TRAFFIC SIGNAL CONTROL IN CONGESTED URBAN NETWORKS:
SIMULATION-BASED OPTIMIZATION APPROACH

Xiao Chen

UNIVERSITY OF COIMBRA

Xiao Chen

TRAFFIC SIGNAL CONTROL IN CONGESTED URBAN NETWORKS: SIMULATION-BASED OPTIMIZATION APPROACH

PhD Thesis in Doctoral Program in Transport System supervised by Carolina Osorio and Bruno Santos, presented to the Department of Civil Engineering of the Faculty of Sciences and Technology of the University of Coimbra

September, 2014

· U C ·

UNIVERSIDADE DE COIMBRA

FCTUC FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Traffic Signal Control in Congested Urban Networks: Simulation-based Optimization Approach

Doctoral thesis

Thesis submitted to the Faculty of Sciences and Technology of the University of Coimbra in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the field of Transport Systems.

Author

Xiao Chen

Supervisors

Carolina Osorio (Massachusetts Institute of Technology, Cambridge, USA)

Bruno Filipe L. Santos (University of Coimbra, Portugal)

Financial support

This research work was financed by Foundation for Science and Technology (FCT) through the MIT Portugal Program and PhD grant SFRH/BD/51296/2010, and was co-financed by project EMSURE - Energy and Mobility for Sustainable Regions (CENTRO-07-0224-FEDER-002004).

Acknowledgements

I would like to take this opportunity to show my greatest honor and appreciation to Professor Carolina Osorio for her patient guidance, enthusiastic encouragement and valuable critiques for all research topics we are working on. Her brilliant ideas always inspire me to conquer every difficulty we are facing. Without her constant help I will never finish this thesis.

I also express my deep gratitude to Professor Bruno Santos for his valuable and constructive suggestions during the planning and development of this research work. He constantly encourages me when I am depressed, I am always infected by his optimism.

Words are limited to express my very great appreciation to Professor Antonio Pais Antunes. I am bothering him not only with research problems but also with all the troubles I am confronting during my whole PhD: housing issue, VISA issue, computer problems, etc.. He provides his help and time generously with every trouble I have. He is a wisdom mentor and a good friend, he makes me feel I am not alone abroad.

I acknowledge the valuable suggestions from Professor Alvaro Seco and Professor Jose Viegas at the planning stage of this thesis.

I sincerely thank FCT, MIT Portugal Program and Energy and Mobility for Sustainable Regions project (EMSURE) at University of Coimbra for the financial support

for my studies both in Portugal and at MIT. Moreover, it offers me the opportunity to present my work in major conferences around the world.

My thankfulness goes also to Dr. Emmanuel Bert and Prof. André-Gilles Dumont (LAVOC, EPFL) for providing the Lausanne simulation model. I would like also to express my deep gratitude to New York City Department of Transport (NYCDOT) for providing the Manhattan simulation model. I really appreciate the supports from the NYCDOT folks: Michael Marsico, Mohamad Talas, Jingqin Gao and Shitao Zhang, for their patient help and brilliant suggestions.

I want to acknowledge all my colleagues and friends at the Department of Civil Engineering. Thanks to my dear officemates Diego, Melissa, Ashenafi, Joana, and Joao. A special thank to Diana, thanks for her help during my whole PhD.

My deepest gratitude goes to my dear friends and colleagues at MIT. A specially thank goes to Linsen, we share ideas, thoughts, happiness and codes. I still remember the trips we made together. Thanks to all the lab members at Carolina's group, I learned a lot from their brilliant thoughts and ideas. To my cute roommates Fei and Caly at MIT, we have spent so much excited and crazy time together, thanks for the delicious food they made for me. I really miss our weekend parties.

Words are limited to express my sincere gratitude to my parents and grandmother for their supports and endless love. They are always there for me no matter what happens.

Finally, to Gang, I really appreciate all the things you have done for me. Your selfless supports make my life much easier. Your love and encouragements inspire me all the time. We all love traveling, and we have completed so many journeys together. Future is unknown, but I am grateful to have you in my journey of life.

Contents

List of Figures

List of Tables

Abstract

Congestion has become a global phenomenon. In particular in great urban areas, daily traffic jams are in most cases a major concern. Managing signal plans efficiently is one of the most cost-effective methods. However, existing signal control strategies are less powerful in handling congested network with spillbacks and grid-type topology.

Enhancing the reliability of our networks is currently recognized as a critical goal in the US and in Europe. There is extensive evidence that indicates that travel time reliability is accounted by travelers in a variety of travel decisions, such as departure time and route choice. Hence, operating our networks such as to reduce both the average and the variability of trip travel times would be highly valued by travelers. However, urban traffic management strategies are typically formulated such as to improve first-order performance metrics (e.g. expected trip travel times, expected link speeds). The main challenge in addressing reliability in traditional transportation optimization problems is the need to provide an accurate analytical and tractable approximation of trip travel time distribution, or of its first- and second-order moments.

The complex between-link spatial-temporal dependency patterns makes accurate analytical modeling of urban road networks a challenge. In particular when the aim is to model metrics related to the paths chosen by the drivers, in order to reflect driver experiences. Thus, this work proposes new signal control strategies for large-scale congested urban networks that can tackle these challenges.

In this thesis, a simulation-based optimization (SO) approach is used to address traffic signal control problems. Microscopic simulators describe in detail the interactions between vehicle performance, traveler behavior and the underlying transportation infrastructure. They can ultimately contribute to the design of traffic management strategies, providing detailed system performance estimates to infer the design and operations of urban networks. To ensure the computational efficiency, an analytical approximation of objective function is needed. We develop different formulations of travel time reliability based on both link travel time and path or trip travel time distributional information, and then use those formulations in signal design strategies to fulfill the reliability requirements. We also design a simulation-based adaptive traffic signal control algorithm to adjust signals plans dynamically according to real-time traffic conditions.

We apply the reliable signal control strategy to both city center and the full city of Lausanne. The proposed simulation-based adaptive traffic signal control algorithm is applied to a grid-type urban network with heavy traffic in east Manhattan area (New

York City, USA). In both cases, proposed methods lead to signal plan with better performance in terms of various performance metrics.

Keywords: signal control, reliability, Little's law, adaptive traffic signal control, simulation-based optimization.

Resumo

O congestionamento tornou-se um fenómeno global, com particular relevância no caso das grandes áreas urbanas onde os engarrafamentos rodoviários são por norma uma preocupação diária. A gestão eficiente de planos semafóricos é certamente um das formas mais rentáveis de lidar com este fenómeno. No entanto, as estratégias existentes para o controle de sinais semafóricos são por norma pouco poderosas na manipulação de redes congestionadas de tipologia reticulada e que sofrem de efeitos spillback.

Melhorar a fiabilidade das nossas redes é atualmente reconhecido como um objetivo fundamental, tanto nos EUA como na Europa. Há uma ampla evidência sobre como a fiabilidade tempo de viagem é considerada por viajantes em uma variedade de decisões de viagem, tais como na escolha do horário de saída e da rota de viagem. Assim, operar as redes rodoviárias por forma a reduzir tanto a média como a variabilidade dos tempos de viagem seria muito valorizado pelos viajantes. No entanto, as atuais estratégias de gestão do tráfego urbano são normalmente formuladas de modo a apenas melhorar os indicadores de desempenho de primeira ordem (como é o casa

dos tempos de viagem esperados ou as velocidades esperadas nos eixos). O principal desafio na abordagem de introduzir objetivos de fiabilidade em problemas de otimização de transporte tradicionais é a necessidade de encontrar uma aproximação analítica útil e precisa para a distribuição do tempo de viagem, ou seja, para os seus momentos de primeira e de segunda ordem. A complexidade das dependências entre os eixos da rede e as próprias relações espaço-temporais, torna a modelação analítica exata da rede rodoviária um desafio. Em particular quando se pretende modelar indicadores de performance ao nível dos percursos tomados pelos condutores, de forma a refletir as experiências de viagem dos condutores. Este trabalho propõe por isso novas estratégias de controlo semafórico, que conseguem lidar com os desafios indicados e que são particularmente úteis para redes urbanas congestionadas de grande escala.

Nesta tese, um modelo de otimização baseada em simulação (SO) é usado para tratar problemas semafóricos de controle de tráfego. Os simuladores microscópicos descrevem em detalhe as interações entre o desempenho do veículo, o comportamento dos condutores e a infraestrutura de transporte subjacente. Eles podem contribuir para o desenvolvimento de estratégias de gestão de tráfego, proporcionando estimativas detalhadas do desempenho do sistema que podem ser usadas tanto no planeamento como na avaliação da performance de redes urbanas. No entanto, para assegurar a eficiência computacional, é necessária lidar com uma aproximação analítica da função objetivo. Para isso, desenvolvemos diferentes formulações de fiabilidade do tempo

de viagem, com base em distribuições tanto do tempo de viagem nos eixos como do tempo de viagem no percurso ou viagem. Essas formulações foram posteriormente usadas no desenvolvimento de estratégias que preenchem os requisitos em termos da fiabilidade dos tempos de viagem. Desenvolveu-se ainda um algoritmo de controle reactivo de semáforos baseado em simulação, de modo a ajustar os planos semafóricos dinamicamente de acordo com as condições de tráfego em tempo real.

A estratégia de fiabilidade para o controle semafórico proposta nesta tese é aplicada tanto à rede do centro da cidade de Lausanne como à rede completa da mesma cidade. O algoritmo de controle semafórico reactivo com base em simulação é aplicado a uma rede urbana reticulada, com elevados níveis de tráfego, na zona leste da Ilha de Manhattan (Nova York, EUA). Em ambos os casos, os métodos propostos obtêm planos semafóricos com melhor desempenho em termos das várias medidas de desempenho usadas.

Palavras-chave: planos semafóricos, fiabilidade, Little's law, controle semafórico adaptativo, otimização baseada em simulação.

Chapter 1

Introduction

1.1 Background

In the future, given the growing number of population living in the urban area, the urban road systems are facing more demand. Given urban space constraints, road systems capacity cannot develop at the same rhythm. In fact, road systems seem confronting a bottleneck that commonly leads to major congestion experiences. Road congestion happens in many large cities. It is characterized by slower speeds, longer trip times, and queueing phenomena. It may result in late arrival for work or school, reducing travelers' productive time, increasing fuel consumption and air pollution; increasing individuals stress, and limit regional economic growth. In the year 2001, the external costs caused by road traffic congestion reached almost 0.5% of the EU Community GDP (gross domestic product) (EuropeanComission, 2001). These costs

were expected to grow by 142% until 2010, reaching 1% of the community GDP, and to increase more 50% in the next four decades (EuropeanComission, 2011).

Building new infrastructures, such as roads, tunnels and more interchanges, are usually the most direct approach to increase road capacity. However, in high-density urban areas, especially in the historical areas, it is hard to build new infrastructures. Other approach is to explore more efficiently use of existing infrastructures by adopting efficient and effective traffic management solutions. Intersection is an important component of urban road network, and one of the most common types is the signalized intersection. Traffic signal controls are implemented to reduce or eliminate conflicts among multiple traffic streams at intersections. Signals manage these conflicts by controlling access to the intersection, allocating green time to some movements while showing the red signal to the conflicting movements. Generally speaking, there are two types of traffic signal setting strategies, namely the fixed-time, and adaptive traffic signal control systems (ATCSs). Compared to the traditional fixed-time signal control strategies, adaptive traffic control systems provide a more flexible option for adjusting signal timings to accommodate changing traffics. Fixed-time traffic signal control is used in majority intersections because it is easier to deploy and maintain.

Traffic signal optimization has been a major topic of research in the last 50 years since the work of Webster (1958). However, it is widely accepted that traffic signal benefits are not fully realized, there is plenty of room for improvement (Lo et al., 2001).

The most common signal control design objectives only account for first-order distributional information such as average or total travel time, system throughput, number of vehicle stops. Very limited efforts have been done to account for higher-order distributional information in signal control design objectives such as travel time reliability. The main challenge in addressing reliable signal control problem is to provide accurate and tractable analytical approximation of the trip travel time distribution accounting for between-link spatial-temporal dependency. Besides the reliable traffic signal control problem, another problem for current traffic control strategies is the limited ability to provide signal plans that could improve the system performance efficiently under very congested traffic condition, especially for congested grid-type urban networks. It is still a challenge for designing signal control strategy to handle congested and oversaturated traffic conditions.

One way of solving these issues is to incorporating the detailed and accurate system performance estimates obtained from microscopic simulator to inform the design and operations of traffic signal controls. Stochastic microscopic traffic simulators are widely used in signal control analysis. The stochastic modeling of demand and supply improves our ability to understand complex traffic and behavioral phenomena. Simulators provide a detailed description of the underlying supply (e.g. road capacity and traffic management strategies), demand (e.g. drivers behavior models), as well as of their interaction. One of the most important system performance measure is travel

time or delay. Taking travel time as an example, analytical techniques are computationally tractable and efficient, yet rely on strong distributional assumptions, such as the choice of a given parametric distribution for link or path delay (e.g.: normal or lognormal). More importantly, they fail to account for the complex spatial-temporal dependencies between links, which are due, for instance, to vehicle-to-vehicle interactions and vehicle-to-supply interactions. Such interactions highlight the complex that the travel time distribution may take. The use of microscopic simulators, which account for local dynamics and for the complex local and network-wide supply-demand interactions, can yield a more detailed representation of between-link dependencies, and travel time distributions. The direct use of these stochastic and computationally intensive simulators for control purposes is a challenging task. In order to derive computationally efficient methods that embed non-efficient simulators, information from other more tractable traffic models is used throughout the optimization process. The role of these auxiliary models is to provide analytical structural information to the algorithm, which enables the identification of well performing alternatives with very small samples. Osorio (2010) presented a metamodel simulation-based (SO) optimization method that combines the information from a microscopic traffic simulation model with an analytical queueing network model. In that approach, a fixed-time signal control problem that accounts for first-order travel time information (expected trip travel time) is solved. Although the results obtained from this metamodel proved

to be suitable for congestion urban networks, providing better signal plans with improved performance, there is still much room for improvements such as incorporating reliability concerns in traffic signal design objectives and extent the fixed-time signal control problem to adaptive signal control.

Nowadays, enhancing the reliability of transportation networks is recognized as a critical goal. Recent London and U.S. reports have demonstrated the importance of improving the reliability of our transportation systems (Transport for London, 2010; Texas Transportation Institute, 2012; Department of Transportation, 2008). Increased reliability yields a more stable and less disruptive transportation service. Past work has emphasized that traffic signal control has the potential to improve travel time reliability (Robert L. Gordon, 2005). Furthermore, for grid-type networks with heavy traffic, the formation of queues can cause spillback effects, blocking nearby intersections, and spreading the congestion phenomena across the road network for a longer time period. As a result, this research aims to extend the work developed by Osorio to solve signal control problems that are important but receive less attention or have limitations. We are aiming to design signal control strategy taking into consideration of travel time reliability, and develop both fixed-time and adaptive traffic signal control strategies for very congested urban network especially for grid-type networks with spillbacks.

1.2 Objectives

In this thesis, two problems are of particular interests: one is designing signal control strategy that offers reliable service for congested urban networks; the other is developing traffic signal control strategy that can be controlled and dynamically adjusted according to the travel demand.

For the first problem, the major challenge in improving travel time reliability is the approximation of the network travel time distribution. An analytical and accurate expression for the full joint network distribution is difficult to derive given the intricate between-link dependencies. For the second problem, the major challenge is to deal with grid-type congested urban network. The studied area contains numerous intersections, multimodal traffic and short links. In the literature, either fixed-time, or adaptive traffic signal control systems have limited ability in handling such type of network. Within this context, the objectives of this thesis are:

1. To incorporate tractable link travel time distributional information in signal design objectives within the SO framework to enhance travel time reliability for urban network.

2. To propose an analytical and accurate expression of trip travel time distribution considering intricate between-link dependency to overcome the limitation of using independent link travel times.

3. To incorporate the proposed trip travel time distributional information within the SO framework to solve reliable signal control problems.

4. To investigate the performance of the signal plans derived by using different travel time reliability formulations.

5. To design adaptive traffic signal control algorithm to improve system performance under various traffic conditions for congested grid-type urban area.

1.3 Thesis structure and contributions

To the best of our knowledge, this thesis constitutes the first attempt to 1) derive analytical and tractable approximation of second-order distributional information of link, path and trip travel time that can be used to solve transportation optimization problems and simulation-based optimization problems; 2) use higher-order distributional information (both analytical and simulation-based) to solve urban traffic signal control problem, and 3) design simulation-based adaptive traffic signal control algorithm for highly congested grid-type network using queue management techniques in design objective.

Chapter 2 considers a simulation-based reliable signal control problem. In this chapter, first-order and second-order link travel time distributional information are combined in the signal design objective function to derive fixed-time signal plans.

This formulation is used to address signal plans for both city center and full city of a Swiss city Lausanne. The signal plans derived are compared with the signal plans that only consider the first-order or second-order travel time distributional information to address the added value of combing both expectation and variability of travel time in signal control problems.

The results of Section 2.4.2 have been presented and published as:

Chen, X., and Osorio, C., and Santos, B. (2012). *A Simulation-Based Approach to Reliable Signal Control.* Proceedings of the International Symposium on Transportation Network Reliability (INSTR), Dec. 18-19, 2012.

The results of Section 2.4.2-2.4.5 have been presented and published as:

Chen, X., and Osorio, C., and Santos, B. (2013). *Travel Time Reliability in Signal Control Problem: Simulation-Based Optimization Approach.* Proceedings of the Transportation Research Board (TRB) Annual Meeting January 13-17, 2013.

Chen, X., and Osorio, C., and Santos, B. (2013). *Simulation-based reliable signal control.* Proceedings of the Triennial Symposium on Transportation Analysis (TRISTAN VIII), June 9-14, 2013.

The full chapter has been submitted to *Transportation Science*.

Chapter 3 proposes an analytical and tractable formulation of trip travel time variability that explicitly considers between-link dependency. This formulation is compared with the formulations that ignore the between-link dependency for different toy

networks to verify the accuracy of the formulation. The proposed formulation of trip travel time variability is used to solve a reliable signal control problem for the city center of Lausanne. The performance of the signal plans derived in **Chapter 3** are compared with the signal plans derived in **Chapter 2** to address the added value of accounting between-link dependency in reliable signal control problems.

The results of Section 3.3, Section 3.4 and Section 3.5 have been presented and published as:

Chen, X., and Osorio, C. (2014). *Analytical formulation of trip travel time distribution.* Proceedings of the EURO Working Group on Transportation (EWGT) July 2-4, 2014.

The results of Section 3.3, Section 3.4 and Section 3.5 have been published as:

Chen, X., and Osorio, C. (2014). *Analytical formulation of trip travel time distribution.* Transportation Research Procedia Special Issues.

The full chapter has been submitted to *Transportation Science*.

Chapter 4 proposes fixed-time signal control strategy for a grid-type urban network with heavy traffic in east Manhattan. Based on the signal plan control strategy proposed in **Chapter 4**, **Chapter 5** further proposes an adaptive traffic signal control algorithm for the same area of Manhattan. The performance of the proposed signal plans are compared with the existing signal plan in use for that area.

Chapter 4 has been presented and published as:

Osorio, C., Chen, X., Marsico, M., Talas, M., Gao, J., Zhang, S. (2014). *Reducing gridlock probabilities via simulation-based signal control.* Proceedings of the International Symposium of Transport Simulation (ISTS) June 1-4, 2014.

The preliminary results of **Chapter 5** have been accepted by Transportation Research Board (TRB) Annual Meeting, 2015.

Chapter 6 presents conclusions and future development of research that are related to this thesis.

Chapter 2

A Simulation-Based Approach to Reliable Signal Control

2.1 Introduction

Traffic signal control is a cost-effective way to make better use of the existing potential capacity of an urban transportation network, and more generally of the existing infrastructure. It is widely accepted that traffic signal benefits are not fully realized and there is plenty of room for improvement (Lo et al., 2001).

Travel time reliability is an important metric used to evaluate the performance of a transportation system. It can be defined as travel time variability. According to a stated preference survey, it is considered to be either the most important or second most important reason for the commuting route choices of 54% of morning

commuters in Los Angeles(Abdel-Aty and Jovanis, 1996). Bates et al. (2001) showed that some transportation users value more the reduction of travel time variability than the expected travel time.

Enhancing the reliability of transportation networks is currently recognized as a critical goal. A recent Transport for London report identifies trip travel time reliability improvements as their primary objective (Transport for London, 2010). U.S. reports have also emphasized the importance of improving the reliability of our transportation systems (Texas Transportation Institute, 2012; Department of Transportation, 2008). Increased reliability yields a more stable and less disruptive transportation service.

Transport network can be model by urban traffic simulation models. There are three main families of urban traffic simulation models: macroscopic, mesoscopic and microscopic (for a review see Barceló (2010)). Microscopic models embed the most detailed representation of both demand and supply. They explicitly consider vehicle-specific attributes for each individual vehicle. They also represent individual travelers and embed detailed disaggregate behavioral models (e.g. departure-time choice, route choice, lane-changing, car-following, re-routing). Since they account for complex local traffic dynamics and demand-supply interactions, they capture the between-link spatial-temporal dependencies of the main performance measures, and can thus yield accurate estimates of the full distribution of the main performance measures. These distributions can be used to inform the design and operations of transportation sys-

tems by, for instance, addressing reliable formulations of traditional transportation problems.

The direct use of these stochastic and computationally intensive simulators for control purposes is a challenging task. In order to derive computationally efficient methods that embed non-efficient simulators, information from other more tractable traffic models is used throughout the optimization process. The role of these auxiliary models is to provide analytical structural information to the algorithm, which enables good short-term algorithmic performance to be achieved.

This chapter proposes a methodology that enables the use of detailed stochastic traffic simulators to efficiently address higher-order simulation-based optimization (SO) problems. Additionally, we focus on the development of computationally efficient SO techniques, the objective is to identify within the pre-specified computational budget signal plans that improve both first- and second-order distributional information (defined as a maximum number of simulation runs). In order to achieve efficiency, information from the (inefficient) simulator is coupled with information from an efficient (i.e. tractable and differentiable) analytical approximation of the objective function. The role of the simulator is to provide a highly detailed approximation of the distributions of interest, whereas that of the analytical model is to provide structural information to the SO algorithm, enhancing its efficiency. This chapter is structured as follows. Section 2.2 presents a review of reliability metrics, and their use for signal

control. We then present the proposed methodology (Section 2.3). Empirical results based on case studies in the Swiss city of Lausanne are presented in Section 2.4. We conclude with a brief discussion in Section 2.5.

2.2 Literature Review

There are four types of reliability measures presented in transportation studies. The early-proposed reliability measures are connectivity and travel time reliability. Connectivity reliability is defined as the probability that the network nodes are still connected if one or more links fail to connect due to incidents (Wakabayashi and Iida, 1991). Travel time reliability is used to account for the stochastic travel time variations. Capacity reliability is defined as the maximum traffic volume that a network can accommodate (Chen et al., 1999a). More recently, the concept of potential reliability or vulnerability is proposed (D'Este and Taylor, 2003). It can be defined as the exposure of the road system to incidents that can result in significant reductions in the system capacity. Santos et al. (2010) integrated vulnerability in a network design problem by focusing on the potential consequences on overall network performance if some links are closed. For a more detailed description of network reliability metrics, see Clark and Watling (2005).

This chapter focuses on travel time reliability. The two most common metrics used to address travel time reliability are trip travel time variability and trip travel time

percentiles (e.g., 95^{th} percentile) (OECD, 2010).

In order to account for travel time reliability, there is a need to go beyond the approximation of expected travel times, and use higher-order distributional information (e.g. variance or full distributional information). Nonetheless, a major challenge in improving travel time reliability is the approximation of the network travel time distribution. An analytical and accurate expression for the full joint network distribution is difficult to derive given the intricate between-link spatial-temporal dependencies. A variety of analytical approximations have been proposed based on distributional assumptions such as functional form of the full joint distribution (Mirchandani and Soroush, 1987), functional form of marginal link distributions (Fu and Hellinga, 2000), and moments of the marginal distribution (Ng et al., 2011). Empirical (non-parametric) analysis of link travel time distributions have also been proposed (van Lint and van Zuylen, 2005; Chen et al., 2003). On the other hand, simulators can yield distributional estimates that account for such complex dependencies. The use of these simulators is mostly limited to what-if (i.e., scenario-based) analysis (as in, for instance, Bullock et al. (2004); Ben-Akiva et al. (2003)). Their use within simulation-based optimization (SO) algorithms is rare, and limited to the use of first-order distributional information (Li et al., 2010a; Stevanovic et al., 2009, 2008; Branke et al., 2007; Yun and Park, 2006; Hale, 2005; Joshi et al., 1995).

This chapter focuses on reducing travel time variability. In general, spatial-temporal

variations, in both demand and supply, can lead to increased variability (see Clark and Watling (2005) or Noland and Polak (2002) for details on common underlying causes of supply and demand variability). Increased variability leads to increased uncertainty for travelers, and increased travel cost (Noland and Polak, 2002). There is a substantial body of research that studies the behavioral impacts of travel time variability. Noland and Polak (2002) provide a review. Carrion and Levinson (2012) review methodologies to quantify the value of travel time reliability. Such studies highlight that travel time variability is accounted for in numerous travel decisions, and that its reduction is of high value to travelers. Thus, there is a need to design and operate transportation systems to account for it.

The importance of accounting for travel time variability in signal control has been emphasized by Yin (2008). The traditional signal control objectives are network efficiency maximization, such as throughput maximization (Abu-Lebdeh and Benekohal, 1997), travel time minimization (Osorio and Bierlaire, 2009b), and number of vehicle stops or delay minimization (Wong et al., 2002).

To the best of our knowledge, the few studies that have accounted for travel time variability in the design of signal plans are based on analytical methods. Yin (2008) proposes an analytical technique to reduce the standard deviation of delay and, ultimately, enhance the robustness of signal plans to fluctuations in demand. The demand fluctuation is represented by different demand scenarios. The technique is applied to

an isolated intersection. Zhang et al. (2010) extend the work of Yin (2008) to account for multiple intersections along an arterial. Another extension is proposed by Li (2011), which illustrates the method on an isolated intersection.

Park and Kamarajugadda (2007) and Kamarajugadda and Park (2003) develop an analytical approximation of delay variance. Parametric distributions are assumed for link volumes and the corresponding parameters are estimated with traffic count data. The analytical delay variances are then used to address a signal control problem for an isolated intersection and then for a set of two adjacent intersections.

In this chapter, we use simulated travel time distributional estimates. The use of detailed microscopic traffic simulators allows for the complex vehicle-to-vehicle and vehicle-to-infrastructure interactions to be accounted. The simulated travel time distributional estimates are then embedded within a simulation-based optimization algorithm and are used to identify signal plans with reduced expectation and standard deviation of travel time metrics.

2.3 Methodology

2.3.1 Simulation-based optimization framework

For recent reviews of SO methods, see Hachicha et al. (2010), Barton and Meckesheimer (2006) and Fu et al. (2005). We use the SO framework proposed by Osorio (2010).

This SO method is a metamodel method. Metamodels are deterministic functions that used to approximate objective functions. Comparing to other simulation-based optimization techniques such as genetic algorithm, the metamodel method is a computational efficient method because the deterministic optimization techniques can be used.

The method has been used to successfully address complex constrained simulation-based problems in a computationally efficient manner (Osorio et al., 2013; Osorio and Chong, 2012; Osorio and Nanduri, 2012). This section briefly presents the framework.

This algorithm can address continuous nonlinear generally constrained optimization problems where the objective function is derived from a stochastic simulator, i.e. a closed-form expression is not available for the objective function, whereas closed-form analytical expressions are available for all constraints. Such problems can be formulated as:

$$\min_{x} f(x, z; p) \tag{2.1}$$

subject to

$$g(x, z; p) = 0. \tag{2.2}$$

The feasible space is defined by g which is a set of general, typically nonconvex, de-

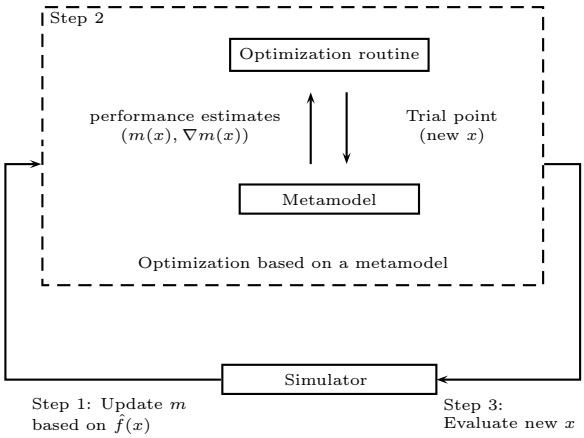


Figure 2-1: Metamodel simulation-based optimization methods. Adapted from Alexandrov et al. (1999).

terministic, analytical and differentiable constraints. The objective function f can be, for instance, the expected value of a given stochastic performance measure F: $f(x, z; p) = E[F(x, z; p)]$. The decision vector x is real-valued (e.g., green splits), z denotes other endogenous variables (e.g., departure-time/mode/route choice probabilities), and p denotes the deterministic exogenous parameters (e.g., network topology).

A metamodel is an analytical approximation of the objective function f. The main ideas of metamodel SO methods are given in Figure 2-1. At a given iteration k, the SO algorithm iterates over the following steps: 1) fit the metamodel, m_k, based on the set of simulation observations collected so far, 2) use m_k to perform optimization and derive a trial point x_k, 3) evaluate the performance of this trial point with the simulator, which leads to new simulation observations. As new simulated observations become available, the accuracy of the metamodel can be improved (Step 1), leading

to trial points with improved performance (Step 2). These steps are iterated until, for instance, the computational budget is depleted. The SO algorithm is given in detail in Appendix A.2.

Metamodels are classified in the literature as either physical or functional metamodels (Søndergaard, 2003; Serafini, 1998). Physical metamodels consist of application-specific metamodels, their functional form and parameters have a physical or structural interpretation. Functional metamodels are general-purpose (i.e. generic) 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.

The Osorio (2010) framework proposes a metamodel that combines a functional and a physical component and has the following functional form:

$$m(x, y; \alpha, \beta, q) = \alpha f_A(x, y; q) + \phi(x; \beta), \tag{2.3}$$

where ϕ (the functional component) is a quadratic polynomial in x (green split), f_A (the physical component) represents the approximation of the objective function (f of Equation (2.1)) as derived by an analytical macroscopic traffic model, y are endogenous macroscopic model variables (e.g., queue-length distributions), q are exogenous macroscopic parameters (e.g., total demand), α and β are parameters of the metamodel. The metamodel is fitted based on simulation observations of objective function

via regression.

We define the functional component ϕ as a quadratic polynomial in x with diagonal second-derivative matrix:

$$\phi(x;\beta) = \beta^1 + \sum_{j=1}^{d} \beta^{j+1} x^j + \sum_{j=1}^{d} \beta^{j+d+1} (x^j)^2, \tag{2.4}$$

where d is the dimension of x, x^j and β^j are the j^{th} components of x and β, respectively. At each iteration, the simulator and the queueing model are evaluated at one or two points, and then the metamodel parameters α and β are fitted by solving a least square problem based on both the current iteration observations and all the pervious observations.

The physical component f_A is derived by evaluating an analytical macroscopic traffic model, which is an analytical and differentiable macroscopic traffic model formulated based on finite capacity queueing network theory. It provides an approximation of the objective function across the entire feasible region. It enables the identification of well performing alternatives (often called trial points e.g.: green split) with very small samples. The metamodel is therefore a linear combination of an analytical approximation of the objective function and a quadratic error term. By resorting to a metamodel approach, the stochastic response of the simulation is replaced by a deterministic metamodel response function, m, such that efficient deterministic optimization techniques can be used.

In this chapter, we use this SO framework to address traffic signal control problems that are formulated based on higher-order (i.e., beyond first-order) distributional information.

2.3.2 Reliable signal control problem

The most common approach to account for both expected travel time and travel time standard deviation information is to use a linear combination: $t_E + rt_V$, where t_E denotes the expected trip travel time, t_V denotes a measure of trip travel time variability, and r is a weight parameter known as the reliability ratio. Such an approach is used in various studies, such as in Yin (2008) and in the traditional "mean-variance" approach (Jackson and Jucker, 1982). The reliability ratio can be interpreted as the relative importance that either the travelers' or the network operators' valuation of travel time variability. Normally, the variability metric t_V is trip travel time standard deviation.

The objective function of this chapter combines expectation and standard deviation information of a given travel time performance metric. The travel time metric used is the total link travel time (i.e., the sum of travel times over all links in the network of interest). Link travel time metrics are easier to measure in the field, and to approximate analytically, compared to trip travel time metrics.

In order to formulate the problem, we introduce the following notation:

b_i	available cycle ratio of intersection i;
T_i	Travel time along link i;
$x(j)$	green split of phase j;
x_L	vector of minimal green splits;
$\mathscr{L}$	set of links within the area of interest;
$\mathscr{Q}$	set of queues that represent the links of $\mathscr{L}$;
$\mathscr{I}$	set of intersection indices;
$\mathscr{P}_I(i)$	set of phase indices of intersection i;
r	reliability ratio.

Note that cycle ratio is calculated as available green time over cycle time.

The signal control problem is formulated as follows:

$$\min_x f(x,z;p) = E[\sum_{i\in\mathscr{L}} T_i(x,z;p)] + rSD[\sum_{i\in\mathscr{L}} T_i(x,z;p)], \tag{2.5}$$

subject to

$$\sum_{j\in\mathscr{P}_I(i)} x(j) = b_i,\ \forall i \in \mathscr{I} \tag{2.6}$$

$$x \geq x_L. \tag{2.7}$$

This problem is a fixed-time signal control problem, where the decision variables are the green splits. In this problem, the stage structure (e.g.: phase sequence) is given, the offsets, the cycle times and the all-red durations are fixed. The performance metric used, $\sum_{\mathscr{L}} T_i$, is the total link travel time. The objective function of this problem (Equation (2.5)) consists of a linear combination of the expected total link travel time, $E[\sum_{\mathscr{L}} T_i(x, z; p)]$, and the standard deviation of total link travel time $SD[\sum_{\mathscr{L}} T_i(x, z; p)]$. Constraints (2.6) guarantee that for a given intersection the sum of green splits of the endogenous phases equals the available cycle time. Constraints (2.7) correspond to the lower bound value for the green splits. In the case studies of this chapter it is set to 4 seconds following Swiss transportation norms (VSS, 1992).

2.3.3 Physical component

Recall that the metamodel formulation of Equation (2.3) requires an analytical expression for f_A, which is the approximation of the objective function f as derived by the auxiliary traffic model. This section derives the analytical (and differentiable) approximation of the two components of f provided by the auxiliary traffic model. That is, we derive analytical approximations for $E[\sum_{\mathscr{L}} T_i]$ and for $SD[\sum_{\mathscr{L}} T_i]$; or equivalently $E[\sum_{\mathscr{Q}} T_i]$ and $SD[\sum_{\mathscr{Q}} T_i]$.

The auxiliary model used is an analytical queueing network model based on finite capacity queueing theory. Each lane in the road network is modeled as one (or a set

of) queues. Each queue of the model is a finite capacity $M/M/1/k$ queue. k is the space capacity of each queue. The model is based on a stationary regime assumption. It consists of a system of nonlinear equations that relate the arrival and service rates of a queue to the demand and supply of its upstream and downstream queues. It describes spillbacks through the queueing theory notion of *blocking*. We briefly recall the main variables and parameters that define each queue. For a given queue i, we use the following notation.

λ_i	arrival rate;
$\hat{\mu}_i$	effective service rate (accounts for both service and eventual blocking);
k_i	space capacity;
N_i	number of vehicles in queue i;
$P(N_i = k_i)$	probability of queue i being full, known as blocking or spillback probability;
ρ_i	traffic intensity (defined as the ratio of arrival rate and effective service rate).

In traffic engineering, normally degree of saturation (defined as demand over capacity) is used to measure the level of congestion. In our case, the queueing network model is used to represent the studied area. In queueing theory, traffic intensity is a measure of the occupancy of the server, and it is used to measure the congestion level instead of degree of saturation.

