if  $u_k \geq \bar{u}$ , then reduce the trust region radius:  

$$\Delta_{k+1} = \max\{\gamma\Delta_k, \Delta_{min}\}, u_k = 0,$$
  - otherwise,  $\Delta_{k+1} = \Delta_k$ .
- If  $\Delta_{k+1} \leq \bar{d}$ , then switch to *conservative mode*.

Set  $n_{k+1} = n_k, u_{k+1} = u_k$ .

Set  $k = k + 1$ .

If  $n_k < n_{max}$ , then go to Step 1.

### 5.3.4 Algorithmic details

**Criticality step** The criticality step of the algorithm ensures that if the measure of stationarity goes under a given threshold, then the model can be improved so that its stationarity measure can be trusted. The model is then said to be *certifiably fully linear* (i.e. it satisfies Taylor-type bounds). We assume throughout that we cannot certify whether the model is fully linear. If at a given iteration, the measure of stationarity does go under the criticality threshold then a purely quadratic metamodel along with an appropriate sampling strategy (e.g. Monte Carlo, Quasi-Monte Carlo) can be used in order to obtain an accurate gradient estimate and to certify full linearity. This is denoted as the *conservative mode* in the algorithm.

**Step calculation** Details regarding the TR subproblem are given for the traffic signal control problem in Section 5.4.2. .

**Model improvement step** At each iteration, we run the simulator at the trial point,  $x_k + s_k$  (Step 3 of the algorithm). In order to diversify the set of sampled points, we may sample points other than the trial points. This step attempts to improve the accuracy of the model, by improving the geometric properties of the sampled space (e.g. attempting to fully span the feasible space such that a full rank least squares matrix is obtained, or in the case of interpolation methods improving the poisenedness of the sample (Conn et al., 2009b)). We do so if the condition  $\tau_{k+1} < \bar{\tau}$  is satisfied. To sample we draw uniformly from the feasible space.

**TR radius update** In the Conn et al. (2009a) algorithm the TR radius can be reduced if the model is *fully linear* but has not performed well. Since we assume throughout

that we cannot certify whether the model is *fully linear*, we reduce the TR radius after  $\bar{u}$  successive trial points have been rejected. If the TR radius reaches a lower bound  $\bar{d}$ , then a quadratic polynomial with an appropriate sampling strategy is used, and as mentioned previously, we can ensure that within a uniformly bounded number of sampling steps the model will be *fully linear*.

**Algorithmic parameters** The following values are used for the parameters of the TR algorithm:

- $\Delta_{max} = 10^{10}$ ,
- $\Delta_0 = 10^3$ ,
- $\eta_1 = 10^{-3}$ ,
- $\gamma = 0.9$ ,
- $\gamma_{inc} = 1.2$ ,
- $\epsilon_c = 10^{-6}$ ,
- $\bar{\tau} = 0.1$ ,
- $\bar{d} = 10^{-2}$ ,
- $\bar{u} = 10$ ,
- $w_0 = 0.1$ .

Typical values for TR parameters are given in Carter (1986). For the algorithm used to solve the TR subproblem we set the tolerance for relative change in the objective function to  $10^{-3}$  and the tolerance for the maximum constraint violation to  $10^{-2}$ .

## 5.4 Optimization problem

### 5.4.1 Traffic signal control

We illustrate the use of this framework with a signal control problem for a subnetwork of the city of Lausanne. A review of the different formulations, as well as the definitions of the traffic signal terms used hereafter, is given in Section 4.2.2. We consider the same problem as in Section 4.4, i.e. we consider a fixed-time signal control problem where the offsets, the cycle times and the all-red durations are fixed. The stage structure is also given. In other words, the set of lanes associated with each stage as well as the sequence of stages are both known. To formulate this problem we recall the notation introduced in Section 4.4:

- $b_i$  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;
- $\mathcal{P}_I(i)$  set of phase indices of intersection  $i$ .

The problem is traditionally formulated as follows:

$$\min_{x,z} E[f(x, z; p)] \quad (5.7)$$

subject to

$$\sum_{j \in \mathcal{P}_I(i)} x(j) = b_i, \quad \forall i \in \mathcal{I} \quad (5.8)$$

$$x \geq x_L. \quad (5.9)$$

Let us recall that the decision vector  $x$  consists of the green splits for each phase. The objective is to minimize the expected travel time (Equation (5.7)). The linear constraints (5.8) link the green times of the phases with the available cycle time for each intersection. The bounds (5.9) correspond to minimal green time values for each phase. These have been set to 4 seconds according to the Swiss transportation norm (VSS, 1992).

As detailed by Conn et al. (2009b), DF TR methods are a relatively recent topic. The algorithms developed so far are derived based on sound theoretical properties that lead to a solid global convergence theory, but they are mostly formulated for unconstrained problems. Unfortunately, the optimization problems encountered in practice are rarely unconstrained. Conn et al. (2009b) review constrained DF algorithms, and confirm that for constrained problems “currently, there is no convergence theory developed for TR interpolation-based methods”, not to mention TR methods that allow for regression models.

Conn et al. (1998) propose a method to solve problems with general constraints using an unconstrained TR algorithm. The traffic management problems that we are interested in solving fall into the category of what they denote as *easy* constraints. These are general constraints that are continuously differentiable and whose first order partial derivatives can be computed relatively cheaply (with regards to the cost of evaluating the objective function). In their approach, they include such constraints in the TR subproblem, which ensures that all trial points are feasible. Conn et al. (2009b) mention that such an approach is often sufficient in practice.

Here we use the TR algorithm proposed by Conn et al. (2009a) for unconstrained methods, and extend its use to constrained problems as Conn et al. (1998) suggest. That is, we include the constraints in the TR subproblem to ensure that all trial points are feasible.

The next section formulates the TR subproblem.

### 5.4.2 Trust region subproblem

At a given iteration  $k$  the TR subproblem includes three more constraints than the previous problem. It is formulated as follows:

$$\min_{x,y} m_k = \alpha_k T(x, y; q) + \phi(x; \beta_k) \quad (5.10)$$

subject to

$$\sum_{j \in \mathcal{P}_I(i)} x(j) = b_i, \quad \forall i \in \mathcal{I} \quad (5.11)$$

$$h_2(x, y; q) = 0 \quad (5.12)$$

$$\|x - x_k\|_2 \leq \Delta_k \quad (5.13)$$

$$y \geq 0 \quad (5.14)$$

$$x \geq x_L, \quad (5.15)$$

where  $x_k$  is the current iterate,  $\Delta_k$  is the current trust region radius,  $\alpha_k$  and  $\beta_k$  are the current metamodel parameters, and  $h_2$  denotes the queueing model. Equation (5.12) consists of Equations (4.4) and (4.5), the corresponding endogenous variables are subject to positivity constraints (Equation (5.14)). The analytical form of  $T$  is given by Equation (4.8). Constraint (5.13) is the TR constraint. It uses the Euclidean norm (Conn et al., 2009a). Thus the TR subproblem consists of a nonlinear objective function subject to nonlinear and linear equalities, a nonlinear inequality and bound constraints. This problem is solved with the Matlab routine for constrained nonlinear problems, *fmincon*, which resorts to a sequential quadratic programming method (Coleman and Li, 1996; Coleman and Li, 1994).

### 5.4.3 Signal plan features

**Sampling.** The model improvement step of the algorithm attempts to diversify the set of sampled points by drawing points uniformly from the feasible space. A feasible signal plan is defined by Equations (5.8) and (5.9) (or equivalently Equations (5.11) and (5.15)). We draw uniformly from this space, using the code of Stafford (2006). Given this signal plan, we solve the network model (Equation (5.12)) following the procedure described in Section 2.4.3.

**Explanatory/independent variables.** The polynomial component of the metamodel,  $\phi$ , is a quadratic polynomial in the decision variables  $x$ , which are the phase variables of the different intersections. For a given intersection the phase variables are linked

through the linear Equation (5.8). To reduce the correlation between the explanatory variables of the metamodel, we exclude one phase per intersection. Thus for a set of  $i$  intersections and  $p$  phases, the polynomial is a function of  $p - i$  phase variables, and has a total of  $2(p - i) + 1$  coefficients.

## 5.5 Empirical analysis

In this section, we evaluate the performance of the proposed method on two Lausanne city subnetworks. Firstly, we consider a simplified demand distribution. Secondly, we analyze the performance of the method given the demand of the city of Lausanne for the evening peak hour, and control the plans of a larger set of intersections.

To refer to the metamodel or to its components we use the notation of Equation (5.2). In both sections, we compare the performance of the proposed metamodel,  $m$ , to that of two other metamodels:

- a quadratic polynomial with diagonal second derivative matrix, (i.e. the metamodel consists of  $\phi$ ),
- the queueing model (i.e. the metamodel consists of  $T$ ). This is the procedure presented in the previous chapter. Namely, this procedure uses the same algorithm as the one used to solve the TR subproblem.

### 5.5.1 Lausanne subnetwork with simplified demand distribution

We consider the Lausanne road network with a simplified demand distribution. We control a set of two adjacent signalized intersections. Demand arises at the nine centroids nearest to these two intersections. The simulation setup considers a 20 minute scenario, preceded by a 15 minute warm-up time.

A total of 13 phases are considered variable (i.e. the dimension of the decision vector is 13). This leads to a polynomial with 23 coefficients. The queueing model considers 12 roads that are connected to either of these two intersections. These roads are modeled as a set of 21 queues. The corresponding TR subproblem consists of 131 endogenous variables with their corresponding lower bound constraints, 84 nonlinear and 36 linear equalities.

Firstly, we consider the performance of the proposed metamodel  $m$  and of the polynomial  $\phi$ . For both metamodels, we run the TR algorithm and allow for a total of 750 simulation runs. To initialize both methods we consider a uniformly drawn initial signal plan. To generate these plans we use the method of Stafford (2006). Initially, no simulated observations are available, i.e. we start off with an empty sample.

We compare the performance of these methods for increasing sample sizes. To evaluate the performance of a given signal plan, we run 50 replications of the simulation model. All

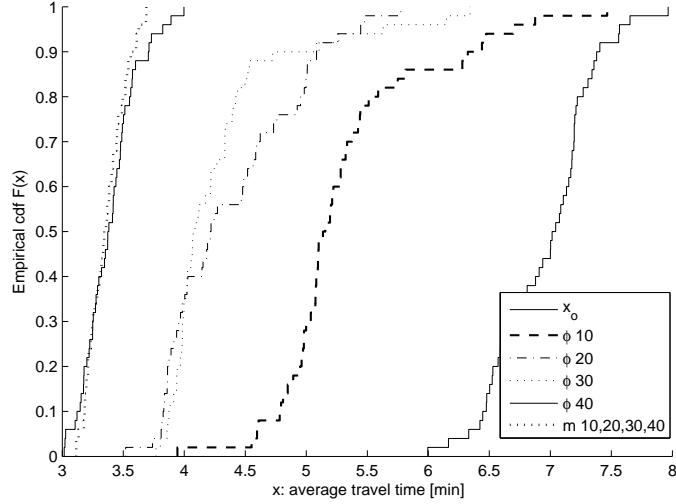


Figure 5.2: Empirical cumulative distribution functions of the average travel times considering an initial random signal plan and evaluating the performance as the sample size increases from 10 to 40.

simulations are preceded by a 15 minute warm-up period. We then compare the distribution of the average travel times.

Figure 5.2 considers the signal plans derived by the polynomial method and the proposed method (i.e. metamodels  $\phi$  and  $m$ , respectively) at sample sizes 10, 20, 30 and 40. For each signal plan, the figure displays the empirical cumulative distribution functions (cdf's) of the average travel times over the 50 replications. The plans  $m$ ,  $\phi$  and  $x_0$  denote, respectively, the plans derived by the proposed metamodel, the polynomial metamodel and the initial random plan. The numbers denote the corresponding sample sizes, e.g. the cdf denoted “ $\phi$  10” denotes the signal plan proposed by the polynomial with a sample of size 10.