Expected total travel time

The expected total travel time is obtained by summing the expected travel times of the queues (or equivalently links) of interest:

$$E[\sum_{i\in\mathscr{Q}} T_i] = \sum_{i\in\mathscr{Q}} E[T_i]. \tag{2.8}$$

The expected travel time of a given queue i is derived by applying Little's law (Little, 2011, 1961b):

$$E[T_i] = \frac{E[N_i]}{\lambda_i(1 - P(N_i = k_i))}, \tag{2.9}$$

where the expected queue-length of queue i, $E[N_i]$, is derived in Osorio and Chong (2012) and given by:

$$E[N_i] = \rho_i \left(\frac{1}{1-\rho_i} - (k_i+1)\frac{\rho_i^{k_i}}{1-\rho_i^{k_i+1}} \right). \tag{2.10}$$

Total travel time standard deviation

We now describe the approximation for $SD[\sum_{\mathscr{Q}} T_i]$. By definition:

$$SD[\sum_{\mathscr{Q}} T_i] = \sqrt{VAR[\sum_{\mathscr{Q}} T_i]}. \tag{2.11}$$

In order to derive a tractable analytical expression, we make the following approximation:

$$Var[\sum_{i\in\mathcal{Q}} T_i] \approx \sum_{i\in\mathcal{Q}} Var[T_i]. \tag{2.12}$$

The latter expression is exact only if all queues have independent travel times. This may be an inaccurate approximation in various congestion regimes. Nonetheless, recall that the main role of the physical component is to provide a tractable approximation of the objective function. Given the difficulty of accurately modeling between-link dependencies while preserving tractability (Flötteröd and Osorio, 2013; Osorio and Wang, 2012), this independence approximation ensures tractability. By definition:

$$Var[T_i] = E[T_i^2] - E[T_i]^2. \tag{2.13}$$

Equation (2.9) gives the expression for $E[T_i]$. An expression for $E[T_i^2]$ is derived in Section 2.3.4 and is given by:

$$E[T_i^2] = \frac{1}{\hat{\mu_i}^2}\left(\frac{4\rho_i - 2\rho_i^2}{(1-\rho_i)^2} - \frac{2k_i\rho_i^{k_i+1}}{(1-\rho_i^{k_i})(1-\rho_i)} + \frac{2-(k_i+1)(k_i+2)\rho_i^{k_i}}{1-\rho_i^{k_i}}\right). \tag{2.14}$$

$Var[T_i]$ is therefore given by:

$$Var[T_i] = \frac{1}{\hat{\mu_i}^2}\left(\frac{4\rho_i - 2\rho_i^2}{(1-\rho_i)^2} - \frac{2k_i\rho_i^{k_i+1}}{(1-\rho_i^{k_i})(1-\rho_i)} + \frac{2-(k_i+1)(k_i+2)\rho_i^{k_i}}{1-\rho_i^{k_i}}\right) - \left(\frac{\rho_i\left(\frac{1}{1-\rho_i} - (k_i+1)\frac{\rho_i^{k_i}}{1-\rho_i^{k_i+1}}\right)}{\lambda_i\left(1-P(N_i=k_i)\right)}\right)^2 . \quad (2.15)$$

The approximation of the objective function given in Equation (2.5) is a differentiable closed-form expression that depends on three endogenous variables per queue: ρ_i, λ_i and $P(N_i = k_i)$. Appendix A.1 gives the formulation of two auxiliary traffic models used in this chapter to approximate (2.5). That of Appendix A.1.1 is derived in Osorio and Bierlaire (2009b) and is used in this chapter to address a signal control problem for the Lausanne city-center (Section 2.4.2). That of Appendix A.1.2 is a formulation that is more efficient for large-scale problems (Osorio and Chong, 2012). It is used in this chapter to address a signal control problem for the full Lausanne city (Section 2.4.3).

2.3.4 Analytical approximation of $E[T^2]$

We derive the expression for $E[T^2]$, where T denotes the sojourn time at a given queue. We represent an urban road network as a finite capacity queueing network as in Osorio and Bierlaire (2009b). Each lane is modeled as one (or a set of) $M/M/1/k$ queue(s). For an $M/M/1/k$ queue the cumulative distribution function $F(t)$ of the sojourn time

is given by (cf. Gross et al. (1998), pages 587-641):

$$F(t) = \frac{1-\rho}{1-\rho^k} \sum_{n=0}^{k-1} \rho^n \left(1 - \sum_{m=0}^{n} \frac{(\hat{\mu}t)^m e^{-\hat{\mu}t}}{m!} \right), t \geq 0, \tag{2.16}$$

with $\hat{\mu}, \rho$ and λ defined in Section 2.3.3. The probability density function $f(t)$ is obtained as follows:

$$f(t) = \frac{dF(t)}{dt} = -\frac{1-\rho}{1-\rho^k} \sum_{n=0}^{k-1} \rho^n \sum_{m=0}^{n} \frac{\hat{\mu}^m}{m!} \frac{dg(t)}{dt}, \tag{2.17}$$

where $g(t)$ is defined by:

$$g(t) = t^m e^{-\hat{\mu}t}, t \geq 0. \tag{2.18}$$

Since:

$$\frac{dg(t)}{dt} = mt^{m-1}e^{-\hat{\mu}t} - \hat{\mu}t^m e^{-\hat{\mu}t}, \tag{2.19}$$

then:

$$f(t) = \frac{1-\rho}{1-\rho^k} \sum_{n=0}^{k-1} \rho^n \sum_{m=0}^{n} \frac{\hat{\mu}^m}{m!} \left(\hat{\mu}t^m e^{-\hat{\mu}t} - mt^{m-1}e^{-\hat{\mu}t} \right). \tag{2.20}$$

By definition:

$$E[T^2] = \int_0^\infty t^2 f(t)dt = \int_0^\infty \frac{1-\rho}{1-\rho^k} \sum_{n=0}^{k-1} \rho^n \sum_{m=0}^{n} \frac{\hat{\mu}^m}{m!} \left(\hat{\mu} t^{m+2} e^{-\hat{\mu}t} - m t^{m+1} e^{-\hat{\mu}t}\right) dt. \quad (2.21)$$

$$E[T^2] = \frac{1-\rho}{1-\rho^k} \sum_{n=0}^{k-1} \rho^n \sum_{m=0}^{n} \frac{\hat{\mu}^m}{m!} \int_0^\infty \left(\hat{\mu} t^{m+2} e^{-\hat{\mu}t} - m t^{m+1} e^{-\hat{\mu}t}\right) dt. \quad (2.22)$$

According to Gradshteyn and Ryzhik (2007) (pages 247-386):

$$\int_0^\infty t^a e^{-ct^b} dt = \frac{\Gamma(\frac{a+1}{b})}{b c^{(a+1)/b}}, \quad (2.23)$$

where Γ denotes the gamma function defined as $\Gamma(x) = (x-1)!$.

Using the expression of Equation (2.23), we obtain the following two equalities:

$$\int_0^\infty \hat{\mu} t^{m+2} e^{-\hat{\mu}t} dt = \hat{\mu} \frac{\Gamma(m+3)}{\hat{\mu}^{m+3}} = \frac{(m+2)!}{\hat{\mu}^{m+2}} \quad (2.24)$$

$$\int_0^\infty m t^{m+1} e^{-\hat{\mu}t} dt = m \frac{\Gamma(m+2)}{\hat{\mu}^{m+2}} = m \frac{(m+1)!}{\hat{\mu}^{m+2}}. \quad (2.25)$$

Inserting the expressions of Equations (2.24) and (2.25) into (2.22), leads to:

$$\begin{aligned}
E[T^2] &= \frac{1-\rho}{1-\rho^k}\sum_{n=0}^{k-1}\rho^n\sum_{m=0}^{n}\frac{\hat{\mu}^m}{m!}\left(\frac{(m+2)!}{\hat{\mu}^{m+2}} - m\frac{(m+1)!}{\hat{\mu}^{m+2}}\right) \\
&= \frac{1-\rho}{1-\rho^k}\sum_{n=0}^{k-1}\rho^n\sum_{m=0}^{n}\left(\frac{(m+1)(m+2)}{\hat{\mu}^2} - \frac{m(m+1)}{\hat{\mu}^2}\right) \\
&= \frac{1-\rho}{1-\rho^k}\sum_{n=0}^{k-1}\rho^n\sum_{m=0}^{n}\frac{2(m+1)}{\hat{\mu}^2} \\
&= \frac{1-\rho}{1-\rho^k}\sum_{n=0}^{k-1}\rho^n\frac{2}{\hat{\mu}^2}\left(\frac{n(n+1)}{2} + (n+1)\right) \\
&= \frac{1}{\hat{\mu}^2}\frac{1-\rho}{1-\rho^k}\sum_{n=0}^{k-1}(n+1)(n+2)\rho^n. \qquad (2.26)
\end{aligned}$$

The above summation can be further simplified, for $\rho \neq 1$, as follows:

$$\sum_{n=0}^{k-1}(n+1)(n+2)\rho^n = \sum_{n=0}^{k-1}\frac{d^2(\rho^{n+2})}{d\rho^2} = \frac{d^2\left(\sum_{n=0}^{k-1}\rho^{n+2}\right)}{d\rho^2} = \frac{d^2\left(\rho^2\frac{1-\rho^k}{1-\rho}\right)}{d\rho^2}. \qquad (2.27)$$

We first calculate the first derivative with regards to ρ:

$$\frac{d\left(\rho^2\frac{1-\rho^k}{1-\rho}\right)}{d\rho} = \frac{2\rho-(k+2)\rho^{k+1}}{1-\rho} + \frac{\rho^2-\rho^{k+2}}{(1-\rho)^2}. \qquad (2.28)$$

We then take the first derivative of (2.28) with regards to ρ:

$$\begin{aligned}\frac{d\left(\frac{2\rho-(k+2)\rho^{k+1}}{1-\rho}+\frac{\rho^2-\rho^{k+2}}{(1-\rho)^2}\right)}{d\rho} &= \frac{2-(k+1)(k+2)\rho^k}{1-\rho}+\frac{2\rho-(k+2)\rho^{k+1}}{(1-\rho)^2} \\ &\quad +\frac{2\rho-(k+2)\rho^{k+1}}{(1-\rho)^2}+\frac{2(1-\rho)(\rho^2-\rho^{k+2})}{(1-\rho)^4} \\ &= \frac{2(\rho^2-\rho^{k+2})}{(1-\rho)^3}+\frac{4\rho-2(k+2)\rho^{k+1}}{(1-\rho)^2}+\frac{2-(k+1)(k+2)\rho^k}{1-\rho}.\end{aligned} \tag{2.29}$$

Inserting the above expression into (2.26), we obtain:

$$\begin{aligned}E[T^2] &= \frac{1-\rho}{\hat{\mu}^2(1-\rho^k)}\left(\frac{2(\rho^2-\rho^{k+2})}{(1-\rho)^3}+\frac{4\rho-2(k+2)\rho^{k+1}}{(1-\rho)^2}+\frac{2-(k+1)(k+2)\rho^k}{1-\rho}\right) \\ &= \frac{1}{\hat{\mu}^2}\left(\frac{2\rho^2}{(1-\rho)^2}+\frac{4\rho}{1-\rho}-\frac{2k\rho^{k+1}}{(1-\rho^k)(1-\rho)}+\frac{2-(k+1)(k+2)\rho^k}{1-\rho^k}\right) \\ &= \frac{1}{\hat{\mu}^2}\left(\frac{4\rho-2\rho^2}{(1-\rho)^2}-\frac{2k\rho^{k+1}}{(1-\rho^k)(1-\rho)}+\frac{2-(k+1)(k+2)\rho^k}{1-\rho^k}\right).\end{aligned} \tag{2.30}$$

2.4 Case studies

2.4.1 General description

We evaluate the performance of this framework based on a calibrated microscopic traffic simulation model of the Lausanne city center developed by Dumont and Bert (2006). It is calibrated for the Lausanne city road network during evening peak period (17h-18h). It is implemented in Aimsun (TSS, 2011). We address signal control

problems within two networks: 1) the Lausanne city center (Section 2.4.2), 2) the full city network (Section 2.4.3). The Lausanne city center contains 48 roads and 15 intersections, 9 of which are signalized and control the traffic on 30 roads. The full network contains 603 roads and 231 intersections. 17 signalized intersections are controlled by the algorithm. During the peak period, around 12,000 vehicles pass through this area. During the simulated period, congestion increases as time goes by.

We compare the performance of the following SO metamodel approaches:

- the proposed metamodel, m (of Equation (2.3));
- a quadratic polynomial with diagonal second derivative matrix, (i.e. the metamodel consists of ϕ as defined in Equation (2.4)). In this approach, the metamodel consists of only a functional component, there is no physical component.

We evaluate the performance of both metamodel methods by addressing three different signal control problems that vary according to their objective function.

- P1: this is a traditional signal control problem which uses only expectation information in the objective function, which is given by $E[\sum_{\mathscr{L}} T_i(x, z; p)]$.
- P2: this is the reliable signal control problem, with the objective function given by Equation (2.5).
- P3: this signal control problem uses only standard deviation information in the objective function, which is given by $SD[\sum_{\mathscr{L}} T_i(x, z; p)]$.

Problem P2 requires the estimation of the reliability parameter r. Recall that the mean-variance approach considers functions of the form $t_E + rt_V$, where t_E denotes the expected trip travel time and t_V usually denotes the standard deviation of trip travel time.

In order to identify a suitable r value, we resort to travel time and travel time variability valuation studies. The estimates for r of this parameter vary according to, for instance, the traveler population and the trip purpose. In past work, where t_V is defined as the standard deviation of trip travel time, estimates of r have varied between 0.1 (Hollander, 2006) and 2.1 (Batley and Ibáñez, 2009). Black and Towriss (1997) estimate an r value of 0.79 for commuters traveling with a car. More recently, Li et al. (2010b) derived a value of 1.43 for car commuters.

We consider evening peak period traffic, where most trips consist of commuters. Additionally, the simulation model that we use represents only car traffic. Thus, we use the value of 1.43, which was estimated for car commuters by Li et al. (2010b). Additionally, the largest r value found in the literature (value of 2.1) is used to evaluate the sensitivity of our approach to r (Section 2.4.4).

Note that the r estimates derived from these surveys are obtained by using trip travel time as the travel time metric, whereas in this chapter we use total link travel time. Thus, the actual r value derived from an analysis that would consider total link travel time for the evening peak period of Lausanne, may differ from the value of 1.43

that we use.

For all experiments the computational budget is set to 150 runs, i.e., a signal plan with improved performance needs to be identified within 150 simulation runs. Given the stochasticity of the simulation outputs as well as the large-scale problems that we are addressing, these are considered very tight computational budgets.

When evaluating the performance of a given method, we need to account for the fact that the outputs of the simulator are stochastic. Thus, for a given experiment (i.e., a given combination of: metamodel, objective function, network, initial point and computational budget) we run the SO algorithm five times. Each run yields an "optimal" (or proposed) signal plan. Thus a given experiment yields five signal plans. We then compare the performance of the signal plans across experiments. In order to evaluate the performance of a proposed signal plan, 50 simulation replications are run. This yields 50 observations of the expected total link travel time and total link travel time standard deviation. We then plot the empirical cumulative distribution function (cdf) of each of these 2 performance metrics, and compare the cdf's obtained by different methods.

2.4.2 Lausanne city center

The Lausanne city network is represented in Figure 2-2. The city center of interest is delimited by an oval.

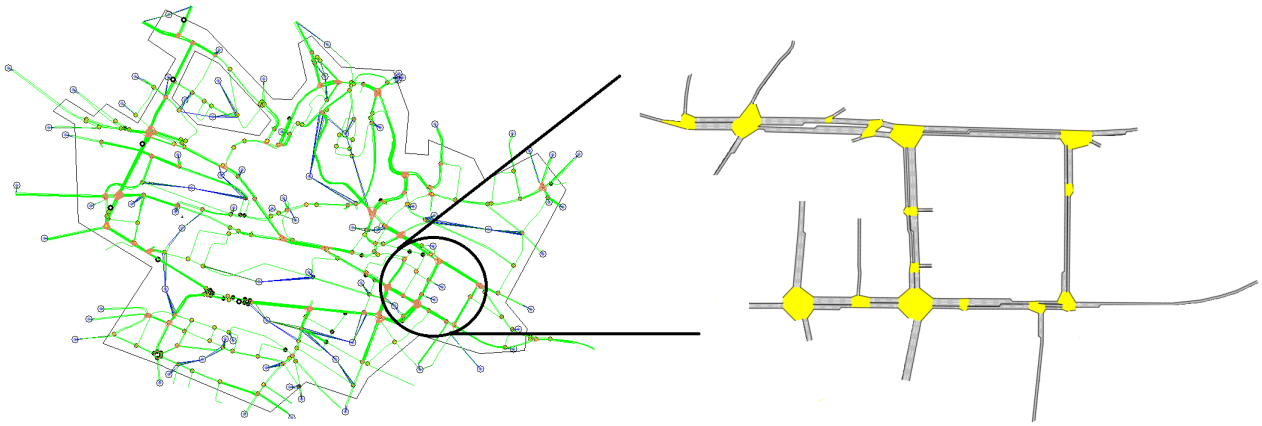


Figure 2-2: Lausanne city network model with city center delimited by a circle (left), city center of interest (right).

A total of 51 signal phases are endogenous. The queueing model of this network consists of 102 queues. The trust region subproblem that is solved at each iteration of the SO algorithm consists of 621 endogenous variables with their corresponding lower bound constraints, 408 nonlinear equality constraints and 171 linear equality constraints.

Figure 2-3 displays 6 plots with the results obtained from all the methods we use. The plots in a given column correspond to a given initial point. The plots of a row correspond to a given performance measure. The upper (resp. lower) row displays the cdf's of the standard deviation (resp. expectation) of total link travel time (within the city center). Each plot displays 7 cdf's: the solid blue cdf corresponds to the cdf of the initial signal plan (denoted x_0), the remaining 6 cdf's correspond to solving a given problem (P1, P2 or P3) with a given metamodel method (m or ϕ). The red (resp. black) cdf's correspond to the signal plans obtained when using m (resp. ϕ).

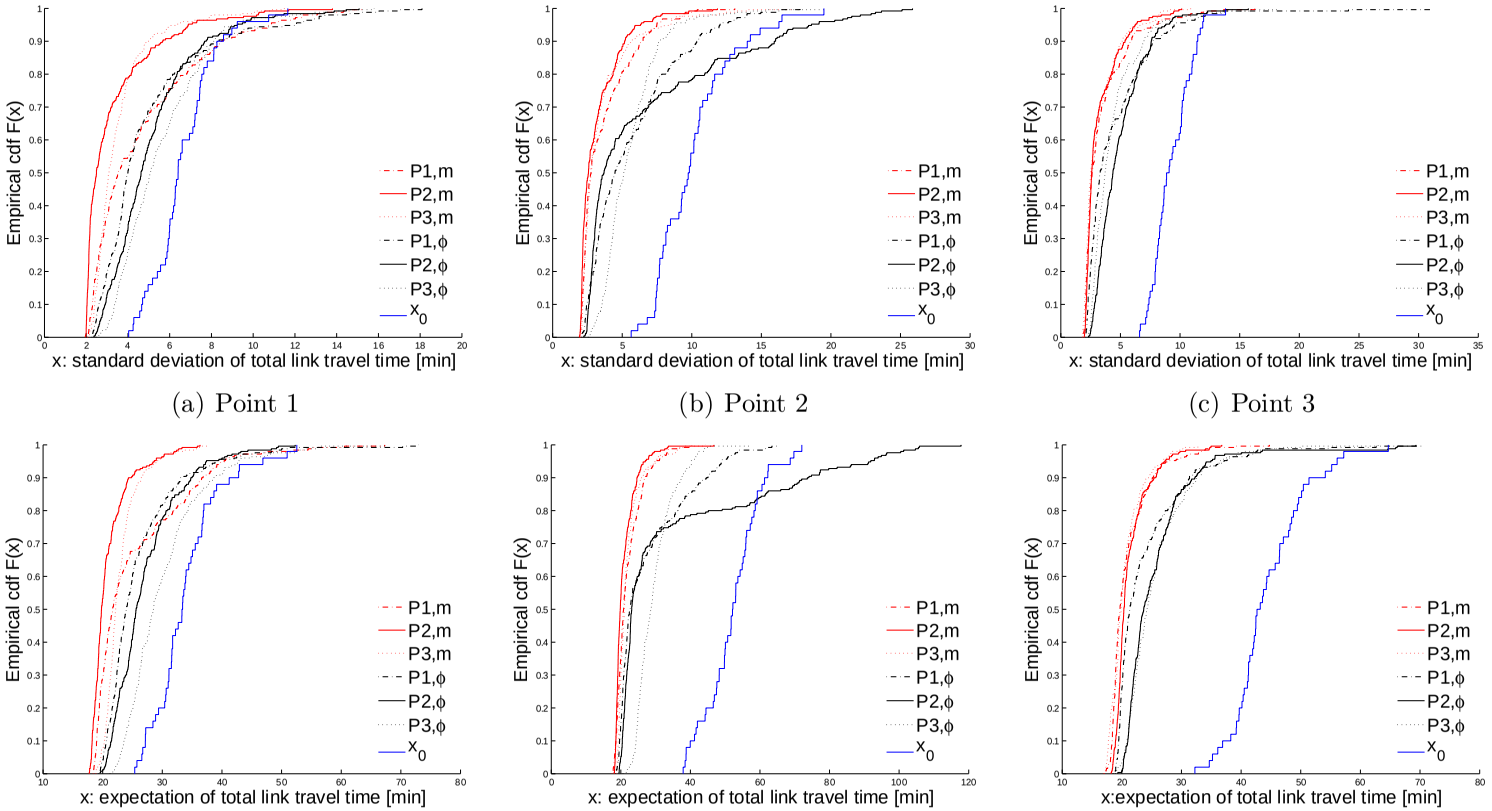
The initial points are uniformly drawn from the feasible space (Equations (2.6) and (2.7)) using the code of Stafford (2006).

Recall that when solving a given problem with a given metamodel, we run the SO algorithm 5 times, yielding 5 signal plans, and then evaluate each of the 5 proposed signal plans by running 50 simulation replications. The cdf's displayed in Figure 2-3 are obtained by aggregating (for a given problem and a given metamodel) the observations from all 5 signal plans, i.e. they consist of 5*50 observations.

For the first initial point (column 1), the signal plans with best performance both in terms of expectation and standard deviation are obtained by solving P2 (i.e., a problem that combines expectation and standard deviation information) and using the proposed metamodel, m. The signal plans derived by using m outperform those derived by the traditional metamodel ϕ regardless of the problem formulation (i.e., for all P1, P2 and P3). Similar conclusions hold for both the second initial point (column 2) and the third initial point (column 3).

All plots of Figure 2-3 indicate that using metamodel m to solve problem P2 leads to signals plans with the lowest average standard deviation, and the lowest variance across-replications. Both contribute to a more reliable and predictable system performance.

Figure 2-3 also indicates that when using ϕ, the best signal plans are obtained by using only expected total travel time (P1), and the performance deteriorates when



(a) Point 1 (b) Point 2 (c) Point 3

(d) Point 1 (e) Point 2 (f) Point 3

Figure 2-3: Performance of the signal control methods when applied to the Lausanne city center. These plots consider various initial points and various problem formulations.

higher-order information is included (P2 and P3). This illustrates the added value of using auxiliary traffic models to approximate complex objectives functions, such as accounting for higher-order distributional information.

When comparing the use of ϕ to address the formulation that includes only standard deviation (P3) with the formulation that includes both expectation and standard deviation (P2), the latter leads to standard deviations that are either similar or better, which is counterintuitive. This may be explained as follows. Firstly, formulation P1 (only expectation information) leads to low standard deviation values, thus the expectation and standard deviation metrics may be correlated. Second, the expectation metric has less variability across replications, thus it will be estimated more accurately with few replications, leading to a better algorithmic performance for tight computational budgets. For these 2 reasons, the formulations that include expectation information (P1 and P2) seem to lead to improved standard deviation. Particularly when considering tight computational budgets.

The cdf's presented so far display the performance aggregated across all links of the city center. Figure 2-4 illustrates the performance at the link level. This figure displays two plots of the city center network. The links of the network are color coded according to their link travel time standard deviation. The colors green, yellow and red correspond, respectively, standard deviations that are lower than 20 seconds, are between 20 and 40 seconds, and are greater than 40 seconds. These standard deviation

estimates are obtained by running 50 replications of a given signal plan. The top network considers the initial plan (that of column 1 of Figure 2-3), the bottom network considers one of the plans proposed by using that initial plan and the metamodel m to solve the reliable signal control problem P2. Figure 2-4 shows that there is an improvement across the entire city center. This illustrates that the proposed approach leads to both improvements when aggregating across links (e.g., total link travel time), as well as systematic improvements at the link level.

2.4.3 Lausanne city

In this section, we address a signal control problem that controls intersections across the entire city of Lausanne. Figure 2-5 displays the road network of the city, Figure 2-6 displays the corresponding network model. We determine the plans for 17 intersections, which are represented as filled rectangles in Figure 2-6.

A total of 99 signal phases are endogenous. The queueing model consists of 902 queues. The trust region subproblem that is solved at each iteration of the SO algorithm consists of 2805 endogenous variables with 1821 nonlinear equality constraints and 902 linear equality constraints. The problem we address in this section is considered a large-scale traffic signal control problem and a complex simulation-based optimization problem.

In order to compare the performance of the methods across various problems,

Figure 2-4: Link based travel time standard deviation for initial plan (top plot) and plan obtained by solving problem P2 with metamodel m (standard deviation estimates are obtained by averaging over 50 replications).

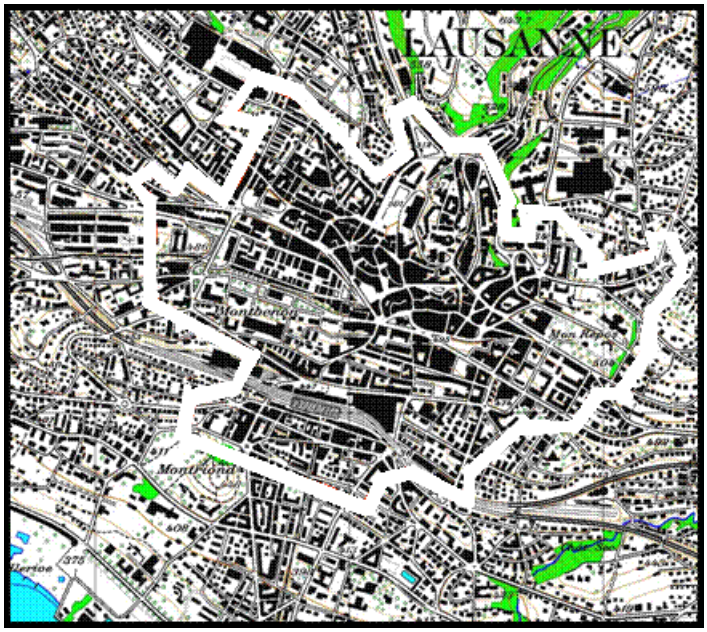


Figure 2-5: Lausanne city road network (adapted from Dumont and Bert (2006)).

Figure 2-6: Lausanne network model with the 17 controlled intersections displayed as grey rectangles.

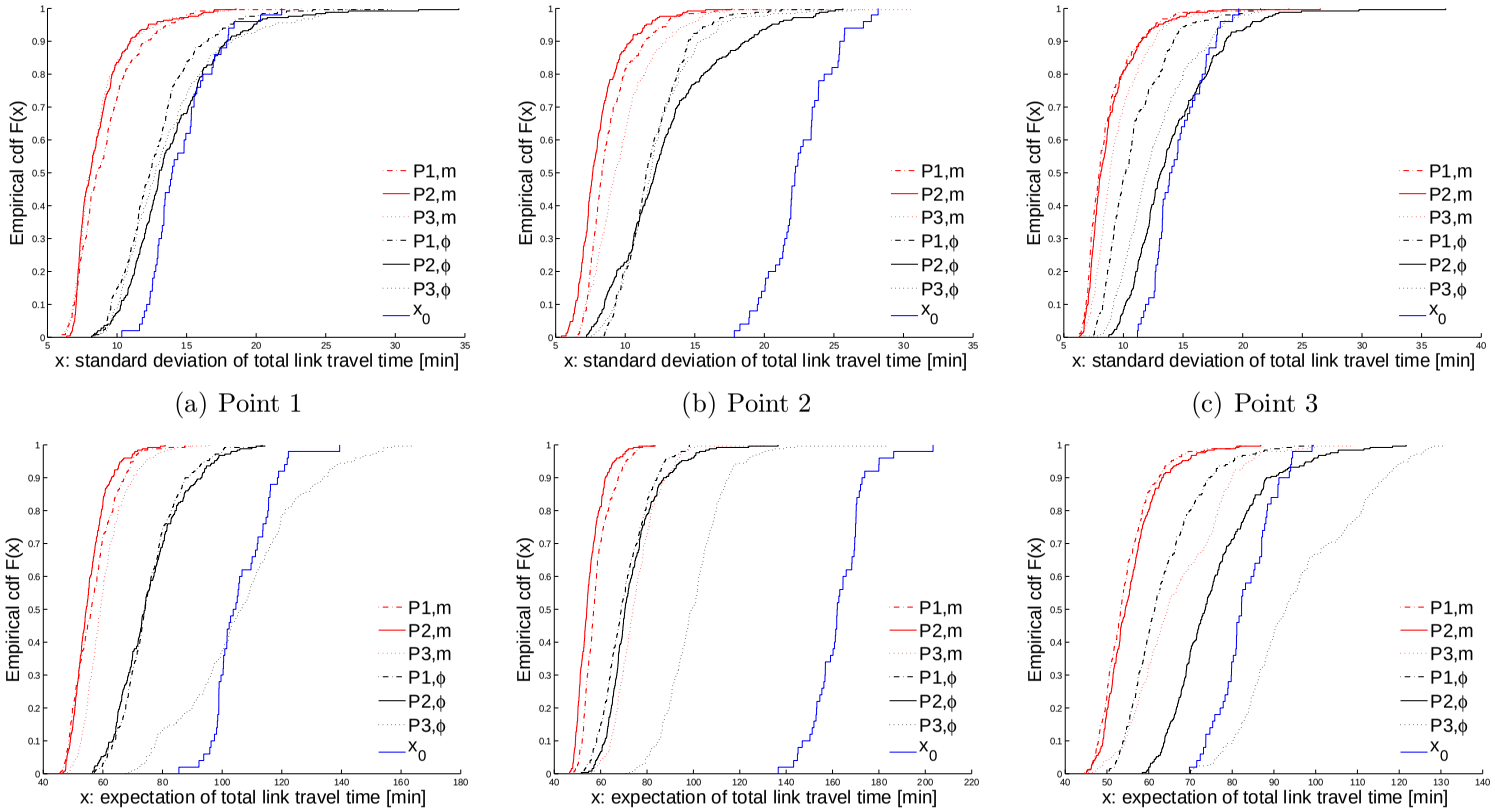
we proceed as for the city center (i.e., Section 2.4.2). Figure 2-7 displays 6 plots: each column corresponds to a given initial point, each row corresponds to a given performance measure. The upper (resp. lower) row displays the cdf's of the standard deviation (resp. expectation) of total link travel time within the full city network. Each cdf aggregates 250 (i.e., 5*50) observations.

For the first initial point (column 1), the signal plans with best performance both in terms of expectation and standard deviation are obtained by solving P2 (i.e., a problem that combines expectation and standard deviation information) and using the proposed metamodel, m. The signal plans derived by using m, solving any of the three problems, outperform those derived by the traditional metamodel ϕ. Similar conclusions hold for initial points 2 (column 2) and 3 (column 3).

The plans obtained by using only standard deviation information (i.e. solving P3) with metamodel m still provide improvement in terms of expected travel time (see row-wise plots) when compared to the initial point, whereas those derived by ϕ fail to do so for initial points 1 and 3.

As for the city center case study, the plots of Figure 2-7 indicate that using metamodel m to solve problem P2 leads to signals plans with low average and variance of the standard deviation. Both indicators enhance the travel time reliability of the network.

Figure 2-8 displays the link-level results for a part of the city network. Each plot



(a) Point 1 (b) Point 2 (c) Point 3

(d) Point 1 (e) Point 2 (f) Point 3

Figure 2-7: Performance of the signal control methods when applied to the full city of Lausanne. These plots consider various initial points and various problem formulations.

displays the link standard deviation (averaged over 50 simulation replications). The top plot considers initial point 2, and the bottom plot considers a signal plan proposed by solving P2 and using the metamodel m, given initial point 2. The colors green, yellow and red correspond, respectively, to values smaller than 20 seconds, from 20 to 40 seconds, and greater than 40 seconds. Just as for the city center, there is a systematic improvement at the link level. This shows that the proposed plan reduces both the total variability as well as the individual link travel time variability within the interval analyzed.

2.4.4 Sensitivity to reliability ratio

In this section, we evaluate the sensitivity of our proposed approach to the value of the reliability ratio parameter r. We choose the highest r value found in the literature, namely 2.1. We address the reliable signal control problem P2 with the proposed metamodel m. We compare the performance of an approach that sets r to 1.43 to one that sets r to 2.1.

We proceed as in Sections 2.4.2 and 2.4.3: we consider an initial point, and run each approach 5 times, deriving 5 signal plans. We then evaluate the performance of each of these signal plans by running 50 simulation replications.

Figure 2-9 displays two plots. The left (resp. right) plot displays the cdf's of the standard deviation (resp. expectation) of total link travel time. Each cdf consists of

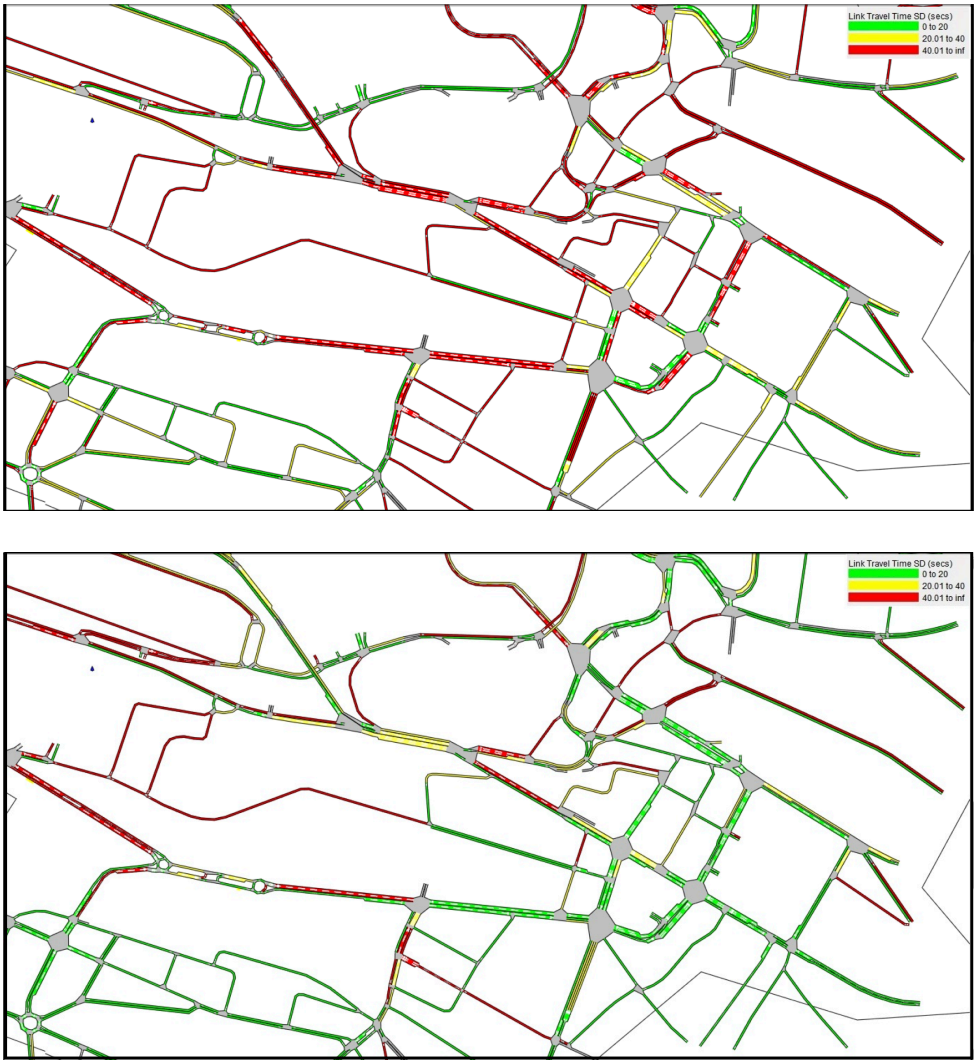


Figure 2-8: Link travel time standard deviation for initial plan (top plot) and plan obtained by solving problem P2 with metamodel m (standard deviation estimates are obtained by averaging over 50 replications).

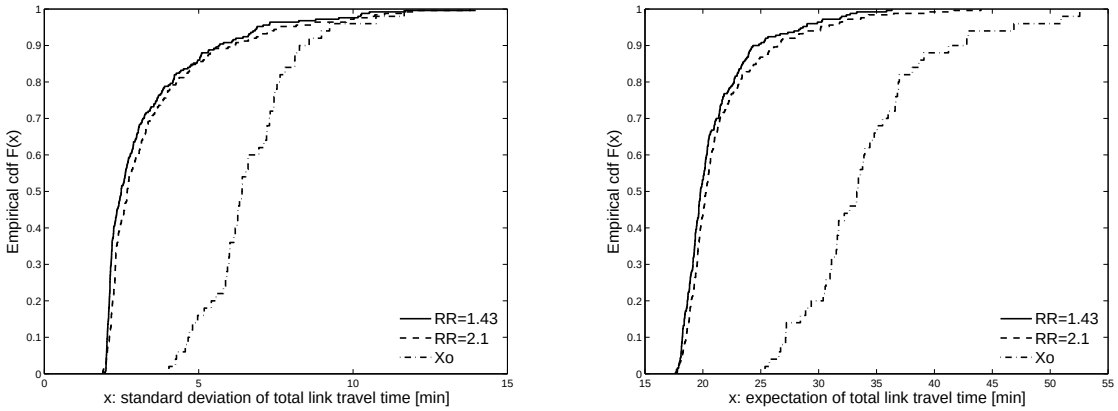


Figure 2-9: Empirical cdf's of the total link travel time standard deviation (left plot) and expected total link travel time (right plot) with different reliability ratio values.

5*50 simulation observations (i.e., 5 signal plans with 50 simulation replications for each signal plan). The cdf of the initial signal plan corresponds to the dash-dotted curve, the cdf for the signal plans derived with $r = 1.43$ (resp. $r = 2.1$) is the solid (resp. dashed) curve. Solving the problem P2 with these two different reliability ratio values leads to signal plans with similar performance. The methodology seems insensitive to such changes in the reliability ratio values. The reason is that average link travel time and link travel time standard deviation is correlated, the curves correspond to expected link travel time and link travel time standard deviation follow similar trends.

2.4.5 Computational Efficiency

Each iteration of the SO algorithm involves two computational intensive tasks: 1) running the simulator; 2) solving the trust region subproblem. In this section, we compare the run time needed for each of these tasks. We solve the subproblem with the Matlab (Mathworks, Inc., 2011) fmincon routine for constrained nonlinear problems, and use its sequential quadratic programming algorithm (Coleman and Li, 1996, 1994).

For a given initial point, we solve problem P2 5 times allowing each time for 150 SO iterations. The computer used for calculation has a processor of Intel Core i7, 3.50 Ghz and RAM of 8GB. Figure 2-10 displays the cdf of all 5*150 computational run time observations. The left (resp. right) plot displays the run times for the Lausanne city center (resp. full Lausanne city). The solid cdf curve displays the run time needed for the convergence of the trust region subproblem, whereas the dashed cdf curve displays the run time for one simulation replication. The simulation run time is relatively constant across iterations, with run times of the order of 30 seconds, and not exceeding 60 seconds. The trust region subproblem is solved quicker than a single simulation run in the city center case study. For the full city case study, it can be of the order of several minutes (i.e., equivalent to several simulation replications). This illustrates the computational efficiency of the overall SO framework at each iteration.

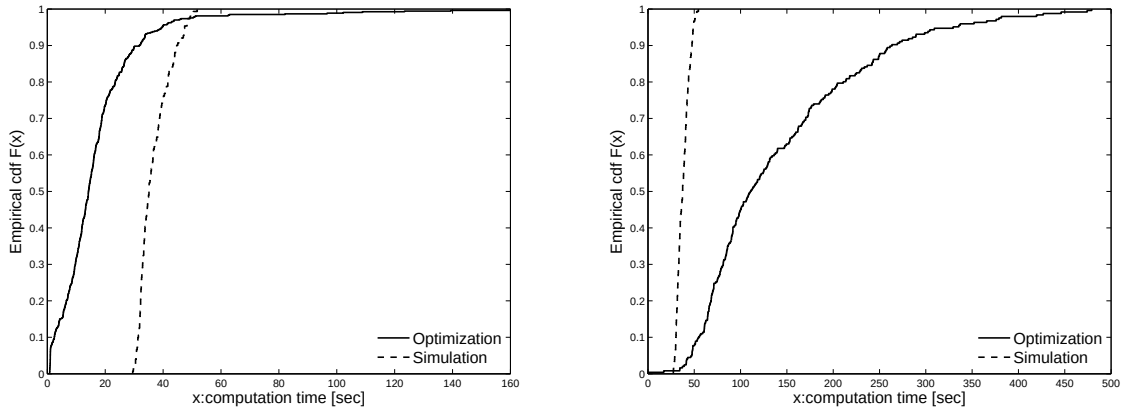


Figure 2-10: Computational run time for Lausanne city center (left) and full Lausanne city (right).

2.5 Conclusion

This chapter presents a method to address a reliable signal control problem by using higher-order distributional information derived from a stochastic simulator. The objective function is a linear combination of the expectation and the standard deviation of total link travel time. Distributional travel time estimates are derived from a detailed stochastic microscopic urban traffic simulator. They are combined with analytical approximations, which are obtained from differentiable probabilistic macroscopic traffic models. A metamodel simulation-based optimization (SO) algorithm is used.

The SO approach used is compared to a traditional SO approach. Three different signal control formulations are considered. Experiments on the Lausanne city center network and the full city network are carried out. The SO methods are evaluated

within tight computational budgets, where the simulator can only be evaluated a total of 150 times.

The use of the proposed method to solve a reliable signal control problem leads to signal plans with the lowest expected total link travel time and the lowest standard deviation of total link travel time. These signal plans also have the lowest across-replication variability of the travel time standard deviation. The proposed approach systematically outperforms the traditional approach. It leads to aggregate improvements (total link metrics), as well as link-level improvements. The proposed method is not sensitive to the changes in reliability ratio values.

The proposed method enables the use of highly detailed distributional information provided by these stochastic simulators to inform the design and operations of urban transportation networks. Such an approach can be used to efficiently address other reliable and robust formulations of traditional transportation problems.

Chapter 3

Analytical Approximation of Trip Travel Time Distribution and its Application in Reliable Signal Control Problem

3.1 Introduction

In Chapter 2, we incorporate the second-order link travel time distributional information in a signal control problem, and successfully reduced the travel time variability. Given the difficulty of analytically modeling dependency between link travel times,

we assumed that all links have independent travel times. This assumption might be true when the network is not congested. However, congestion has become a global phenomenon, affecting most urban areas in the world, which not only affects the local links but also propagates through adjacent links, and thus affects a larger area in the network. Previous researches have addressed that providing an analytical and tractable approximation of the distribution of the main network performance measures (e.g.: travel time) is a major challenge (see for instance, Osorio and Flötteröd (2012); Peterson et al. (1995)), and is often achieved by simplifying, or even omitting, spatial-temporal dependencies. When traffic congestion propagates both temporally and spatially, Rakha et al. (2006) show that this independence assumption underestimates path travel time variance significantly for freeway or signalized arterial road using field link travel time data.

To overcome this limitation and incorporate path travel time variability information in a more realistic way, in this chapter, we derive an analytical tractable approximation of trip travel time variability from link travel time distribution. The main challenge is to incorporate spatial dependencies between links in a tractable manner. To achieve this, we derive a tractable extension of Little's law for finite capacity Markovian queueing networks. In this approach, given the topology of any road network, each lane in the road network is modeled as one (or a set of) queues. We assume that travel time of non-adjacent queues are independent, spatial dependencies are explicitly considered

for any sets of adjacent queues. We validate this analytical approximation of path travel time variance in a general queueing network. This expression is then used to address an urban traffic signal control problem.

Section 3.2 presents a literature review. Section 3.3 formulates the proposed queueing model. Section 3.4 derives an analytical expression for the first- and second-order moments of path and trip sojourn time. Section 3.5 validates the proposed method, and compares its approximations to those obtained via simulation and to those obtained by other approximate analytical methods. Section 3.6 uses the proposed method to address an analytical urban traffic signal control problem. Section 3.7 uses the method to address a simulation-based traffic signal control problem. Section 3.8 presents the main conclusions.

3.2 Literature Review

Accurate path/trip travel time estimation is a challenging topic in the field of transportation. Empirical studies have indicated the importance of path travel time variability in departure-time, mode and route choices (Xing and Zhou, 2011).

Most work that approximates path travel time metrics has assumed independence across links (Noland and Polak, 2002). He et al. (2002) illustrates the inadequacy of the independence approximation through a simulation study of a congested corridor. In particular, they observe that traditionally used distributions with closed-form ex-

pressions do not provide a suitable fit for the path/trip travel time distribution. The most popular approach to approximate higher-order path travel time metrics is the use of link travel time distributional metrics.

Several data-driven approaches that use vehicle probe data to infer both link and path travel time metrics have been proposed. Xing and Zhou (2011) propose a sampling-based algorithm in order to taking spatial dependencies among links into consideration using available historical travel time data from traffic monitoring systems. They take path travel time of several days from the historical traffic database and use them to calculate sample mean and variance directly. Charle et al. (2010) use the link travel time information from a historical dataset, for a specific number of observation days, each day is divided into several time intervals, average link travel time over each time interval is calculated. The path travel time variability is derived using these historical link travel time observations. To simplified the problem, they propose clustering algorithm to build artificial link which combines successive links that have correlated travel time fluctuations. The correlation of travel time fluctuations between any two subsequent links is calculated using historical link travel time observations. Westgate (2013) uses Global Positioning System data (GPS) vehicle data particularly for ambulance, they proposed Whole Trip(WT) method to predict trip travel time distribution along an arbitrary route in a road network. The travel time of each trip is modeled with a lognormal distribution conditional on the path the ambulance trav-

elled. Mean and variance of trip travel time can depend on time, weather and other explanatory variables, then they use a Bayesian formulation to estimate the parameters of the WT model from the observations. For a recent review of the data-driven approaches, see Zheng and Van Zuylen (2013).

Besides the data-driven approaches, analytical approximations based on the use of link travel time data are used (Rakha et al., 2006; Fu and Rilett, 1998). In order to ensure tractability, most analytical methods assume between-link independence (Noland and Polak, 2002; He et al., 2002). An extensive recent review of trip travel time estimation methods is presented by Vlahogianni et al. (2014). We focus here on methods that approximate higher-order moments (i.e., go beyond first-order moments) or full distributions of trip or path travel times. Most methods have focused on highway networks. The analysis for urban networks is more intricate due to more complex dynamic demand-supply interactions (e.g., due to signalized intersections, short links, high-dimensional routing alternatives). This independence assumption tends to underestimate the path travel time variance (Rakha et al., 2006). He et al. (2002) also illustrate the inadequacy of the independence approximation through a simulation study of a congested corridor. Few studies have proposed analytical and tractable approaches while also accounting for spatial dependencies (Chen et al., 2012a; Xing and Zhou, 2011).

In this chapter, we model an urban road network as a network of finite (space)

capacity queues. In queueing theory, the time in a queueing system (e.g., total time, delay time) is referred to as the sojourn time. We propose an analytical and tractable description of the between-queue interactions, and derive expressions for the first- and second-order moments of path sojourn times and trip sojourn times. These expressions are based on an extension of Little's law.

In queueing theory, Little's law (Little, 1961a) states that for a given queueing system (e.g., a single queue or a queueing network) the expected number of jobs (e.g., vehicles) in the system, $E[L]$, and the expected sojourn time in the system, $E[W]$, are related as follows:

$$E[L] = \lambda E[W], \tag{3.1}$$

where λ represents the arrival rate to the system. Little's law is a simple relationship that is valid for a general class of queueing systems: from single queues to networks of queues, from single-class to multi-class systems, for any type of arrival and service processes. Hence, it is considered a fundamental relationship in queueing theory, and has been extensively used in a variety of application fields.

Numerous extensions of Little's law have been proposed and this continues to be an active field of research (Wolff and Yao, 2013; Whitt, 2012). A more general law known as $H = \lambda G$ (Brumelle, 1972) relates the arrival rate λ to a more general time-average metric H and an associated customer-average metric G. Little's law can be seen as a

special case of $H = \lambda G$. Laws that relate the distributions of L and W have also been proposed (Bertsimas and Nakazato, 1995; Keilson and Servi, 1988; Haji and Newell, 1971). Since they relate full distributions, they lack generality.

Past work has also focused on the formulation of higher-order Little's laws that relate higher-order moments (i.e., beyond first-order moments) of L and W (or similarly of H and G). Expressions exist for single queues with general arrival and service distributions (Brumelle, 1972; Marshall and Wolff, 1971), as well as for product-form queueing networks (McKenna, 1989; Heffes, 1982).

Extensions of Little's law have been derived mostly for a single infinite capacity queue, or for a network of infinite capacity queues where there is no overtaking. No overtaking means that the first-in-first-out (FIFO) principle holds at the network level. This is a strong assumption which may not hold for simple networks such as multi-server tandem (i.e., series) queueing networks with stochastic service times.

Finite capacity queueing networks (FCQNs) have received less attention than their infinite capacity counterparts, this is arguably due to the analytical complexity involved in the analysis of FCQNs. The latter can accurately mimic the limited space capacity in urban networks, and hence describe the spillback effects in congested urban traffic, where the queue of vehicles on a road spills back to its upstream roads. In finite capacity queueing theory, spillback is referred to as blocking. Spillback is at the origin of complex spatial between-road dependencies. Providing an analytical, let

alone tractable, description of this dependency is intricate.

In this chapter, the proposed method focuses on finite capacity queueing networks, with single server Marokovian queues. We propose an analytical and tractable approximate expression for the second-order moments of L and W. We then provide an analytical approximation of path and trip sojourn time expectation and variance. This expression is then used to address an urban traffic signal control problem. The proposed model provides a simple, stationary and highly-tractable description of interrupted vehicular traffic. It goes beyond existing models by providing a more detailed description of between-queue interactions. As is shown in Section 3.7, it can be used to efficiently address a variety of SO problems, where a detailed description of between-link dependencies is needed.

3.3 Finite capacity queueing network model

We consider a general topology finite capacity queueing network (FCQN) with single server queues. In the urban traffic case studies of Sections 3.6 and 3.7, we represent a road network as a queueing network, where each road is represented by one or multiple single server finite capacity queues. Thus, the models presented in Sections 3.3.1 and 3.3.2 consider single server finite capacity queueing networks.

3.3.1 State-independent queueing network model

This section presents an FCQN model, hereafter referred to as the "state-independent model". Section 3.3.2 then describes how we build upon this state-independent model in order to formulate an FCQN model that provides a more detailed description of between-queue dependencies.

The state-independent model is derived from Osorio and Bierlaire (2009a), which is formulated for multi-server queues. The equivalent formulation for single-server queues is derived in Chapter 4 of Osorio (2010). This chapter focuses on this single-server FCQN model formulation. This model is formulated as a differentiable and tractable system of nonlinear equations. Given its tractability, this model has been used to enhance the computational efficiency of simulation-based optmization algorithms for various urban transportation problems (Osorio and Bierlaire, 2013; Osorio and Chong, 2013; Osorio and Nanduri, 2013).

Here, we briefly present its formulation. For a given queue i, we use the following

notation.

γ_i	external arrival rate;
$\hat{\lambda}_i$	effective arrival rate;
μ_i	service rate;
$\hat{\mu}_i$	effective service rate;
ρ_i	traffic intensity;
k_i	space capacity, i.e., upper bound of the queue length;
N_i	number of vehicles in queue i;
p_{ij}	transition probability from queue i to queue j;
$\mathcal{DS}_i$	set of downstream queues of queue i;
$\mathcal{Q}$	set of queues.

We consider a network of finite capacity queues. For each queue, external arrivals arise following a Poisson process. Upon arrival to a queue, a job (e.g., a vehicle) waits in the physical queue if there are other jobs already undergoing or waiting for service. Jobs are processed in FIFO manner. Service times are independent and identically distributed exponential random variables. Upon service completion, a job in queue i transitions to queue j with probability p_{ij}. If upon service completion queue j is full and hence cannot recieve new jobs, the job at queue i remains at the server of queue i. It is said to be blocked, and is also blocking the use of the underlying server.

This blocking mechanism, known as blocking-after-service, mimics spillback effects in vehicular traffic. The job at queue i is unblocked once there is space available at queue j. Unblocking is also carried out in FIFO manner (i.e., first blocked, first unblocked). In FCQNs, the actual time a job occupies a server is composed of a traditional service time and potentially a blocked time. This actual time is known as the *effective* service time.

The main challenge in the formulation of an FCQN model is the analytical description of the blocking and unblocking mechanisms. These induce intricate dependencies between adjacent queues. Additionally, this chapter focuses on the formulation of tractable (i.e., computationally efficient) models, which can be efficiently used for optimization and more specifically for simulation-based optimization. The formulation of FCQN models that describe between-queue dependencies in an analytical and tractable manner is a challenge.

The state-independent model is formulated as follows.

$$\hat{\lambda}_i = \gamma_i(1 - P(N_i = k_i)) + \sum_{j \in \mathcal{Q}} p_{ji}\hat{\lambda}_j \tag{3.2a}$$

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + \left(\sum_{j \in \mathcal{Q}} p_{ij} P(N_j = k_j)\right)\left(\sum_{j \in \mathcal{DS}_i} \frac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j}\right) \tag{3.2b}$$

$$P(N_i = k_i) = \frac{1 - \rho_i}{1 - \rho_i^{k_i+1}}\rho_i^{k_i} \tag{3.2c}$$

$$\rho_i = \frac{\hat{\lambda}_i}{\hat{\mu}_i(1 - P(N_i = k_i))}. \tag{3.2d}$$

Equation (3.2a) is a flow conservation equation that relates flow at a given queue i to flow that arises from either external arrivals or from upstream queues. Equation (3.2b) yields the expected effective service time (which is denoted $1/\hat{\mu}_i$), i.e., the expected time a job occupies a server, this accounts for both an expected service time (represented by the term $1/\mu_i$) and an expected blocked time (represented by the second term on the right-hand side of the equation). Equation (3.2c) defines the probability that a queue is full, this is also known as the blocking probability in queueing theory or the spillback probability in vehicular traffic. This expression is derived from finite capacity queueing theory (Bocharov et al., 2004). The queue-length distribution of an isolated M/M/1/k queue is given by:

$$P(N = n) = \frac{(1 - \rho)\rho^n}{1 - \rho^{k+1}}. \tag{3.3}$$

Equation (3.2c) assumes that the functional form of the mariginal queue-length distribution of queue i is that of an isolated M/M/1/k queue. Equation (3.2d) defines the traffic intensity of a queue (denoted ρ_i), which is the ratio of expected demand to expected supply. In the System of Equations 3.2, the exogenous parameters are the external arrival rates γ (which in urban traffic can be obtained from network demand estimates such as an origin-destination matrix), the service rates μ (e.g., lane flow capacities), the space capacities k and the transition probabilities $\{p_{ij}\}$ (e.g., routing or turning probabilities). All other variables are endogenous. The system of equations is

solved simultaneously for all queues. One of the main outputs is the traffic intensities ρ that account for blocking (i.e., spillbacks). Given ρ the queue-length distribution of each queue is approximated via Equation (3.3). A variety of queue performance metrics can be derived based on this marginal distribution (e.g., expected queue-length, expected delay).

In order to approximate path or trip metrics a more accurate description of between-queue dependencies is needed. Section 3.3.2 presents an FCQN model that builds upon the state-independent model while describing these between-queue dependencies in greater detail.

3.3.2 State-dependent queueing network model

The purpose of the state-dependent model is to provide a more detailed description of between-queue interactions. The state-dependent model describes these interactions through the use of state-dependent rates.

Queues interact through the transmission of jobs across nodes. In the case of urban traffic these transmissions represent vehicles turning from one road to another or vehicles changing lanes. These across-node interactions are mainly goverened by:

(i) the downstream traffic conditions of the upstream queues,

(ii) the upstream traffic conditions of the downstream queues.

In a queueing model, the upstream traffic conditions are described by the arrivals, while

the downstream traffic conditions are described by the service completions. Hence, for a given node, the proposed model considers:

(i) state-dependent effective service rates of its upstream queues,

(ii) state-dependent effective arrival rates of its downstream queues.

The remaining rates (i.e., the effective arrival rates of the upstream queues and the effective service rates of the downstream queues) are considered state-independent.

We introduce the following notation.

$\mathcal{D}_m$	set of downstream queues of node m;
$\mathcal{U}_m$	set of upstream queues of node m;
$\mathcal{US}_i$	set of upstream queues of queue i;
$\mathscr{S}(m)$	state space of node m;
S_m	random variable that describes the state of node m;
s_m	a given realization of S_m;
$\hat{\lambda}_{i,s_m}$	effective arrival rate for queue i and node state s_m;
$\hat{\mu}_{i,s_m}$	effective service rate for queue i and node state s_m.

Node m consists of a set of upstream queues $\mathcal{U}_m$ and a set of downstream queues $\mathcal{D}_m$. The state of node m indicates for each downstream queue i, whether the queue is full or not, i.e., whether $N_i = k_i$ or $N_i < k_i$. Consider the following binary random

variable:

$$A_i = \begin{cases} 1, & \text{if } N_i = k_i, \\ 0 & \text{if } N_i < k_i. \end{cases} \tag{3.4}$$

Indexing the set of downstream queues by $i_1, i_2, \ldots, i_m$, then the state of node m is the random tuple: $S_m = (A_{i_1}, A_{i_2}, \ldots, A_{i_m})$. The set of all states of node m, known as the state space, is then defined as:

$$\mathscr{S}(m) = \left\{ s_m = (a_{i_1}, a_{i_2}, \ldots, a_{i_m}) \in \{0,1\}^{i_m} \right\}. \tag{3.5}$$

The state-dependent model is formulated as follows.

$$\begin{cases} \hat{\lambda}_{i,s_m} = \hat{\lambda}_i & \forall i \in \mathcal{U}_m \quad (3.6a) \\ \dfrac{1}{\hat{\mu}_{i,s_m}} = \dfrac{1}{\mu_i} + \displaystyle\sum_{j \in \mathcal{DS}_i} \mathbb{1}(s_m, j) \frac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j} & \forall i \in \mathcal{U}_m \quad (3.6b) \\ \hat{\lambda}_{i,s_m} = (1 - \mathbb{1}(s_m, i)) \cdot \left(\displaystyle\sum_{j \in \mathcal{US}_i} p_{ji} \hat{\mu}_{j,s_m} \left(1 - P(N_j = 0 | S_m = s_m)\right) + \gamma_i \right) + \ldots & \\ \quad \mathbb{1}(s_m, i) \cdot \hat{\mu}_i \left(1 - \displaystyle\prod_{j \in \mathcal{US}_i} P(N_j = 0 | S_m = s_m) \right) & \forall i \in \mathcal{U}_m \quad (3.6c) \\ \hat{\mu}_{i,s_m} = \hat{\mu}_i & \forall i \in \mathcal{U}_m \quad (3.6d) \end{cases}$$

The indicator function $\mathbb{1}(s_m, i)$ is defined as:

$$\mathbb{1}(s_m, i) = \begin{cases} 1, & \text{if in state } s_m\text{: } a_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

This indicator function describes whether downstream queue i is full in state s_m.

Equations (3.6a)-(3.6b) describe the rates of the upstream queues. Equation (3.6a) assumes that the arrival rate of an upstream queue i is state-independent. This arrival rate equals $\hat{\lambda}_i$, which is defined by the state-independent Equation (3.2a). Equation (3.6b) gives a state-dependent expression for the effective service rate. In this equation the terms $\hat{\lambda}_j$, $\hat{\lambda}_i$ and $\hat{\mu}_j$ are state-independent rates given, respectively, by Equations (3.2a), (3.2a) and (3.2b). Equation (3.6b) states that the (state-dependent) effective service time of upstream queue i (represented by $1/\hat{\mu}_{i,s_m}$) is composed of an exogenous expected service time (represented by $1/\mu_i$) and an expected blocked time (represented by the second term on the right hand side of the equation). Overall this equation is similar to Equation (3.2b). The main difference is in the calculation of the expected blocked time. In state s_m, we know which downstream queues of queue i are full, i.e., we know which queues can actually block jobs at queue i. Hence, the expected blocked time is a function of the effective service time of downstream queues that are indeed full. A detailed derivation of Equation (3.6b) is presented in Appendix B.1.

Equations (3.6c)-(3.6d) describe the rates of the downstream queues. Equation (3.6c)

gives a state-dependent expression for the arrival rate. In this equation the terms $\hat{\mu}_i$ and $\hat{\mu}_{j,s_m}$ are given by Equations (3.2b) and (3.6b), respectively. The approximation for the term $P(N_j = 0|S_m = s_m)$ is derived below and given by Equation (3.10). The right-hand side of Equation (3.6c) consists of a summation over 2 lines. The first line considers the case where downstream queue i is not full. In this case, the effective arrival rate to queue i is determined by the sum of the flow arising from upstream queues and the flow from external arrivals. The flow from upstream queue j to queue i is given by the departure rate from upstream queue j (represented by the term $\hat{\mu}_{j,s_m}$) and the probability that there are jobs in queue j (represented by the term $(1 - P(N_j = 0|S_m = s_m))$). The external arrivals arise with a rate of γ_i. The second line considers the case where downstream queue i is full. In this case, the effective arrival rate to queue i is determined by the departure rate from queue i (term $\hat{\mu}_i$) and by the probability that there are jobs upstream that would like to proceed to queue i (term $(1 - \prod_{j \in \mathcal{US}_i} P(N_j = 0|S_m = s_m))$).

Equation (3.6d) assumes that the effective service rate of a downstream queue i is state-independent, it equals $\hat{\mu}_i$ which is defined by the state-independent Equation (3.2b).

The System of Equations (3.6) defines the rates of all queues. Given these rates, we can define the state-dependent traffic intensity of a queue i adjacent to node m:

$$\begin{cases} \rho_{i,s_m} = \dfrac{\hat{\lambda}_i}{\hat{\mu}_{i,s_m}(1 - P(N_i = k_i))} & \forall i \in \mathcal{U}_m \quad (3.8a) \\ \rho_{i,s_m} = \dfrac{\hat{\lambda}_{i,s_m}}{\hat{\mu}_i} & \forall i \in \mathcal{D}_m. \quad (3.8b) \end{cases}$$

This state-dependent traffic intensity is an extension of the state-independent traffic intensity of Equation (3.2d). They differ in that: (i) for upstream queues the state-dependent effective service rate is used, rather than the state-independent rate, and (ii) for downstream queues the state-dependent effective arrival rate is used, rather than the state-independent rate.

This state-dependent traffic intensity allows us to define conditional queue-length distributions for a given queue. For an upstream queue i of node m, the conditional distribution is given by:

$$P(N_i = n | S_m = s_m) = \frac{(1 - \rho_{i,s_m})\rho_{i,s_m}^n}{1 - \rho_{i,s_m}^{k_i+1}}, \quad (3.9)$$

where ρ_{i,s_m} is defined by Equation (3.8). This expression assumes that conditional on the node state, the functional form of the queue-length distribution of queue i is that of an isolated M/M/1/k_i queue. This expression is used to evaluate the conditional

probability of an upstream queue i being empty by setting n to zero:

$$P(N_i = 0|S_m = s_m) = \frac{1 - \rho_{i,s_m}}{1 - \rho_{i,s_m}^{k_i+1}}. \tag{3.10}$$

Equation (3.10) is used in the expression of the state-dependent arrival rate defined by Equation (3.6c).

For a downstream queue i of node m, the state s_m indicates whether the queue i is full or not. Hence, we have the two following conditional distributions:

$$P(N_i = k_i|S_m = s_m) = \begin{cases} 1, & \text{if } \mathbb{1}(s_m, i) = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

For $n < k_i$:

$$P(N_i = n|S_m = s_m) = \begin{cases} 0, & \text{if } \mathbb{1}(s_m, i) = 1, \\ \dfrac{(1 - \rho_{i,s_m})\rho_{i,s_m}^n}{1 - \rho_{i,s_m}^{k_i}}, & \text{otherwise.} \end{cases} \tag{3.12}$$

Equation (3.11) expresses that if in node state s_m downstream queue i is full, then the conditional probability $P(N_i = k_i|S_m = s_m)$ is equal to 1; otherwise it is equal to 0. In other words, given the node state s_m, we have full certainty about whether or not the downstream queue i is full. Equation (3.12) expresses that if in node state s_m downstream queue i is full, then the conditional probability $P(N_i = n|S_m = s_m)$ is

equal to 0; otherwise it is equal to a conditional distribution. The latter is assumed to have the same functional form than the marginal queue-length distribution of an isolated M/M/1/$k-1$ queue (Equation (3.3)).

In summary, the proposed model consists of Equations (3.2), (3.6), (3.8) and (3.10). These equations are solved simultaneously, and can be used to derive first-order and second-order moments of network performance measures. In particular in the following section, we show how they are used to derive an approximate first- and second-order Little's law for finite capacity networks.

3.4 First- and second-order sojourn time moments

Section 3.4.1 considers a path within a network, and derives an analytical approximation of the first- and second-order moments of the path sojourn time. Section 3.4.2 considers a network with multiple paths, and derives an analytical approximation of the first- and second-order moments of the trip sojourn time.

3.4.1 First- and second-order moments of the path sojourn time

In this section we present the analytical approximation of the first- and second-order moments of the path sojourn times. We introduce the following notation.

$\tilde{W}_p$	sojourn time of path p;
W_i	sojourn time of queue i;
$\mathcal{Q}_p$	set of queues in path p.

By definition:

$$\tilde{W}_p = \sum_{i \in \mathcal{Q}_p} W_i. \tag{3.13}$$

Thus, the first-order moment is given by:

$$E[\tilde{W}_p] = \sum_{i \in \mathcal{Q}_p} E[W_i]. \tag{3.14}$$

Similarly, the second-order moment is given by:

$$
\begin{aligned}
VAR[\tilde{W}_p] &= VAR[\sum_{i \in \mathcal{Q}_p} W_i] && (3.15)\\
&= \sum_{i \in \mathcal{Q}_p} VAR[W_i] + \sum_{(i,j) \in \mathcal{Q}_p^2, i \neq j} COV[W_i, W_j] && (3.16)\\
&= \sum_{i \in \mathcal{Q}_p} (E[W_i^2] - E[W_i]^2) + \sum_{(i,j) \in \mathcal{Q}_p^2, i \neq j} (E[W_i W_j] - E[W_i]E[W_j]) && (3.17)
\end{aligned}
$$

Thus, in order to approximate $E[\tilde{W}_p]$ (Equation (3.14)) and $VAR[\tilde{W}_p]$ (Equation (3.17)), we need to approximate the following three types of terms: $E[W_i]$, $E[W_i^2]$ and $E[W_i W_j]$. We present their approximation in what follows.

Approximation of $E[W_i]$

We can apply Little's law and obtain:

$$E[W_i] = E[N_i]/\hat{\lambda}_i, \tag{3.18}$$

where $\hat{\lambda}_i$ is given by Equation (3.2a).

For queue i with traffic intensity ρ_i and capacity k_i, $E[N_i]$ is given by:

$$E[N_i] = \rho_i \left(\frac{1}{1-\rho_i} - (k_i+1)\frac{\rho_i^{k_i}}{1-\rho^{k_i+1}} \right), \tag{3.19}$$

where ρ_i is given by Equation (3.2d). This closed-form expression of Equation (3.19)

is derived in Appendix A of Osorio and Chong (2013).

Approximation of $E[W_i^2]$

For a given queue i, the following relationship holds for an isolated infinite capacity M/G/m queue (see Equation (12) of Marshall and Wolff (1971)):

$$E[N_i(N_i-1)\dots(N_i-r+1)] = (\hat{\lambda}_i)^r E[W_i^r], \quad r \in \mathbb{N}^*. \tag{3.20}$$

We use Equation (3.20) to approximate the higher-order moments of the sojourn time for a finite capacity queue within a general topology queueing network.

Note that Equation (3.20) uses the effective arrival rate to queue i, $\hat{\lambda}_i$, rather than the total arrival rate (which is given by $\hat{\lambda}_i/(1-P(N_i=k_i))$). In the case of an infinite capacity queue, the total arrival rate is equivalent to the effective arrival rate. In networks that contain finite capacity queues, there may be losses. For these networks Little's law (and its extensions) holds for the effective arrival rate. For a more detailed description of how to apply Little's law to finite capacity queues, we refer the reader to Tijms (2003) (pages 52-53).

Equation (3.20) for $r = 2$ yields:

$$E[N_i(N_i-1)] = (\hat{\lambda}_i)^2 E[W_i^2], \tag{3.21}$$

which is equivalent to:

$$E[W_i^2] = \frac{E[N_i^2] - E[N_i]}{(\hat{\lambda}_i)^2}, \tag{3.22}$$

where $E[N_i]$ is given by Equation (3.19) and $\hat{\lambda}_i$ is given by Equation (3.2a). For queue i with traffic intensity ρ_i (given by Equation (3.2d)) and capacity k_i, $E[N_i^2]$ is given by:

$$E[N_i^2] = \frac{2\rho_i^2}{(1-\rho_i)^2} - \frac{k_i(k_i+1)\rho_i^{k_i+1}}{1-\rho_i^{k_i+1}} - \frac{2(k_i+1)\rho_i^{k_i+2}}{(1-\rho_i^{k_i+1})(1-\rho_i)} + E[N_i]. \tag{3.23}$$

The expression of Equation (3.23) is derived in Appendix B.2.

Approximation of $E[W_iW_j]$

A set of queues is referred to as adjacent queues if they share a common node. In other words, adjacent queues include all upstream queues and downstream queues connected to the same node.

For non-adjacent queues, we approximate $E[W_iW_j]$ with:

$$E[W_iW_j] = E[W_i]E[W_j], \tag{3.24}$$

where $E[W_i]$ and $E[W_j]$ are given by Equation (3.18).

For adjacent queues, we use the state-dependent model to account for the between-

queue interactions. Let queues i and j be adjacent queues with common node m. Conditional on the node state, the variables W_i are approximated as independent W_j. That is:

$$E[W_i W_j] = \sum_{s_m \in \mathscr{S}(m)} P(S_m = s_m) E[W_i | S_m = s_m] E[W_j | S_m = s_m]. \tag{3.25}$$

An expression for the conditional expectation is obtained by applying Little's law.

$$E[W_i | S_m = s_m] = \frac{E[N_i | S_m = s_m]}{\hat{\lambda}_{i,s_m}}, \tag{3.26}$$

where the rates $\hat{\lambda}_{i,s_m}$ are given by Equations (3.6a) and (3.6c). The conditional expected queue-length is given by:

$$E[N_i | S_m = s_m] = \sum_{n=0}^{k_i} n P(N_i = n | S_m = s_m), \tag{3.27}$$

where the conditional queue-length probabilities are defined by Equations (3.9), (3.11) and (3.12).