The signal plan derived by the proposed method is the same at sample sizes 10, 20, 30 and 40. At a sample size of 10 both  $m$  and  $\phi$  lead to improved average travel times, when compared to the initial plan. As the sample size increases, the polynomial leads to plans with improved performance. At sample size 40, its performance is similar to that of the signal plan proposed by  $m$ .

Figure 5.3 considers the signal plans proposed at sample sizes 40, 50, 100, 250, 500 and 750. For each signal plan, the cdf of the average travel times over the 50 replications are displayed. This figure shows that the performance of the plans is similar for sample sizes larger than 40.

Figure 5.4 considers the signal plans proposed by  $m$  and  $\phi$  at sample size 750, as well as the initial signal plan and the signal plan proposed by the queueing model  $T$ . It displays for each method the cdf of the average travel times. All three methods,  $m$ ,  $\phi$  and  $T$ , lead to improved performance compared to the random signal plan. The methods that use simulated observations throughout the optimization process,  $m$  and  $\phi$ , lead to improved signal plan performance when compared to the queueing method  $T$ .

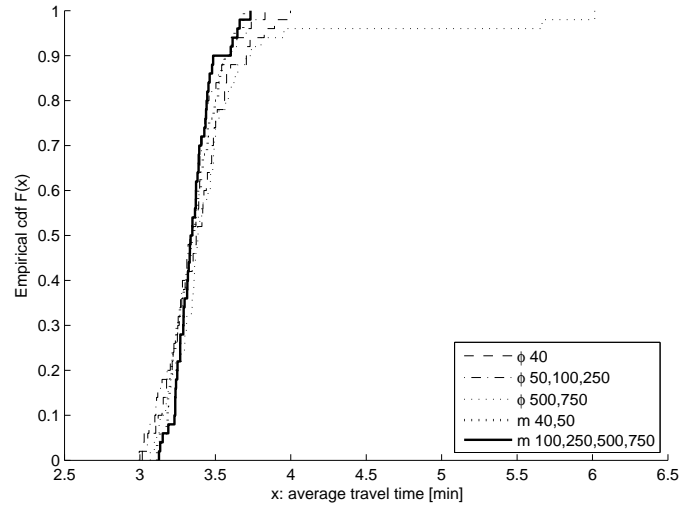


Figure 5.3: Empirical cumulative distribution functions of the average travel times considering an initial random signal plan and evaluating the performance as the sample size increases from 40 to 750.

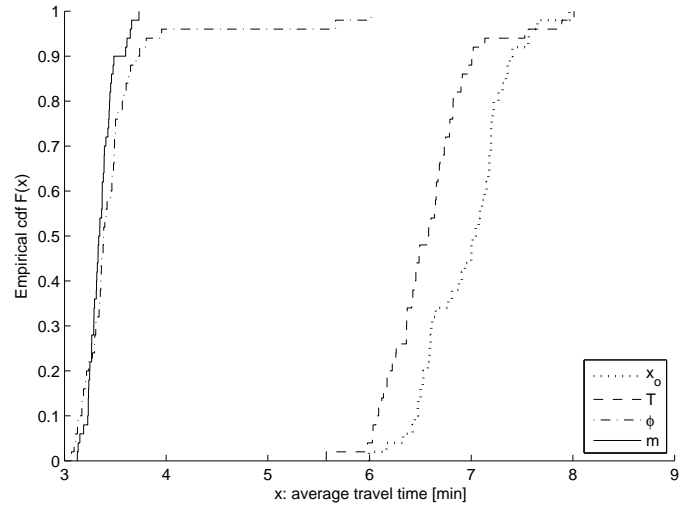


Figure 5.4: Empirical cumulative distribution functions of the average travel times considering an initial random signal plan and evaluating the performance at sample size 750.

We consider the full sample (750 observations) and test whether the metamodel parameters of the proposed model are significantly different from zero. To do so we perform a t-test. The null hypothesis assumes that the parameters are equal to zero, whereas the alternative hypothesis assumes they differ. We set the confidence level of the test to 0.05. The corresponding critical value is 1.96. Recall that there are 23 parameters. Nine are significantly different from zero. These 9 parameters concern 5 linear terms, 2 quadratic terms, the queueing model parameter ( $\alpha$ ) and the intercept ( $\beta_1$ ). This indicates that the proposed metamodel indeed captures information about the relationship between the observed travel times and the phase variables.

The results of this section indicate that for small to moderate sample sizes (compared to the dimension of the decision vector) the structural information provided by the queueing model leads to well-performing signal plans. As the sample size increases the polynomial metamodel improves its accuracy, and achieves with a moderate sample size similar performance to that of the proposed metamodel. By comparing the proposed model to the queueing model, the results indicate that the simulated observations indeed improve the accuracy of the model, leading to signal plans that reduce the average travel times.

### 5.5.2 Lausanne subnetwork with evening peak hour demand

We evaluate the performance of the proposed method by considering a subnetwork of the Lausanne city center. The subnetwork was presented in Section 4.5.1. The considered scenario consists of the evening peak period (17h-18h). The simulation outputs used both to fit the metamodel and to evaluate the performance of the derived signal plans are the subnetwork average travel times.

The queueing model of this subnetwork consists of 102 queues. The TR subproblem consists of 621 endogenous variables with their corresponding lower bound constraints, 408 nonlinear equality constraints, 171 linear equality constraints and 1 nonlinear inequality constraint.

Note that this problem is considered a large-scale problem for existing unconstrained DF methods, not to mention the added complexity of the nonlinear constraints. In particular, the problem has 51 decision variables. Thus if one were to resort to a classical interpolation-based quadratic polynomial surrogate, 1378 function evaluations would be necessary to fit the full polynomial. This is because for a problem with  $n$  decision variables  $(n + 1)(n + 2)/2$  suitably sampled points (i.e. well poised (Conn et al., 2000; Conn et al., 2009b)) are necessary to fit the full quadratic.

We consider a tight computational budget, which is defined as a maximum number of simulation runs that can be carried out, and no initial observation available. The computational budget is set to 150 runs. For a given initial signal plan, we run the corresponding algorithm 10 times, deriving 10 signal plans. We then evaluate the performance of each

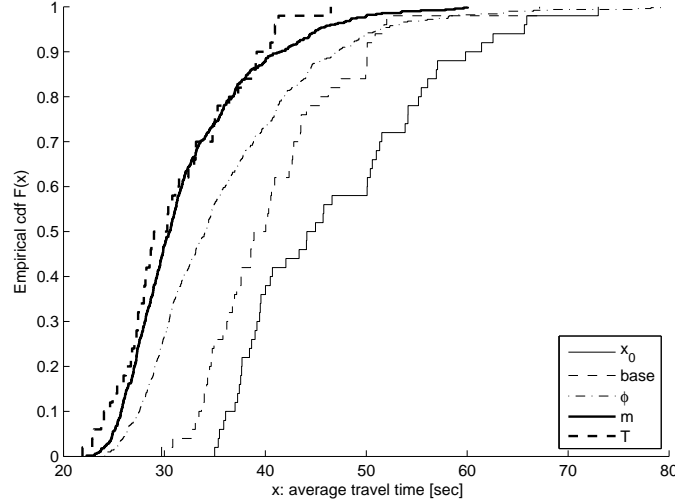


Figure 5.5: Empirical cumulative distribution functions of the average travel times considering an initial random signal plans and running 10 instances of each method. Each instance consists of 150 simulation runs.

	Average [sec]	Standard deviation
$m$	32.16	6.55
$\phi$	35.96	8.57
$T$	31.08	5.76
$x_0$	46.8	9.50

Table 5.1: Travel time statistics for the different signal plans, obtained based on 500 observations. The methods have been initialized considering the initial plan  $x_0$ .

of these signal plans by running 50 replications of the simulation model. All simulations are preceded by a 15 minute warm-up period. To compare the methods, we consider the empirical cumulative distribution function (cdf) of the average travel times over all 10 signal plans and 50 replications, i.e. each cdf consists of a set of 500 observations.

Firstly, we consider the performance of these methods given a uniformly drawn initial signal plan, which we generate with the method of Stafford (2006). The plot of Figure 5.5 considers a random initial plan and presents the cdf's of the average travel times. The plans  $m$ ,  $T$ ,  $\phi$  and  $x_0$  denote, respectively, the plans derived by the proposed metamodel, the queueing model, the polynomial metamodel and the initial random plan. The plan denoted by *base plan* is an existing signal plan for the city of Lausanne.

Figure 5.5 indicates that all methods have an improved performance when compared to both the base and the initial plans. Both the proposed metamodel and the queueing model derive signal plans with improved performance compared to the polynomial.

The travel time statistics for the different methods are displayed in Table 5.1. To test whether the difference in the average travel times is significant we perform t-test's. The assumptions of these tests are described in Section 4.5.2. The t-statistics are given in

	$m$	$\phi$	$x_0$
$\phi$	7.88		
$x_0$	28.38	18.96	
$T$	-2.76	-10.57	-31.66
degrees of freedom: 998			
confidence level: 0.05			
hypothesized mean difference: 0			

Table 5.2: T-statistics assuming equal variance. Each statistic is calculated based on the difference between the corresponding row and column methods. This scenario considers a random initial plan.

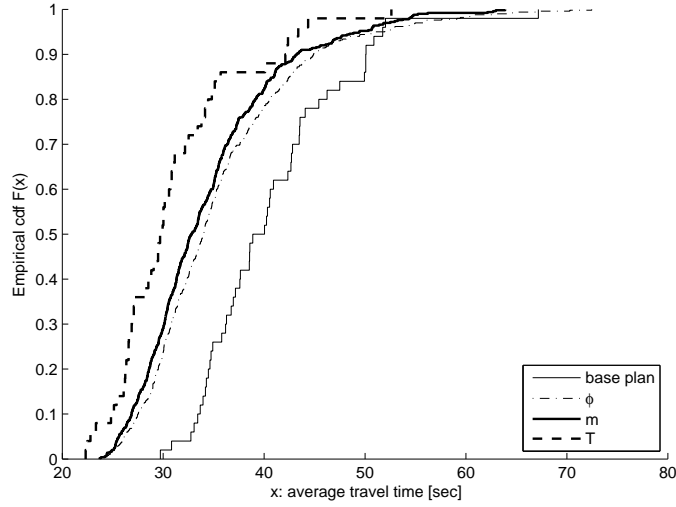


Figure 5.6: Empirical cumulative distribution functions of the average travel times considering the base plan as the initial point and running 10 instances of each method. Each instance consists of 150 simulation runs.

Table 5.2. Each statistic is calculated based on the difference between the row-wise method and the column-wise method. The critical value is 1.96. This table indicates that all differences are significant. That is, the queueing approach leads to signal plans with the best performance, followed by the proposed approach, the polynomial and the initial plan.

Secondly, we use the existing signal plan for the city of Lausanne (the base plan) as the initial plan. Once again, we run each method 10 times, and allow for 150 simulations each time. Figure 5.6 gives the cdf's of the different methods. Here the proposed method and the polynomial lead to signal plans with similar performance, and improved performance compared to the base plan. The queueing model leads to a signal plan with the best performance in terms of the distribution of the average travel times.

The corresponding travel time statistics are displayed in Table 5.3. We perform t-test's to evaluate whether the difference in the average travel times is significant. The t-statistics are given in Table 5.4. Each statistic is calculated based on the difference between the row-wise method and the column-wise method. The critical value is 1.96. Once again, all

	Average [sec]	Standard deviation
$m$	34.35	7.09
$\phi$	35.5	7.78
$T$	31	6.29
$base$	40.65	6.95

Table 5.3: Travel time statistics for the different signal plans, obtained based on 500 observations. The methods have been initialized considering the base plan as the initial plan.

	$m$	$\phi$	$base$
$\phi$	2.44		
$base$	14.18	11.04	
$T$	-7.91	-10.06	-23.01
degrees of freedom: 998			
confidence level: 0.05			
hypothesized mean difference: 0			

Table 5.4: T-statistics assuming equal variance. Each statistic is calculated based on the difference between the corresponding row and column methods. This scenario considers the base plan as the initial plan.

differences are significant and we have the same ranking between the methods, that is: the queueing approach, the proposed metamodel, the polynomial and the base plan.