The state probability $P(S_m = s_m)$ of Equation (3.25) is given by:

$$P(S_m = s_m) = \prod_{i \in \mathcal{D}_m; \mathbb{1}(s_m, i) = 1} P(N_i = k_i) \cdot \prod_{i \in \mathcal{D}_m; \mathbb{1}(s_m, i) = 0} \left(1 - P(N_i = k_i)\right), \tag{3.28}$$

where the probabilities $P(N_i = k_i)$ are given by Equation (3.2c). The first (resp.

second) product considers the set of downstream queues that are (resp. are not) full in state s_m.

The next section (Section 3.4.2) considers a network with multiple paths, it uses the moments of path sojourn time derived in this section Equations (3.14) and (3.17)) to approximate the moments of trip sojourn times.

3.4.2 First- and second-order moments of the trip sojourn time

In this section, we present the analytical approximation of the first- and second-order moments of the trip sojourn time. Let TT denote the trip sojourn time random variable. The first-order moment can be obtained by a direct application of Little's law:

$$E[TT] = \begin{cases} \dfrac{\sum_{i \in \mathcal{Q}} E[N_i]}{\sum_{i \in \mathcal{Q}} \gamma_i (1 - P(N_i = k_i))} & \text{blocking exists,} \quad (3.29a) \\ \dfrac{\sum_{i \in \mathcal{Q}} E[N_i]}{\sum_{i \in \mathcal{Q}} \gamma_i} & \text{no blocking.} \quad (3.29b) \end{cases}$$

For the proposed method and the method states in Section 3.5.2, in which blocking occurs, we use Equation (3.29a). For the method states in Section 3.5.2, since no blocking occurs, we use the formulation in Equation (3.29b).

where $E[N_i]$ is given by Equation (3.19). The second-order moment is given by:

$$VAR[TT] = \sum_{p \in \mathcal{P}} P(X = p) VAR[\tilde{W}_p], \tag{3.30}$$

where $\mathcal{P}$ represents the set of paths in the network, X represents the path choice of a traveler, and $P(X = p)$ represents the probability of choosing path p, and $VAR[\tilde{W}_p]$ is the path sojourn time variance (given by Equation (3.17)). We approximate $P(X = p)$ the path choice by the expected proportion of network demand that travels along path p. That is:

$$VAR[TT] = \sum_{p \in \mathcal{P}} \frac{\bar{\lambda}_p}{d} VAR[\tilde{W}_p], \tag{3.31}$$

where $\bar{\lambda}_p$ denotes the expected flow on path p, and d represents the expected total travel demand in the network.

The evaluation of Equation (3.31) requires the enumeration of all paths, which can be a high-dimensional set in general topology networks. Hence, we now show how we can evaluate this expression without path enumeration.

Inserting Equation (3.17), we obtain:

$$VAR[TT] = \sum_{p \in \mathcal{P}} \frac{\bar{\lambda}_p}{d} \left\{ \sum_{i \in \mathcal{Q}_p} (E[W_i^2] - E[W_i]^2) + \sum_{(i,j) \in \mathcal{Q}_p^2, i \neq j} (E[W_i W_j] - E[W_i] E[W_j]) \right\}. \tag{3.32}$$

For non-adjacent queues, Equation (3.24) holds, so the second inner summation of Equation (3.32) equals zero. Thus, we can limit this second inner summation to adjacent queues along a path. This leads to:

$$VAR[TT] = \sum_{p \in \mathcal{P}} \frac{\bar{\lambda}_p}{d} \left\{ \sum_{i \in \mathcal{Q}_p} (E[W_i^2] - E[W_i]^2) + \sum_{i \in \mathcal{Q}_p} \sum_{j \in \mathcal{Q}_p \cap \{\mathcal{DS}_i \cup \mathcal{US}_i\}} (E[W_i W_j] - E[W_i]E[W_j]) \right\}. \tag{3.33}$$

Let $\mathcal{G}_i$ denote the set of paths that go through queue i, and let $\mathcal{G}_{ij}$ denote the set of paths that go through adjacent queues i and j. We can exchange the order of the summations in (3.33), this leads to:

$$VAR[TT] = \sum_{i \in \mathcal{Q}} \sum_{p \in \mathcal{G}_i} \left(\frac{\bar{\lambda}_p}{d} (E[W_i^2] - E[W_i]^2) \right) + \sum_{i \in \mathcal{Q}} \sum_{j \in \mathcal{DS}_i \cup \mathcal{US}_i} \sum_{p \in \mathcal{G}_{ij}} \frac{\bar{\lambda}_p}{d} (E[W_i W_j] - E[W_i]E[W_j]). \tag{3.34}$$

This expression can be further rearranged to:

$$VAR[TT] = \sum_{i \in \mathcal{Q}} (E[W_i^2] - E[W_i]^2)(\sum_{p \in \mathcal{G}_i} \frac{\bar{\lambda}_p}{d}) + \sum_{i \in \mathcal{Q}} \sum_{j \in \mathcal{DS}_i \cup \mathcal{US}_i} (E[W_i W_j] - E[W_i]E[W_j])(\sum_{p \in \mathcal{G}_{ij}} \frac{\bar{\lambda}_p}{d}) \tag{3.35}$$

The term $\sum_{p \in \mathcal{G}_i} \bar{\lambda}_p / d$ is the ratio of the expected demand along queue i, and the expected network demand. Similarly, the term $\sum_{p \in \mathcal{G}_{ij}} \bar{\lambda}_p / d$ is the ratio of the expected

demand that goes through both queues i and j and, and the expected network demand. These terms are approximated as follows.

$$\sum_{p \in \mathcal{G}_i} \frac{\bar{\lambda}_p}{d} = \frac{\tilde{\lambda}_i}{\sum_{i \in \mathcal{Q}} \gamma_i} \tag{3.36}$$

$$\sum_{p \in \mathcal{G}_{ij}} \frac{\bar{\lambda}_p}{d} = \frac{p_{ij}\tilde{\lambda}_i}{\sum_{i \in \mathcal{Q}} \gamma_i}, \tag{3.37}$$

where the expected demand along queue i is denoted $\tilde{\lambda}_i$ and is obtained by solving the following flow-conservation equations:

$$\tilde{\lambda}_i = \gamma_i + \sum_j p_{ji}\tilde{\lambda}_j. \tag{3.38}$$

This leads to:

$$VAR[TT] = \sum_{i \in \mathcal{Q}} (E[W_i^2] - E[W_i]^2) \frac{\tilde{\lambda}_i}{\sum_{i \in \mathcal{Q}} \gamma_i} + \sum_{i \in \mathcal{Q}} \sum_{j \in \mathcal{DS}_i \cup \mathcal{US}_i} (E[W_i W_j] - E[W_i]E[W_j]) \frac{p_{ij}\tilde{\lambda}_i}{\sum_{i \in \mathcal{Q}} \gamma_i}, \tag{3.39}$$

Equation (3.39) is used to evaluate the second-order moment of the trip sojourn time. The terms $E[W_i]$, $E[W_i^2]$ and $E[W_i W_j]$ are given by Equations (3.18), (3.22) and (3.25), respectively.

In the following sections the proposed approximations of the second-order moments of both path and trip sojourn time will be validated and then used to address urban

traffic management problems.

3.5 Validation

This section validates the approximations of the second-order moments of both the path sojourn time and the trip sojourn time. We compare the analytical approximations to both simulated esimates and to the approximations obtained by other approximate analytical methods. Section 3.5.1 presents the considered networks and scenarios. Section 3.5.2 presents the simulation-based and analytical methods that are compared. A computational run time comparison is presented in Section 3.5.3. The validation of the second-order moments of the path (resp. trip) sojourn times is discussed in Section 3.5.4 (resp. Section 3.5.5).

3.5.1 Validation scenarios

We consider two networks. Network 1 is displayed in Figure 3-1. The queues are depicted as circles, and the possible turnings or transitions are depicted with arrows. This network consists of 8 queues. External arrivals arise only to queue 1 (i.e., $\forall i \neq 1, \gamma_i = 0$). Departures from the network arise only from queues 7 and 8. There are a total of 3 paths: path 1 goes from queue 1 to queue 7 via queue 2; path 2 goes from queue 1 to queue 7 via queue 3; path 3 goes from queue 1 to queue 8. Jobs at queue 1 proceed to queue 2 with probability 0.3 and to queue 3 with probability 0.7. Jobs

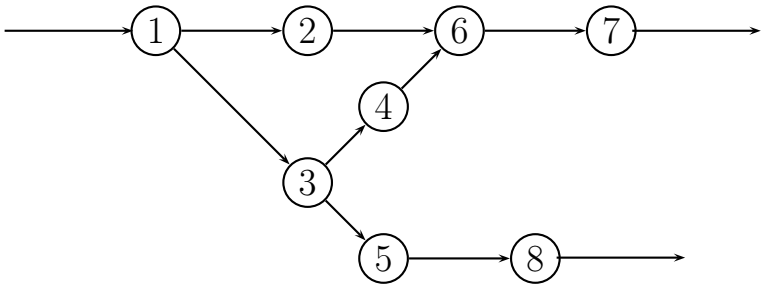


Figure 3-1: Topology of network 1.

i :	1	2	3	4	5	6	7	8
μ_i	10	4	7	4	5	6	6	5
k_i	8	8	8	8	8	8	8	8

Table 3.1: Configuration of network 1.

at queue 3 proceed to queue 4 with probability 0.4 and to queue 5 with probability 0.6. The service rates and space capacities of the queues are defined in Table 3.1. We consider a set of 5 demand scenarios with increasing levels of congestion (i.e., increasing external arrival rate to queue 1). These 5 scenarios are defined in Table 3.2.

Network 2 is displayed in Figure 3-2. It consists of 10 queues. External arrivals arise only to queue 1 and queue 7. Departures from the network arise only from queues 4, 6 and 10. This leads to a total of 5 paths: path 1 goes from queue 1 to queue 4; path 2 goes from queue 1 to queue 10; path 3 goes from queue 1 to queue 6; path 4

scenario	1	2	3	4	5
γ_1	6	7	8	9	9.99

Table 3.2: Demand scenarios for network 1.

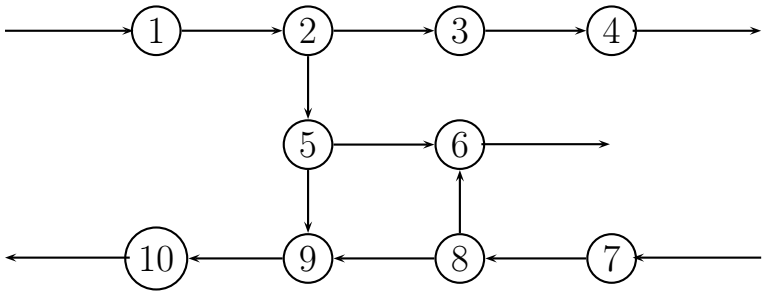


Figure 3-2: Topology of network 2.

$i:$	1	2	3	4	5	6	7	8	9	10
μ_i	10	10	6	6	6	10	10	9	9	9
k_i	8	8	8	8	8	8	8	8	8	8

Table 3.3: Configuration of network 2.

goes from queue 7 to queue 10; path 5 goes from queue 7 to queue 6. Jobs at queue 2 proceed to queue 3 (resp. queue 5) with probability 0.4 (resp. 0.6). Jobs at queue 5 proceed to queue 6 (resp. queue 9) with probability 0.5 (resp. 0.5). Jobs at queue 8 proceed to queue 6 (resp. queue 9) with probability 0.5 (resp. 0.5). The service rates and space capacities of the queues are defined in Table 3.3. We consider a set of 5 demand scenarios with increasing levels of congestion (i.e., increasing external arrival rates to queues 1 and queue 7). These 5 scenarios are defined in Table 3.4.

scenario	1	2	3	4	5
γ_1	6	7	8	9	9.99
γ_7	6	7	8	9	9.99

Table 3.4: Demand scenarios for network 2.

3.5.2 Benchmark methods

Stochastic simulation model

We use a stochastic discrete-event simulation model of finite capacity Markovian networks (Meier, 2007). For each scenario, the simulation estimates are obtained from 1000 replications, each with a total run time of 10,000 time units including a warm up period of 1,000 time units. We display 95% confidence intervals, which are given by: $\bar{s} \pm 1.96\hat{s}/\sqrt{1000-1}$, where $\bar{s}$ represents the estimated average sojourn time, and $\hat{s}$ represents the estimated standard deviation of the sojourn time.

State-independent model

The state-independent model is defined by the System of Equations (3.2). The comparison of the state-independent model with the proposed state-dependent model serves to illustrate the added value of using state-dependent arrival and service rates to yield a more detailed description of between-queue dependencies and ultimately a more accurate description of path and trip sojourn time metrics.

We describe here how the first- and second-order moments of the path and the trip sojourn times are calculated for the state-independent model. For the path sojourn time, the first-order moment is given by Equations (3.14), (3.18) and (3.19). For this model, the sojourn times of all queues are assumed independent, hence all the covariance terms $COV[W_i W_j]$ of Equation (3.16) equal zero. Thus, the second-order

moment of the path sojourn time is given by:

$$VAR[\tilde{W}_p] = \sum_{i \in \mathcal{Q}_p} (E[W_i^2] - E[W_i]^2), \tag{3.40}$$

where $E[W_i]$ is given by Equation (3.18) and $E[W_i^2]$ is given by Equations (3.22) and (3.23).

In summary, the differences with the proposed state-dependent model are: (i) the use of state-independent rates (rather than state-dependent rates), (ii) the assumption of independent queue sojourn times (whereas the state-dependent model assumes that the sojourn time of adjacent queues along a path are dependent).

State-independent model without blocking

This model differs from the model of Section 3.5.2 in that it does not account for any blocking (i.e., spillback) effects between queues. The description of blocking events is necessary to describe the spatial propagation of congestion. Nonetheless, blocking events are the main reason why an analytical analysis of finite capacity networks is challenging. The comparison of the model of Section 3.5.2 with this model illustrates

the added value of accounting for blocking. The model is formulated as follows.

$$\tilde{\lambda}_i = \gamma_i + \sum_j p_{ji}\tilde{\lambda}_j \tag{3.41a}$$

$$\tilde{\rho}_i = \frac{\tilde{\lambda}_i}{\mu_i}, \tag{3.41b}$$

where $\tilde{\lambda}_i$ represents the arrival rate to queue i, and $\tilde{\rho}_i$ represents the traffic intensity of queue i. Equation (3.41a) is a flow conservation equation that assumes that no blocking occurs, and hence no losses at the entries of the network occur. For a network where for all queues there is a zero probability of blocking, then Equation (3.41a) is equivalent to Equation (3.2a). Equation (3.41b) defines the traffic intensity as the ratio of the arrival rate to the service rate. Note that the denominator is the service rate rather than the effective service rate that is used in Equation (3.2d). If there is a zero probability of jobs at queue i getting blocked, then the service rate, μ_i, equals the effective service rate, $\hat{\mu}_i$.

For this model, the sojourn times of all queues are assumed independent, hence all the covariance terms $COV[W_iW_j]$ of Equation (3.16) equal zero. This leads to the following expressions for the first- and second-order moments of the path and the trip

sojourn times.

$$E[\tilde{W}_p] = \sum_{i \in \mathcal{Q}_p} E[W_i] \tag{3.42a}$$

$$VAR[\tilde{W}_p] = \sum_{i \in \mathcal{Q}_p} (E[W_i^2] - E[W_i]^2) \tag{3.42b}$$

$$E[W_i] = E[N_i]/\tilde{\lambda}_i \tag{3.42c}$$

$$E[W_i^2] = \frac{E[N_i^2] - E[N_i]}{(\tilde{\lambda}_i)^2} \tag{3.42d}$$

$$E[N_i] = \tilde{\rho}_i \left(\frac{1}{1-\tilde{\rho}_i} - (k_i+1)\frac{\tilde{\rho}_i^{k_i}}{1-\tilde{\rho}^{k_i+1}} \right), \tag{3.42e}$$

$$E[N_i^2] = \frac{2\tilde{\rho}_i^2}{(1-\tilde{\rho}_i)^2} - \frac{k_i(k_i+1)\tilde{\rho}_i^{k_i+1}}{1-\tilde{\rho}_i^{k_i+1}} - \frac{2(k_i+1)\tilde{\rho}_i^{k_i+2}}{(1-\tilde{\rho}_i^{k_i+1})(1-\tilde{\rho}_i)} + E[N_i]. \tag{3.42f}$$

This model is equivalent to assuming infinite capacity queues (where no blocking can occur) to calculate the arrival rate and traffic intensity of a queue, while assuming a finite capacity queue to calculate the marginal queue-length distribution. In summary, the differences of this model with the proposed state-dependent model are: (i) the assumption of no blocking occurring at any queue, (ii) the use of state-independent rates (rather than state-dependent rates), (iii) the assumption of independent queue sojourn times (whereas the state-dependent model assumes that the sojourn time of adjacent queues along a path are dependent).

Scenario	1	2	3	4	5
Network 1	0.40	0.45	0.43	0.41	0.39
Network 2	0.61	0.47	0.51	0.46	0.49

Table 3.5: Computational time to evaluate the proposed analytical model (in seconds).

Scenario	1	2	3	4	5
Network 1	114	147	120	159	186
Network 2	54	75	120	159	144

Table 3.6: Average computational time to run one simulation replication (in seconds).

3.5.3 Computational run times

For all scenarios of network 1and network 2, the computational time needed to evaluate the proposed analytical model in Matlab for each demand scenario is shown in Table 3.5. For each scenario and each network, the analytical model can be solved instantly.

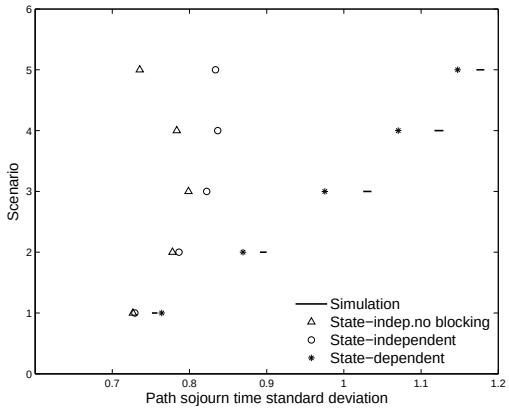
The computational time for one replication of the simulator is displayed, for each scenario and each network, in Table 3.6. For the analysis of this paper, 1000 simulation replications are run, hence the total computational savings for each network and each scenario are three orders of magnitude larger than the values displayed in Table 3.6. This table shows that even for very small networks, the analytical model is significantly more computationally efficient than a simulation-based model.

3.5.4 Second-order moment of path sojourn time

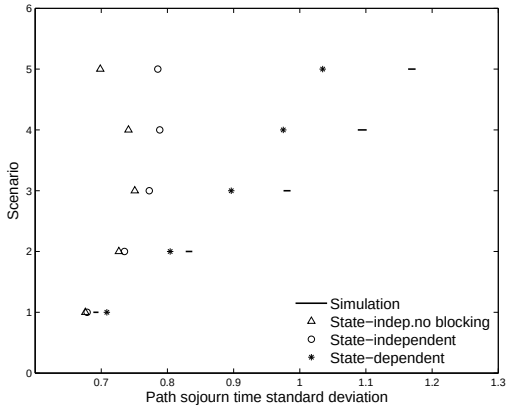
In this section, we validate the second-order moment of the path sojourn time, i.e., the standard deviation of the path sojourn time. The results for network 1 are displayed in Figure 3-3. Figures 3-3(a), 3-3(b) and 3-3(c) represent the results for paths 1, 2 and 3, respectively. Each plot displays the path sojourn time standard deviation along the x-axis. The y-axis considers each of the 5 demand scenarios. Note that as the scenario index increases, so does the network demand.

The 95% confidence intervals of the simulation estimates are displayed as solid lines. The estimates for the state-independent-no-blocking method are displayed as triangles. Those for the state-independent (resp. state-dependent) method are displayed as circles (resp. stars).

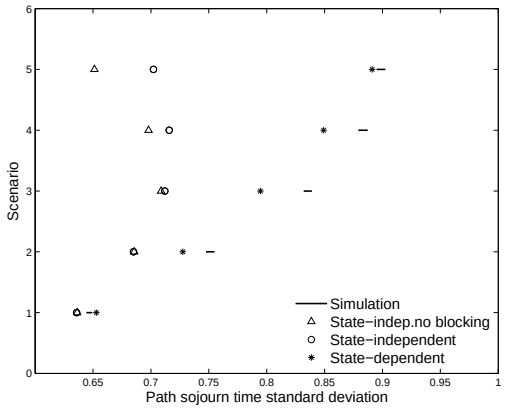
For all paths and all scenarios, the method with the least accurate performance is the state-independent-no-blocking method. It is followed by the state-independent method. Both of these methods only perform well under light traffic conditions (e.g., scenario 1), and both fail to capture the trend of the performance measure as congestion increases. Under congested conditions, they both significantly underestimate the standard deviation of path sojourn time. The proposed state-dependent yields accurate standard deviation approximations for both uncongested and highly congested conditions. As congestion increases, the proposed method indeed captures the trends of the simulation estimates. This is particularly important if the model is to be used



(a) Path 1



(b) Path 2



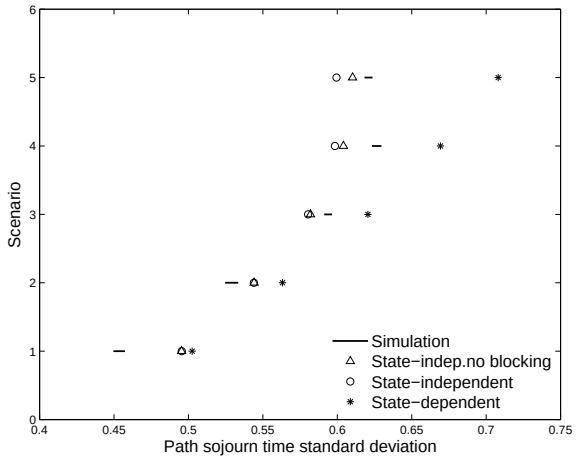
(c) Path 3

Figure 3-3: Path sojourn time standard deviation for each of the 3 paths in network 1.

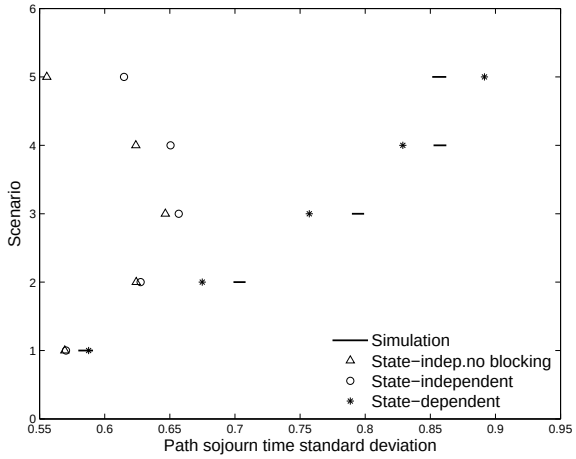
for optimization purposes.

Note that the state-independent method outperforms the state-independent-no-blocking method, and the level of outperformance increases with congestion. This indicates the added value of analytically accounting for blocking.

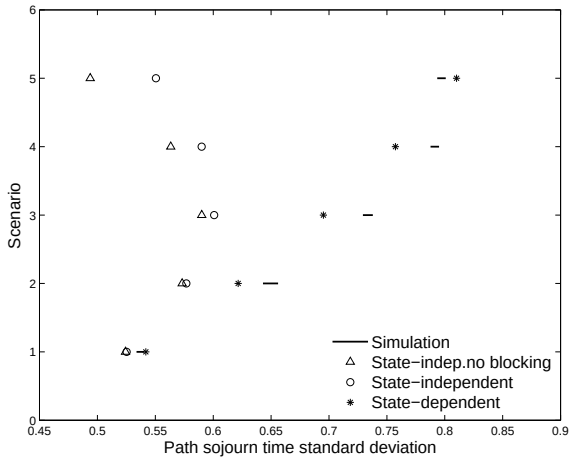
Figure 3-4 displays the results for the scenarios of network 2. Again, as congestion increases, so does the scenario index. Similar conclusions hold. For paths 2-5, the following conclusions hold. The state-independent-no-blocking method and the state-independent methods: (i) yield accurate estimates only for light traffic conditions, (ii) have decreasing accuracy with increasing congestion, (iii) fail to capture the trends of the simulation estimates as congestion increases. The state-independent-no-blocking method is outperformed by the the state-independent method, and the level of outperformance increases with congestion. This illustrates the added value of analytically describing blocking, in particular for congested traffic conditions. The proposed method leads to consistently accurate estimates, and it captures the trends of the simulated estimates as congestion increases. For path 1, the most accurate estimates are those obtained by the state-independent-no-blocking method, followed by the state-independent method, and then the state-dependent method. The proposed method does not capture the trends of the simulated estimates, it overestimates path sojourn time standard deviation. This overestimation increases with congestion.



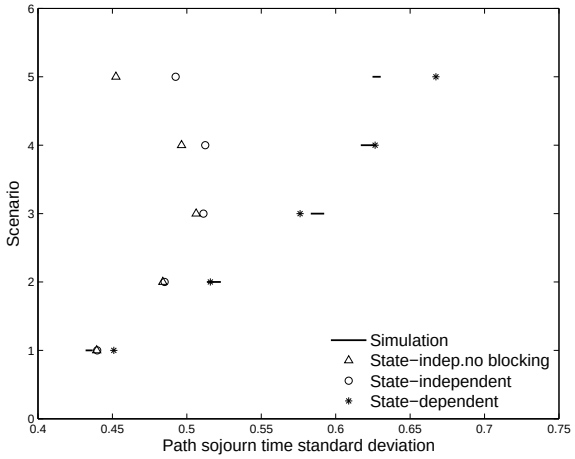
(a) Path 1



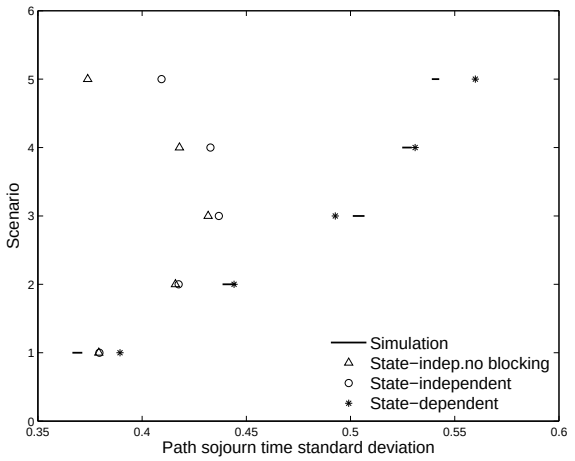
(b) Path 2



(c) Path 3



(d) Path 4

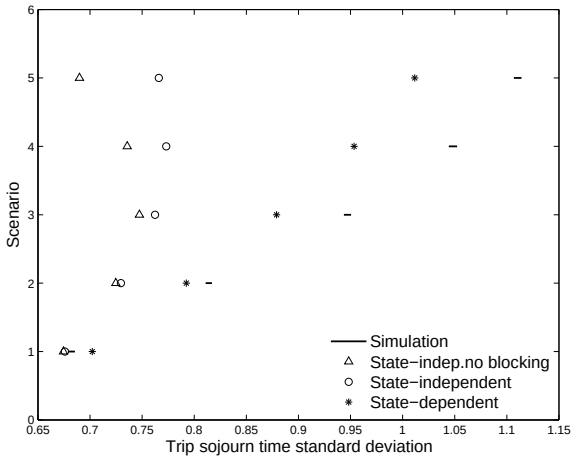


(e) Path 5

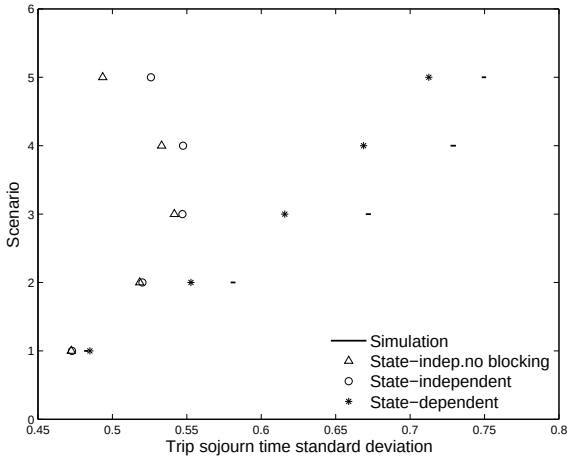
Figure 3-4: Path sojourn time standard deviation for each of the 5 paths in network 2.

3.5.5 Second-order moment of trip sojourn time

In this section, we validate the second-order moment of the trip sojourn time, i.e., the standard deviation of the trip sojourn time. Figure 3-5(a) (resp. Figure 3-5(b)) displays the results for network 1 (resp. network 2). The x-axis displays the trip sojourn time standard deviation, the y-axis displays the scenario index. Just as in the figures of Section 3.5.4, as the scenarios index increases, so does congestion. Similar conclusions as for Section 3.5.4 hold. In particular, the most accurate method is the proposed state-dependent method, followed by the state-independent method, and then the state-independent-no-blocking method. The state-independent-no-blocking method and the state-independent methods yield accurate estimates only for light traffic conditions, have decreasing accuracy with increasing congestion, and they both fail to to capture the trends of the simulation estimates as congestion increases. The state-independent-no-blocking method is outperformed by the the state-independent method, and the level of outperformance increases with congestion, i.e., it increases as the occurence of blocking increases. The proposed method leads to consistently accurate estimates, and it captures the trends of the simulated estimates as congestion increases.



(a) Network 1



(b) Network 2

Figure 3-5: Trip sojourn time standard deviation for networks 1 and 2.

3.6 Analytical optimization case study

In this section we evaluate the ability of the proposed method to address an analytical urban traffic signal control problem. The road network of interest is presented in Section 3.6.1. The traffic signal control problem is formulated in Section 3.6.2, and the results are discussed in Section 3.6.3.

3.6.1 Road network

We consider a signal control problem for the city center of the Swiss city of Lausanne (same as the network used in Chapter 2). We consider the evening peak period 17h-18h, where congestion gradually increases.

3.6.2 Traffic signal control problem

For a review of traffic signal control terminology and formulations, we refer the reader to Appendix A of Osorio (2010) or to Lin (2011). The signal control problem that we consider is known as a fixed-time, also known as pre-timed control strategy. A fixed-time signal plan is a periodic plan defined by a cycle time (the period). For a given intersection, the cycle time is typically of the order of 60, 90 or 120 seconds. The green times of the cycle are allocated to signal phases. These phases consider a set of non-conflicting traffic movements. For a given phase, the green time to cycle time ratio is known as the green split. Offset variables are often used to synchronize

the signal plans of adjacent intersections.

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

b_i	available cycle ratio of intersection i;
$x(j)$	green split of phase j;
x_L	vector of minimal green splits;
z	endogenous queueing model variables;
q	exogenous queueing model parameters;
$\mathscr{I}$	set of intersection indices;
$\mathscr{P}_I(i)$	set of phase indices of intersection i;
r	reliability ratio.

The signal control problem is formulated as follows:

$$\min_x E[T(x,z;q)] + rSD[T(x,z;q)] \tag{3.43}$$

$$\sum_{j\in\mathscr{P}_I(i)} x(j) = b_i,\ \forall i \in \mathscr{I} \tag{3.44}$$

$$x \geq x_L, \tag{3.45}$$

where T is the trip sojourn time in the city center, x is the decision vector, which represents the green splits, z are endogenous queueing model variables (e.g., blocking

probabilities, queue-length distributions) and q are exogenous queueing model parameters (e.g., network topology, travel demand).

In Problem (3.43)-(3.45), the decision variables x are the green splits. All other signal plan parameters (e.g., cycle time, offsets, all-red durations, stage structure) are considered fixed. The objective function (Equation (3.43)) consists of a linear combination of the expected trip travel time, $E[T(x, z; p)]$, and the standard deviation of trip travel time $SD[T(x, z; p)]$. The latter is weighted with weight r, which is known as the reliability ratio. The term reliability refers to the interpretation of travel time variability (as measured by the standard deviation) as a metric for travel time reliability. The reliability ratio r is set to 1.43, as given in Li et al. (2010b). Constraints (3.44) guarantee that for a given intersection the sum of green splits of the endogenous phases equals the available cycle time. Constraints (3.45) ensure lower bounds for the green splits. In the case studies of this chapter, the lower bounds are set to 4 seconds following the Swiss transportation norms (VSS, 1992).

3.6.3 Results

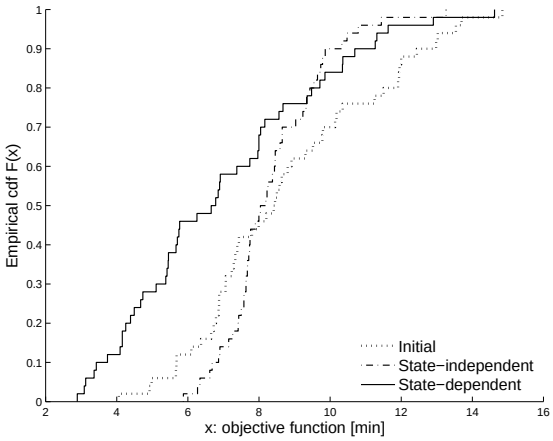
Problem (3.43)-(3.45) is solved with each of the following two queueing models: (i) the state-dependent queueing model , and (ii) the state-independent queueing model. The objective function (Equation (3.43)) is calculated via Equations (3.29a) and (3.30) as:

$$E[TT] + r\sqrt{VAR[TT]}. \tag{3.46}$$

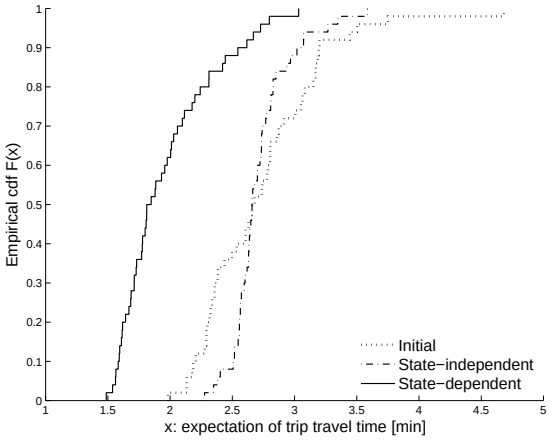
The initial point is a random initial signal plan sampled uniformly from the feasible region, which is defined by Equations (3.44)-(3.45). Uniform sampling is done with the method of Stafford (2006).

In order to evaluate the performance of an optimal signal plan, we use a stochastic simulator of urban traffic. We use the calibrated microscopic traffic simulation model of the Swiss city Lausanne developed by Dumont and Bert (2006). It is calibrated for the evening peak period of Lausanne. It is implemented in Aimsun (TSS, 2013). For a given signal plan, we embed it within the traffic simulator and run 50 simulation replications. For each replication, we evaluate the average trip travel time (i.e., average travel time within the city center) and the trip travel time standard deviation. We then plot the cumulative distribution function (cdf) of these 50 simulation replications. The x-axis of each plot of Figure 3-6 considers a given performance measure (e.g., average trip travel time). For a given x value, the y-axis gives the proportion of simulation replications (out of the 50 replications) that yield a performance measure smaller or equal to x. Hence, the more the cdf curves are shifted to the left, the higher the occurence of low values of the performance measures, i.e., the better the performance.

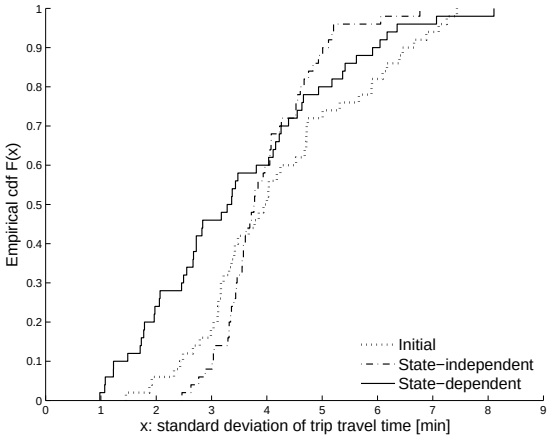
Each plot of Figure 3-6 considers a given performance measure. Figure 3-6(a) considers the objective function (Equation (3.43)). The individual components of the objective function are displayed in Figures 3-6(b) and 3-6(c). Figure 3-6(b) considers the average trip travel time, Figure 3-6(c) considers the trip travel time standard



(a)



(b)



(c)

Figure 3-6: Cumulative distribution functions of the objective function, the average trip travel time and the trip travel time standard deviation.

deviation.

Figure 3-6(a) displays three cdf curves of the objective function. The solid curve is that of the proposed state-dependent method, the dashed curve is that of the state-independent method. The dotted curve is the cdf of the intial signal plan. The signal plan proposed by the state-independent method has slightly better performance than the initial signal plan. Both plans are outperformed by the plan of the state-dependent method. The same conclusions hold for each component of the objective function (Figures 3-6(b) and 3-6(c)). The proposed method leads to lower average trip travel times as well as lower trip travel time standard deviation.

We test whether the objective function performance of the signal plan proposed by the state-dependent method is statistically lower than that of the state-independent method. We perform a paired t-test, where the null hypothesis states that the objective function of the signal plan proposed by the state-dependent method is statistically lower than that of the state-independent method. We perform a paired t-test. When running the 50 simulation replications that evaluate the performance of a given signal plan, we used the same set of replication seeds for each signal plan. The paired t-tests are carried out by pairing observations that have common seeds. Let $\overline{Y}$ denote the average paired difference, let $\hat{s}$ denote its standard deviation, and let O denote the sample size. Then the paired t-statistic is given by (see, for instance, Hogg et al. (1977, p. 486)): $t = \sqrt{O}\ \overline{Y}/\hat{s}$. The average paired difference is $\overline{Y} = 1.40$, its standard

deviation is $\hat{s} = 3.13$, and the sample size $O = 50$. Hence, the t-statistic is 3.15. The critical value at the 2.5% significance level is $t_{0.025}(49) = 2.01$. Hence, the null hypothesis is rejected. The signal plan proposed by the state-dependent method leads to statistically significantly lower objective function values.

3.7 Simulation-based optimization (SO) signal control problem

In this section, we use the proposed analytical traffic model to address a simulation-based traffic signal optimization problem. We consider the same city center network and the same peak-period demand scenario as those of Section 3.6. The only difference with Section 3.6 is the objective function of the signal control problem. In this section, the objective function is a simulation-based objective function. It can be written as:

$$E[T(x;\tilde{q})] + rSD[T(x;\tilde{q})], \tag{3.47}$$

where x represents the decision vector (i.e., the green splits), and $\tilde{q}$ represents the exogenous parameters of the simulator (e.g., network topology, network demand, etc.). The first term (resp. second term) of Equation (3.47) represents the simulation-based expected trip travel time (resp. simulation-based trip travel time standard deviation). The weight parameter r is the reliability ratio, we use the same numerical value as in

Section 3.6. The SO problem has the same constraints as the problem of Section 3.6, these are given by Equations (3.44)-(3.45). These are analytical constraints.

In summary, the problem considered in this section consists of a simulation-based objective function (Equation (3.47)) and analytical constraints (Equations (3.44)-(3.45)).

The state-dependent analytical model proposed in this chapter is used to construct a metamodel. A metamodel is an analytical approximation of the (unknown) simulation-based objective function. For details on metamodel formulations and metamodel SO literature, see Osorio and Bierlaire (2013). We use the metamodel SO algorithm of Osorio and Bierlaire (2013). This algorithm considers a metamodel that combines information from an analytical traffic model and from the simulation-based traffic model. In this section, we use the proposed state-dependent analytical traffic model as the analytical traffic model. We call this method the "State-dependent SO" method.

In order to benchmark the performance of this SO approach, we compare its performance to 2 other methods. The first considers the SO algorithm of Osorio and Bierlaire (2013), yet uses the state-independent analytical traffic model (defined/discussed in AFAF) to construct the metamodel. We call this method the "State-independent SO" method. The comparison of "State-dependent SO" method to the "State-independent SO" method illustrates the added value of accounting *analytically* for detailed between-

queue dependencies when performing SO.

The second method that is benchmarked is that proposed in (Chen et al., 2012b). We call this method the "Reliability SO" method. This is an SO method that has also been used to design signal plans with reduced travel time variability. Since an analytical and computationally tractable approximation of the trip travel time standard deviation (i.e., term $SD[T(x;\tilde{q})]$ of Equation (3.47)) was not available at the time, the SO problem was formulated considering the link travel time standard deviation. The SO problem used the following objective function:

$$E[\sum_{i\in\mathscr{L}} T_i(x;\tilde{p})] + rSD[\sum_{i\in\mathscr{L}} T_i(x;\tilde{p})], \tag{3.48}$$

where $\mathscr{L}$ denotes the set of queues within the network of interest, and $T_i(x;\tilde{p})$ denotes the (simulation-based) travel time along queue i. In other words, the objective function considers the first- and second-order moments of the total link travel time rather than that of the trip travel time.

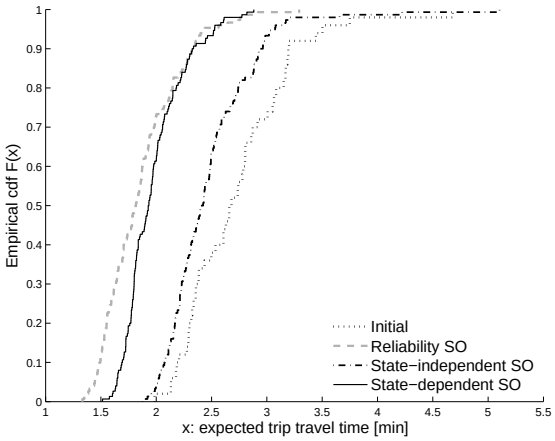
The signal plans identified by the "Reliability SO" method are designed based on a different objective function than that of Equation (3.47). Nonetheless, they aim to achieve the same goal: reducing the average travel time and the travel time variability for travelers. The comparison of "State-dependent SO" method to the "Reliability SO" method illustrates the ability of our proposed method to actually achieve this goal.

The computational budget is set to 450 simulation runs. The performance of a given

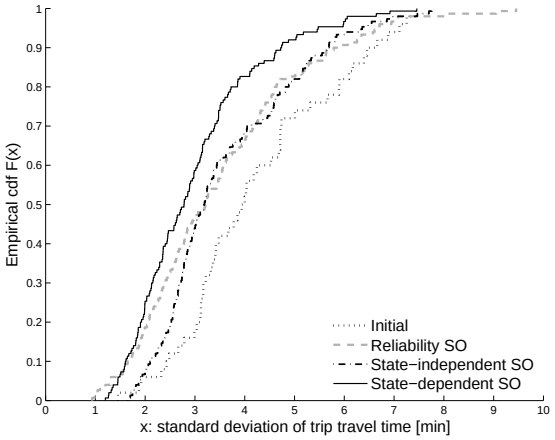
point (i.e., a signal plan) is evaluated by running 3 simulation replications. Hence, the computational budget of 450 allows for a maximum of 150 points to be evaluated. Once the computational budget is depleted, the current iterate is considered as the final solution, i.e., it is the "proposed" signal plan.

We consider the same initial signal plan in Section 3.6.2. This plan is sampled uniformly from the feasible region, which is defined by Equations (3.44)-(3.45). Uniform sampling is done with the method of Stafford (2006). For each of the 3 SO methods mentioned above, we consider the given initial plan and the given computational budget, and we run the SO method three times. We run it three times since the output of the algorithm is now stochastic. For each SO method, this leads to 3 proposed signal plans. The performance of a proposed signal plan is evaluated just as in Section 3.6: i.e., we run 50 simulation replications with the same set of random seeds for each signal plan. For each SO method, we aggregate the results for all 3 proposed signal plans, and construct a single cumulative distribution function (cdf). In other words, the cdf curve consists of 50*3 observations.

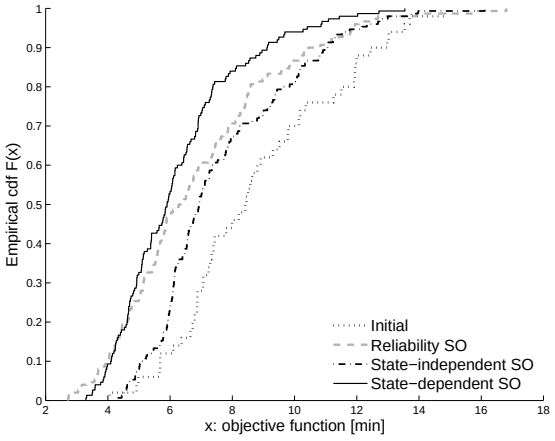
We first plot the empirical cumulative distribution function (cdf) of the performance measures obtained by solving different methods aggregately over all 3 signal plans in Figure 3-7. Because of the stochastic feature of the simulator, we then plot the cdf of the performance metrics for each signal plan in Figure 3-8. The objective function for the reliability method is not the same as the other 2 methods, to be com-



(a)



(b)



(c)

Figure 3-7: Expected trip travel time, trip travel time SD and objective function of the signal control methods when applied to the Lausanne city center. These plots consider various problem formulations.

parable with other two methods, we also calculate the summation of expected trip travel time and 1.43 times travel time SD for the reliability method.

In Figure 3-7, each subfigure shows 4 cdf curves. Figure 3-7(a) shows the plots for expected trip travel time. The black dotted curve shows the expected trip travel time obtained from the initial signal plan which contains 50 observations. The other 3 cdf curves represent the expected trip travel time of the signal plans obtained by solving different signal control problems. Each of the curve contains 3*50 observations from 3 signal plans because we run the SO algorithm for each problem 3 times. The x-axis represents the value of the expected trip travel time. The signal plan derived by "reliability" (displayed as grey dashed line) lead to signal plans with smallest expected trip travel time. The "state-independent" method (displayed as dash-dot line) yields the signal plan with largest expected trip travel time. Figure 3-7(b) shows the plots for trip travel time SD. The "state-dependent" method (displayed as solid line) leads to the signal plans with smallest value of trip travel time SD. Figure 3-7(c) shows the plots for objective function. The "state-dependent" method leads to the signal plans with smallest value of objective function, the "state-independent" method yields the signal plan with largest objective function value. Although the "reliability" method does not optimize the summation of expected trip travel time and 1.43 times travel time SD, it leads to signal plan with smaller values comparing to the "state-independent" method. All signal plans are better than the initial signal plan for all measures.

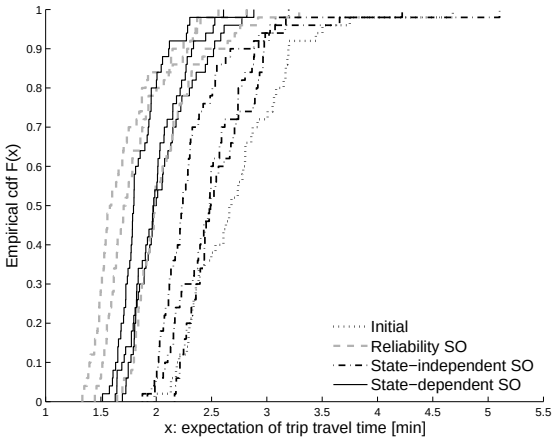
To test if the performance of the signal plan derived by solving different methods are statically different from each other. We perform a paired t-test to test the hypothesis that the expected trip travel time, trip travel time SD, and objective function values derived from the “reliability" formulation are equal to the expected trip travel time, trip travel time SD, and objective function values derived by “state-dependent" method. Since each curve contains three signal plans and we use the same set of random seeds when we evaluate all signal plans in the simulator, we take the average value of average trip travel time, trip travel time SD and objective function over all signal plans derived by the same method.

The mean of the paired difference between “reliability" and “state-dependent" for average trip travel time, average trip travel time SD and objective function are -0.1116, 0.5162 and 0.6265 respectively; the corresponding standard deviation of the paired difference are 0.2100, 1.1185 and 1.7798. The t values are -3.7582, 3.2631 and 2.4890 for expected trip travel time, trip travel time SD and objective function respectively. For expected trip travel time, the mean of the paired difference between “reliability" and “state-dependent" method is negative, which means “reliability" yields signal plan with smaller average trip travel time. The t-value for expected trip travel time is -3.7582, thus the null hypothesis is rejected. The mean differences are positive for other two metrics, the t-values are all larger than $t_{0.025}(49) = 2.01$, thus null hypothesis for trip travel time SD and objective function are rejected. Signal plans

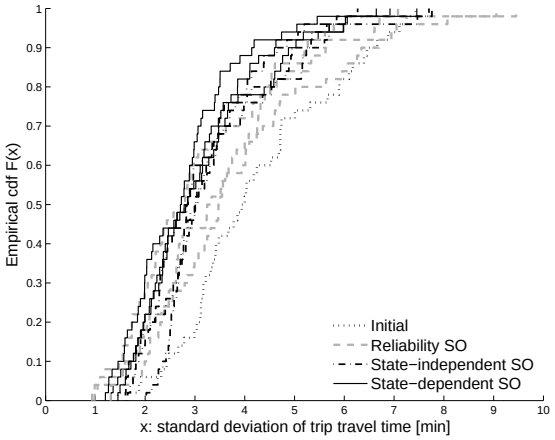
derived by “reliability" method have smaller expected trip travel time than the signal plans derived by “state-dependent" method. Signal plans derived by ‘state-dependent" method have better performance in terms of trip travel time SD.

We also perform a paired t-test to test the hypothesis that the expected trip travel time, trip travel time SD, and objective function values derived from the ’state-independent’ formulation are equal to the expected trip travel time, trip travel time SD, and objective function values derived by “state-dependent" method. The mean of the paired difference between “state-dependent" and “state-independent" for average trip travel time, average trip travel time SD and objective function are 0.5031, 0.6290 and 1.4026 respectively; the corresponding standard deviation of the paired difference are 0.2748, 1.0402 and 1.6974. The t values are 12.9461, 4.2758 and 5.8430 for expected trip travel time, trip travel time SD and objective function respectively.They are all larger than $t_{0.025}(49) = 2.01$, thus null hypothesis are rejected for all performance metrics. Signal plans derived by “state-dependent" method have better performance for all performance metrics than the signal plans derived by “state-independent" method.

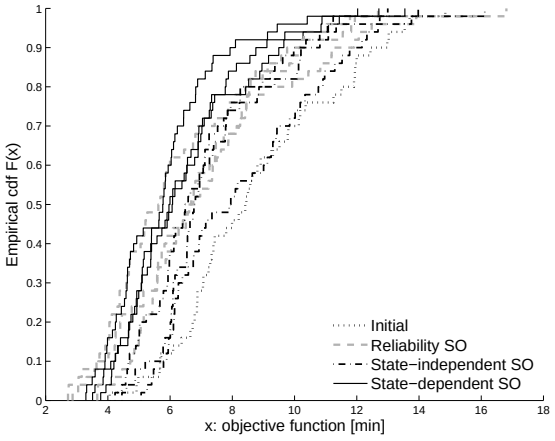
Figure 3-8 shows the performance comparison of each signal plan obtained by solving different problems with different objective functions. Figure 3-8(a), Figure 3-8(b) and Figure 3-8(c) display 10 cdf curves of expected trip travel time, trip travel time SD and objective function for different methods respectively. In each of the figure, the black dotted curve shows the expected trip travel time, trip travel time SD and



(a)



(b)



(c)

Figure 3-8: Performance of the signal control methods when applied to the Lausanne city center. These plots consider various problem formulations.

objective function obtained from the initial signal plan which contains 50 observations. The other 9 cdf curves represent the performance of the signal plans obtained by solving different signal control problems. Each of the curve contains 50 observations. In Figure 3-8(a), 2 signal plans obtained by solving the “reliability" problem (displayed as grey dashed line) have smaller expected trip travel time than all other plans. In Figure 3-8(b), 2 signal plans derived by “state-dependent" method (displayed as solid line) have the smaller trip travel time SD than all other plans. Signal plans derived by “state-independent" method (displayed as dash-dot line) have better performance than 2 out of 3 signal plans derived by 'reliability' method. In Figure 3-8(c), 1 signal plan obtained by 'reliability' method has similar performance to the signal plans derived by 'state-dependent' method. Signal plans obtained by 'state-independent' method have the worst performance. In all figures, the signal plans derived by using different methods have better performance than the initial signal plan.

The proposed formulation that accounts for between link dependency yields signal plans with smaller expected trip travel time and trip travel time SD comparing to the one assuming independent link travel time. Furthermore, signal plans obtained by minimizing average total link travel time and total ink travel time SD has smaller expected trip travel time but larger trip travel time SD comparing to the signal plans derived by proposed formulation.

3.8 Conclusions

In this chapter, we derive path travel time SD that could be used for finite capacity queueing networks based on first- and second- order Little's law which is originally derived for infinite capacity queues. We take into consideration the interactions between adjacent queues and model the link dependency. The results obtained from our approach are compared with the formulation that ignores the link dependency and the simulated results for a toy network with 10 queues and 5 different paths. We then use this model to address a traffic signal control problem analytically to account the added value of accounting link dependencies. Further more, the proposed method is used to solve a simulation-based optimization signal control problem. The results show that accounting link dependency helps to reduce trip travel time variability comparing to the approach that does not account for that. We also compare the performance of the signal plan derived from the proposed method with the signal plan obtained in Chapter 2. The method proposed in Chapter 3 reduces trip travel time variability significantly at the expenses of increasing average trip travel time comparing to the signal plan derived in Chapter 2. This represents a tradeoff between expectation and SD of trip travel time, a balance point between efficiency (first-order information) and reliability (second-order information) can be discussed in future research. It is of interest to performance a sensitive test with respect to different values of reliability ratio.

Chapter 4

Limiting the spatial propagation of congestion via simulation-based signal control

4.1 Introduction

The occurrence, dynamics and impact of urban network spillbacks have received attentions. For uncongested network, there is no significant queue formation, but for congested network, demand approaches or even exceeds capacity, queues build up. The propagation of congestion may have major impacts in the vicinity of major arterials. In a recent FHWA report, it states that different signal control strategies are appropriate

for different traffic conditions (e.g. peak, off-peak) (Gettman et al., 2013). Thus for uncongested and congested network, signal design strategy differs. The control strategy which is suitable for uncongested network might not be appropriate for congested network. For a highly congested urban network with multimodal traffic, numerous signalized intersections, short links and a grid-type topology, the design of signal plans that indeed improve traffic conditions is a real challenge. The grid-type topology leads to high-dimensional route alternatives, and may lead to complex behavior of travelers as they react to the formation and propagation of congestion. Furthermore, congested networks with grid-type topologies and short links are highly prone to the occurrence of spillbacks. If spillback happens in certain links, congestion propagates quickly and affects larger areas.

In this chapter, we propose a method to design signal control strategies that can be used for highly congested urban road network with grid-type topology. In particular, we propose signal settings for an area in eastern Manhattan (New York City, USA) around the highly congested Queensboro bridge. It is the busiest bridge in New York City with around 178,000 vehicles crossing during each normal weekdays in the year of 2010 (NYCDOT, 2014). Morning peak period vehicular traffic in this area is in the order of 11,000 vehicles per hour. The traffic conditions around Queuesboro area have a large impact on the traffic access to/egress from the highly congested corridor. Currently, only fixed-time signal plan is used in this area, it is of particular interest

to New York City Department of Transport (NYCDOT) to explore the potentials of using novel signal control strategy in that area. Traditional signal control strategies are difficult to tailor to the specific needs of such networks, this is due to the following reasons. First, they embed low-resolution macroscopic traffic models which do not provide a detailed description of traveler behavior or of the underlying network supply (e.g. prevailing traffic operations). Second, they most often have pre-determined objective functions to be used for optimization. The simulation-based optimization algorithm stated in Chapter 2 is used to identify traffic signal plans tailored to the context and needs of the specific underlying networks.

4.1.1 Network topology

Figure 4-1 shows the topology of the studied Queensboro bridge area. The network consists of a total of 134 roads, 313 lanes, 27 signalized intersections and 5 non-signalized intersections.

This chapter is structured as follows: Section 4.2 presents a review of traffic signal control strategy for congested urban road network. We then present the methodology in Section 4.3. We evaluate the performance of the proposed signal plan in Section 4.4. We conclude with a brief discussion in Section 4.5.

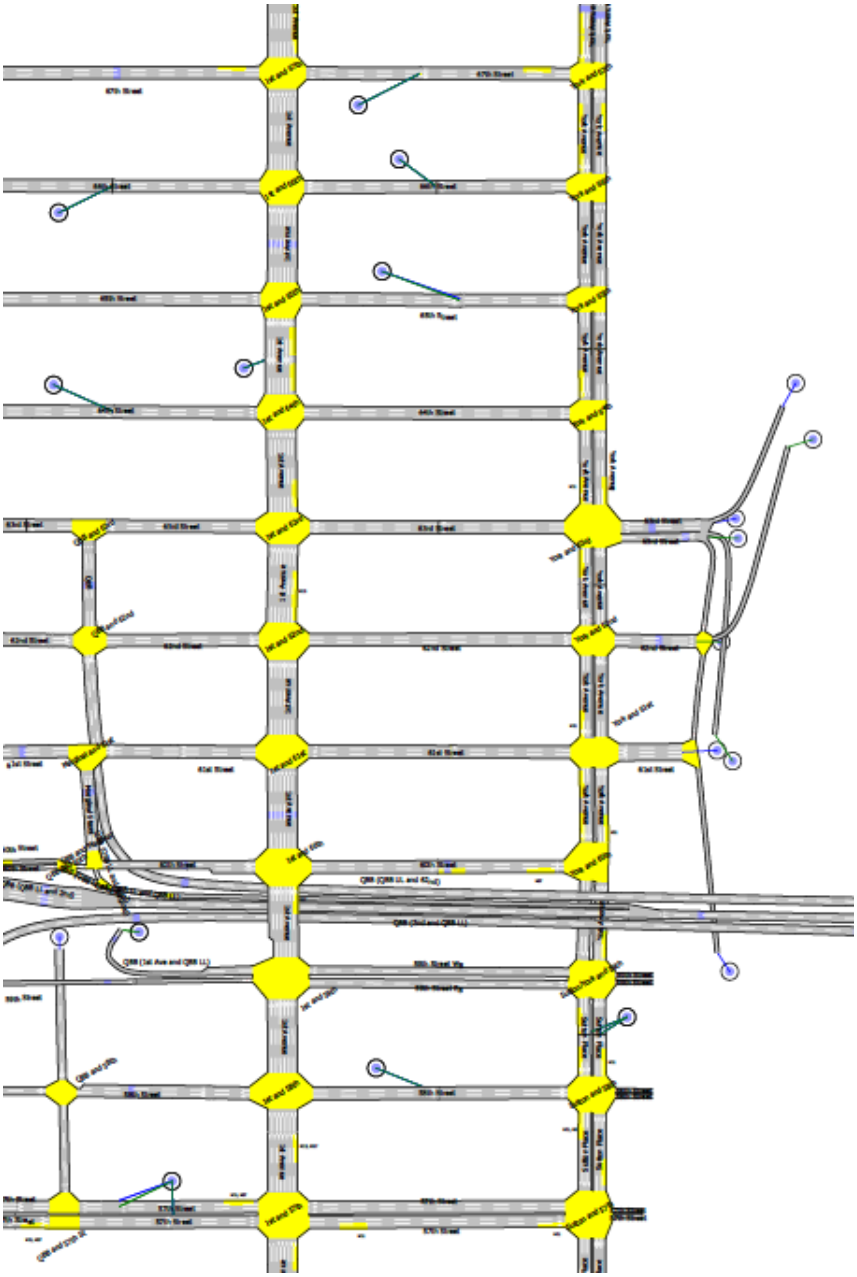


Figure 4-1: Topology of Queensboro bridge area.

4.2 Literature Review

Queue management is often used in congested network to form up control strategy. Michalopoulos and Stephanopoulos (1977) optimize signal plans for two congested intersections, they minimize delay at intersections subject to queue-length constraints. Abu-Lebdeh and Benekohal (1997) maximize system throughput for a three-intersection congested system. They use state equations to manage queue formulation and dissipation, in which the number of vehicles in the queues, and the number of vehicles arriving to and departing from the queues for each cycle are explicitly considered in order to ensure the upstream queues are not blocked when downstream queues build up; offsets and green splits change when demand and queue status change dynamically. Aboudolas et al. (2010) minimize links occupancy which is defined as the ratio of queue-length over time and maximum admissible queue-length subject to maximum admissible queue-length. They prove that considering queue-length in signal control problems helps to reduce the risk of queue spillback. Liu and Chang (2011) model dynamic evolution of physical queues as a function of signal timing, arrivals, departures over time, the control objective can be either minimizing total travel time or maximizing throughput. For a detailed review of queue management method and signal control strategy for congested network, see Quinn (1992) and Hajbabaie et al. (2011).

4.3 Methodology

The SO algorithm described in Chapter 2 is used to identify the signal plan for this area. In this SO framework, signal design objectives can be adjusted according to the needs of transport agencies such as incorporating reliability concerns described in Chapter 2 and 3, or enhancing system efficiency. The SO algorithm looks for signal plans that could improve the system performance for the whole studied area rather than individual intersection. The advantage of using such strategy is that under different traffic conditions, considering all links in calculating signal plans give us more potential to achieve an improvement for the whole area of interest. Signal changing might influence drivers' routing behavior; vehicle re-routing might influence the travel time again. Looking at the signal plan for an area rather than a set of intersections along a major road would help us to address the influence of the signal plan on drivers' behavior and the consequences it might bring back to the overall system performance.

In order to use the SO framework, the queuing model needs to be calibrated according to the network topology and flow level associated. For a detailed description of the calibration techniques, see Appendix C. All the links in the study area are represented as 284 queues.

To illustrate the congestion level of the studied area, we present a few details regarding the queue-lengths of the network of interest.

Figure 4-2 displays for each queue in the network its spillback probability under the

signal plan currently used in the field for morning peak demand. These probabilities are calculated as follows. We run 50 replications of the simulation model. For each replication and each queue, every three seconds we evaluate the vehicular queue-length. We use these queue-length measurements to estimate over the 8am-9am hour the proportion of time where spillback occurred. This proportion is obtained as an average over both the 8am-9am period of interest and over the 50 simulation replications. These proportions are used as estimates of the spillback probabilities. What we can see from Figure 4-2 is that there are various queues where spillback happens more than 50% of the time. More importantly, even for the queues where the spillback probability is low, the occurrence of spillback may have a significant effect on congestion propagation upstream due to the existence of short links. Once spillback happens, it spreads out quickly. This motivates the use of a signal control formulation that explicitly accounts for queue-length metrics.

In order to formulate the signal control problem, we introduce the following nota-

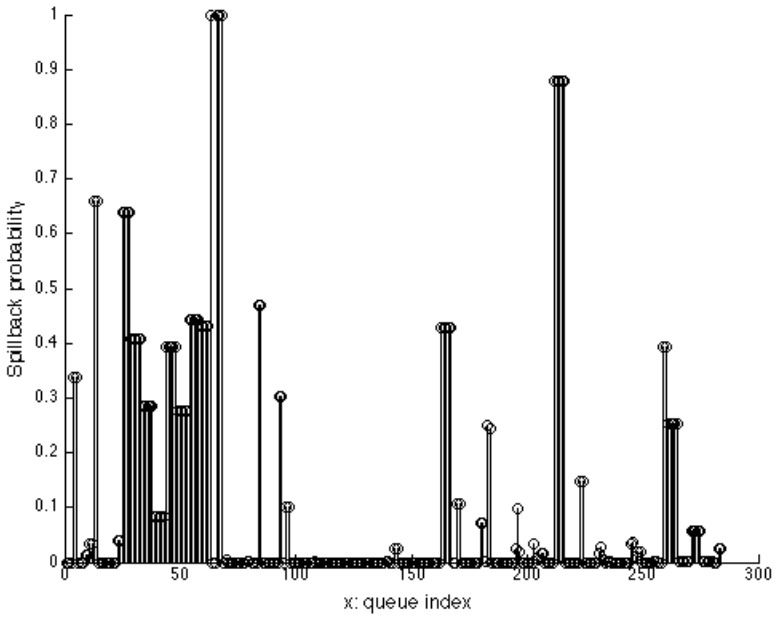


Figure 4-2: Spillback probability for each queue under the existing signal plan.

tion:

b_i	available cycle ratio of intersection i;
QL_l	queue-length of link l;
T	average trip travel time;
$x(j)$	green split of phase j;
x_L	vector of minimal green splits;
$\mathscr{L}$	set of links within the area of interest;
$\mathscr{I}$	set of intersection indices;
$\mathscr{P}_I(i)$	set of phase indices of intersection i;

For congested network, the signal control problem is formulated as follows:

$$\min_{x} f(x) = \sum_{l \in \mathscr{L}} E[QL_l(x, z; p)] \tag{4.1}$$

subject to

$$\sum_{j \in \mathscr{P}_I(i)} x(j) = b_i, \ \forall i \in \mathscr{I} \tag{4.2}$$

$$x \geq x_L. \tag{4.3}$$

This problem is a fixed-time signal control problem, where the decision variables x are the green splits. In this problem, the stage structure (e.g. phase sequence) is given, the offsets, the cycle times and the all-red durations are fixed. The performance metric used, $\sum_{l \in \mathscr{L}} E[QL_l(x, z; p)]$, is the sum of expected queue-lengths over all links. Constraints (4.2) guarantee that for a given intersection the available cycle time is distributed across all endogenous phases. Constraints (4.3) ensure lower bounds for the green splits. They are set to 5 seconds, and are based on current New York City Department of Transportation (NYCDOT) practices.

Recall that the metamodel formulation (described in Chapter 2) requires an analytical expression, which is the approximation of the objective function f as derived by the analytical queueing-theoretic model. Here, we present the analytical and differentiable expressions for f for congested and uncongested networks. We first derive

the objective function for congested network.

Let $\mathscr{Q}$ denotes the set of queues that represent the links, $\mathscr{L}$. Then, the objective function for congested network can be rewritten as a function of queue metrics, rather than link metrics:

$$\sum_{l \in \mathscr{L}} E[QL_l(x, z; p)] = \sum_{i \in \mathscr{Q}} E[N_i]. \tag{4.4}$$

We now present how an analytical expression for the expected queue-length of a queue, $E[N_i]$, is derived. We use the the same analytical queueing-theoretic traffic model described in Chapter 2, and the same notations for all variables (detailed in Appendix A.1.1). We recalled the notations for each variable:

γ_i	external arrival rate;
λ_i	total arrival rate;
$\hat{\mu}_i$	effective service rate;
k_i	space capacity;
$P(N_i = k_i)$	probability of queue i being full, known as blocking or spillback probability;
ρ_i	traffic intensity (defined as the ratio of arrival rate and effective service rate);
$E[N_i]$	expected queue-length.

In order to approximate the objective function f (of Equation (4.1)), we proceed

as follows. For a given queue i, its expected queue-length is defined as:

$$E[N_i] = \sum_{n=0}^{k_i} nP(N_i = n). \tag{4.5}$$

The stationary marginal queue-length probabilities $P(N_i = n)$ are obtained when evaluating the traffic model, they are given by:

$$P(N_i = n_i) = \frac{1 - \rho_i}{1 - \rho_i^{n_i+1}} \rho_i^{n_i}, n \in [0, k_i] \tag{4.6}$$

Combining ideas from Equations (4.5) and (4.6), we can obtain the following closed-form expression for $E[N_i]$:

$$E[N_i] = \rho_i \left(\frac{1}{1 - \rho_i} - (k_i + 1) \frac{\rho_i^{k_i}}{1 - \rho_i^{k_i+1}} \right). \tag{4.7}$$

Hence, the analytical approximation of the objective function (Equation (4.1)) is given by:

$$\sum_{i \in \mathcal{Q}} E[N_i] = \sum_{i \in \mathcal{Q}} \rho_i \left(\frac{1}{1 - \rho_i} - (k_i + 1) \frac{\rho_i^{k_i}}{1 - \rho_i^{k_i+1}} \right). \tag{4.8}$$

4.4 Performance of the proposed fixed-time signal plan

We start from the existing NYC signal plan and run the SO five times. The computational budget is set to 150 simulation runs each time. In total we derive five signal plans. Signal plan with the smallest total queue-length without deteriorating the system throughput is selected as the new signal plan. To evaluate the performance of the new signal plan derived by SO, we run the signal plan derived by SO and existing signal plan 50 replications respectively.

In order to provide a more detailed analysis of the performance of the derived signal plans and the existing signal plan, we consider both temporal evolution of a set of performance measures every 15 minutes and the aggregated performance over the simulation period.

After a warm-up period of 20 minutes, we consider the temporal evolution of the following 4 performance metrics every 15 minutes:

- average network queue-length over every 15 minutes;
- average trip travel time (including all finished and unfinished trips for that time period) over every 15 minutes;
- entry flow every 15 minutes;
- average spillback probability over every 15 minutes.

For each 15 minutes, average network queue-length is calculated as the average of the queue-length over all links in the network. To calculate average trip travel time, total network travel time experienced by all users (both finished and unfinished trips) is obtained over every 15 minutes, then average trip travel time is calculated as the ratio between total network travel time and total number of vehicles entered the network during 15 minutes. Entry flow is calculated as the total number of vehicle entered network for each 15 minutes, to be comparable with the total demand level, the entry flow every 15 minutes is then transformed to entry flow per hour (multiplied by 4). Average spillback probability is calculated as the average of the spillback probabilities over all queue in the network. In order to estimate the spillback probability, we measure queue-length for each link every 3 seconds. Then each link is mapped into a queue or a set of queues, unused links are not modeled in the queueing network. Note that the spillback probability of a queue can be interpreted as the proportion of time the queue remains full.

We then study the following performance measures over the whole studied period (1 hour):

- average network queue-length.
- average spillback probability;
- average trip travel time of finished trips;

- total number of finished trips;
- average trip travel time of unfinished trips;
- total number of unfinished trips;

Average network queue-length and average spillback probability are calculated in the same way as the performance measures mentioned above for temporal evolution study. Instead of calculating average network flow without distinguishing finished and unfinished trips, we calculate number of finished trips and unfinished trips respectively at the end of simulation period. The number of unfinished trips represents the number of vehicles blocked in the network, together with the number of vehicles that just enter the network and do not have enough time to finish the trip at the end of simulation. It is hard to distinguish these two types of vehicles in the simulator. Assume that under the same demand level, number of vehicles enter the network but do not have enough time to finish their trips at the end of simulation are similar for different signal plan, larger number of unfinished trips means more blocked vehicles.

Comparing to the average travel time of those blocked vehicles, average travel time for those vehicles that just enter the network and do not have time to finish the trip is very small. Normally, the average travel time for blocked vehicles are larger than average travel time of finished trips. When we compare the average travel time of unfinished trips obtained by different signal plan, we first compare the average travel time of unfinished trips with average travel time of finished trips to justify if blocking

happens.

For each signal plan and each performance metric mentioned above, we use the 50 observations obtained from 50 simulation replications. We use these 50 observations to construct a cumulative distribution function (cdf). We then perform a paired t-test to test the hypothesis that the performance measures obtained from the signal plan derived by SO are better than the performance measures obtained from the existing signal plan for each signal plan. The paired t-test is performed using script coded in Matlab.

Figure 4-3 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure 4-3(a), Figure 4-3(b), Figure 4-3(c), and Figure 4-3(d) show the comparison of average queue-length of the adaptive signal settings and the existing signal plan for each 15 minutes; Figure 4-3(e), Figure 4-3(f), Figure 4-3(g), and Figure 4-3(h) show the the comparison of average trip travel time.

Figure 4-4 shows the comparison of average spillback probability and entry flow for all time period. Figure 4-4(a), Figure 4-4(b), Figure 4-4(c), and Figure 4-4(d) show the comparison of the average spillback probability Figure 4-4(e), Figure 4-4(f), Figure 4-4(g), and Figure 4-4(h) show the comparison of entry flow.