We now consider a scenario with a higher computational budget. We allow for 1000 simulation runs and consider a random initial point. In this case, we run the algorithm once. We then evaluate the performance of the derived plans by running 50 replications of the simulation model.

Figure 5.7 presents the cdf's of the average travel times across the 50 replications, considering the initial plan, and the plans derived by both the proposed and the polynomial method at sample sizes 10, 20 and 30. The plan proposed by the polynomial method at sample size 10 has similar performance compared to the initial plan, whereas the plans at sample sizes 20 and 30 perform less well than the initial plan. The proposed method leads to the same plan for all three sample sizes. This plan has improved performance when compared to the initial plan and to the plans derived by the polynomial method.

Figure 5.8 presents the cdf's of the average travel times across the 50 replications, considering the initial plan, and the plans derived by both the proposed and the polynomial method at sample sizes 40 and 50. The signal plans derived by the polynomial do not provide improvement. The proposed method leads to the same plan for both sample sizes. This plan improves the distribution of average travel times compared to both the polynomial and the initial plan. Figure 5.9 considers sample sizes 250, 500 and 750. With a sample of size 250 the polynomial leads to reduced travel times compared to the proposed method. For the other sample sizes the signal plans of both methods have similar performance.

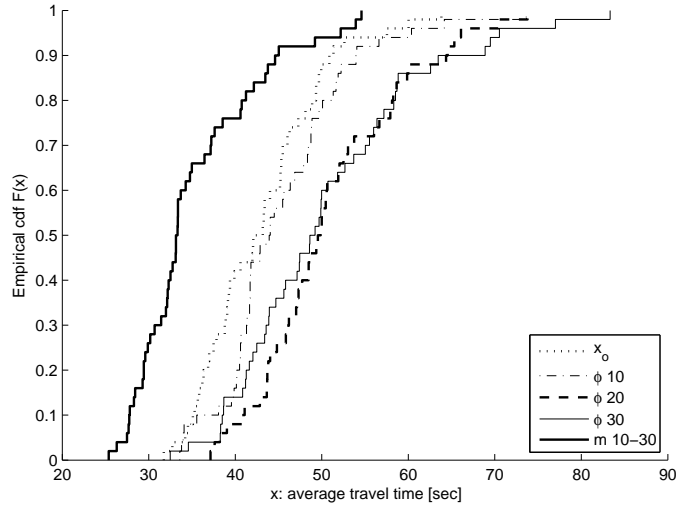


Figure 5.7: Empirical cumulative distribution functions of the average travel times considering a random signal plan as the initial point. The signal plans displayed are those derived at sample sizes 10, 20 and 30.

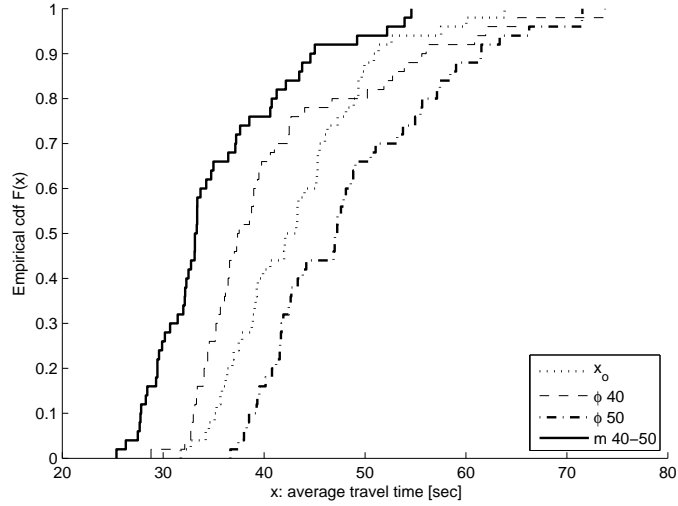


Figure 5.8: Empirical cumulative distribution functions of the average travel times considering a random signal plan as the initial point. The signal plans displayed are those derived at sample sizes 40 and 50.

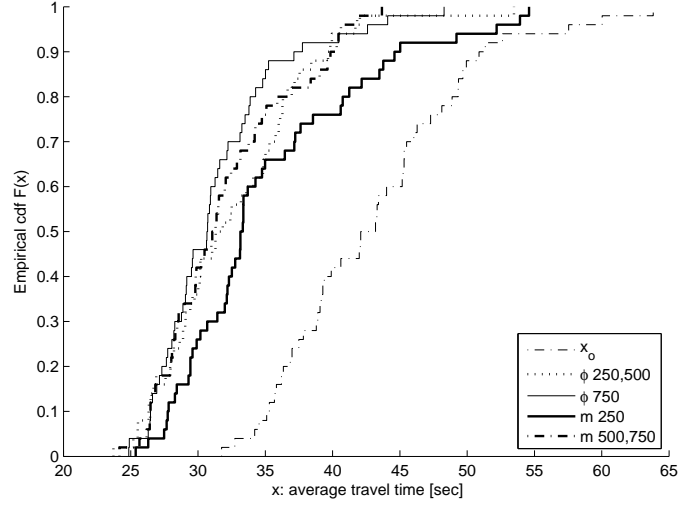


Figure 5.9: Empirical cumulative distribution functions of the average travel times considering a random signal plan as the initial point. The signal plans displayed are those derived at sample sizes 250, 500 and 750.

	Average [sec]	Standard deviation
$m$	32.01	4.9
$\phi$	32.47	6.47
$T$	30.84	6.42
$x_0$	43.06	7

Table 5.5: Travel time statistics for the different signal plans, obtained based on 50 replications. The methods have been initialized considering a uniformly drawn random initial plan.

Figure 5.10 considers the plans proposed at sample size 1000, and compares them to the initial plan, as well as to the plan proposed by the queueing method. All methods have an improved performance when compared to the initial plan. The plans derived by the proposed and the polynomial methods have similar performance, while the queueing model performs best.

The travel time statistics are displayed in Table 5.5. We perform t-test's to evaluate whether the difference in the average travel times is significant. The t-statistics are given in Table 5.6. Each statistic is calculated based on the difference between the row-wise method and the column-wise method. The critical value is 1.98. This table indicates that all three methods lead to improved performance compared to the initial plan. The difference in performance between the three methods is not significant.

We consider this sample of 1000 observations and test whether the metamodel parameters of the proposed model are significantly different from zero. We perform the corresponding t-tests as described in Section 5.5.1. In this case there are 86 model parameters. Seven are significantly different from zero. These 7 parameters concern 3 linear terms, 2 quadratic

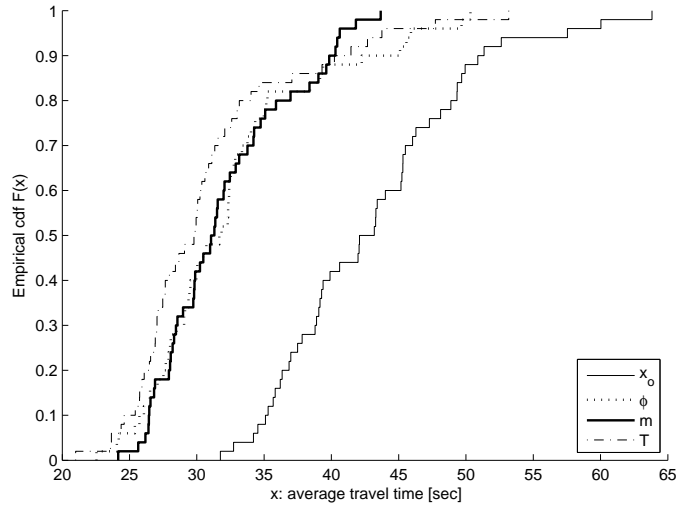


Figure 5.10: Empirical cumulative distribution functions of the average travel times considering a random signal plan as the initial point. The signal plans displayed are those obtained with a sample of size 1000.

	$m$	$\phi$	$x_0$
$\phi$	0.39		
$x_0$	9.14	7.86	
$T$	-1.04	-1.27	-9.1
degrees of freedom: 98			
confidence level: 0.05			
hypothesized mean difference: 0			

Table 5.6: T-statistics assuming equal variance. Each statistic is calculated based on the difference between the corresponding row and column methods. This scenario considers a random plan as the initial point.

terms, the queueing model parameter ( $\alpha$ ) and the intercept ( $\beta_1$ ). This indicates that the proposed metamodel indeed captures information regarding the relationship between the observed travel times and the decision variables.

The results of this section indicate that for small sized samples (compared to the dimension of the decision vector) limiting the metamodel of this framework to a quadratic polynomial fails to provide well-performing trial points, whereas providing structural information analytically via the queueing model allows for improvement. For small sized samples the queueing approach has the best performance, followed by the proposed method, the polynomial and the initial plans. We also considered one instance with a larger sample. Here all three methods yield similar performance, and improved performance compared to the initial plan.

## 5.6 Conclusions and future work

This chapter presents a simulation-based optimization framework for the management of congested networks. It proposes a metamodel that combines information from a traffic simulation tool and an analytical network model. It integrates this metamodel within a derivative-free trust region optimization algorithm.

Firstly, the performance of this framework is evaluated by solving a fixed-time signal control problem for two intersections of the Lausanne city center, and a simplified demand distribution. For small to moderate sized samples (small compared to the dimension of the decision vector), the metamodel provides improvement when compared to the polynomial method. For larger samples, the proposed metamodel leads to signal plans that perform similarly to those of the polynomial. Both methods yield reduced travel times, when compared to the plan proposed by the queueing model and to the initial plan.

Secondly, we consider the demand of the Lausanne city network for the evening peak period (17-18h). A larger set of intersections is considered endogenous. The performance is evaluated considering as initial plans: random plans and also an existing plan for the city of Lausanne.

In this case we ran two types of experiments. Firstly, we assumed a tight computational budget (approximately three times the dimension of the decision vector) and ran 10 instances of the algorithm. We initialized the algorithm both with a random plan and with an existing plan for the city of Lausanne. In both cases, the queueing approach leads to the signal plans with best performance, followed by the proposed metamodel, the polynomial and the initial plans.

We then allowed for a large computational budget and initialized with a random initial plan. We analyzed the performance of the signal plans with increasing sample sizes. For small samples, the proposed method yields well-performing plans, unlike the polynomial. For

large samples, the proposed and the polynomial methods derive signal plans with similar performance. Their performance is also similar to that of the queueing approach.

Efficiently tackling unconstrained high dimensional problems (e.g. more than 100 variables) is one of the main limitations of existing derivative-free methods, not to mention the added complexity of constrained problems. The generic metamodels used in these algorithms, e.g. quadratic polynomials, require a moderate to large sample to initially fit the metamodel of interest. By combining these generic metamodels with application-specific models that analytically capture the structure of the underlying problem, these algorithms can be used to tackle high dimensional problems under tight computational budgets. This added structure overcomes the need for a substantial initial sample, and provides meaningful trial points since the very first iterations.

For the Lausanne subnetwork scenario and a tight computational budget, it is the queueing model that leads to the best performance. Nonetheless, it is of interest to preserve the functional or generic component of the metamodel, since it ensures the asymptotic convergence of the algorithm. More research is needed to investigate whether other functional forms (e.g. splines, radial basis functions), or other interaction terms can improve the accuracy of the metamodel. We will also investigate the sensitivity of the method to the numerous algorithmic parameters.

Given the good performance of the queueing model, a natural extension would be to use the simulated observations to improve the accuracy of the exogenous queueing model parameters. This can be done by iteratively calibrating these parameters as the observations are collected. In particular, we expect the calibration of the exogenous parameters that depend on the decision vector, such as the turning proportions, to further improve the models accuracy.



# Chapter 6

## Conclusions

This chapter reviews the main contents and contributions of this thesis. It then indicates directions for future research.

**Chapter 2** formulates an analytical network model based on finite capacity queueing theory, and validates it versus existing methods, exact results and simulation results.