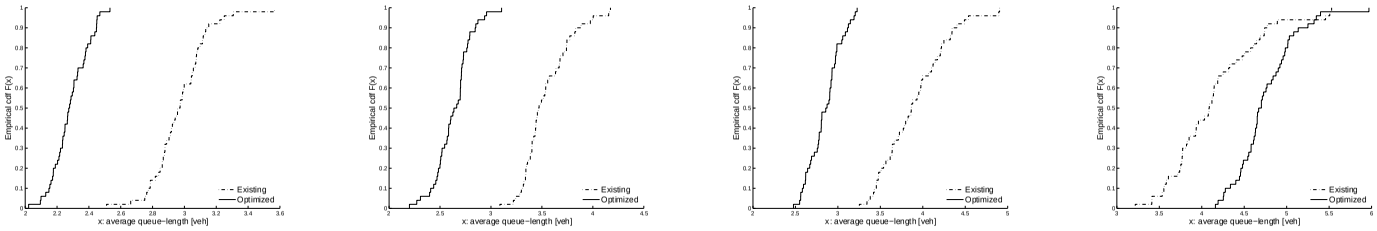
The performance of the proposed signal plan is displayed in solid line (indicated as optimized); the performance of the existing signal plan is displayed in dashed line

(indicated as existing). For all time intervals, the signal plan derived by SO yields smaller average trip travel time, and smaller average spillback probability. For the first and second time interval, proposed signal plan leads to similar entry flow to existing signal plan, and for the third and fourth time interval, proposed signal plan leads to significantly larger entry flow with smaller variability. For the first, second and third time intervals, the signal plan derived by SO yields smaller average queue-length, for the fourth time interval, proposed signal plan leads to larger queue-length.

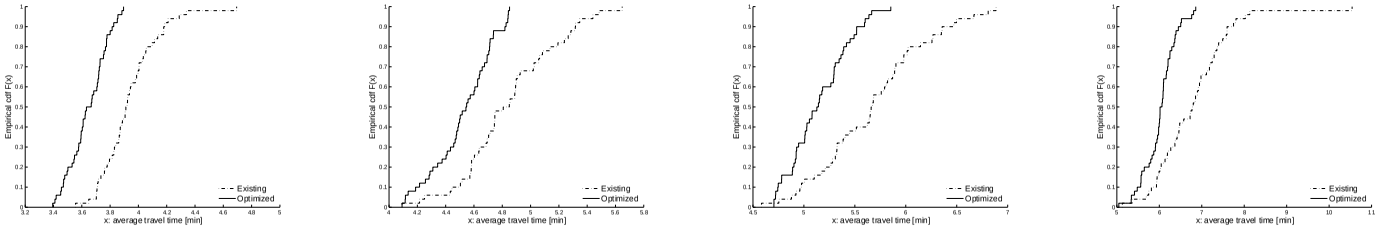
The aggregated performance over the simulating period is shown in Figure 4-5. Figure 4-5(a) shows the average queue-length; Figure 4-5(b) shows the average spillback probability; Figure 4-5(c) shows the average travel time of unfinished trips; Figure 4-5(d) shows the average travel time of all finished trips; Figure 4-5(e) shows the number of unfinished trips; Figure 4-5(f) shows the number of finished trips. proposed signal plan leads to better performance in terms of all performance measures.

We use across-replication variability to represents the day-to-day variability in performance metrics. For average finished trip travel time, number of finished trips, and average spillback probability, the cdf curves correspond to proposed signal plan are steeper, which means it leads to more stable system performance.

Table 4.1 shows the statistics of each performance measure over 50 replications. TT represents average trip travel time of finished trips; TP is the number of finished trips; TTun is the average trip travel time for unfinished trips; TPun is the number of

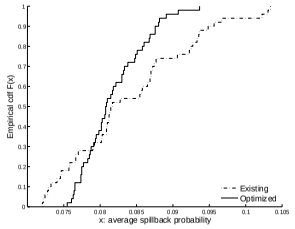


(a) average queue-length for the first 15 mins (b) average queue-length for the second 15 mins (c) average queue-length for the third 15 mins (d) average queue-length for the fourth 15 mins

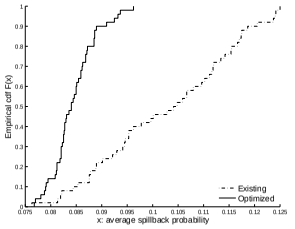


(e) average trip travel time for the first 15 mins (f) average trip travel time for the second 15 mins (g) average trip travel time for the third 15 mins (h) average trip travel time for the fourth 15 mins

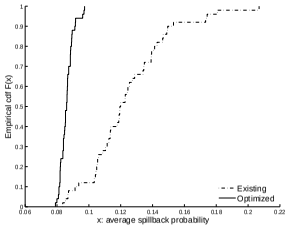
Figure 4-3: Comparison of the average queue-length and average trip travel time of proposed signal plan and the existing signal plan.



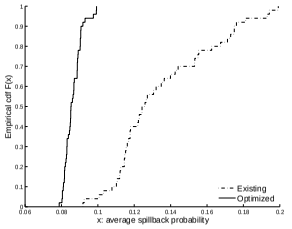
(a) average spillback probability for the first 15 mins



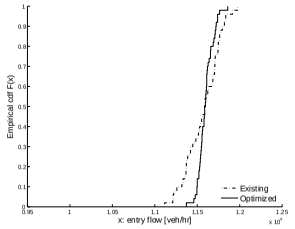
(b) average spillback probability for the second 15 mins



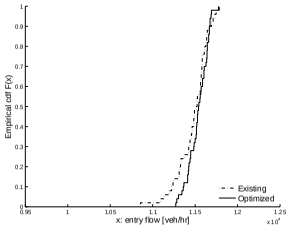
(c) average spillback probability for the third 15 mins



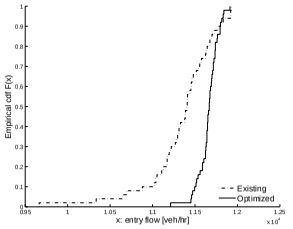
(d) average spillback probability for the fourth 15 mins



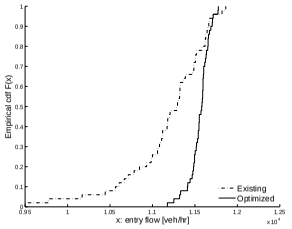
(e) entry flow for the first 15 mins



(f) entry flow for the second 15 mins

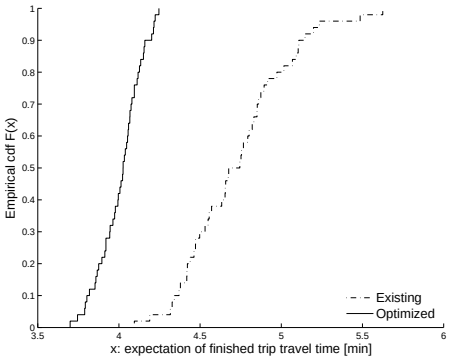


(g) entry flow for the third 15 mins

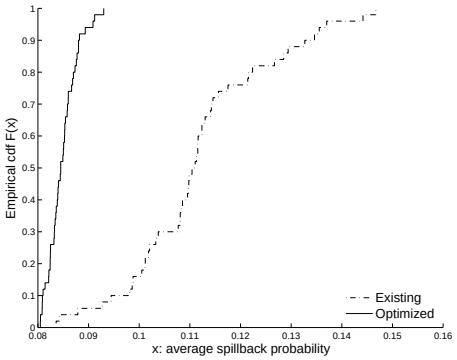


(h) entry flow for the fourth 15 mins

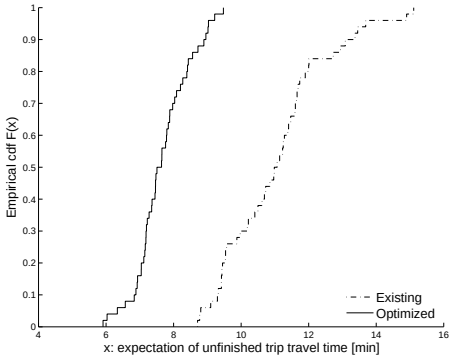
Figure 4-4: Comparison of the average spillback probability and entry flow of proposed signal plan and the existing signal plan.



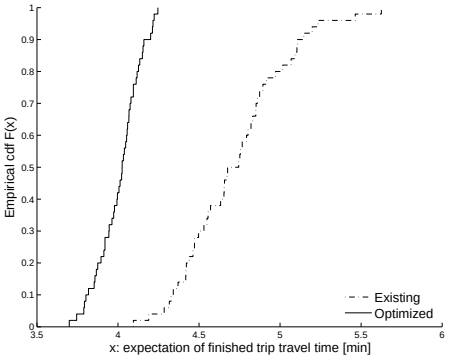
(a) average queue-length



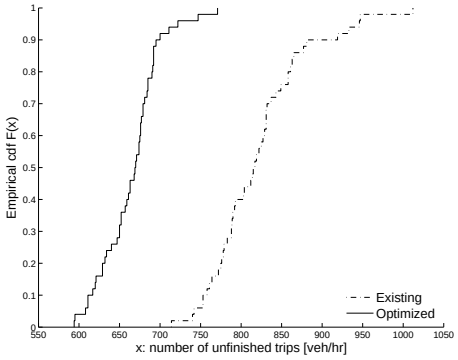
(b) average spillback probability



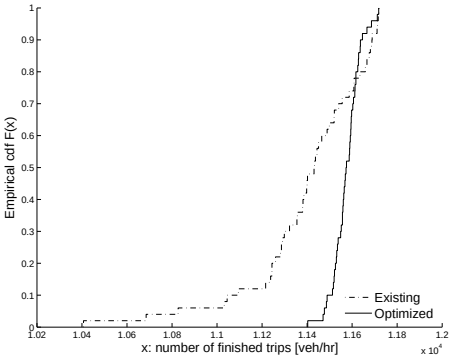
(c) expectation of unfinished trip travel time



(d) expectation of finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure 4-5: Comparison of the performance of the proposed signal plan and the existing signal plan.

	Existing plan				New plan			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	4.10	4.72	5.63	0.32	3.70	4.01	4.25	0.13
TP	10407	11400	11719	267.42	11403	11576	11718	60.86
TTun	8.71	11.03	15.11	1.52	5.91	7.67	9.48	0.80
TPun	714	821	1012	58.81	594	664	771	35.93
QL	3.07	3.67	5.03	0.32	2.41	2.69	2.99	0.15
SP	0.0836	0.1119	0.1467	0.0141	0.0805	0.0848	0.0930	0.0029

Table 4.1: Performance metrics statistics for proposed signal plan and existing signal plan.

unfinished trips; QL is the average network queue-length; SP is the average spillback probability. Average trip travel time for unfinished trips is larger than average trip travel time for finished trips, thus blocking happens. The new plan derived by SO reduces the average travel time for those travelers that are blocked in the network. The new plan also yields smaller average queue-length, smaller average spillback probability, smaller average finished trip travel time, smaller number of unfinished trips, and larger number of finished trips.

Table 4.2 shows the paired t-test for each performance measure. For each aggregated performance measure, the null hypothesis states that the performance of the signal plan proposed by SO method is equal to that of the existing signal plan, the alternative hypothesis states that the performance of the signal plan proposed by SO method is better (e.g. smaller average queue-length; smaller average spillback probability; smaller average travel time for unfinished trips; smaller average travel time for finished trips; smaller number of unfinished trips and lager number of finished trips)

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	0.7153	0.3572	14.1572
TP	176.4800	275.4490	4.5304
TTun	3.3602	1.6915	14.0472
TPun	156.5600	66.3224	16.6919
QL	0.9831	0.3528	19.7036
SP	0.0271	0.0151	12.6513

Table 4.2: Paired t-test for proposed signal plan and existing signal plan.

than that of the existing signal plan. When running the 50 simulation replications to evaluate the performance of a given signal plan, we use the same set of 50 replication seeds for each signal plan. The paired t-tests are carried out by pairing observations that have common seeds. Let $\overline{Y}$ denote the average paired difference between any two aggregated performance measures , let $\hat{s}$ denote its standard deviation, and let O denote the sample size. Then the paired t-statistic is given by (see, for instance, Hogg et al. (1977, p. 486)): $t = \sqrt{O}\ \overline{Y}/\hat{s}$.

Taking the average finished trip travel time as an example, we test the hypothesis that the average finished trip travel time from the proposed signal plan is equal to the average finished trip travel time obtained from existing signal plan. The mean of the paired differences $\overline{Y}$ is around 0.7153 minutes. The standard deviation of the paired differences $\hat{s}$ is around 0.3572 minutes. The sample size O is 50. The critical value at the 2.5% significance level is $t_{0.025}(49) = 2.01$. The t values is 14.1572. Thus the null hypothesis is rejected. For all the other performance metrics, the t-values are larger than $t_{0.025}(49) = 2.01$, thus the null hypothesis is rejected. Proposed signal

plan derived by SO leads to significant smaller queue-length, smaller average spillback probability, smaller average trip travel time for both finished and unfinished trips, smaller number of unfinished trip, and larger number of finished trips.

Improving total system throughput (number of finished trips) while reducing number of travelers being blocked is not a simple task. The proposed signal plan leads to larger number of finished trips which means the system throughput is increased, more travelers could pass the network with a reduced average trip travel time. Compared to the existing signal plan, the proposed signal plan also reduces the number of unfinished trips which includes the travelers that are blocked in the network, and reduces the time spent in the network for them.

Since both major and minor links are important in this area, to visualize the improvements obtained by proposed signal plan under normal morning peak demand for each link, we proceed as follows: for the existing signal plan and the new signal plan, we estimate for each link the following performance metrics:

- average link queue-length;
- average link travel time.

Average link queue-length is calculated as the average queue-length for each link over the whole simulation period, we then calculate the average over 50 simulation replications. Similarly, the average link travel time is calculated as the average link travel time over the simulation period, then for 50 replications.

We use the ratio of performance measure obtained from the proposed signal plan and the existing signal plan. In terms of both measures, a smaller ratio means a larger improvement. We classify the ratio into 4 levels:

- more than 20% reduction (green);
- less than 20% reduction (dark green);
- increased less than 20% (orange);
- increased more than 20% (red).

Figure 4-6 displays the results for the average link queue-lengths. The majority of the links, and in particular almost all cross street links (minor streets), are marked by green and dark green. This indicates a reduction in their average queue-length.

Figure 4-7 displays the results for the link travel times. Almost all links are marked by green and dark green. This indicates a reduction in their average travel time.

4.5 Conclusion

In this chapter, we address a signal control problem for a highly-congested area in eastern Manhattan (New York City, USA), where spillbacks frequently occur. The network has complex traffic dynamics due to its multimodal congested traffic, short links, numerous signalized intersections and grid-type topology. For such networks it is a great challenge to design signal plans that mitigate the spatial and temporal

Figure 4-6: Average queue-length: ratio between proposed plan and existing plan.

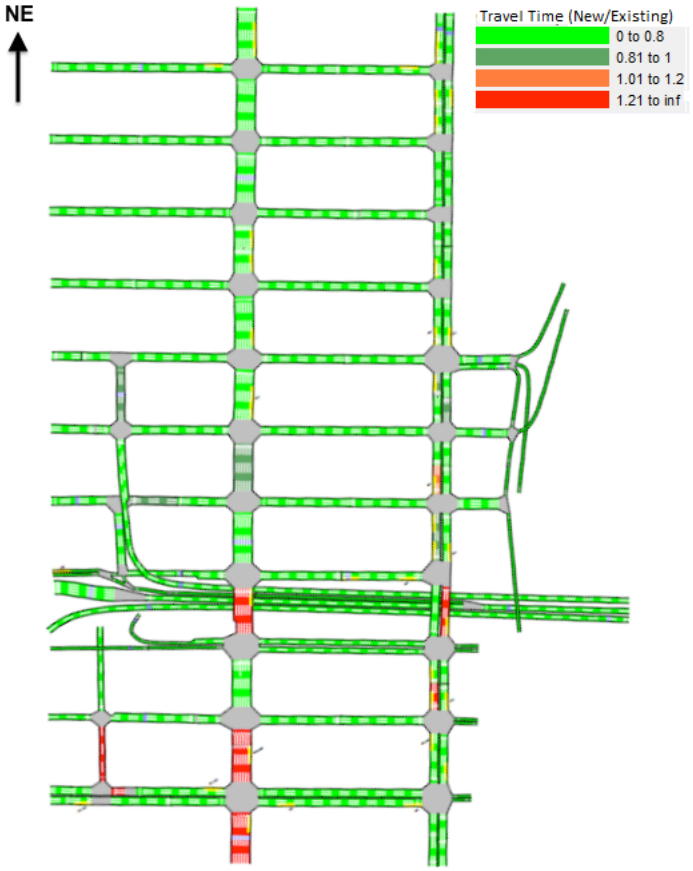


Figure 4-7: Average link travel time: ratio between proposed plan and existing plan.

propagation of congestion. The performance of the proposed plan is compared to that of the existing plan for that area. The proposed plan yields significant improvements when evaluated with various performance metrics. In future research, queue-length can be scaled by link length for each queue in the objective function to limit spillback probability directly.

Chapter 5

Simulation-based adaptive traffic signal control algorithm

5.1 Introduction

In Chapter 4, fixed-time signal plan control strategy has been proposed for the congested Queensboro bridge area. As what has been stated before, design signal control strategy for congested grid-type network is a very challenging task. Compared to the traditional pre-timed signal control strategies, adaptive traffic control systems (ATCSs) provide a more flexible option for adjusting signal timings to accommodate changing traffic variations.

The purpose of this chapter is to design an adaptive simulation-based optimization algorithm for such type of network. Different traffic conditions that vary from

light traffic to oversaturated condition are explicitly modeled in the highly detailed stochastic microscopic traffic simulators via a set of demand scenarios. We first consider the design of fixed-time signal plans under each demand scenario to form up a set of candidate signal plans. We then propose a simulation-based adaptive traffic control algorithm to select signal plan for different time periods (e.g. every 15 minutes) based on simulation observations. Two case studies are used to evaluate the performance of the proposed algorithm. To apply the proposed algorithm in reality, real-time field data can be used instead of the simulation observations in adjusting signal plans.

This chapter is structured as follows: Section 5.2 presents a review of current adaptive traffic signal control systems (ATSCs). We then present the methodology in Section 5.3. We explain how the proposed algorithm can be applied to the subnetwork of Manhattan in Section 5.4. We evaluate the performance of the algorithm in Section 5.5 through two case studies. We conclude with a brief discussion in Section 5.6.

5.2 Literature Review

Adaptive traffic control systems (ATCSs) adjust signal timings according to real time traffic information. In most cases, ATCSs are considered as effective ways of reducing travel time, delays and number of stops (Stevanovic, 2010). Usually it utilizes measurements of traffic volume and occupancy data. In reality, the signal plans are adjusted by using softwares such as SCOOT (Split Cycle Offset Optimization Technique)

(Hunt et al., 1982), SCATS (Sydney Coordinated Adaptive Traffic System) (Sims and Dobinson, 1979), OPAC (Optimization Policies for Adaptive Control) (Gartner, 1983), and RHODES (Real Time Hierarchical Optimized Distributed and Effective System) (Head et al., 1992).

SCOOT and SCATS are the most widely used adaptive traffic control softwares. SCOOT adjusts signal timing with small (several seconds each time) changes, the loop detectors measure traffic volume and occupancy information each second and send them to the central controller to estimate the real-time flow. SCOOT has three optimizers that optimize green split, offset and cycle time. Each optimizer estimates the impact of a small change on the overall performance (a weighted measure of delay, stops at individual link level) of the area of interest to decide if the signal plan will be adjusted. SCOOT adjusts the cycle time to maintain degree of saturation (flow/capacity) below 90% for each movement. SCATS groups intersections into a few subsystems with a critical intersection in each. The signal plans for each subsystem are mainly determined by the signal setting at the critical intersection. Then each subsystem will be coordinated with adjacent subsystems to maintain traffic platoons of vehicles. There are two levels of control: strategic and tactical. Strategic control adjusts green split, offset and cycle time for each subsystem; tactical control determines if the phase needs to be terminated earlier or even omitted at each individual intersection. These two softwares share around 67% of the ATCSs market in the

US (Stevanovic, 2010). More recently, new ATCSs use mathematical programming techniques to calculate signal plans such as OPAC. Based on predefined stages, these strategies calculate the optimal values of next switching times (red-green switching) that minimize the overall vehicle delays obtained by simple traffic model. For the global optimization of the performance function (total delay), OPAC uses a complete enumeration (red-green switching time) method.

ATCSs require extensive amount of detectors, and the infrastructures need to allow communications between central and/or local processors. Due to the high operating and maintenance costs of these equipments. In a recent survey, it has been found that in the US, less than 1% of existing traffic signals are using ATCSs based on real time information (Hagemann et al., 2010). Thus simpler adaptive control software ACS-Lite (Luyanda et al., 2003) emerges. ACS-Lite adjusts coordinated signal plans along corridors by changing phase duration and offset every 5-10 mins (cycle times are fixed). For a more detailed description of the characteristics and operating logic behind all methods discussed above, see Stevanovic (2010).

Although it has been widely accepted that the deployment of ATCSs helps to reduce delay, there are still some challenging situations for those systems to handle. One is their application in grid-topology networks; the other is their ability to cope with oversaturated traffic condition.

After the deployment of the ATCS in the grid-topology network, overall system

performance (e.g. total delay) is better than before: improvements are obtained for the major streets whereas the delay on minor streets increased (Hutton et al., 2010). Furthermore, in urban grid-type network with high volume of pedestrians, the deployment of ATCS results in pedestrian delays (green time duration assigned to pedestrian is reduced in order to assign more time to motorized vehicles) which might offset the benefit the ATCS brings (Hu, 2014).

In oversaturated networks, the benefit of using ATCS is more controversial: in a survey carried out by Stevanovic (2010), only 3% of the interviewed agencies that operate ATCSs consider such systems could help to prevent and eliminate oversaturated situations, over one third of the interviewed users thought that it worsens the traffic. When some links or a set of intersections are oversaturated, the ATCSs might skip stages or extend phases to allow more green time for those links with large flow. The delay for the main street is reduced at the expenses of the side streets, and the overall delay might increase (Martin, 2007).

In reality, to investigate the performance of ATCS, 89% of the ATCSs are evaluated on field through before-and-after study. Using microscopic simulation to evaluate the performance of ATCS before install it is very rare due to the complexity of incorporating ATCS software with detailed simulator and the high cost of modeling traffic conditions in microscopic simulation (Stevanovic, 2010), let alone using microscopic simulator to design ATCSs.

In this chapter, we propose a simulation-based adaptive traffic signal control algorithm to design signal setting for highly congested grid-type urban networks without imposing extra delay for minor street users. Due to the high volume of pedestrian traffic, the green time assigned to them are fixed, thus we cannot assign more green times to motorized traffic by reducing the green time assigned to pedestrians. Unlike the traditional ATCSs that adjust signal plan based on the flow observed at individual intersection, the proposed algorithm looks for signal plans that could improve the system performance for the whole studied area, which give us more potential to achieve an improvement for the whole area of interest.

5.3 Methdology

In this section, we propose a simulation-based adaptive traffic control algorithm that can be used for highly congested grid-type urban networks. The conceptual structure of the algorithm can be described as follow:

- Step 1: specify traffic condition into different levels according to historical data (e.g. flow, speed) from light traffic to heavy traffic;
- Step 2: derive signal plans using the simulation-based optimization (SO) framework described in Chapter 2 for each traffic condition;
- Step 3: build look-up tables using simulator for each proposed signal plan un-

der different traffic conditions. The look-up table includes the information of the performance metrics (e.g. link travel time, speed) under different traffic condition;

- Step 4: use the proposed adaptive traffic signal control algorithm to adjust signal plans. The proposed algorithm divides the studied time period (e.g. morning peak) into several time periods, and selects signal plans according to the observations we obtained from the simulator for each time period. Based on the selection, we forecast the influence of changing to a new plan for the next time period. If switching plan results in worse system performance than not switching, no changes will be made, otherwise, signal plans are switched.

In the next sub-sections, we describe each step in greater details to show how it can be applied to a congested urban road network with grid-type topology.

5.3.1 Specify traffic conditions

Normally, traffic conditions are classified into different congestion levels based on historical data of flow or travel time. Thus in step 1, for different traffic levels such as light traffic (e.g. weekend), moderate traffic (e.g. off-peak period of weekdays), heavy traffic (e.g. peak period) and very severe congestion (e.g. demand grows and congestion lasts, spillbacks happen, congestion propagates spatially and blocks the adjacent streets), a fully calibrated microscopic simulator is needed.

5.3.2 Derive signal plans for each traffic condition

In step 2, the SO algorithm described in Chapter 2 is used to identify the best signal timing for each demand level. For uncongested and congested network, signal design objective function differs. The objective function that is suitable for uncongested network (minimization of average travel time is the most common design objective) might not be appropriate for congested network. When we are facing a set of traffic conditions with different demand levels, different signal design objectives should be used.

A general simulation-based signal control problem can be formulated as follows:

$$\min_{x \in \Omega} f(x) = E[F(x, y; p)], \tag{5.1}$$

where the decision vector x represents the signal control variables (e.g. green times), and the objective function is the expected function of a stochastic network performance metric F (e.g. link speeds, trip travel time), which depends on x as well as on other endogenous simulation variables y (e.g. link flow capacities, route choice probabilities) and exogenous (i.e., fixed) simulation parameters p (e.g. dynamic origin-destination matrices, network topology, transit network). The feasible region Ω is typically a set of analytical differentiable constraints and bound constraints.

5.3.3 Look-up table creation

A set of look-up tables will be built. For each signal plan, there is a look-up table associated. The reason we build one look-up table for each signal plan is that the performance metrics such as link travel time are influenced by demand and supply. Not only demand levels but also signal settings will influence the link travel time. Under the same demand level, different signal plans will have different performance in terms of link travel time. To infer the traffic condition from system performance measure such as link travel time, both demand and supply play an important role on it. We will further justify the use of multiple look-up tables in Section 5.4 with an example.

For each traffic condition, we use microscopic simulator to reproduce the day-to-day variability in traffic dynamics. By running the simulator, we obtain detailed performance measures such as average trip/link travel time, average queue length, etc..

Let $\mathscr{S}$ denote the set of links of interest for the studied network. It could be the set of links that has detection equipments.

Taking average link travel time as an example, performance measure can be calculated as the average value over the studied period (e.g. one hour morning peak).

$$\sum_{l \in \mathscr{S}} E[TT_l(x, z; p)]. \tag{5.2}$$

$E[TT_l(x, z; p)]$ is the average link travel time for link l. Assume we have G traffic conditions and G signal plans, traffic conditions and their corresponding signal plans are ordered from demand scenario with lowest demand (e.g. demand scenario 1) to highest demand (e.g. demand scenario G). For each signal plan t, a look-up table is constructed.There are several steps of building the table:

- Step a. For signal plan t, run simulator 300 replications to obtain a vector PM_j^t for each traffic condition j ($j \in [1, G]$), PM_j^t contains 300 observations of the selected performance measure such as average link travel time (given in Equation 5.2) under traffic condition j;

- Step b: Define boundary value b_j^t of the performance measure between traffic condition j and traffic condition $j+1$ according to:

  $\min[P(X_j^t > b_j^t) + P(X_{j+1}^t < b_j^t)], j \in [1, G-1]$;

  in which, X_j^t and X_{j+1}^t are the variables of average link travel time associated with the j^{th} and $j+1^{th}$ demand scenarios;

- Step c: Set the lower and upper bound of average total link travel time for traffic condition j as:

  $[b_{j-1}^t, b_j^t), j \in [1, G]$,

  in which $b_0^t = 0$, $b_G^t \approx \infty$.

  When there are G signal plans, there will be $G+1$ boundary values from b_0^t to b_G^t, and G link travel time intervals.

In step b, $P(X_j^t > b_j^t)$ represents the probability that the variable X_j^t takes a value larger than b_j^t; $P(X_{j+1}^t < b_j^t)$ represents the probability that the variable X_{j+1}^t takes a value less than b_j^t. $P(X_j^t > b_j^t)$ can be calculated as the number of the observations that are larger than b_j^t over the total observation number 300. $min[P(X_j^t > b_j^t) + P(X_{j+1}^t < b_j^t)]$ minimizes the summation of the probability that these two curves overlap each other in interval j and $j+1$.

We show a simple example here to further explain the algorithm. In Figure 5-1, there are two cdf (cumulative distribution function) curves, the x-axis shows the total average link travel time. The two cdf curves are obtained from the simulator by using the same signal plan t under different demand levels. The vertical line classifies the boundary between the first and second interval. The length of the line marked by red on top shows the probability that variable X_1^t on first curve takes a value greater than b_1^t: $P(X_1^t > b_1^t)$ (the probability that the first curve enters the second interval). The length of the line marked by red at bottom shows the probability that variable X_2^t on second curve takes a value smaller than b_1^t: $P(X_2^t < b_1^t)$ (the probability that the second curve enters the first interval). $\min[P(X_1^t > b_1^t) + P(X_2^t < b_1^t)]$ minimizes the summation of the probability that these 2 curves overlap each other in the first and second interval.

To calculate b_1^t, we move the vertical line from the smallest value on curve 1 to the largest value on curve 2 to calculated the summation of $P(X_1^t > b_1^t)$ and $P(X_2^t < b_1^t)$.

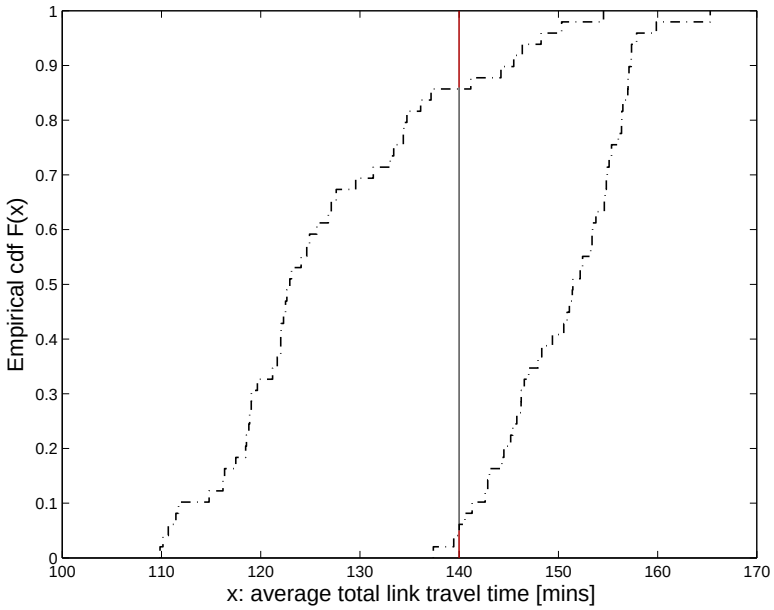


Figure 5-1: Obtaining boundary values.

To calculate b_j^t, we follow the steps as follows:

- Step a. set up a small step size s, which depends on the magnitude of X_j^t;
- Step b. set $n = 0$;
- Step c. start from $minPM_j^t$, calculate the summation of $P(X_j^t > minPM_j^t + n * s) + P(X_{j+1}^t < minPM_j^t + n * s)$,
- Step d: $n = n + 1$;
- Step e: if $minPM_j^t + n * s < maxPM_{j+1}^t$, go back to step c; otherwise, continue;
- Step f: find $minPM_j^t + n * s$ that has the smallest value of $P(X_j^t > minPM_j^t + n * s) + P(X_{j+1}^t < minPM_j^t + n * s)$;

- Step g: set b_j^t to $minPM_j^t + n * s$

For each signal plan, under each traffic condition there is a lower bound and an upper bound of the performance measure to classify the boundary values. The values between the lower bound and upper bound represent the day-to-day variability of the performance measure.

5.3.4 Simulation-based adaptive traffic signal control algorithm

In step 4, a simulation-based adaptive traffic signal control algorithm is used. In our algorithm, we divide the simulation period into several time intervals with 15 minutes each. Take the one hour morning peak period (8am-9am) as an example, it is divided into 4 time periods: TP_1, TP_2, TP_3, and TP_4. TP_1 represents the first time period from 8:00am to 8:15am; TP_2 represents the second time period from 8:15am to 8:30am; TP_3 represents the third time period from 8:30am to 8:45am; TP_4 represents the last time period from 8:45am to 9:00am. SP_i is the signal plan selected for each time period. TB_i is the look-up table associated with signal plan SP_i.

If the studied network uses fixed-time signal control strategy, there will be a signal plan in hand. The algorithm starts from this existing signal plan (SP_1). Then the proposed algorithm selects a competing signal plans (CP_1) by matching the observations we obtained from the simulator with the look-up table. Based on the selection, we forecast the influence of changing to a new plan (CP_1) for the second time period.

If switching plan results in worse system performance than not switching, no changes will be made. This process will be iterated until the end of the studied period.

We denote the signal plan selected for time period i as SP_i, we evaluate its performance several times for TP_i to obtain a vector of performance measure under signal plan SP_i, and denote it as $PM_i^{SP_i}$. Then the average value $\overline{PM}_i^{SP_i}$ of $PM_i^{SP_i}$ is used to judge the traffic condition j and select a competing plan CP_i to be used for that traffic condition j. To forecast the performance of SP_i and CP_i under traffic condition j, we evaluate their performance several times respectively in the simulator to obtain vectors of performance measure $PM_{i+1}^{SP_i}$ under signal plan SP_i and $PM_{i+1}^{CP_i}$ under signal plan CP_i. $\overline{PM}_{i+1}^{SP_i}$ is the average value of $PM_{i+1}^{SP_i}$ for time period $i+1$, $\overline{PM}_{i+1}^{CP_i}$ is the average value of $PM_{i+1}^{CP_i}$ for time period TP_{i+1}.

Taking the the one hour morning peak (8am-9am) as an example, we have:

0. Initialization.

- Set a demand level in the simulator;
- set i=1;
- for the first 15 minutes, signal plan SP_i is set to existing signal plan.

1. Select CP_i.

- Run the simulator 50 replications;
- calculate the average value of $PM_i^{SP_i}$ over 50 replications for TP_i as: $\overline{PM}_i^{SP_i} = \sum_{n=1}^{50} PM_i^{SP_i}(n)/50$;
- go to look-up table associated with signal plan SP_i;
- find traffic condition j such that $\overline{PM}_i^{SP_i} \in [b_{j-1}, b_j)$;
- select the signal plan j as CP_i (competing plan);
- if CP_i is different from SP_i, go to step 2; otherwise, go to step 3.

2. Forecast the performance of CP_i and SP_i under traffic condition j, and select SP_{i+1} for TP_{i+1}.

- Set signal plan SP_i and traffic condition j in the simulator, and run the simulator 50 replications;
- set signal plan CP_i and traffic condition j in the simulator, and run the simulator 50 replications;
- calculate $\overline{PM}_{i+1}^{SP_i}$ under traffic condition j as: $\overline{PM}_{i+1}^{SP_i} = \sum_{n=1}^{50} PM_{i+1}^{SP_i}(n)/50$;
- calculate $\overline{PM}_{i+1}^{CP_i}$ for TP_{i+1} as: $\overline{PM}_{i+1}^{CP_i} = \sum_{n=1}^{50} PM_{i+1}^{CP_i}(n)/50$;
- if $\overline{PM}_{i+1}^{SP_i} <= \overline{PM}_{i+1}^{CP_i}$, set SP_{i+1} to SP_i; otherwise, set SP_{i+1} to CP_i;

- set $i = i + 1$;
- if $i < 4$, go to step 1; otherwise, stop.

3. Set SP_i for TP_{i+1}.

- Set SP_{i+1} to SP_i;
- set $i = i + 1$;
- if $i < 4$, go to step 1; otherwise, stop.

For any network that uses fixed time signal control strategy, we start from the existing signal plan. Based on the simulation observations obtained every 15 minutes, the proposed algorithm classifies the current traffic condition and suggests the signal plan to use for the next time period.

5.4 Case study

We apply the proposed algorithm to the highly congested area around Queensboro Bridge in east Manhattan. We consider part of the morning peak-period 8am-9am. Since the historical data of different levels of congestion is not available, we set up a set of demand scenarios to represent different traffic conditions from light traffic to

oversaturated traffic. For each demand scenario, we calculate the optimal signal plan that could provide significant reduction in travel time, queue length, and spillback probability without deteriorating the system throughput. Then we select signal plans according to the observations we obtained from the simulator for each time period. Based on the selection, we forecast the influence of changing to a new plan for the next time period. If switching plan results in worse system performance than not switching, no changes will be made.

The topology of the studied Queensboro bridge area is shown in Figure 5-2. To optimize the signal plan for each demand level, all links and intersections inside this area are considered. The links marked by red rectangular are particularly important (identified by NYCDOT and in the future detection equipments might be installed for those links). Travel times along those links inside the red rectangular are used to create the look-up table to classify the traffic condition for each signal plan.

Demand scenarios

In simulator, demand is scaled into different levels to represent different levels of congestion. Besides the morning peak demand (scenario 4), we build 6 additional demand scenarios that are calculated as:

- Scenario 1: 70% of morning peak demand;
- Scenario 2: 80% of morning peak demand;

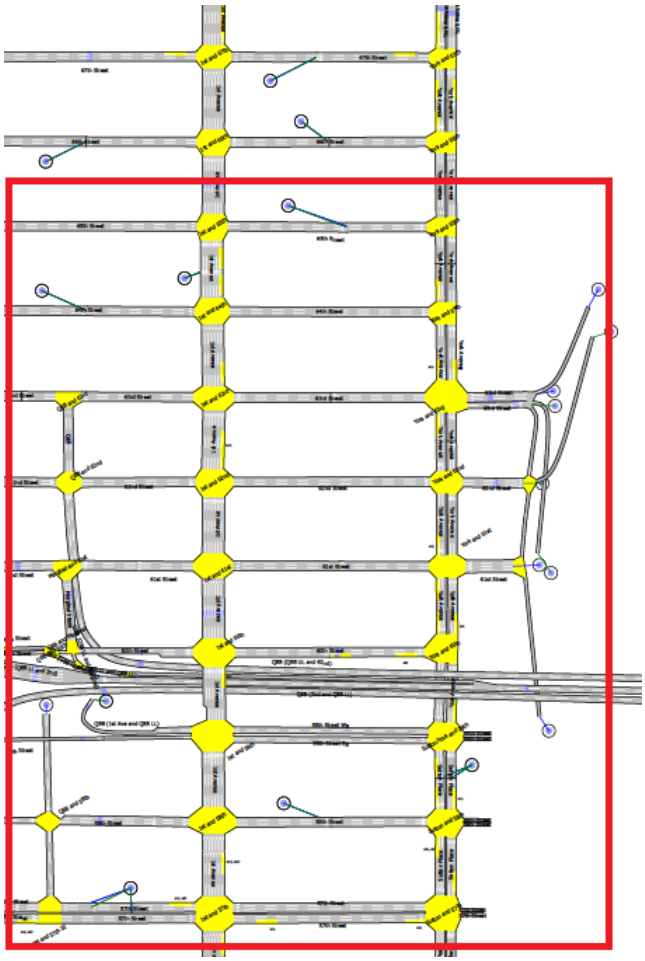


Figure 5-2: Topology of Queensboro bridge area.

- Scenario 3: 90% of morning peak demand;
- Scenario 4: morning peak demand;
- Scenario 5: 110% of morning peak demand;
- Scenario 6: 120% of morning peak demand;
- Scenario 7: 130% of morning peak demand;

Signal design for different demand scenarios

The study area of interest is a Manhattan subnetwork that consists of a total of 134 roads, 313 lanes, 27 signalized intersections and 5 non-signalized intersections. We consider part of the morning peak-period 8am-9am. For each demand scenario, using the existing signal plan, we recalibrate the queueing model. For a detailed description of the queueing model calibration, see Appendix C.

In this part, we observed that different signal control design objectives are suitable for different demand levels. For demand scenarios with less demand than the normal morning peak demand (scenario 4), minimizing average trip travel time yields signal plan with best performance in terms of average trip travel time, average queue length, system throughput and spillback probability. Signal plans derived by minimizing total queue length do not outperform the existing signal plan for demand scenarios with light traffic. For demand scenario 4 and scenarios with higher demand levels, minimizing

average trip travel time results in a significant reduction in the system throughput. In this case study, for congested network, minimizing total average queue length yields signal plans with better system performance without deteriorating throughput.

For each demand scenario, we start from the existing NYC signal plan and run the SO algorithm 5 times. The computational budget is set to 150 simulation runs each time. In total we derive five signal plans. To evaluate the performance of the signal plans derived by SO, we run each signal plan 50 simulation replications and compare the average trip travel time, system throughput, average queue-length, and spillback probability with existing signal plan. Signal plan with the smallest average trip travel time without deteriorating the system throughput will be selected as the new signal plan. If all signal plans derived by SO yield larger average trip travel time or smaller system throughput, the existing signal plan will be used for that demand level. For a detailed description of the formulation of the objective functions, and the performance of each derived signal plan, see Appendix D.

We have new plans for demand scenario 1, 3, 4, 5, 6. For demand scenario 2 and 7, the signal plans derived deteriorate the system throughput, thus we stick to existing signal plan. We name signal plan for each demand scenario from plan 1 to plan 7, plan 2 and plan 7 are the same.

- Plan 1 (70% of morning peak demand);
- Plan 2 (80% of morning peak demand);

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	29.70	31.00	35.26	1.21	(0, 34)
scenario 2	32.24	39.94	73.10	5.45	[34,61)
scenario 3	63.98	98.17	127.46	15.97	[61,126)
scenario 4	97.68	148.43	207.94	21.27	[126,172)
scenario 5	102.13	182.39	250.47	26.30	[172, 207)
scenario 6	131.58	211.68	301.75	33.15	[207,240)
scenario 7	163.43	257.20	363.97	34.52	[240, inf)

Table 5.1: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 1.

- Plan 3 (90% of morning peak demand);
- Plan 4 (morning peak demand provided by NYCDOT);
- Plan 5 (110% of morning peak demand);
- Plan 6 (120% of morning peak demand);
- Plan 7 (130% of morning peak demand);

Look-up table creation

Under certain signal setting, the performance measure recorded in the look-up table helps us to identify the congestion level. An example is presented later to illustrate the needs of constructing a look-up table for each signal plan.

Let $\mathscr{S}$ denote the set of links of interest for the studied network.

We show look-up tables for signal plan 1 and signal plan 4 in Table 5.1 and Table 5.2, for a detailed analysis of all look-up tables, see Appendix E.

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	36.09	38.50	41.82	1.14	(0, 40)
scenario 2	37.08	40.98	44.82	1.27	[40,43)
scenario 3	40.85	45.76	53.19	1.96	[43,53)
scenario 4	53.54	64.95	85.64	5.49	[53,81)
scenario 5	72.83	93.74	108.93	6.60	[81, 109)
scenario 6	109.86	125.91	154.54	10.41	[109,140)
scenario 7	136.88	150.85	166.87	6.97	[140, inf)

Table 5.2: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 4.

PM_j represents the vector of total average link travel time for demand scenario j, which contains 300 simulation replications. In each table, minimum, maximum, mean and standard deviation of the total average link travel times for traffic condition j are indicated by $minPM_j$, $maxPM_j$, $\overline{PM_j}$ and σ_j, TT intervals shows the lower and upper bounds of total average link travel time specified for each demand scenario. Taking Table 5.1 as an example, from demand scenario 1 to demand scenario 7, the values of $\overline{PM_j}$ and σ_j increase, which indicates the growing across-replication variability.

To further justify the reason that a look-up table is needed for each signal plan, we show an example: if plan 1 is used, and the total average link travel time obtained from the simulator is 100 minutes. Assume we just have one look-up table which is designed based on the performance of plan 4 (Table 5.2), the value of 100 minutes falls into the fifth interval, which corresponds to demand scenario 5. The demand level under this traffic condition is considered to be similar to demand scenario 5. If we have a look-up table designed for plan 1 (Table 5.1), the value of 100 minutes falls to

the third interval which corresponds to demand scenario 3. Thus using a single look-up table under different signal settings might result in different classification of traffic condition. If we just use a generalized look-up table from one signal plan, we ignore the influence of using different signal plans on performance measures. To represent the traffic condition accurately, we build a look-up table for each signal plan under those seven demand scenarios. In total, we have six different signal plans for seven demand scenarios (plan 2 and plan 7 are the same).

In reality, signal plans are adjusted according to real-time information. In our algorithm, since the real-time information is not available, we use simulation outputs to select plans and evaluate the performance of the proposed algorithm. Assume we have both historical data and real-time information, for any agency who would like to apply such adaptive traffic signal control algorithm in reality, a fully calibrated microscopic traffic model is needed for the area of interests. The structure of the algorithm can be adjusted slightly:

- specify traffic condition into different levels according to historical data from light traffic to heavy traffic instead of simply scaling demand in simulator;
- for each traffic condition, initialize the simulator with the demand associated with that traffic condition. The queueing model should be calibrated accordingly, the details of calibrating the queueing model is specified in Appendix C;
- for heavy traffic condition, use queue management techniques to optimize the

signal plan; for light traffic, optimize signal plan by minimizing average trip travel time;

- build look-up tables using simulator for each proposed signal plan under different traffic condition based on historical data;
- classify traffic condition and selecting the competing signal plan each 15 minutes by matching real-time information with the look-up table;
- given the traffic condition classified in previous step, using the simulator and the historical demand data corresponds to that traffic condition (has been defined in the simulator) to approximate the traffic condition. To ensure fast response to real-time traffic information, fewer replications can be used (e.g. 10 replications). Then simulator is used to forecast the performance of the the competing signal plan for next time period to decide if a switch in signal plan is needed.

5.5 Results

We apply the proposed algorithm to the highly congested area around Queensboro Bridge in east Manhattan. We consider part of the morning peak period 8am-9am. Based on the demand scenarios defined, and the look-up tables proposed, we design two case studies.

In the first case study, we use demand scenario 6 with 20% higher of normal demand

for each OD pair. In the second case study, we use another given OD matrix that is different from the OD matrix (7 demand scenarios) we used to derived signal plans. In previous comparison, it has been shown that for each demand level, the signal plan derived by SO is better than existing signal plan in terms of many performance metrics. If we still use the same set of demand scenario to test the performance of the proposed adaptive algorithm, it would be less convincible to show the robustness of algorithm because in real world, traffic patterns are different over time. Furthermore, to study if the framework we developed in this chapter can be applied to various traffic conditions that are different from the demand scenarios defined.

The travel demand of the studied area is extracted from a simulation model that contains broader area. The demand we used to derive the signal plan is based on static traffic assignment for this larger area. The demand we use to validate the performance of the proposed algorithm in the second case study is calculated based on dynamic traffic assignment. They are all provided by NYCDOT. For these two sets of demand data, total demand are similar (around 11,000 trips per hour) but demand for each OD (origin-destination) pair are different.

For the studied area, there is no ATCS and detection equipments, thus fixed-time signal plan is used for morning peak regardless of flow changes. We initialize the algorithm with plan 4 because plan 4 is used to replace the existing fixed-time signal plan in the proposed algorithm.

For the transport agency (NYCDOT), they are interested in to what extent we could improve the system performance comparing with their existing solution, thus we first compare the temporal evolution of the performance measures of adaptive signal setting and existing signal plan for each 15 minutes to illustrate the benefit of using proposed method over time.

Given that plan 4 outperforms the existing signal plan even without adaptive signal setting. To investigate the added value of using adaptive signal setting, we then compare the aggregated performance of adaptive signal setting, existing signal plan and plan 4 over the whole simulation period. In each case study, to compare the performance of different signal plans, we run the adaptive signal setting, plan 4, and the existing fixed time signal plan 50 simulation replications respectively.

We use the same performance measures stated in Chapter 4. After a warm-up period of 20 minutes, we consider the temporal evolution of the following 4 performance metrics every 15 minutes:

- average network queue-length over every 15 minutes;
- average trip travel time (including all finished and unfinished trips for that time period) over every 15 minutes;
- entry flow over every 15 minutes;
- average spillback probability over every 15 minutes.

We then study the following performance measures over the whole studied period (1 hour):

- average network queue-length;
- average spillback probability;
- average trip travel time of finished trips;
- total number of finished trips;
- average trip travel time of unfinished trips;
- total number of unfinished trips;

5.5.1 Case study with severe congestion

In the first case study, the whole simulating period is divided into 4 time intervals, and the demand is 20% higher than the normal peak demand. By applying the adaptive signal control algorithm, several signal plans are selected for different time period. For the first 15 minutes, plan 4 (initial plan) is used; for the second 15 minutes, plan 5 is selected; for the third and fourth 15 minutes, plan 6 is selected.

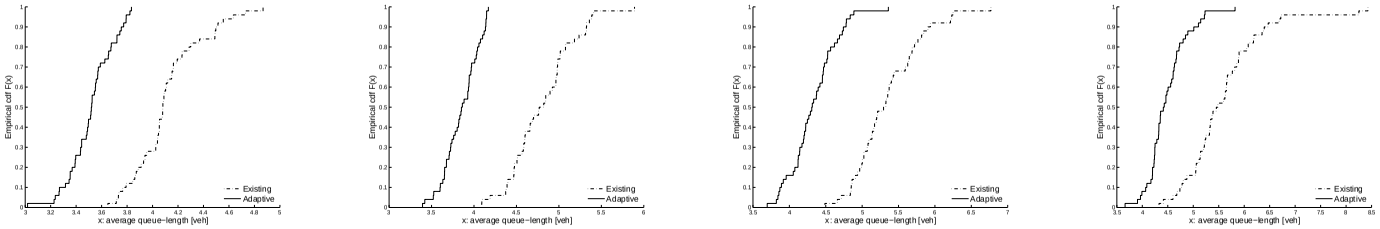
Figure 5-3 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure 5-3(a), Figure 5-3(b), Figure 5-3(c), and Figure 5-3(d) show the comparison of average queue-length of

the adaptive signal settings and the existing signal plan for each 15 minutes; Figure 5-3(e), Figure 5-3(f), Figure 5-3(g), and Figure 5-3(h) show the the comparison of average trip travel time.

Figure 5-4 shows the comparison of average spillback probability and entry flow for all time period. Figure 5-4(a), Figure 5-4(b), Figure 5-4(c), and Figure 5-4(d) show the comparison of the average spillback probability. Figure 5-4(e), Figure 5-4(f), Figure 5-4(g), and Figure 5-4(h) show the comparison of entry flow.

The performance of the adaptive signal setting is displayed in solid line; the performance of the existing signal plan is displayed in dashed line. They show the temporal evolution of each performance measure. The performance of the adaptive signal setting is displayed in solid line; the performance of the existing signal plan is displayed in dashed line. For the first and second 15 minutes, adaptive signal setting yields larger average trip travel time due to the increased number of vehicles entering the network. In the third and fourth time periods, average trip travel time obtained from adaptive signal setting is reduced. Adaptive signal setting yields smaller average queue-length and lower average spillback probability for all time periods. Adaptive signal setting increases entry flow for all time periods. The adaptive signal setting also leads to small across-replication variability in terms of entry flow and average spillback probability, which indicates more stable performances.

The aggregated performance over the simulating period is shown in Figure 5-5. The

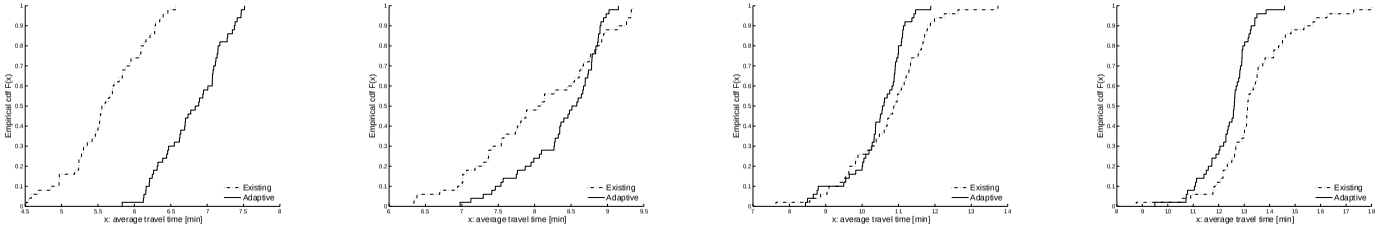


(a) average queue-length for the first 15 mins

(b) average queue-length for the second 15 mins

(c) average queue-length for the third 15 mins

(d) average queue-length for the fourth 15 mins



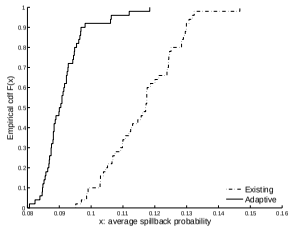
(e) average trip travel time for the first 15 mins

(f) average trip travel time for the second 15 mins

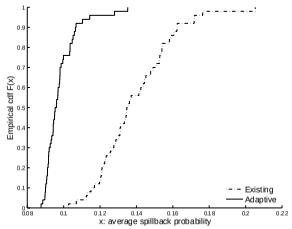
(g) average trip travel time for the third 15 mins

(h) average trip travel time for the fourth 15 mins

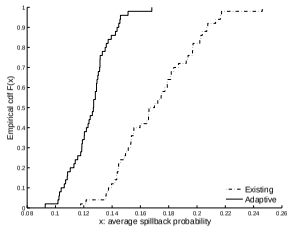
Figure 5-3: Comparison of the average queue-length and average trip travel time of adaptive signal setting and the existing fixed-time signal plan (case study 1).



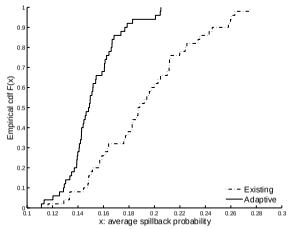
(a) average spillback probability for the first 15 mins



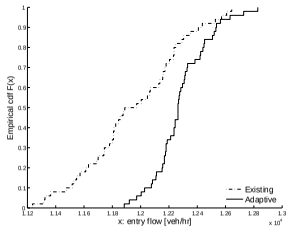
(b) average spillback probability for the second 15 mins



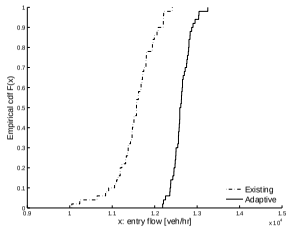
(c) average spillback probability for the third 15 mins



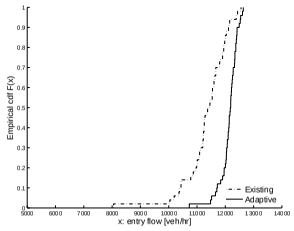
(d) average spillback probability for the fourth 15 mins



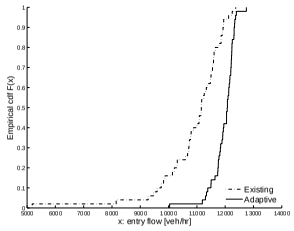
(e) entry flow for the first 15 mins



(f) entry flow for the second 15 mins



(g) entry flow for the third 15 mins



(h) entry flow for the fourth 15 mins

Figure 5-4: Comparison of the average spillback probability and entry flow of adaptive signal setting and the existing fixed-time signal plan (case study 1).

performance of the adaptive signal setting is displayed in solid line; the performance of the existing signal plan is displayed in dashed line; the performance of plan 4 is displayed in dotted line. Figure 5-5(a) shows the average queue-length; Figure 5-5(b) shows the average spillback probability; Figure 5-5(c) shows the average travel time of unfinished trips; Figure 5-5(d) shows the average travel time of all finished trips; Figure 5-5(e) shows the number of unfinished trips; Figure 5-5(f) shows the number of finished trips.

Comparing to the existing signal plan, adaptive signal setting achieves significant improvements for all performance measures: smaller average queue-length; smaller average spillback probability; smaller average travel time for finished and unfinished trips; smaller number of unfinished trips and larger number of finished trips.

Comparing to the newly proposed fixed-time signal plan 4, adaptive signal setting yields smaller average travel time of finished trips; smaller average queue-length and smaller average spillback probability. Both number of finished and unfinished trips obtained by adaptive signal setting are larger than plan 4, expectation of unfinished trip travel time is similar to that of plan 4. The difference between the number of unfinished trips obtained from adaptive signal setting and plan 4 is around 100 vehicle per hour, but the average travel time for these travelers does not increase. Note that 20% higher than normal demand is considered to be highly congested, in which spillback and blocking could easily happen. Adaptive signal setting allows more

	Existing plan				New plan			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	6.74	8.09	9.47	0.60	6.30	7.53	8.18	0.44
TP	8370	11342	12248	719.72	11608	12242	12754	213.96
TTun	18.27	22.19	35.89	2.87	14.33	16.23	17.86	0.95
TPun	896	1015	1249	85.35	800	883	990	40.03
QL	4.36	4.92	6.14	0.33	3.31	3.66	4.16	0.19
SP	0.1153	0.1549	0.1980	0.0180	0.0863	0.0937	0.1175	0.0070

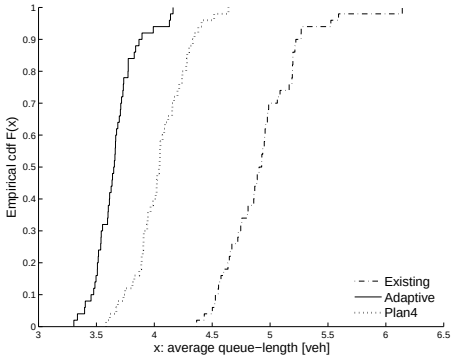
Table 5.3: Performance metrics statistics for adaptive signal setting and existing signal plan (case study 1).

vehicles to enter, and might result in some of these vehicles being blocked in the network.

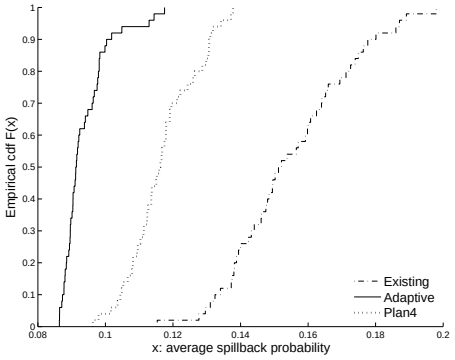
Comparing to the existing signal plan, fixed-time signal plan (plan 4) we proposed leads to better performance for all performance metrics.

To illustrate the benefit of using our algorithm, we compare the performance of the proposed algorithm with the existing signal plan in use in New York City. Table 5.3 shows the statistics of each performance measure over 50 replications. TT represents average trip travel time of finished trips; TP is the number of finished trips; TTun is the average trip travel time for unfinished trips; TPun is the number of unfinished trips; QL is the average network queue-length; SP is the average spillback probability.

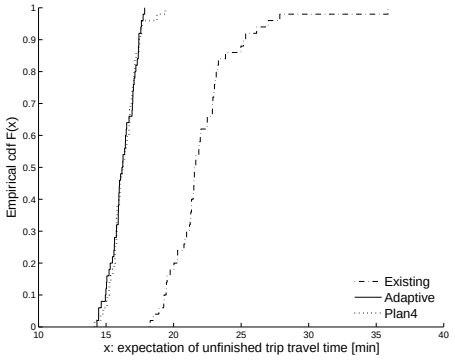
Table 5.4 shows the paired t-test for each performance measure. For each aggregated performance measure, the null hypothesis states that the performance of the adaptive signal setting is equal to that of the existing signal plan, the alternative hypothesis states that the performance of the adaptive signal setting is better (e.g.



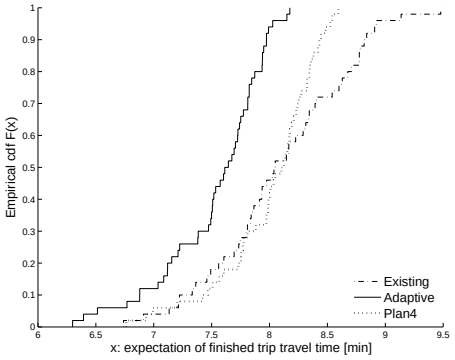
(a) average queue-length



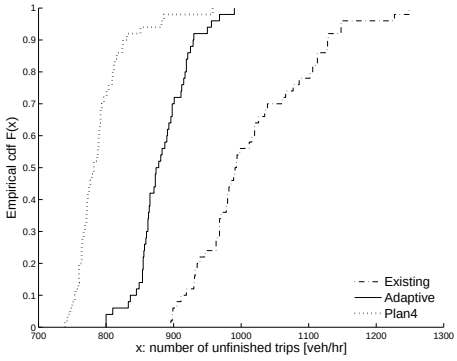
(b) average spillback probability



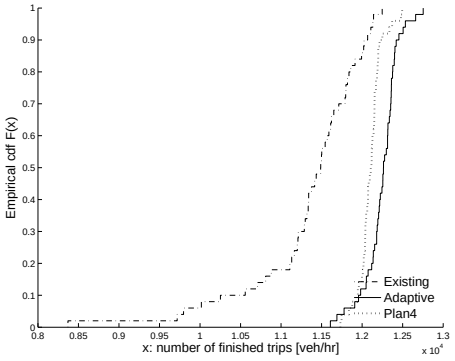
(c) average unfinished trip travel time



(d) average finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure 5-5: Comparison of the performance of adaptive signal setting, existing signal plan, and plan 4 (case study 1).

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	0.5670	0.8903	4.5032
TP	900.1800	790.6408	8.0507
TTun	5.9524	3.1048	13.5563
TPun	132.1400	95.0310	9.8323
QL	1.2623	0.3946	22.6179
SP	0.0611	0.0186	23.2253

Table 5.4: Paired t-test for adaptive signal setting and existing signal plan (case study 1).

smaller average queue-length; smaller average spillback probability; smaller travel time for unfinished trips; smaller average travel time for finished trips; smaller number of unfinished trips and lager number of finished trips) than that of the existing signal plan. The adaptive signal setting yields significant smaller finished, unfinished trip travel time, smaller number of unfinished trips, smaller average queue-length, smaller average spillback probability, and significant larger number of finished trips.

For each performance measure, we compare the mean value over 50 replications for adaptive signal setting and existing signal plan (using the value shown in Table 5.3): the adaptive signal setting reduces average finished trip travel time by 7% from 8.09 minutes to 7.53 minutes; increases the average number of finished trips by 8% from 11342 veh/hr to 12242 veh/hr; reduces the average trip travel time of unfinished trips by 35% from 22.19 minutes to 16.23 minutes; reduces the number of unfinished trips by 21% from 1015 veh/hr to 883 veh/hr; reduces average queue-length by 27% from 4.92 to 3.66; and reduces average spillback probability by 44% from 0.1549 to 0.0937.

5.5.2 Case study with different demand data

In this case study, a different set of OD demand data is used to evaluate the performance of the proposed algorithm. As mentioned before, total demand of these two demand data are similar. In this case study, we use the normal peak hour demand (around 11,000 trips per hour). By applying the proposed algorithm, plan 4 is used in the first 15 minutes; plan 5 is selected in the second and third 15 minutes; plan 6 is selected in the fourth 15 minutes.

Figure 5-6 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure 5-6(a), Figure 5-6(b), Figure 5-6(c), and Figure 5-6(d) show the comparison of average queue-length of the adaptive signal settings and the existing signal plan for each 15 minutes; Figure 5-6(e), Figure 5-6(f), Figure 5-6(g), and Figure 5-6(h) show the the comparison of average trip travel time.

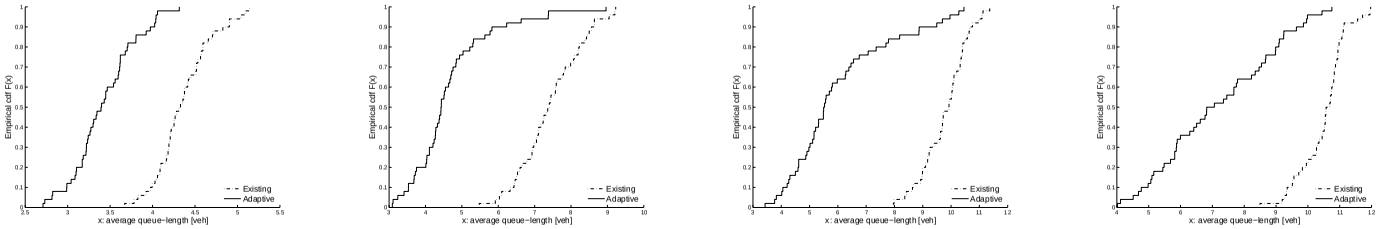
Figure 5-7 shows the comparison of average spillback probability and entry flow for all time period. Figure 5-7(a), Figure 5-7(b), Figure 5-7(c), and Figure 5-7(d) show the comparison of the average spillback probability. Figure 5-7(e), Figure 5-7(f), Figure 5-7(g), and Figure 5-7(h) show the comparison of entry flow. The performance of the adaptive signal setting is displayed in solid line; the performance of the existing signal plan is displayed in dashed line.

Comparing to the normal demand level for morning peak (around 11,000 trips

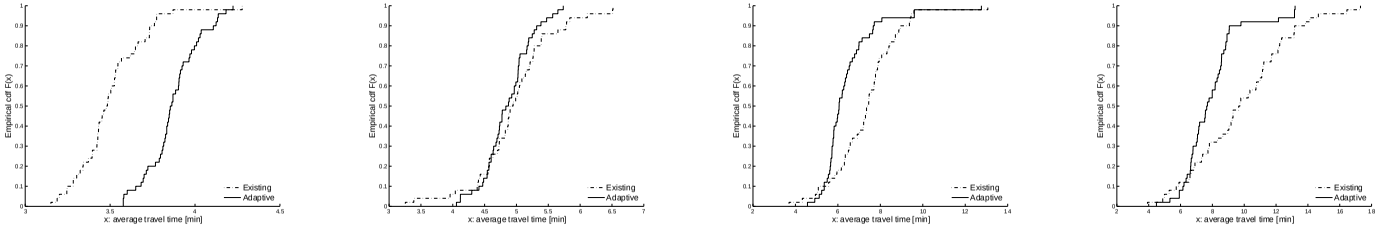
per hour), for all time periods, existing signal plan deteriorates system throughput significantly. Adaptive signal setting increases system throughput significantly for all time periods. For the first 15 minutes, adaptive signal setting yields larger average trip travel time, in the second 15 minutes, adaptive signal setting yields similar average trip travel time to that of existing signal plan. In the third and fourth 15 minutes, adaptive signal setting has smaller average trip travel time. For average queue-length and average spillback probability, adaptive signal setting yields better performance for all time periods.

We then study the performance of the adaptive signal setting, existing signal plan and proposed fixed-time signal plan (plan 4) for the whole simulation period aggregately. The performance of the adaptive signal setting is displayed in solid line; the performance of the existing signal plan is displayed in dashed line; the performance of plan 4 is displayed in dotted line. Figure 5-8(a) shows the average queue-length; Figure 5-8(b) shows the average spillback probability; Figure 5-8(c) shows the average travel time of unfinished trips; Figure 5-8(d) shows the average travel time of all finished trips; Figure 5-8(e) shows the number of unfinished trips; Figure 5-8(f) shows the number of finished trips.

Comparing to existing signal plan, adaptive signal setting reduces average queue-length and average spillback probability. The reason that adaptive signal plan leads to similar average travel time for finished trips is the significantly improved number of

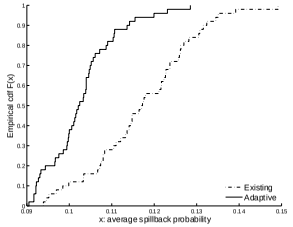


(a) average queue-length for the first 15 mins (b) average queue-length for the second 15 mins (c) average queue-length for the third 15 mins (d) average queue-length for the fourth 15 mins

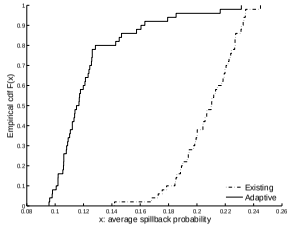


(e) average trip travel time for the first 15 mins (f) average trip travel time for the second 15 mins (g) average trip travel time for the third 15 mins (h) average trip travel time for the fourth 15 mins

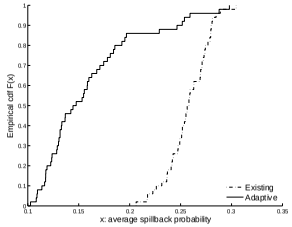
Figure 5-6: Comparison of the average queue-length and average trip travel time of adaptive signal setting and the existing fixed-time signal plan (case study 2).



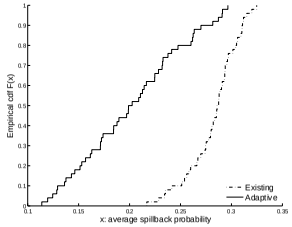
(a) average spillback probability for the first 15 mins



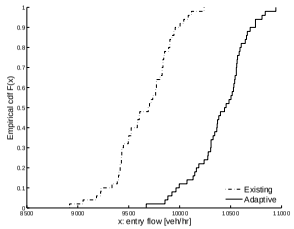
(b) average spillback probability for the second 15 mins



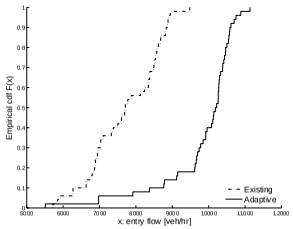
(c) average spillback probability for the third 15 mins



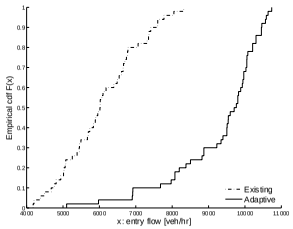
(d) average spillback probability for the fourth 15 mins



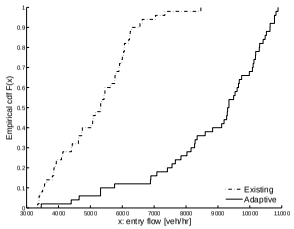
(e) entry flow for the first 15 mins



(f) entry flow for the second 15 mins



(g) entry flow for the third 15 mins



(h) entry flow for the fourth 15 mins

Figure 5-7: Comparison of the average spillback probability and entry flow of adaptive signal setting and the existing fixed-time signal plan (case study 2).

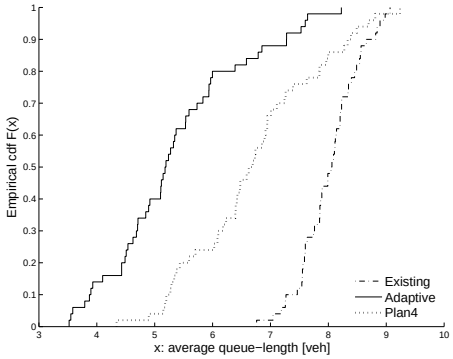
	Existing plan				Adaptive setting			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	3.13	5.15	7.40	0.93	3.85	4.90	7.08	0.60
TP	5778	7135	8500	678.58	6511	9582	10769	962.29
TTun	24.57	32.80	41.74	4.37	10.91	17.97	37.72	5.15
TPun	1256	1505	1772	105.90	776	1188	1619	211.96
QL	6.76	8.02	9.07	0.51	3.52	5.35	8.23	1.16
SP	0.1843	0.2167	0.2505	0.0156	0.1050	0.1482	0.2364	0.0308

Table 5.5: Performance metrics statistics for adaptive signal setting and existing signal plan (case study 2).

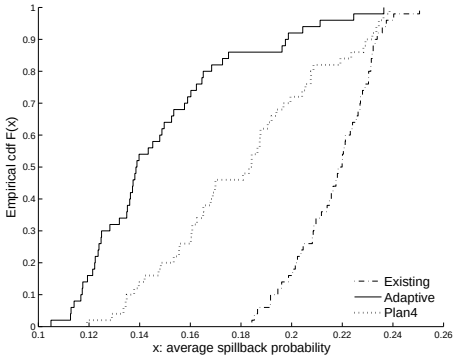
finished vehicles. The adaptive signal setting increases the number of finished trips and reduces their travel times. Meanwhile, the number of unfinished trips are reduced, the travel time experienced by the those vehicles are also reduced. Furthermore, adaptive signal setting leads to smaller across-replication variability in terms of number of finished trips. Comparing to the newly proposed fixed-time signal plan 4, the adaptive signal settings leads to improved performance for all performance metrics.