The method is an approximate method that decomposes the network into single queues, it solves the global balance equations for each queue and derives estimates for their marginal stationary distributions. The interactions between the queues are captured by structural parameters, which describe congestion in terms of its sources, its propagation and dissipation rates as well as its frequency. This method is suitable for networks with an arbitrary topology, and allows for queues with multiple servers.

The main novelty of this method lies in the definition of the state space, which explicitly accounts for blocked units in the network. In this approach, congestion is modeled based on the blocking mechanism known as blocking-after-service. Unlike existing methods, the proposed method preserves the queue capacities and the network topology exogenous, thus avoiding a posteriori validations.

**Chapter 3** formulates this model for two applications. Firstly, we use it to investigate the phenomenon of bed blocking in a network of operative and post-operative units of the Geneva University Hospitals. We validate the distributional estimates with those obtained via simulation.

This application responds to a recently recognized need for methods that estimate the number of blocked patients. Furthermore, the blocking-after-service mechanism of the model accurately mimics in-patient bed blocking. We go beyond existing analytical methods that have been used in the health care sector by allowing for networks with an arbitrary topology and with an arbitrary number of queues with finite capacity.

This case study illustrates in detail how the endogenous variables of the model can be

used to describe congestion in detail, and in particular identify its causes and quantify its impact on the network performance.

Secondly, Chapter 3 considers a protein synthesis network. We show how the model of Chapter 2 simplifies substantially for single server bufferless queues in a tandem topology. The model consists of a system of linear and quadratic equations. This simplicity and tractability overcomes the main limitations of the protein synthesis model of Mehra and Hatzimanikatis (2006): poor numerical conditioning for congested scenarios, limited scalability and computational efficiency.

**Chapter 4** proposes a surrogate model to perform urban traffic control. This chapter gives a detailed formulation of an urban traffic model, which is based on the approach proposed in Chapter 2. A traffic signal problem is solved for a subnetwork of the city of Lausanne. The performance of the signal plans derived by this method is compared to that of other signal plans, including an existing signal plan for the city of Lausanne. Several performance measures are analyzed such as average flow, density and travel time. The proposed method delays the propagation of congestion, and leads to improved performance measures.

In this chapter, we also present a formulation of the traffic model suitable for large-scale networks. We use it solve a signal control problem for the road network of the entire city of Lausanne. We compare the performance of the proposed signal plans to that of other plans. The new signal plans lead to reduced average travel times, and delay both the increase of density and the decrease of flow at the network scale.

This traffic model innovates by analytically capturing the interactions between adjacent intersections. The results demonstrate the relevance of accounting for these interactions under congested conditions. Operational signal control methodologies that can capture the complex features of congested traffic flows, such as spillbacks, are needed. This model contributes to the development of such methodologies, by achieving a suitable trade-off between describing the intricate interactions of congested traffic flows and analytical tractability.

**Chapter 5** presents a framework to perform simulation-based optimization for urban networks. The framework is based on a metamodel which combines the traffic model of Chapter 4 with a calibrated simulation model of the city of Lausanne. The proposed metamodel innovates by combining two families of metamodels known as physical and functional metamodels.

This metamodel is integrated within a derivative-free trust region algorithm. We evaluate the performance of this method considering two subnetworks of the city of Lausanne. The first scenario considers a simplified demand distribution. The second

considers the demand of the city of Lausanne for the evening peak hour and a larger set of controlled intersections.

In both cases, for small sample sizes (compared to the dimension of the decision vector), the proposed metamodel derives well performing trial points, unlike the polynomial. For large samples both methods have similar performance. We have also compared the performance of this metamodel to that of the queueing approach presented in Chapter 4. For the first scenario, the proposed model leads to improved performance. For the second scenario, the queueing model has improved performance for moderate-sized samples and similar performance for large samples.

This method yields well-performing trial points since the very first iterations. It is therefore particularly suitable to tackle problems given a tight computational budget, which is the main motivation of derivative-free methods. Furthermore, derivative-free methods are currently efficient methods for unconstrained small-scale problems. The traffic control problem solved is a constrained and large-scale problem. The results indicate that the structural information provided by the traffic model allows constrained large-scale problems to be tackled more efficiently.

### **Future research perspectives**

The proposed queueing method decomposes the network into subnetworks, which consist of single queues. We expect decompositions that consist of larger subnetworks (e.g. pairs or triplets of queues) to provide more accurate distributional estimates. This is particularly appealing for specific topologies such as tandem networks, where tractable approximate closed form expressions for the joint stationary distribution of two or three queues in tandem can be derived. The key difficulty here is to preserve a numerically efficient methodology.

We believe that the detailed blocking information provided by our approach can be used to quantify the occurrence of deadlock (i.e. gridlock) within a network. Furthermore, it would be interesting to derive analytical expressions for deadlock probabilities. Such performance measures would be relevant, for instance, for network design problems.

Extending the queueing network methodology, presented in this thesis, to account for several classes of jobs is methodologically interesting, and also appropriate for both the urban traffic and the health care applications.

The proposed queueing model describes the spatial propagation of congestion. It is interesting to extend this static approach to a dynamic model, where the temporal propagation of congestion is also captured. For the finite capacity queueing model considered in this thesis there are closed form expressions for the transient distributions. It is worth investigating whether these expressions can be combined with the

decomposition method presented in this work to derive a tractable dynamic methodology. This would also allow us to evaluate the impact of the transient regimes, e.g. the build-up and dissipation of congestion, on the networks performance.

The metamodel proposed in Chapter 5 combines a calibrated analytical traffic model with a polynomial model. The simulated observations are used to improve the metamodel by iteratively calibrating its polynomial component. An extension of this approach, consists of iteratively calibrating the exogenous parameters of the traffic model and, in particular, those that depend on the decision variables of the underlying optimization problem.

The metamodel of Chapter 5 consists of a physical (or structural) surrogate and a functional surrogate. The functional surrogate is a quadratic polynomial, which is the most commonly used functional surrogate for trust region methods. Investigating the performance of other families of functional surrogates is an interesting extension.

Furthermore, the use of quadratic polynomials ensures that asymptotically the model can satisfy Taylor-type bounds. This is necessary to ensure the global convergence of the underlying trust region algorithm. On the other hand, the queueing model provides global information via the structural parameters. An interesting framework would be to switch between these two models. This can be investigated following the ideas of the trust region algorithms, known as *model selection* and *model switching* algorithms, presented in Carter (1986).

# Appendix A

## Review of traffic signal control methodologies

### A.1 Fixed-time isolated strategies

These strategies can be stage-based such as SIGSET (Allsop, 1971) and SIGCAP (Allsop, 1976). SIGSET minimizes delay using Webster's nonlinear formulation (Webster, 1958), whereas SIGCAP maximizes reserve capacity. Both methods consider a set of linear constraints. A phase-based method formulated as a mixed-integer linear program is considered by Improtà and Cantarella (1984), where formulations for both delay minimization and reserve capacity maximization problems are given.

### A.2 Fixed-time coordinated strategies

Optimizing a set of signals along an arterial is the focus of the arterial progression schemes MAXBAND (Little et al., 1981) and MULTIBAND (Gartner et al., 1991). These methods aim at maximizing the bandwidth of through traffic along an arterial. MULTIBAND is an extension of MAXBAND allowing, among others, for different bandwidths for each link of the arterial. These problems are formulated as mixed-integer linear programs. They have been extended to consider a set of intersecting arterials (Gartner and Stamatiadis, 2002). Heuristics have also been specifically developed to solve this problem (Pillai et al., 1998). Nevertheless, under congested scenarios where there is a strong interaction among the different queues, the calculated bands fail to grasp this complexity. Furthermore, in dense urban networks with complex traffic movements bandwidth has little meaning (Robertson and Bretherton, 1991).

Several phase-based strategies have been proposed (Wong et al., 2002; Wong, 1997; Wong, 1996). The phase-based approach, although more general, is limited due to the exponential number of integer variables needed to describe the precedence constraints of

incompatible phases.

Chaudhary et al. (2002) compares the performance of three fixed-time coordinated stage-based methods: TRANSYT, PASSER and SYNCHRO. TRANSYT is the most widely used signal timing optimization package. It is a macroscopic model that aims at minimizing both delay and stops. A descriptive figure of its underlying methodology is given by Papageorgiou et al. (2003). SYNCHRO and TRANSYT have similar traffic models. SYNCHRO seeks to minimize stops and queues, by using an exhaustive search technique to determine the optimal signal timings. PASSER determines the green splits (also known as the green ratios), stage structure, cycle length and offsets that maximize arterial progression (i.e. it is a bandwidth-based method) for signalized arterials. PASSER performs an exhaustive search over the range of cycle lengths provided by the user, and sets the green splits using Webster's method (Webster, 1958). These splits are then adjusted to improve progression. Boillot et al. (1992) highlight that in congested conditions, TRANSYT and PASSER do not grasp the queue length appropriately. Traditionally TRANSYT's traffic model considered vertical queueing (i.e. the spatial extension of the queue is ignored), thus not capturing spillbacks, making this software suitable only for undersaturated scenarios. Although, more recent versions now take into account the effects of queue formation using horizontal queueing models (Abu-Lebdeh and Benekohal, 2003), Chow and Lo (2007) emphasize that the use of TRANSYT is appropriate only for low to moderate degrees of saturation.

### A.3 Traffic-responsive methods

Traffic-responsive methods use real-time measurements to drive the underlying optimization algorithm. The signal plans of these methods are derived either by making small adjustments to a predefined plan, by choosing between a set of pre-specified plans or by deciding when to switch to the next stages over a future time horizon (Boillot et al., 1992). The trend of real-time methods is the latter, where the optimization parameters are no longer cycle time, splits or offsets, but rather the switching times. These methods are referred to as non-parametric methods by Sen and Head (1997). Nevertheless, these methods are limited by the exponential size of the search space, due to the introduction of the integer variables used to describe the switching times.

The British software SCOOT (Bretherton, 1989) is considered to be the traffic-responsive version of TRANSYT. A description of how TRANSYT evolved into SCOOT is given by Robertson and Bretherton (1991). SCOOT seeks to minimize the total delay by carrying out incremental changes to the off-line timings derived by TRANSYT. It therefore makes a large number of small optimization decisions (typically over 10000 per hour in a network of 100 junctions (Robertson and Bretherton, 1991)). The Australian method SCATS (Lowrie, 1982) modifies signal timings on a cycle-by-cycle basis by minimizing stops and delay while

constraining the formation of queues. Both SCOOT and SCATS are widely used strategies suitable for undersaturated conditions, but as Aboudolas et al. (2007) and Dinopoulou et al. (2006) both describe, their performance deteriorates under congested conditions.

Dynamic programming methods are used in the French system PRODYN (Henry and Farges, 1989) as well as in the US systems OPAC and RHODES. RHODES (Mirchandani and Head, 2001) uses the COP algorithm (Sen and Head, 1997) to determine the switching times at a given intersection. This method does not react to traffic conditions just observed but rather proactively sets phase durations for predicted traffic conditions. A description of the OPAC model and algorithm, as well as its implementation are given by Gartner et al. (2001) and Gartner et al. (1991). The Italian method UTOPIA is yet another method that has been evaluated and implemented (Mauro and Di Taranto, 1989). As Dinopoulou et al. (2006) describe, the exponential complexity of these methods does not allow for network-wide optimization. This is also emphasized by Boillot et al. (1992): “the existing systems are not capable of controlling a zone of several junctions in a complete and coordinated manner. The chosen compromise is to control only one junction as OPAC or to use a decentralized optimization method as UTOPIA, PRODYN or to make little changes of the fixed-time signal plan as SCOOT and SCATS.” Acknowledging the importance and lack of efficient control strategies under saturated conditions has lead to the development of the French system CRONOS (Boillot et al., 2006; Boillot et al., 1992) and of the TUC method (Dinopoulou et al., 2006).



# Appendix B

## Webster and Highway Capacity Manual methods

### B.1 Webster's method

Webster's method is based on an estimate of the average delay per vehicle at a signalized intersection. It determines cycle times and green-splits of pre-timed signals that minimize delay. These green splits are used in signal setting software packages such as SYNCHRO and PASSER V (Chaudhary et al., 2002); and the delay estimate is one of the best known (Cascetta, 2001). The analysis is based on isolated intersections under the assumption of the number of arrivals following a Poisson distribution and undersaturated conditions (traffic intensity  $\rho < 1$ ). To present Webster's method we use the following notation:

$a_i$	available cycle time of intersection $i$ (cycle time minus the all-red times of intersection $i$ ) [seconds];
$b_i$	available cycle ratio of intersection $i$ (ratio of $a_i$ and the cycle time of intersection $i$ );
$g_p$	green split of phase $p$ (green time of phase $p$ divided by the cycle time of its corresponding intersection);
$Y_p$	the maximum ratio of flow to saturation flow among the lane groups that belong to phase $p$ ;
$\mathcal{P}_I(i)$	set of phase indices of intersection $i$ .

In this approach, each phase is represented by one approach only: the one with the highest degree of saturation (ratio of flow to saturation flow), denoted  $Y_p$ . More specifically, assuming no yellow times and no lost times per phase, Webster's method leads to:

$$g_p = \frac{Y_p}{\sum_{j \in \mathcal{P}_I(i)} Y_j} b_i \quad \forall p \in \mathcal{P}_I(i). \quad (\text{B.1})$$

This method requires as input the flows and saturation flows for each approach. In the experiments of this thesis, these have been derived as follows. For a signalized intersection the saturation flow is set to a common value for all approaches, this value is based on the standards VSS (1999b). The approach flows are set using the observed flows derived by the simulation model.

## B.2 Highway Capacity Manual method

The following notation, taken from the HCM (TRB, 1994), will be used in this section:

- $\left(\frac{\nu}{s}\right)_p$  the maximum ratio of flow to saturation flow among the lane groups that belong to phase  $p$ ;
- $C$  cycle length [sec];
- $L$  lost time per cycle [sec];
- $X_p$  desired flow to capacity ratio for the lane groups of phase  $p$ , also known as the degree of saturation;
- $X_c$  critical ratio of flow to capacity for the intersection of interest.

The method suggested in the HCM (2000 and 1994) determines the green splits as:

$$g_p = \left(\frac{\nu}{s}\right)_p \frac{1}{X_p}. \quad (\text{B.2})$$

As suggested in the Appendix 2 of the HCM (TRB, 1994) the desired degrees of saturation of the different phases,  $X_p$ , may be set so that they are all equal to the critical ratio  $X_c$  of the intersection  $i$ , where:

$$X_c = \sum_p \left(\frac{\nu}{s}\right)_p \frac{C}{C - L}. \quad (\text{B.3})$$

This leads to:

$$g_p = \left(\frac{\nu}{s}\right)_p \frac{1}{X_c}. \quad (\text{B.4})$$

## B.3 Equivalence between both methods

The mapping between the notations of both methods is given by:

$$Y_p = \left(\frac{\nu}{s}\right)_p \quad (\text{B.5})$$

$$b_i = \frac{C - L}{C}. \quad (\text{B.6})$$

Thus Equations (B.3) and (B.4) are also given by:

$$X_c = \frac{1}{b_i} \sum_{j \in \mathcal{P}_I(i)} Y_p \quad (\text{B.7})$$

$$g_p = \frac{Y_p}{\sum_{j \in \mathcal{P}_I(i)} Y_j} b_i. \quad (\text{B.8})$$

Thus we have retrieved Equation (B.1). By allocating the green splits such that the flow to capacity ratios for the critical movements of each phase,  $X_p$ , are equal to the critical ratio of the intersection  $X_c$ , the HCM method leads to the same green splits as Webster's method.



# Appendix C

## Trust region algorithms

### C.1 Basic trust region algorithm

This algorithm is presented in detail in Conn et al. (2000).

0. **Initialization.** Set an initial point  $x_0$ , an initial trust region radius  $\Delta_0$  and the parameters  $\eta_1, \eta_2, \gamma_1$  and  $\gamma_2$  such that

$$0 < \eta_1 \leq \eta_2 < 1 \text{ and } 0 < \gamma_1 \leq \gamma_2 < 1.$$

Set  $k = 0$ .

1. **Model definition.** Define a model  $m_k$  in  $B(x_k; \Delta_k)$ , where  $B(x_k; \Delta_k)$  is a ball centered at  $x_k$  and of radius  $\Delta_k$ , and is referred to as the trust region.
2. **Step calculation.** Compute a step  $s_k$  that “sufficiently reduces the model”  $m_k$  and such that  $x_k + s_k \in B(x_k; \Delta_k)$ . The point  $x_k + s_k$  is referred to as the trial point.
3. **Acceptance of the trial point.** Compute  $f(x_k + s_k)$  and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

If  $\rho_k \geq \eta_1$ , then  $x_{k+1} = x_k + s_k$ ; otherwise  $x_{k+1} = x_k$ .

4. **Trust region radius update.** Set

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, +\infty) & \text{if } \rho_k \geq \eta_2 \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2) \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1. \end{cases} \quad \begin{array}{l} \text{(C.1)} \\ \text{(C.2)} \\ \text{(C.3)} \end{array}$$

Increment  $k$  by 1 and go to Step 1.

## C.2 Derivative-free trust region algorithm

This algorithm is presented in detail in Conn et al. (2009a) and in Conn et al. (2009b).

### 0. Initialization. Set

- a fully-linear class of models  $\mathcal{M}$  and a corresponding model-improvement algorithm,
- an initial point  $x_0$ ,
- an upper bound for the trust region radius  $\Delta_{max} > 0$ ,
- an initial trust region radius  $\Delta_0^{icb} \in (0, \Delta_{max}]$ ,
- an initial model  $m_0^{icb}$  (with gradient and possibly the Hessian at  $s = 0$  given by  $g_0^{icb}$  and  $H_0^{icb}$ , respectively),
- the parameters  $\eta_0, \eta_1, \gamma, \gamma_{inc}, \epsilon_c, \beta, \mu$  and  $\alpha$  such that
  - $0 \leq \eta_0 \leq \eta_1 < 1$ ,
  - $\eta_1 \neq 0$ ,
  - $0 < \gamma < 1 < \gamma_{inc}$ ,
  - $\epsilon_c > 0$ ,
  - $\mu > \beta > 0$ ,
  - $\alpha \in (0, 1)$ ,
- positive constants  $\kappa_{ef}, \kappa_{eg}, \kappa_{blg}$  that are used to test if a model is fully linear,
- $k = 0$ .

### 1. Criticality step.

- If  $\|g_k^{icb}\| > \epsilon_c$ , then  $m_k = m_k^{icb}$  and  $\Delta_k = \Delta_k^{icb}$ .
- Otherwise, call the model-improvement algorithm to attempt to certify if the model  $m_k^{icb}$  is fully linear on  $B(x_k; \Delta_k^{icb})$ .
  - If the model  $m_k^{icb}$  is not certifiably fully linear on  $B(x_k; \Delta_k^{icb})$ , or if  $\Delta_k^{icb} > \mu\|g_k^{icb}\|$ , then
    - (i) apply the *Criticality Step Algorithm* (described below) to construct a model  $\tilde{m}_k(x_k + s_k)$  (with gradient and possibly the Hessian at  $s = 0$  given by  $\tilde{g}_k$  and  $\tilde{H}_k$ , respectively), which is fully linear (for constants  $\kappa_{ef}, \kappa_{eg}, \kappa_{blg}$ ) on the ball  $B(x_k; \tilde{\Delta}_k)$ , for some  $\tilde{\Delta}_k \in (0, \mu\|\tilde{g}_k\|]$  given by the *Criticality Step Algorithm*.
    - (ii) Set  $m_k = \tilde{m}_k$  and  $\Delta_k = \min\{\max\{\tilde{\Delta}_k, \beta\|\tilde{g}_k\|\}, \Delta_k^{icb}\}$ .
  - Otherwise set  $m_k = \tilde{m}_k^{icb}$  and  $\Delta_k = \tilde{\Delta}_k^{icb}$ .

2. **Step calculation.** Compute a step  $s_k$  that “sufficiently reduces the model”  $m_k$  and such that  $x_k + s_k \in B(x_k; \Delta_k)$ .

3. **Acceptance of the trial point.** Compute  $f(x_k + s_k)$  and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

- If  $\rho_k \geq \eta_1$  or if both  $\rho_k \geq \eta_0$  and the model is fully linear (for constants  $\kappa_{ef}, \kappa_{eg}, \kappa_{blg}$ ) on  $B(x_k; \Delta_k)$ , then  $x_{k+1} = x_k + s_k$  and the model is updated to include the new iterate into the sample set, resulting in a new model  $m_{k+1}^{icb}$  (with gradient and possibly the Hessian at  $s = 0$  given by  $g_{k+1}^{icb}$  and  $H_{k+1}^{icb}$ , respectively).
- Otherwise the model and the iterate remain unchanged ( $m_{k+1}^{icb} = m_k$  and  $x_{k+1} = x_k$ ).

4. **Model improvement.**

- If  $\rho_k < \eta_1$ , then use the model-improvement algorithm to attempt to certify that  $m_k$  is fully linear on  $B(x_k; \Delta_k)$ . If such a certificate is not obtained, we say that  $m_k$  is not certifiably fully linear and make one or more suitable improvement steps.

Define  $m_{k+1}^{icb}$  to be the (possibly improved) model.

5. **Trust region radius update.** Set

$$\Delta_{k+1}^{icb} \in \begin{cases} [\Delta_k, \min\{\gamma_{inc}\Delta_k, \Delta_{max}\}] & \text{if } \rho_k \geq \eta_1 \\ \{\gamma\Delta_k\} & \text{if } \rho_k < \eta_1 \text{ and } m_k \text{ is fully linear} \\ \{\Delta_k\} & \text{otherwise.} \end{cases} \quad \begin{matrix} (C.4) \\ (C.5) \\ (C.6) \end{matrix}$$

Increment  $k$  by 1 and go to Step 1.

## Criticality step algorithm

1. **Initialization.** Set  $i = 0$ ,  $m_k^{(0)} = m_k^{icb}$ .
2. Repeat until  $\tilde{\Delta}_k \leq \mu \|g_k^{(i)}\|$ 
  - (a)  $i = i + 1$ .
  - (b) Use the model-improvement algorithm to improve the previous model  $m_k^{(i-1)}$  until it is fully linear on  $B(x_k; \alpha^{i-1}\Delta_k^{icb})$ .
  - (c) Denote the new model by  $m_k^{(i)}$ . Set  $\tilde{\Delta}_k = \alpha^{i-1}\Delta_k^{icb}$  and  $\tilde{m}_k = m_k^{(i)}$ .



# Bibliography

- Aboudolas, K., Papageorgiou, M. and Kosmatopoulos, E. (2007). Control and optimization methods for traffic signal control in large-scale congested urban road networks, *American Control Conference*, pp. 3132–3138.
- Abu-Lebdeh, G. and Benekohal, R. (1997). Development of traffic control and queue management procedures for oversaturated arterials, *Transportation Research Record* **1603**: 119–127.
- Abu-Lebdeh, G. and Benekohal, R. (2003). Design and evaluation of dynamic traffic management strategies for congested conditions, *Transportation Research Part A: Policy and Practice* **37**(2): 109–127.
- Akyildiz, I. F. and von Brand, H. (1994). Exact solutions to networks of queues with blocking-after-service, *Theoret. Comput. Sci.* **125**(1): 111–130.
- Alexandrov, N. M., Dennis, J. E., Lewis, R. M. and Torczon, V. (1998). A trust region framework for managing the use of approximation models in optimization, *Structural Optimization* **15**: 16–23.
- Alexandrov, N. M. and Lewis, R. M. (2001). An overview of first-order model management for engineering optimization, *Optimization and Engineering* **2**: 413–430.
- Alexandrov, N. M., Lewis, R. M., Gumbert, C. R., Green, L. L. and Newman, P. A. (1999). Optimization with variable-fidelity models applied to wing design, *Technical Report CR-1999-209826*, NASA Langley Research Center, Hampton, VA, USA.
- Alexandrov, N. M., Lewis, R. M., Gumbert, C. R., Green, L. L. and Newman, P. A. (2001). Approximation and model management in aerodynamic optimization with variable-fidelity models, *Journal of Aircraft* **38**(6): 1093–1101.
- Alexandrov, N. M., Nielsen, E. J., Lewis, R. M. and Anderson, W. K. (2000). First-order model management with variable-fidelity physics applied to multi-element airfoil optimization, *Proceedings of the 8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Long Beach, CA, USA.
- Alfa, A. S. and Liu, B. (2004). Performance analysis of a mobile communication network: the tandem case, *Comp. Comm.* **27**(3): 208–221.
- Allsop, R. (1971). SIGSET: A computer program for calculating traffic signal settings, *Traffic Engineering and Control* **13**(2).

- Allsop, R. (1976). SIGCAP: A computer program for assessing the traffic capacity of signal-controlled road junctions, *Traffic Engineering & Control* **17**: 338–341.
- Allsop, R. (1992). Evolving application of mathematical optimisation in design and operation of individual signal-controlled road junctions, in J. D. Griffiths (ed.), *Mathematics in Transport Planning and Control*, Institute of Mathematics and its Applications, University of Wales College of Cardiff, Oxford Clarendon.
- Altioek, T. (1982). Approximate analysis of exponential tandem queues with blocking, *European Journal of Operational Research* **11**(4): 390–398.
- Altioek, T. (1989). Approximate analysis of queues in series with phase-type service times and blocking, *Oper. Res.* **37**(4): 601–610.
- Altioek, T. and Perros, H. G. (1987). Approximate analysis of arbitrary configurations of open queuing networks with blocking, *Annals of Operations Research* **9**(1): 481–509.
- Artalejo, J. R. (1999). Accessible bibliography on retrial queues, *Math. Comput. Modelling* **30**(3-4): 1–6.
- Atkeson, C. G., Moore, A. W. and Schaal, S. (1997). Locally weighted learning, *Artificial Intelligence Review* **11**: 11–73.
- Balsamo, S., De Nitto Persone, V. and Inverardi, P. (2003). A review on queueing network models with finite capacity queues for software architectures performance prediction, *Perf. Evaluation* **51**(2-4): 269–288.
- Balsamo, S., De Nitto Persone, V. and Onvural, R. (2001). *Analysis of Queueing Networks with Blocking*, Vol. 31 of *International Series in Operations Research and Management Science*, Kluwer Academic Publishers, Boston.
- Balsamo, S. and Donatiello, L. (1989). On the cycle time distribution in a two-stage cyclic network with blocking, *IEEE Trans. Software Eng.* **15**(10): 1206–1216.
- Bandler, J. W., Cheng, Q., Dakroury, A., Mohamed, A., Bakr, M., Madsen, K. and Søndergaard, J. (2004). Space mapping: The state of the art, *IEEE Transactions on Microwave Theory and Techniques* **52**(1): 337–360.
- Bandler, J. W., Koziel, S. and Madsen, K. (2006). Space mapping for engineering optimization, *SIAG/Otimization Views-and-News* **17**(1): 19–26.
- Barton, R. R. and Meckesheimer, M. (2006). Metamodel-based simulation optimization, in S. G. Henderson and B. L. Nelson (eds), *Handbooks in operations research and management science: Simulation*, Vol. 13, Elsevier, Amsterdam, chapter 18, pp. 535–574.
- Bell, P. C. (1982). Use of decomposition techniques for the analysis of open restricted queueing networks, *Operations Res. Letters* **1**(6): 230–235.
- Ben-Akiva, M., Bierlaire, M., Burton, M., Koutsopoulos, H. and Mishalani, R. (2001). Network state estimation and prediction for real-time transportation management applications, *Networks and Spatial Economics* **1**(3-4): 293–318.
- Bierlaire, M. and Frejinger, E. (2008). Route choice modeling with network-free data,

- Transportation Research Part C: Emerging Technologies* **16**(2): 187–198.
- Bocharov, P. P., D’Apice, C., Pechinkin, A. V. and Salerno, S. (2004). *Queueing theory*, Modern Probability and Statistics, Brill Academic Publishers, Zeist, The Netherlands, chapter 3, pp. 96–98.
- Boillot, F., Blosseville, J., Lesort, J., Motyka, V., Papageorgiou, M. and Sellam, S. (1992). Optimal signal control of urban traffic networks, *Road Traffic Monitoring (IEE Conf. Pub. 355)*.
- Boillot, F., Midenet, S. and Pierrelée, J. (2006). The real-time urban traffic control system CRONOS: Algorithm and experiments, *Transportation Research Part C: Emerging Technologies* **14**(1): 18–38.
- Booker, A. J., Dennis, J. E., Frank, P. D., Serafini, D. B., Torczon, V. and Trosset, M. W. (1999). A rigorous framework for optimization of expensive functions by surrogates, *Structural Optimization* **17**: 1–13.
- Boxma, O. J. and Konheim, A. J. (1981). Approximate analysis of exponential queueing systems with blocking, *Acta Inform.* **15**(1): 19–66.
- Brandwajn, A. and Jow, Y. (1985). Tandem exponential queues with finite buffers, in T. Hasegawa, H. Takagi and Y. Takahashi (eds), *Comp. Networking and Perf. Evaluation*, Amsterdam, The Netherlands: North Holland, pp. 245–258.
- Brandwajn, A. and Jow, Y. (1988). An approximation method for tandem queues with blocking, *Operations Res. Letters* **36**(1): 73–83.
- Bretherton, R. D. (1989). SCOOT - urban traffic control system - philosophy and evaluation, *IFAC Symposium of Control Communications in Transportation*, Pergamon Press, Oxford, pp. 237–239.
- Burke, P. J. (1976). Proof of a conjecture on the interarrival-time distribution in an  $m/m/1$  queue with feedback, *IEEE transactions on communications* **24**(5): 575–576.
- Carter, R. G. (1986). *Multi-model algorithms for optimization*, PhD thesis, Rice University.
- Cascetta, E. (2001). *Transportation Systems Engineering: theory and methods*, Vol. 49 of *Applied Optimization*, Kluwer academic publishers, Dordrecht, chapter 2, pp. 50–51.
- Cascetta, E., Gallo, M. and Montella, B. (2006). Models and algorithms for the optimization of signal settings on urban networks with stochastic assignment models, *Annals of Operations Research* **144**(1): 301–328.
- CEC (2007). *Green Paper. Towards a new culture for urban mobility*. COM (2007) 551. Office for Official Publications of the European Communities, Luxembourg.
- Chaudhary, N. A., Kovvali, V. G. and Alam, S. M. (2002). Guidelines for selecting signal timing software, *Technical Report 0-4020-P2*, Texas Transportation Institute, U.S. Department of Transportation, Federal Highway Administration.
- Cheah, J. Y. and Smith, J. M. (1994). Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows, *Queueing Syst.* **15**(1-4): 365–386.

- Chow, A. H. F. and Lo, H. K. (2007). Sensitivity analysis of signal control with physical queuing: Delay derivatives and an application, *Transportation Research Part B: Methodological* **41**(4): 462–477.
- Cochran, J. and Bharti, A. (2006). Stochastic bed balancing of an obstetrics hospital, *Health Care Management Sci.* **9**(1): 31–45.
- Coleman, T. F. and Li, Y. (1994). On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds, *Mathematical Programming* **67**(2): 189–224.
- Coleman, T. F. and Li, Y. (1996). An interior, trust region approach for nonlinear minimization subject to bounds, *SIAM Journal on Optimization* **6**: 418–445.
- Conn, A. R., Gould, N. I. M. and Toint, P. L. (2000). *Trust-region methods*, MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadelphia, PA, USA.
- Conn, A. R., Scheinberg, K. and Toint, P. L. (1998). A derivative free optimization algorithm in practice, *Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, St. Louis, MO, USA.
- Conn, A. R., Scheinberg, K. and Vicente, L. N. (2009a). Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points, *SIAM Journal on Optimization* **20**(1): 387–415.
- Conn, A. R., Scheinberg, K. and Vicente, L. N. (2009b). *Introduction to derivative-free optimization*, MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadelphia, PA, USA.
- Daganzo, C. F. (1996). The nature of freeway gridlock and how to prevent it, in J. B. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Pergamon Press, pp. 629–646.
- Dennis, J. E. and Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*, Vol. 16 of *Classics in Applied Mathematics*, SIAM, Philadelphia.
- Dinopoulou, V., Diakaki, C. and Papageorgiou, M. (2006). Applications of the urban traffic control strategy TUC, *European Journal of Operational Research* **175**(3): 1652–1665.
- Dion, F., Rakha, H. and Kang, Y. (2004). Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections, *Transportation Research Part B: Methodological* **38**(2): 99–122.
- Dumont, A. G. and Bert, E. (2006). Simulation de l’agglomération Lausannoise SIMLO, *Technical report*, Laboratoire des voies de circulation, ENAC, Ecole Polytechnique Fédérale de Lausanne.
- Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G. and Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery, *Journal of*

- Public Health Medicine* **25**(4): 325–335.
- Fu, M. C. (2006). Gradient estimation, in S. G. Henderson and B. L. Nelson (eds), *Handbooks in operations research and management science: Simulation*, Vol. 13, Elsevier, Amsterdam, chapter 19, pp. 576–616.
- Fu, M. C., Glover, F. W. and April, J. (2005). Simulation optimization: a review, new developments, and applications, in M. E. Kuhl, N. M. Steiger, F. B. Armstrong and J. A. Joines (eds), *Proceedings of the 2005 Winter Simulation Conference*, Piscataway, New Jersey, USA, pp. 83–95.
- Garber, N. J. and Hoel, L. A. (2002). *Traffic and Highway Engineering*, 3rd edn, Books Cole, Thomson Learning, chapter 6, pp. 204–210.
- Gartner, N. H., Assman, S. F., Lasaga, F. and Hou, D. L. (1991). A multi-band approach to arterial traffic signal optimization, *Transportation Research Part B: Methodological* **25**(1): 55–74.
- Gartner, N., Pooran, F. and Andrews, C. (2001). Implementation of the OPAC adaptive control strategy in a trafficsignal network, *Intelligent Transportation Systems, IEEE*, pp. 195–200.
- Gartner, N. and Stamatidis, C. (2002). Arterial-based control of traffic flow in urban grid networks, *Mathematical and Computer Modelling* **35**(5): 657–671.
- Grassman, W. and Derkic, S. (2000). An analytical solution for a tandem queue with blocking, *Queueing Syst.* **36**(1-3): 221–235.
- Gupta, S. M. and Kavusturucu, A. (2000). Production systems with interruptions, arbitrary topology and finite buffers, *Annals of Operations Research* **93**(1-4): 145–176.
- Heidemann, D. (1994). Queue length and delay distributions at traffic signals, *Transportation Research Part B: Methodological* **28**(5): 377–389.
- Heidemann, D. (1996). A queueing theory approach to speed-flow-density relationships, *Proceedings of the 13<sup>th</sup> International Symposium on Transportation and Traffic Theory*, Lyon, France, pp. 103–118.
- Heidemann, D. and Wegmann, H. (1997). Queueing at unsignalized intersections, *Transportation Research Part B: Methodological* **31**(3): 239–263.
- Henry, J. J. and Farges, J. L. (1989). PRODYN, *IFAC Symposium of Control Communications in Transportation*, Pergamon Press, Oxford, pp. 253–255.
- Hershey, J. C., Weiss, E. N. and Cohen, M. A. (1981). A stochastic service network model with application to hospital facilities, *Oper. Res.* **29**(1): 1–22.
- Hillier, F. S. and Boling, R. W. (1967). Finite queues in series with exponential or Erlang service times—a numerical approach, *Oper. Res.* **15**(2): 286–303.
- Improta, G. and Cantarella, G. E. (1984). Control system design for an individual signalized junction, *Transportation Research Part B: Methodological* **18**(2): 147–167.
- Inman, R. (1999). Empirical evaluation of exponential and independence assumptions

- in queuing models of manufacturing systems, *Production and Operations Management* **8**(4): 409–432.
- 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.
- Jain, R. and Smith, J. M. (1997). Modeling vehicular traffic flow using M/G/C/C state dependent queueing models, *Transportation science* **31**(4): 324–336.
- Jun, J. B., Jacobson, S. H. and Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey, *J. Oper. Res. Soc.* **50**(2): 109–123.
- Jun, K. P. and Perros, H. G. (1989). Approximate analysis of arbitrary configurations of queueing networks with blocking and deadlock, in H. G. Perros and T. Altioek (eds), *Queueing Networks with Blocking: Proceedings of the First international workshop*, North-Holland, Amsterdam, pp. 259–279.
- Jun, K. P. and Perros, H. G. (1990). An approximate analysis of open tandem queueing networks with blocking and general service times, *European Journal of Operational Research* **46**(1): 123–135.
- Kerbache, L. and Smith, J. M. (1987). The generalized expansion method for open finite queueing networks, *European Journal of Operational Research* **32**(3): 448–461.
- Kerbache, L. and Smith, J. M. (1988). Asymptotic behaviour of the expansion method for open finite queueing networks, *Comp. and Operations Res.* **15**(2): 157–169.
- Kerbache, L. and Smith, J. M. (2000). Multi-objective routing within large scale facilities using open finite queueing networks, *European Journal of Operational Research* **121**(1): 105–123.
- Koizumi, N., Kuno, E. and Smith, T. E. (2005). Modeling patient flows using a queueing network with blocking, *Health Care Management Sci.* **8**(1): 49–60.
- Konheim, A. G. and Reiser, M. (1976). A queueing model with finite waiting room and blocking, *Journal of the Association for Computing Machinery* **23**(2): 328–341.
- Konheim, A. G. and Reiser, M. (1978). Finite capacity queueing systems with applications in computer modeling, *SIAM J. Comput.* **7**(2): 210–229.
- Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: An introduction, *Annals of Operations Research* **113**(1-4): 41–59.
- Korporaal, R., Ridder, A., Klopogge, P. and Dekker, R. (2000). An analytic model for capacity planning of prisons in the Netherlands, *J. Oper. Res. Soc.* **51**(11): 1228–1237.
- Langaris, C. and Conolly, B. (1984). On the waiting time of a two-stage queueing system with blocking, *J. Appl. Probab.* **21**(3): 628–638.
- Latouche, G. and Neuts, M. F. (1980). Efficient algorithmic solutions to exponential tandem queues with blocking, *SIAM Journal on Algebraic and Discrete Methods* **1**(1): 93–106.
- Lee, H. S., Bouhchouch, A., Dallery, Y. and Frein, Y. (1998). Performance evaluation of open queueing networks with arbitrary configuration and finite buffers, *Annals of Operations*

- Research* **79**(0): 181–206.
- Little, J., Kelson, M. and Gartner, N. (1981). MAXBAND: a program for setting signals on arteries and triangular networks, *Transportation Research Record* **795**: 40–46.
- Lowrie, P. (1982). SCATS: The sydney co-ordinated adaptive traffic system, *IEE International conference on road traffic signaling*, pp. 67–70.
- Mackay, M. (2001). Practical experience with bed occupancy management and planning systems: an Australian view, *Health Care Management Sci.* **4**(1): 47–56.
- Mandelbaum, A. (2001). Call centers (centres): Research bibliography with abstracts, *Electronically available: <http://iew3.technion.ac.il/serveng/References/ccbib.pdf>*.
- Mauro, V. and Di Taranto, C. (1989). UTOPIA, *IFAC Symposium of Control Communications in Transportation*, Pergamon Press, Oxford, pp. 245–252.
- McNeil, D. R. (1968). A solution to the fixed-cycle traffic light problem for compound poisson arrivals, *Journal of Applied Probability* **5**: 624–635.
- Mehra, A. and Hatzimanikatis, V. (2006). An algorithmic framework for genome-wide modeling and analysis of translation networks, *Biophysical Journal* **90**: 1136–1146.
- Melamed, B. (1979). Characterizations of poisson traffic streams in jackson queueing networks, *Advances in Applied Probability* **11**(2): 422–438.
- Mier-y-Teran-Romero, L., Silber, M. and Hatzimanikatis, V. (2009). The origins of time-delay in template biopolymerization processes, *Technical report*, Laboratory of Computational Systems Biotechnology, SB, Ecole Polytechnique Fédérale de Lausanne.
- Miller, A. J. (1963). Settings for fixed-cycle traffic signals, *Operational Research Quarterly* **14**(4): 373–386.
- Mirchandani, P. and Head, L. (2001). A real-time traffic signal control system: architecture, algorithms, and analysis, *Transportation Research Part C: Emerging Technologies* **9**(6): 415–432.
- More, J. and Wild, S. (2009). Benchmarking derivative-free optimization algorithms, *SIAM Journal on Optimization* **20**(1): 172–191.
- Nagel, K. (2002). Traffic networks, in S. Bornholdt and H. G. Schuster (eds), *Handbook of Graphs and Networks*, Wiley VCH, LinkWeinheim, pp. 248–272.
- Newell, G. (1965). Approximation methods for queues with application to the fixed-cycle traffic light, *SIAM Review* **7**(2): 223–240.
- Newell, G. F. (1979). *Approximate behavior of tandem queues*, Vol. 171 of *Lecture notes in economics and mathematical systems*, Springer-Verlag, Berlin.
- Obaidat, M. S. (1990). Simulation of queueing models in computer systems, in S. Oezekici (ed.), *Queueing Theory and Applications*, Taylor & Francis/Hemisphere, New York, pp. 111–151.
- Ouvray, R. and Bierlaire, M. (2009). Boosters: a derivative-free algorithm based on radial basis functions, *International Journal of Modelling and Simulation* **29**(1): 26–36.

- Osorio, C. and Bierlaire, M. (2008a). A multiple model approach for traffic signal optimization in the city of lausanne, *Proceedings of the Seventh Swiss Transport Research Conference, 8<sup>th</sup> STRC*, Ascona, Switzerland.
- Osorio, C. and Bierlaire, M. (2008b). Network performance optimization using a queueing model, *Proceedings of the European Transport Conference (ETC)*, Noordwijkerhout, The Netherlands.
- Osorio, C. and Bierlaire, M. (2008c). A queueing network approach to the traffic signal optimization of the lausanne city center, *Proceedings of the Latin-Ibero-American Conference on Operations Research (CLAIO)*, Cartagena, Colombia.
- Osorio, C. and Bierlaire, M. (2008d). Signal control optimization with a queueing network model capturing congestion, *Proceedings of the Pan-American Conference on Traffic and Transportation Engineering (PANAM)*, Cartagena, Colombia.
- 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 metamodel approach for simulation optimization of congested urban road networks, *Computational Management Science Conference (CMS)*, Geneva, Switzerland.
- Osorio, C. and Bierlaire, M. (2009c). A multi-model algorithm for the optimization of congested networks, *Proceedings of the European Transport Conference (ETC)*, Noordwijkerhout, The Netherlands.
- Osorio, C. and Bierlaire, M. (2009d). A simulation optimization framework for the management of congested urban road networks, *Proceedings of the Seventh Swiss Transport Research Conference (STRC)*, Ascona, Switzerland.
- Osorio, C., Weibel, C., Perez, P., Bierlaire, M. and Garnerin, P. (2006). Patient flow simulation as a tool for estimating policy impact, *Swiss Medical Informatics* **58**: 33–36.
- Papadopoulos, H. T. and Heavey, C. (1996). Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines, *European Journal of Operational Research* **92**(1): 1–27.
- Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A. and Wang, Y. (2003). Review of road traffic control strategies, *Proceedings of the IEEE* **91**(12): 2043–2067.
- Perros, H. (1984). Queueing networks with blocking: A bibliography, *ACM SIGMETRICS Performance Evaluation Review* **12**(2): 8–12.
- Perros, H. (1994). *Queueing networks with blocking: Exact and Approximate Solutions*, Oxford University Press, New York, NY, USA.
- Perros, H. (2003). Open queueing networks with blocking - a personal log, in G. Kotsis (ed.), *Performance Evaluation - Stories and Perspectives*, Austrian Computer Society, Vienna, Austria, pp. 105–115.

- Pillai, R., Rathi, A. and L. Cohen, S. (1998). A restricted branch-and-bound approach for generating maximum bandwidth signal timing plans for traffic networks, *Transportation Research Part B: Methodological* **32**(8): 517–529.
- Powell, M. J. D. (1970). A fortran subroutine for solving systems of nonlinear algebraic equations, in P. Rabinowitz (ed.), *Numerical Methods for Nonlinear Algebraic Equations*, Gordon & Breach, London, chapter 7.
- Powell, M. J. D. (2003). On trust region methods for unconstrained minimization without derivatives, *Mathematical Programming* **97**(3): 605–623.
- ProModel (1997). ProModel user’s guide. ProModel Corporation.
- Robertson, D. and Bretherton, R. (1991). Optimizing networks of traffic signals in real time - the SCOOT method, *Vehicular Technology, IEEE Transactions on* **40**(1): 11–15.
- Sadoun, B. (2000). Applied system simulation: a review study, *Information Sciences* **124**(1-4): 173–192.
- Schmidt, L. C. and Jackman, J. (2000). Modeling recirculating conveyors with blocking, *European Journal of Operational Research* **124**(2): 422–436.
- Sen, S. and Head, K. (1997). Controlled optimization of phases at an intersection, *Transportation science* **31**(1): 5–17.
- Serafini, D. B. (1998). *A Framework for Managing Models in Nonlinear Optimization of Computationally Expensive Functions*, PhD thesis, Rice University.
- Shepherd, S. (1994). Traffic control in over-saturated conditions, *Transport Reviews* **14**(1): 13–43.
- Singh, A. and Smith, J. M. (1997). Buffer allocation for an integer nonlinear network design problem, *Comp. and Operations Res.* **24**(5): 453–472.
- Søndergaard, J. (2003). *Optimization using surrogate models - by the Space Mapping technique*, PhD thesis, Technical University of Denmark.
- Stafford, R. (2006). *The Theory Behind the ‘randfixedsum’ Function*. <http://www.mathworks.com/matlabcentral/fileexchange/9700>.
- Stewart, W. J. (2000). Numerical methods for computing stationary distributions of finite irreducible Markov chains, in W. Grassmann (ed.), *Computational Probability*, Kluwer Academic Publishers, Boston, chapter 4.
- Tahilramani, H., Manjunath, D. and Bose, S. K. (1999). Approximate analysis of open network of GE/GE/m/N queues with transfer blocking, *MASCOTS 0*: 164–172.
- Takahashi, Y., Miyahara, H. and Hasegawa, T. (1980). An approximation method for open restricted queuing networks, *Oper. Res.* **28**(3): 594–602.
- Tanner, J. C. (1962). A theoretical analysis of delays at an uncontrolled intersection, *Biometrika* **49**: 163–170.
- The Mathworks, I. (2008). *Optimization Toolbox Version 4. User’s Guide Matlab*, Natick, MA, USA.

- TRB (1994). *Highway capacity manual. Special report 209*, 3rd edn, Transportation Research Board National Research Council. Chapter 9.
- TRB (2000). *Highway capacity manual*, Transportation Research Board, National Research Council, Washington, D.C., USA. Chapter 16.
- TSS (2008). *AIMSUN NG and AIMSUN Micro Version 5.1*, Transport Simulation Systems.
- van Vuuren, M., Adan, I. J. B. F. and Resing-Sassen, S. A. E. (2005). Performance analysis of multi-server tandem queues with finite buffers and blocking, *OR Spectrum* **27**(2-3): 315–338.
- Van Woensel, T. and Vandaele, N. (2007). Modelling traffic flows with queueing models: A review, *Asia-Pacific Journal of Operational Research* **24**(4): 1–27.
- Viti, F. (2006). *The Dynamics and the Uncertainty of Delays at Signals*, PhD thesis, Delft University of Technology. TRAIL Thesis Series, T2006/7.
- 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.
- VSS (1998). *Norme Suisse SN 640017a Capacité, niveau de service, charges compatibles; norme de base*, Union des professionnels suisses de la route, VSS, Zurich.
- VSS (1999a). *Norme Suisse SN 640022 Capacité, niveau de service, charges compatibles; carrefours sans feux de circulation*, Union des professionnels suisses de la route, VSS, Zurich.
- VSS (1999b). *Norme Suisse SN 640023 Capacité, niveau de service, charges compatibles; carrefours avec feux de circulation*, Union des professionnels suisses de la route, VSS, Zurich.
- VSS (2006). *Norme Suisse SN 640024a Capacité, niveau de service, charges compatibles; carrefours giratoires*, Union des professionnels suisses de la route, VSS, Zurich.
- Webster, F. V. (1958). Traffic signal settings, *Technical Report 39*, Road Research Laboratory.
- Weiss, E. N. and McClain, J. O. (1987). Administrative days in acute care facilities: A queueing-analytic approach, *Oper. Res.* **35**(1): 35–44.
- Wild, S. M., Regis, R. G. and Shoemaker, C. A. (2008). ORBIT: Optimization by radial basis function interpolation in trust-regions, *SIAM Journal on Scientific Computing* **30**: 3197–3219.
- Wong, S. (1996). Group-based optimisation of signal timings using the TRANSYT traffic model, *Transportation Research Part B: Methodological* **30**(3): 217–244.
- Wong, S. (1997). Group-based optimisation of signal timings using parallel computing, *Transportation Research Part C: Emerging Technologies* **5**(2): 123–139.
- Wong, S., Wong, W., Leung, C. and Tong, C. (2002). Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control, *Transportation Research Part B: Methodological* **36**(4): 291–312.

- Wong, S. and Yang, H. (1997). Reserve capacity of a signal-controlled road network, *Transportation Research Part B: Methodological* **31**(5): 397–402.
- Yin, Y. (2008). Robust optimal traffic signal timing, *Transportation Research Part B: Methodological* **42**(10): 911–924.
- Ziyou, G. and Yifan, S. (2002). A reserve capacity model of optimal signal control with user-equilibrium route choice, *Transportation Research Part B: Methodological* **36**(4): 313–323.

# Carolina Osorio – Curriculum Vitae

Email: carolina.osoriopizano@epfl.ch  
Phone: +41 21 693 9327

Home page: <http://transp-or2.epfl.ch/osorio>  
Affiliation: TRANSP-OR INTER ENAC EPFL

---

## RESEARCH

---

**INTERESTS :**      **OPERATIONS RESEARCH**, in particular: queueing theory, simulation, optimization.  
Application fields: transportation, health care and biology.

## PUBLICATIONS

---

### INTERNATIONAL JOURNALS

- Osorio, C., and Bierlaire, M. (2009) 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., Weibel, C., Perez, P., Bierlaire, M., and Garnerin, Ph. (2006). Patient flow simulation as a tool for estimating policy impact, *Swiss Medical Informatics* 58:33-36

### CONFERENCE PROCEEDINGS

#### **Osorio, C., and Bierlaire, M. (2009)**

- A multi-model algorithm for the optimization of congested networks. European Transport Conference
- A simulation optimization framework for the management of congested urban road networks. Swiss Transport Research Conference
- A metamodel approach for simulation optimization of congested urban road networks. Computational Management Science Conference

#### **Osorio, C., and Bierlaire, M. (2008)**

- Network performance optimization using a queueing network model. European Transport Conference
- Signal control optimization with a queueing network model capturing congestion. Pan-American Conference on Traffic and Transportation Engineering
- A queueing network approach to the traffic signal optimization of the Lausanne city center. Latin-Ibero-American Conference on Operations Research
- A multiple model approach for traffic signal optimization in the city of Lausanne. Swiss Transport Research Conference

#### **Osorio, C., and Bierlaire, M. (2007)**

- An analytic finite capacity queueing network model capturing congestion and spillbacks. Triennial Symposium on Transportation Analysis
- Describing network congestion and blocking with an analytic queueing network model. Swiss Transport Research Conference

### TECHNICAL REPORTS

- Osorio, C., and Bierlaire, M. (2009). A surrogate model for traffic optimization of congested networks: an analytic queueing network approach. Transport and Mobility Laboratory, ENAC, EPFL
- Osorio, C., and Bierlaire, M. (2008). Accounting for congestion and spillbacks in fixed-time traffic signal optimization: an analytical queueing model approach. Transport and Mobility Laboratory, ENAC, EPFL
- Osorio, C., and Bierlaire, M. (2007). An analytic finite capacity queueing network model capturing blocking, congestion and spillbacks. Transport and Mobility Laboratory, ENAC, EPFL
- Osorio, C. (2005). An auxiliary/latent variables approach to inferring missing haplotype/genotype data. ENSIMAG. Department of Statistical Science, University College London
- Osorio, C. (2004). Analyzing, improving and comparing an outlier robust independent component analysis algorithm. ENSIMAG. Intelligent Data Analysis Group, Fraunhofer Institute FIRST, Berlin Germany

### TALKS

---

#### **2009**

- Intelligent Transportation Systems Research Seminar Series, MIT, Cambridge (USA)
- Computational Management Science Conference, Geneva (Switzerland)
- Swiss Transport Research Conference, Ascona (Switzerland)
- European Transport Conference, Noordwijkerhout (Netherlands)

#### **2008**

- Symposium of the Association of Colombian researchers in Switzerland, Lausanne (Switzerland)
- Latin-Ibero-American Conference on Operations Research, Cartagena (Colombia)

- Pan-American Conference on Traffic and Transportation Engineering, Cartagena (Colombia)
- Industrial Engineering Department Seminar Series, Universidad de los Andes, Bogota (Colombia)
- Swiss Transport Research Conference, Ascona (Switzerland)
- European Transport Conference, Noordwijkerhout (Netherlands)
- Intelligent Transportation Systems Research Seminar Series, MIT, Cambridge (USA)

## 2007

- 3ème Cycle Romand de Recherche Opérationnelle, Zinal (Switzerland)
- Selected chapters from Operations Research, Mathematics Doctoral School, EPFL, Lausanne (Switzerland)
- Triennial Symposium on Transportation Analysis, EPFL, Phuket Island (Thailand)
- Swiss Transport Research Conference, Ascona (Switzerland)

## 2006

- Joint Operations Research Days, EPFL, Lausanne (Switzerland)
- Selected chapters from Operations Research, Mathematics Doctoral School, EPFL, Lausanne (Switzerland)

## RESEARCH PROJECTS

---

- Simulation-based optimization of the performance in hospital operating suites (2005-2009). In collaboration with the Geneva University Hospitals. Funded by the Swiss National Science Foundation (grants 205321-107838 and 205320-117581)
- Location of distribution centers. In collaboration with PostLogistics (2006)

## THESES

---

- **PhD** Mitigating network congestion: analytical models, optimization methods and their applications.  
Dept. of Mathematics, EPFL. Supervised by Prof. Michel Bierlaire.  
Jury: Professors J. Barcelo (UPC), F. Eisenbrand (EPFL), A. Odoni (MIT) and P. Thiran (EPFL)
- **Master of Science** An auxiliary/latent variables approach to inferring missing haplotype/genotype data.  
Dept. of Statistical Science, University College London. Supervised by Dr. David Lunn, Imperial College London.
- **Bachelor Research** Analyzing, improving and comparing an outlier robust independent component analysis algorithm. Intelligent Data Analysis Group, Fraunhofer Institute FIRST, Berlin. Supervised by Dr. Stefan Harmeling.

## OTHER RESEARCH ACTIVITIES

---

- Reviewer of the journals: *European Journal of Operational Research* and *Computers & Operations Research*
- Member of the scientific committee of the Rapid Modeling Conference, Neuchatel, Switzerland (2010)
- Visiting researcher at the Intelligent Transportation Systems laboratory, Massachusetts Institute of Technology (MIT), directed by Prof. Moshe Ben-Akiva (Oct. 2008 - March 2009)
- Involved in the organization of the Sixth Triennial Symposium on Transportation Analysis (TRISTAN VI), Phuket Island, Thailand (June 10-15 2007)
- Former member of the Operations Research laboratory ROSO, EPFL, directed by Prof. Thomas Liebling (2005- 2006)
- Visiting student at the Intelligent Data Analysis laboratory, Fraunhofer Institute FIRST, directed by Prof. Klaus-Robert Müller, in Berlin, Germany (July-Aug. 2004)

## TEACHING

---

### LECTURER

---

- **Simulation tools**  
Two lectures for the undergraduate Mathematics course: *Decision Models*.  
Lausanne Switzerland (Fall 2007)
- **Modeling and simulation in logistics**  
Lecture for the executive master program: *Management of Logistical Systems* (now called *Global Supply Chain Management*), of the International Institute for the Management of Logistics.  
Paris France (December 13, 2007).

### TEACHING ASSISTANT – GRADUATE OR PROFESSIONAL COURSES

---

The following are all one week courses.

- **Discrete choice analysis: predicting demand and market shares**  
Professional course taught by Professors Moshe Ben-Akiva (MIT), Michel Bierlaire (EPFL), Denis Bolduc (University of Laval) and Daniel McFadden (University of California; Nobel Prize Laureate 2000)  
Lausanne Switzerland (2010, 2009, 2008)

- **Modeling and simulation in logistics**

Module in the executive master program: *Management of Logistical Systems* (now called *Global Supply Chain Management*), of the International Institute for the Management of Logistics  
Lausanne Switzerland (2008, 2007, 2006)  
Paris France (2007, 2006)

---

## **TEACHING ASSISTANT – UNDERGRADUATE COURSES**

---

- **Mathematical modeling of behavior** (discrete choice modeling)
  - EPFL Mathematics students (Fall 2007, Fall 2006)
- **Operations Research**
  - EPFL Computer Science students (Spring 2006, Fall 2005)
  - EPFL Communications Systems students and Physics students (Fall 2005)

---

## **STUDENT PROJECT SUPERVISION**

---

Two master projects of students in Mathematics, and four semester projects of students in Mathematics, Computer Science and Civil Engineering.

---

## **AWARDS**

---

2002-2005 Highest Distinction-Scholarship for engineering school studies given by the French government.  
2000-2002 Excellence-Scholarship for undergraduate studies given by the French government.

---

## **EDUCATION - DEGREES**

---

2006-2010 PhD in Mathematics at EPFL. Lausanne Switzerland.  
2004-2005 Master of Science in Statistics with distinction at University College London (UCL). London UK.  
2002-2005 French Engineering Diploma at the National College of Applied Mathematics and Computer Science of Grenoble (ENSIMAG). France.  
2002 Mathematics and Computer Technology applied to Science Diploma, with distinction: DEUG MIAS. Bordeaux I University. France.  
2000-2002 Preparation for the highly competitive entrance examination to French Engineering Colleges, with emphasis on Mathematics and Physics: MPSI-MP at Montaigne School. Bordeaux France.  
2000 Obtained the Colombian high school diploma: ICFES.  
Obtained the French high school diploma with distinction: Baccalauréat.  
1992-2000 Secondary studies at the French high school Louis Pasteur. Bogota Colombia.  
1986-1992 Primary studies at St. Gabriel's School. Toronto Canada.

---

## **LANGUAGES**

---

Spanish, French, English: **trilingual**.

English: TOEFL: 280/300 computer based, 2004.

German: ZDfB: diploma Professional German, with distinction. Goethe Institute (2003).

ZDaF: diploma German as a foreign language, with distinction. Goethe Institute (1999).

1999 July-Aug. Intensive language study in Cologne, Germany.

1998 July-Aug. Intensive language study in Munich, Germany

---

## **INDUSTRIAL INTERSHIP**

---

2003 July-Aug. Internship at Colpatria Bank. Bogota Colombia.  
Assistant for the development of a statistical credit risk model.