Besides the average travel time of finished trips, plan 4 leads to significant better performance for all other performance metrics comparing to existing signal plan. One possible reason is that plan 4 also increases the number of finished trips, when the system throughput increases, it would be hard to reduce the average travel time. This shows that even just use the newly proposed fixed-time signal plan, an achievement in system performance can be obtained.

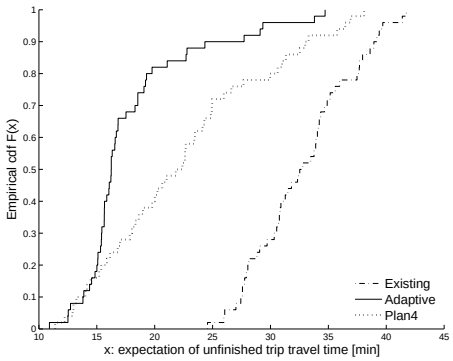
Table 5.5 shows the statistics of each performance measure over 50 replications for existing plan and adaptive signal settings. Table 5.6 shows the paired t-test for each



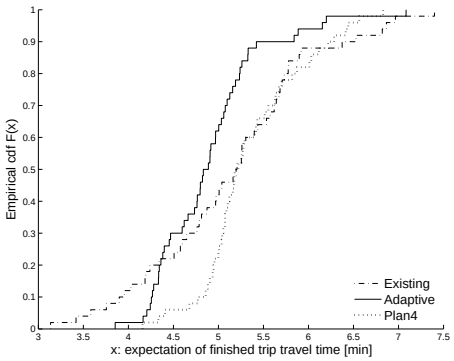
(a) average queue-length



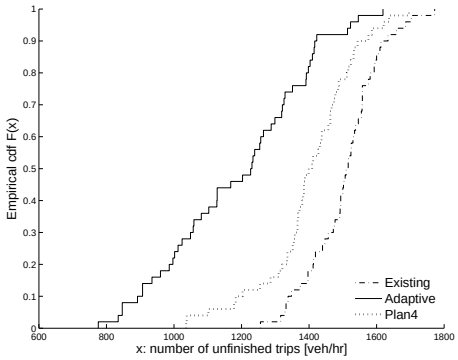
(b) average spillback probability



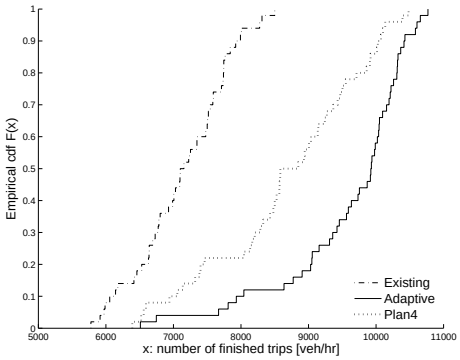
(c) average unfinished trip travel time



(d) average finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure 5-8: Comparison of the performance of adaptive signal setting, existing signal plan, and plan 4 (case study 2).

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	0.2448	0.9875	1.7532
TP	2447	1080	16.0213
TTun	14.8341	6.0573	17.3168
TPun	316.6400	10.3539	10.6439
QL	2.6670	1.2565	15.0088
SP	0.0685	0.0338	14.3541

Table 5.6: Paired t-test for adaptive signal setting and existing signal plan (case study 2).

performance measure. Except average finished trip travel time, adaptive signal setting yields significantly better performance for all other performance metrics. As mentioned before, the adaptive signal setting increases system throughput significantly, thus the average travel time of finished trip is not reduced significantly.

For each performance measure, we compare the mean value over 50 replications for adaptive signal setting and existing signal plan (using the value shown in Table 5.5): the adaptive signal setting reduces average finished trip travel time by 5% from 5.15 minutes to 4.90 minutes, and increases the average number of finished trips by 34% from 7135 veh/hr to 9582 veh/hr. The adaptive signal setting also reduces the average trip travel time of unfinished trips by 45% from 32.80 minutes to 17.97 minutes; reduces the number of unfinished trips by 21% from 1505 veh/hr to 1188 veh/hr; reduces average queue-length by 33% from 8.02 to 5.35; and reduces average spillback probability by 32% from 0.2167 to 0.1482. As mentioned before, increasing system throughput while reducing the number of vehicles being blocked in the network is challenging, furthermore, travelers are experiencing less travel time.

In both case studies, adaptive signal setting does not yield smaller average trip travel time in the first and second time interval comparing to exiting NYC signal plan, one possible reason is that the adaptive signal setting increases the network throughput significantly at the beginning of the simulation period. When more travelers are allowed to enter the network, the network becomes congested and the travel time increases. In this case study, from the temporal evolution study of the average trip travel time for most of the fixed-time signal plans, we learn that normally average travel time keeps increasing from the first time period to the fourth time period when congestion lasts. The adaptive traffic signal control algorithm is able to capture this phenomena. When congestion happens and travel time increases, the algorithm switch signal plans in order to accommodate higher level demand, then travel time does not increase significantly as time goes by. On the contrary, fixed-time plan is not able to tackle this situation. Thus for the last two time periods, the proposed algorithm select signal plans that are suitable for higher demand levels, and lead to smaller average trip travel time.

5.6 Conclusion

In this chapter, we address a simulation-based adaptive traffic signal control problem for a congested grid-type urban network in eastern Manhattan (New York City, USA). To evaluate the performance of the proposed method, two case studies are carried

out. In the first case study, the method is used to address signal setting under severe congestion. In the second case study, the proposed method is used to address signal setting under a different set of demand data with different OD matrices. In both cases, the proposed method leads to signal plans with improved network performance. Furthermore, an improvement could be achieved by just using the newly proposed fixed-time signal plan comparing to the existing signal plan. In this case study, we simply scale demand into different levels to represents different traffic conditions from light traffic to heavy traffic. To apply such algorithm in reality, it is of interest to investigate how to define traffic conditions from the historical data.

Chapter 6

Conclusion

This thesis addressed signal control problems that are important but receive less attentions or have limitations. The main contributions are the development of reliable signal control problems with different formulations, and the adaptive traffic signal control algorithm proposed for congested gird-type urban network.

Chapter 2 and **Chapter 3** address the reliable signal control problems. **Chapter 2** incorporates tractable link travel time distributional information in signal design objectives. Due to the difficulty of approximating link travel time dependency, we assume independent link travel times and derive analytical approximation of link travel time SD. We use two SO metamodel approaches to solve three different signal control problems: reliable signal control problem that combines average and SD of link travel time information; traditional signal control problem that considers average total link travel times, and signal control problem which uses only link travel time SD in the

objective function. We first optimize signal plans for Lausanne city center, and then extend to the full city area. In both cases, the metamodel which combines analytical approximation of objective functions and functional component leads to signal plans with smaller total average link travel time and total link travel time SD. The signal plans derived by solving the reliable signal control problem have the lowest link travel time SD and average link travel time.

Chapter 3 can be considered as an extension of **Chapter 2** with more realistic assumption for between-link dependency and interactions. In **Chapter 3**, an analytical tractable approximation of path travel time SD that accounts for between-link dependency is proposed. The trip travel time SD is then obtained by aggregating the path travel time SD. The formulation that accounts for between-link dependency is compared with the formulation ignores between-link dependency. Taking the simulation observations of path and trip travel time SD as references, we validate the proposed formulation through two toy networks with different topology. It shows that for low demand scenarios, these two methods has similar estimates of path and trip travel time SD. When demand keeps increasing, the proposed formulation that accounts between-link dependency leads to more accurate estimates of path and trip travel time SD. This finding suggests that it is not accurate to ignore between-link dependency for congested networks. However, this is just verified for toy network, it is of interest to test if it still holds for real world network with more complicated

traffic interactions. We then use the formulation of trip travel time SD to address an analytical reliable signal control problem and a simulation-based optimization reliable signal control problem for the city center of Lausanne. For the analytical reliable signal control problem, signal plans derived by using the proposed formulation and the formulation that ignores between link dependency are compared. It shows that the signal plan derived by proposed formulation provides a smaller average trip travel time and trip travel time SD. For the simulation-based optimization reliable signal control problem, besides the formulation that ignores between link dependency, the proposed formulation is compared with the formulation proposed in **Chapter 2**. Comparing to the signal plan derived in **Chapter 2**, it shows that signal plans obtained in **Chapter 3** reduce trip travel time variability at the expense of increasing average trip travel time. Since average trip travel time and trip travel time SD is less correlated, a sensitive test of using different reliability ratios is of interest. Moreover, the reliability ratio used in **Chapter 2** and **Chapter 3** is not estimated for the network we studied, it might not be the right value to represent the trade-off between average travel time travel time SD for the travelers in Lausanne.

In general, **Chapter 2** and **Chapter 3** enable the use of second-order travel time information (both analytical and simulation-based) in large-scale traffic signal optimization problems. Both formulations can be used in signal design objective functions to reduce either link travel time variability or trip travel time variability depending

on the needs of transport agency. The tractable approximation of path travel time SD proposed in **Chapter 3** captures the impacts of the demand changes on trip travel time variability. As mentioned before, path travel time variability is one of the most important factor that would influence the route choices, thus this formulation can be used to study the routing behavior of the drivers in future research.

Besides the reliable signal control problem, **Chapter 4** and **Chapter 5** design signal control strategy for highly congested urban network. In high density urban areas especially in network with short links and grid-type topology, the traditional traffic signal control strategy (both fixed-time and adaptive traffic signal control strategy) has limited ability to reduce the spillbacks and ease congestion. We focus on an area in east Manhattan (New York City, USA) and build a set of demand scenarios to reflect different traffic conditions, from light to heavy traffic. At first, for each traffic condition, we use the SO framework to calculate fixed-time signal plan for that traffic condition. For low demand level, traditional signal design objective is used to calculate the optimal signal plan. Under high demand levels, spillbacks happen frequently. Using traditional signal design techniques yields signal plans that do not capture these spillbacks, deteriorating the system throughput. Under such circumstance, queue management techniques are used to optimize signal plans. For each demand scenario corresponds to each traffic condition, the proposed signal plan outperforms the existing signal plan in terms of different performance measures such as: throughput, average link travel

time, average trip travel time, queue length. We then design a simulation-based adaptive traffic signal control algorithm to select among the set of signal plans designed for different traffic conditions. Comparing to the current signal plan, the proposed algorithm leads to signal plans with less average trip travel time, shorter queue length, smaller spillback probability and higher system throughput. The proposed adaptive traffic signal control algorithm is based on simulation observations instead of real-time information due to the inadequacy of historical and real-time data. From the analysis of the signal plan performance, we observe that travel time variability is larger for demand scenario with higher demand. Incorporating travel time reliability metrics proposed in **Chapter 2** and **Chapter 3** in signal design objective for the proposed adaptive traffic signal control algorithm is of interest in future research. Furthermore, to enhance the performance of the proposed algorithm, more powerful and accurate forecasting sector used to select signal plans can be further investigated.

One limitation of the SO algorithm we used in this thesis is that all the trial points (candidate signal plans) derived by the SO framework is only evaluated once in the simulator, and then the algorithm decides if a trial point would be accepted or rejected. The SO framework optimizes the problem sequentially, in which the trial points derived later are always based on all the previous results. This might lead to the problem of choosing an actually bad design which has good performance in a single run. One way of overcoming this limitation is to perform several replications for the

trial points to verify the performance measures. This way has been verified in **Chapter 3**, we noticed that trip travel time SD has large variability, and evaluating each trial point once sometimes leads to signal plan with even worse performance than the initial signal plan we start. As a results, for each trial point SO derived, we evaluate it three times and calculate the average value over three observations to decide if it should be accepted or rejected. Evaluating each trial point three times leads to signal plans with significant better performance comparing to the method that just evaluates each trial point once. However, this would result in huge computational burden, which violates the aim of solving the signal control problem efficiently. Instead of incorporating the statistical selection techniques during the optimization process, the SO framework can be divided into two stages and the same amount of total computational budget (e.g.:150 simulation runs) can be allocated to them. Based on empirical test of the SO framework, in most of the cases the algorithm converges fast in the first dozens of simulation runs, the following simulation runs do not help to improve the system performance. As a result, the same SO framework is used in the first stage with fewer computational budget and all the accepted trial points will be kept, then a post-processing stage will be added to select the final best solution statistically from the solution sets (all accepted trial points). We have built the post-processing technique upon the widely used optimal computing budget allocation (OCBA) procedure (Chen et al., 1999b, 2000) and applied that technique to **Chapter 2**. The proposed post-

processing techniques are evaluated by probability of selection (PCS) and compared with total equal allocation (TEA) which is defined as allocating computational budget equally to each alternative. Among the solution set, we know which signal plan is the best one with the smallest total link travel time and total link travel time SD, we run OCBA and TEA 1000 times respectively, the PCS is calculated as the how many times that each algorithm chooses the best signal plan over 1000. However, OCBA does not outperform TEA. Comparing to TEA, OCBA takes variance of each alternative into consideration when allocating computational budget (Fu et al., 2008), but average total link travel time and total link travel time SD has small variance (shown in Figure 2-3, the cdf curves are very steep). Thus we cannot benefit from OCBA in this case. It is of interest to apply the proposed method for **Chapter 3** to further investigate this problem, in which the values of the objective function has much larger variance.

Appendix A

Physical components and SO algorithm

A.1 Physical components

A.1.1 Physical component used in Section 2.4.2

Recall from Section 2.3.3 that the analytical approximation of the objective function (Equation (2.5)) provided by the physical component is a function of three endogenous variables per queue: ρ_i, λ_i and $P(N_i = k_i)$. We present below the analytical traffic model that derives these variables. This model is based on the general queueing network model of Osorio and Bierlaire (2009a). Its formulation for an urban traffic network is given in Osorio and Bierlaire (2009b). Each lane of an urban road network

is modeled as one or a set of finite capacity queues. The model describes the between-link interactions (e.g., spillbacks) through the queueing theory notion of blocking. It provides an analytical description of how congestion arises and propagates through the network. In the following notation the index i refers to a given queue.

γ_i	external arrival rate;
λ_i	arrival rate (also referred to as total arrival rate);
μ_i	service rate;
$\tilde{\mu}_i$	unblocking rate;
$\hat{\mu}_i$	effective service rate (accounts for both service and eventual blocking);
ρ_i	traffic intensity;
P_i^f	probability of being blocked at queue i;
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 or spillback probability;
p_{ij}	transition probability from queue i to queue j;
$\mathscr{D}_i$	set of downstream queues of queue i;

The queueing network model is defined through the following system of nonlinear

equations:

$$\begin{cases} \lambda_i = \gamma_i + \dfrac{\sum_j p_{ji}\lambda_j(1 - P(N_j = k_j))}{(1 - P(N_i = k_i))} & \text{(A.1a)} \\ \dfrac{1}{\tilde{\mu}_i} = \displaystyle\sum_{j \in \mathscr{D}_i} \dfrac{\lambda_j(1 - P(N_j = k_j))}{\lambda_i(1 - P(N_i = k_i))\hat{\mu}_j} & \text{(A.1b)} \\ \dfrac{1}{\hat{\mu}_i} = \dfrac{1}{\mu_i} + P_i^f \dfrac{1}{\tilde{\mu}_i} & \text{(A.1c)} \\ P(N_i = k_i) = \dfrac{1 - \rho_i}{1 - \rho_i^{k_i+1}} \rho_i^{k_i} & \text{(A.1d)} \\ P_i^f = \displaystyle\sum_j p_{ij} P(N_j = k_j) & \text{(A.1e)} \\ \rho_i = \dfrac{\lambda_i}{\hat{\mu}_i}. & \text{(A.1f)} \end{cases}$$

The exogenous parameters are γ_i, μ_i, p_{ij} and k_i. All other parameters are endogenous. When used to solve a signal control problem (as in this chapter), the capacity of the signalized lanes become endogenous, which makes the corresponding service rates, μ_i, endogenous.

A.1.2 Physical component used in Section 2.4.3

This model builds upon the model of Osorio and Bierlaire (2009a) and of Osorio and Bierlaire (2009b) (for its detailed derivation see Osorio and Chong (2012)). It approximates the traffic intensity of queue i, ρ_i, by the effective traffic intensity, ρ_i^{eff}, where $\rho_i^{\text{eff}} = \rho_i(1 - P(N_i = k_i))$. Throughout the System of Equations (A.1), ρ is replaced by ρ^{eff}, and the following model is obtained:

$$\begin{cases} \lambda_i = \gamma_i + \dfrac{\sum_j p_{ji}\lambda_j(1 - P(N_j = k_j))}{(1 - P(N_i = k_i))} & \text{(A.2a)} \\ \rho_i^{\text{eff}} = \dfrac{\lambda_i(1 - P(N_i = k_i))}{\mu_i} + \left(\sum_{j \in \mathscr{D}_i} p_{ij} P(N_j = k_j)\right)\left(\sum_{j \in \mathscr{D}_i} \rho_j^{\text{eff}}\right) & \text{(A.2b)} \\ P(N_i = k_i) = \dfrac{1 - \rho_i^{\text{eff}}}{1 - (\rho_i^{\text{eff}})^{k_i+1}}(\rho_i^{\text{eff}})^{k_i}. & \text{(A.2c)} \end{cases}$$

A.2 SO algorithm

This SO algorithm is formulated in detail in Osorio (2010) and is based on the derivative-free trust region algorithm of Conn et al. (2009). The parameters of the algorithm are set according to the values in Osorio (2010).

0. **Initialization.**

   Define for a given iteration k: $m_k(x, y; \alpha_k, \beta_k, q)$ as the metamodel (denoted hereafter as $m_k(x)$), x_k as the iterate, Δ_k as the trust region radius, $\nu_k = (\alpha_k, \beta_k)$ as the vector of parameters of m_k, n_k as the total number of simulation runs carried out up until and including iteration k, u_k as the number of successive trial points rejected, ε_k as the measure of stationarity (norm of the derivative of the Lagrangian function of the trust region (TR) subproblem with regards to the endogenous variables) evaluated at x_k.

   The constants $\eta_1, \gamma, \gamma_{inc}, \varepsilon_c, \bar{\tau}, \bar{d}, \bar{u}, \Delta_{max}$ are given such that: $0 < \eta_1 < 1$, $0 < \gamma < 1 < \gamma_{inc}$, $\varepsilon_c > 0$, $0 < \bar{\tau} < 1$, $0 < \bar{d} < \Delta_{max}$, $\bar{u} \in \mathbb{N}^*$. Set the total

number of simulation runs permitted (across all points) n_{max}, this determines the computational budget. Set the number of simulation replications per point $\tilde{r}$ (here we use $\tilde{r} = 1$).

Set $k = 0, n_0 = 1, u_0 = 0$. Determine x_0 and Δ_0 ($\Delta_0 \in (0, \Delta_{max}]$).

Given the initial point x_0, compute $f_A(x_0)$ (analytical approximation of Equation (2.1)) and $\hat{f}(x_0)$ (simulated estimate of Equation (2.1)), fit an initial model m_0 (i.e., compute ν_0).

1. **Criticality step.** If $\varepsilon_k \leq \varepsilon_c$, then switch to *conservative mode*.

2. **Step calculation.** Compute a step s_k that reduces the model m_k and such that $x_k + s_k$ (the trial point) is in the trust region (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 set of sampled points ($n_k = n_k + \tilde{r}$), and fit the new model m_{k+1}.

4. **Model improvement.** Compute $\tau_{k+1} = \frac{\|\nu_{k+1} - \nu_k\|}{\|\nu_k\|}$. If $\tau_{k+1} < \bar{\tau}$, then improve the model by simulating the performance of a new point x, which is uniformly drawn from the feasible space. Evaluate f_A and $\hat{f}$ at x. Include this new observation in the set of sampled points ($n_k = n_k + \tilde{r}$). Update m_{k+1}.

5. **Trust region radius update.**

$$\Delta_{k+1} = \begin{cases} \min\{\gamma_{inc}\Delta_k, \Delta_{max}\} & \text{if } \rho_k > \eta_1 \\ \max\{\gamma\Delta_k, \bar{d}\} & \text{if } \rho_k \leq \eta_1 \text{ and } u_k \geq \bar{u} \\ \Delta_k & \text{otherwise.} \end{cases}$$

If $\rho_k \leq \eta_1$ and $u_k \geq \bar{u}$, then set $u_k = 0$.

If $\Delta_{k+1} \leq \bar{d}$, then switch to *conservative mode.*

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

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

Appendix B

Derivation of Equation (3.6b) and Equation (3.23)

B.1 Derivation of Equation (3.6b)

We describe the derivation of Equation (3.6b). By definition the expected effective service time conditional on state s_m is the summation of the expected service time and the expected blocked time conditional on state s_m:

$$\frac{1}{\hat{\mu}_{i,s_m}} = \frac{1}{\mu_i} + E[B_i|S_m = s_m], \tag{B.1}$$

where $E[B_i|S_m = s_m]$ is the expected blocked time at queue i given state s_m. We assume that all downstream queues of queue i that are full are also blocking queue i.

This is equivalent to:

$$E[B_i|S_m = s_m] = \sum_{j \in \mathcal{DS}_i} \mathbb{1}(s_m, j) p_{ij} E[B_{i,j}|S_m = s_m], \tag{B.2}$$

where $B_{i,j}$ represents the blocked time at queue i due to blocking by queue j. We approximate $E[B_{i,j}|S_m = s_m]$ by:

$$E[B_{i,j}|S_m = s_m] = \frac{1}{r_{ij}\hat{\mu}_j}, \tag{B.3}$$

where r_{ij} represents the expected proportion of flow that arises to queue j due to queue i, and $\hat{\mu}_j$ is given by Equation (3.2b). Note that $1/\hat{\mu}_j$ represents the expected effective service time of queue j. Since queue j is full, $1/\hat{\mu}_j$ can also be interpreted as the expected time between successive departures from queue j. In other words, it is the expected time between unblockings of jobs blocked by queue j at queues upstream of j. The term $1/(r_{ij}\hat{\mu}_j)$ is used to approximate the time between unblockings of jobs blocked by queue j at queue i. The term r_{ij} accounts for the fact that unblocking events occur in a first-in-first-out manner. The term r_{ij} is approximated as:

$$r_{ij} = \frac{p_{ij}\hat{\lambda}_i}{\hat{\lambda}_j}, \tag{B.4}$$

where $\hat{\lambda}_i$ and $\hat{\lambda}_j$ are given by Equation (3.2a).

Successively inserting Equation (B.2) into (B.1), (B.3) into (B.2), and (B.4) into

(B.3); we obtain:

$$\frac{1}{\hat{\mu}_{i,s_m}} = \frac{1}{\mu_i} + \sum_{j \in \mathcal{DS}_i} \mathbb{1}(s_m, j) p_{ij} E[B_{i,j} | S_m = s_m] \quad \text{(B.5)}$$

$$= \frac{1}{\mu_i} + \sum_{j \in \mathcal{DS}_i} \mathbb{1}(s_m, j) p_{ij} \frac{1}{r_{ij} \hat{\mu}_j} \quad \text{(B.6)}$$

$$= \frac{1}{\mu_i} + \sum_{j \in \mathcal{DS}_i} \mathbb{1}(s_m, j) p_{ij} \frac{\hat{\lambda}_j}{p_{ij} \hat{\lambda}_i \hat{\mu}_j} \quad \text{(B.7)}$$

$$= \frac{1}{\mu_i} + \sum_{j \in \mathcal{DS}_i} \mathbb{1}(s_m, j) \frac{\hat{\lambda}_j}{\hat{\lambda}_i \hat{\mu}_j} \quad \text{(B.8)}$$

Equation (B.8) coincides with Equation (3.6b).

B.2 Derivation of Equation (3.23)

We present the derivation of the expression for $E[N_i^2]$ given by Equation (3.23). Hereafter, we drop the queue index i.

By definition:

$$E[N^2] = \sum_{n=0}^{k} n^2 P(N = n). \quad \text{(B.9)}$$

We can insert the expression for $P(N = n)$ of Equation (3.3) to obtain:

$$E[N^2] = \sum_{n=0}^{k} n^2 \frac{1 - \rho}{1 - \rho^{k+1}} \rho^n. \quad \text{(B.10)}$$

We can rewrite n^2 as $n(n-1)+n$:

$$E[N^2] = \frac{1-\rho}{1-\rho^{k+1}} \sum_{n=0}^{k} \left(n(n-1)\rho^n + n\rho^n \right). \tag{B.11}$$

This is equivalent to:

$$E[N^2] = \frac{1-\rho}{1-\rho^{k+1}} \sum_{n=0}^{k} \left(n(n-1)\rho^n \right) + \sum_{n=0}^{k} n \frac{1-\rho}{1-\rho^{k+1}} \rho^n. \tag{B.12}$$

Notice that the second summation is equal to $E[N]$, hence:

$$E[N^2] = \frac{1-\rho}{1-\rho^{k+1}} \sum_{n=0}^{k} \left(n(n-1)\rho^n \right) + E[N]. \tag{B.13}$$

We now focus on the first summation of the above equation. For a geometric series, such that $\rho \neq 1$, we have:

$$\sum_{n=0}^{k} \rho^n = \frac{\rho^{k+1}-1}{\rho-1}. \tag{B.14}$$

We differentiate the left and the right side of this equation with respect to ρ:

$$\sum_{n=0}^{k} n\rho^{n-1} = \frac{1-\rho^{k+1}}{(1-\rho)^2} - \frac{(k+1)\rho^k}{1-\rho}. \tag{B.15}$$

We differentiate once again with respect to ρ:

$$\sum_{n=0}^{k} n(n-1)\rho^{n-2} = \frac{-(k+1)\rho^k}{(1-\rho)^2} - \frac{(1-\rho^{k+1})2(1-\rho)(-1)}{(1-\rho)^4} - \frac{k(k+1)\rho^{k-1}}{1-\rho} + \frac{(k+1)\rho^k(-1)}{(1-\rho)^2}. \tag{B.16}$$

This can be rearranged to obtain:

$$\sum_{n=0}^{k} n(n-1)\rho^{n-2} = \frac{2(1-\rho^{k+1})}{(1-\rho)^3} - \frac{k(k+1)\rho^{k-1}}{1-\rho} - \frac{2(k+1)\rho^k}{(1-\rho)^2}. \tag{B.17}$$

We insert the above expression into Equation (B.13) to obtain:

$$E[N^2] = \frac{1-\rho}{1-\rho^{k+1}}\rho^2 \left(\frac{2(1-\rho^{k+1})}{(1-\rho)^3} - \frac{k(k+1)\rho^{k-1}}{1-\rho} - \frac{2(k+1)\rho^k}{(1-\rho)^2} \right) + E[N] \tag{B.18}$$

This can be rearranged to obtain:

$$E[N^2] = \frac{2\rho^2}{(1-\rho)^2} - \frac{k(k+1)\rho^{k+1}}{1-\rho^{k+1}} - \frac{2(k+1)\rho^{k+2}}{(1-\rho^{k+1})(1-\rho)} + E[N]. \tag{B.19}$$

Equation (B.19) coincides with Equation (3.23).

Appendix C

Queueing model calibration details

In general, there are three parts of calibration: static, dynamic, and signal plan encoding. The exogenous queueing model parameters stated in Appendix A.1.1(γ_i, μ_i, k_i, p_{ij}, $\mathscr{D}_i$) are obtained via calibration.

The static calibration converts the road network into queueing network. It retrieves information of network topology such as number of links, link length, number of lanes for each link, type of links (e.g.: bus lane, parking lane, entrance link, exit link), type of intersections (e.g.: signalized intersection, non-signalized intersections) from the simulator. Each link is represented by a queue or a set of queues. We are focusing on passenger cars and trucks, thus bus and parking lanes are removed in the queueing network. For exogenous queueing model parameters stated in Appendix A.1.1: upper bound of the queue length k_i, set of downstream queues $\mathscr{D}_i$ of queue i are calculated via static calibration. k_i is calculated from the length of each link, for a detailed

description, see Chapter 4.3.3 of Osorio (2010). $\mathscr{D}_i$ are defined based on the rules of the road (e.g.: turning is allowed) and the topology of the network (e.g.: links that are physically linked together). After performing the static calibration, the queueing network is able to represents the physical structure of road network. Based on the physical representation of the road network, we can proceed to dynamic calibration.

When we perform dynamic calibration for each demand scenario, we run the simulator 10 times for the one hour morning peak with 20 minutes warm-up period. Then we calculate the average flow per hour on each link over 10 replications. Dynamic calibration converts network demand into external arrival rate γ_i based on flow of entrance links, where the vehicles initially originated. The transition probability p_{ij} is calculated as the proportion of flow coming from the one link to another. For each demand scenario, total demand, flow on each link and route choices are different thus some links might not have flow under certain demand scenarios, which means no user selects those links. To ensure the computational efficiency, queues correspond to these links (no flow) are removed in the queueing network.

Based on the information of intersections retrieved from the static calibration, signal plan encoding can be performed to calculate service rate μ_i. For signalized intersections, the service rate is calculated form the green splits correspond to each movement. For non-signalized intersections, each movement is ranked according to HCM (TRB, 2000), all upstream queues linked to that intersection are matched with

certain movement, the service rate of each queue is calculated as the capacity associated with that movement. When queueing model are calibrated, the SO framework can be used to optimize signal plan according to the design objective (e.g.: minimize average queue length, minimize average trip travel time, maximize system throughput, etc.).

Appendix D

Comparison of the performance of signal plans derived and the existing signal plan for different demand levels

We test three signal design objectives for each demand level: 1) traditional signal design objective that minimizes average trip travel time; 2) a formulation that explicitly accounts for queue-length metrics and mitigates the occurrence of urban spillbacks and gridlocks: minimizing average total queue-length in the network; 3) maximize average system throughput.

Based on these experiments, we found that different signal control design objectives are suitable for different demand levels. For demand scenarios with less demand than

the normal morning peak demand (scenario 4), minimizing average trip travel time yields signal plan with best performance in terms of average trip travel time. Signal plans derived by minimizing total queue-length do not outperform the existing signal plan in terms of average trip travel time, and system throughput. For demand scenario 4 and scenarios with higher demand than scenario 4, minimizing average trip travel time results in a significant reduction in the system throughput. For scenarios with higher demand levels, minimizing total average queue-length yields signal plans with better system performance in terms of average queue-length without deteriorating throughput. Maximizing throughout in objective function is also tested for the studied area, comparing to other performance metrics (e.g.: average queue-length, average trip travel time), the value of throughput is large (around 11000 trips per hour), even small fluctuation causes large number change in throughput. For instance, 11100 trips per hour might not be statistically different from 11000 trips per hour due to the variability in demand, but for SO algorithm, it is considered to be a better plan and will be kept. When using system throughput in objective functions, we end up with signal plans that do not outperform existing signal plan for all demand scenarios in terms of both throughput and average trip travel time. In order to use system throughput in SO framework, more replications are needed to evaluate the performance of each derived point, which brings huge computational burden.

For each demand scenario, we start from the existing NYC signal plan and run

the SO five times. The computational budget is set to 150 simulation runs each time. In total we derive five signal plans. To evaluate the performance of the signal plans derived by SO, we run the signal plan derived by SO and existing signal plan under each demand scenario 50 times respectively. We then compare their average trip travel time, system throughput, average queue-length, and spillback probability. For demand scenario 1, 2 and 3, in which minimizing average trip travel time is used as objective function, signal plan with the smallest average trip travel time without deteriorating the system throughput will be selected as the new signal plan. For demand scenario 4, 5, 6, and 7, in which minimizing total queue-length is used as objective function, signal plan with the smallest total queue-length without deteriorating the system throughput will be selected as the new signal plan.

For demand scenario 1, 2 and 3, the objective function is minimizing average trip travel time. The signal control problem is formulated as:

$$\min_{x} f(x) = T(x, z; p) \tag{D.1}$$

subject to

$$\sum_{j \in \mathscr{P}_I(i)} x(j) = b_i, \ \forall i \in \mathscr{I} \tag{D.2}$$

$$x \geq x_L. \tag{D.3}$$

This problem is a fixed-time signal control problem, where the decision variables x are the green splits. In this problem, the stage structure (e.g.: phase sequence) is given, the offsets, the cycle times and the all-red durations are fixed. The performance metric used, $T(x, z; p)$, is the average trip travel time. Constraints (D.2) guarantee that for a given intersection the available cycle time is distributed across all endogenous phases. Constraints (D.3) ensure lower bounds for the green splits.

For demand scenario 4, 5, 6, and 7, the objective function is minimizing total average queue-length. It has been discussed in Chapter 4.

For certain demand scenario, if all signal plans derived by SO yield larger average trip travel time, larger average queue-length, or smaller system throughput, the existing signal plan will be used for that demand level.

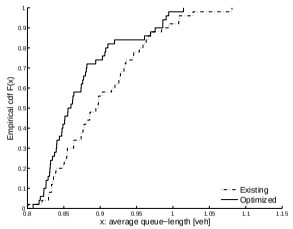
We have new plans for demand scenario 1, 3, 4, 5, and 6. For demand scenario 2 and 7, the signal plans derived do not yields better performance than existing signal plan, thus we stick to existing signal plan. We name signal plan for each demand scenario from plan 1 to plan 7, plan 2 and plan 7 are the same.

The performance comparison of signal plans derived by SO (expect plan 4 that has been shown in Chapter 4) and existing signal plan used in NYC are shown in this Appendix. Under each demand scenario, we study the performance of the signal plan (denoted as "optimized" and displayed in solid line) derived by SO and the existing signal plan (denoted as "existing" and displayed in dashed line). Recall that we have four performance measures to evaluate the temporal evolution of the system performance: average queue length, average trip travel time, average spillback probability and entry flow; and six aggregated performance measures: average queue length, average travel time of finished trips, average travel time of unfinished trips, average spillback probability, number of finished trips, and number of unfinished trips.

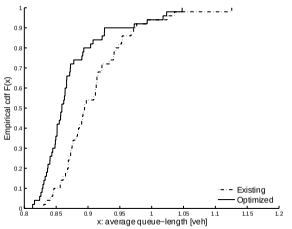
Comparison of the performance of plan 1 and the existing signal plan

This section shows the comparison of the performance of the signal plan derived by SO and the existing signal plan under demand scenario 1. We first show the temporal evolution of each performance measure.

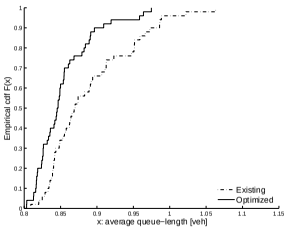
Figure D-1 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure D-1(a), Figure D-1(b), Figure D-1(c), and Figure D-1(d) show the comparison of average queue-length of the adaptive signal settings and the existing signal plan for each 15 minutes; Figure D-1(e), Figure D-1(f), Figure D-1(g), and Figure D-1(h) show the the comparison of average trip travel time.



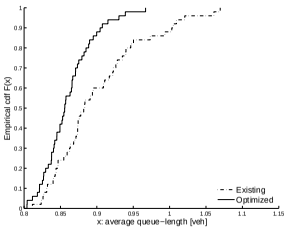
(a) average queue-length for the first 15 mins



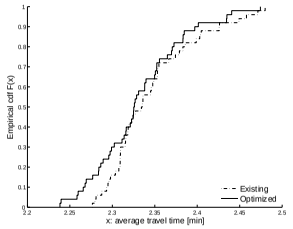
(b) average queue-length for the second 15 mins



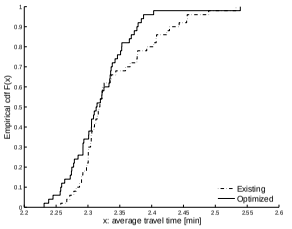
(c) average queue-length for the third 15 mins



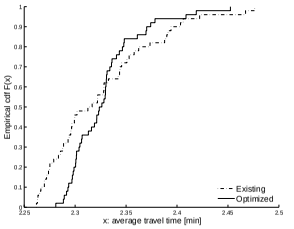
(d) average queue-length for the fourth 15 mins



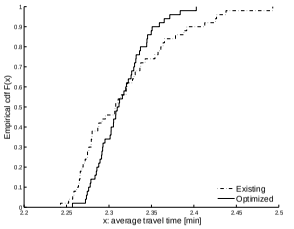
(e) average trip travel time for the first 15 mins



(f) average trip travel time for the second 15 mins

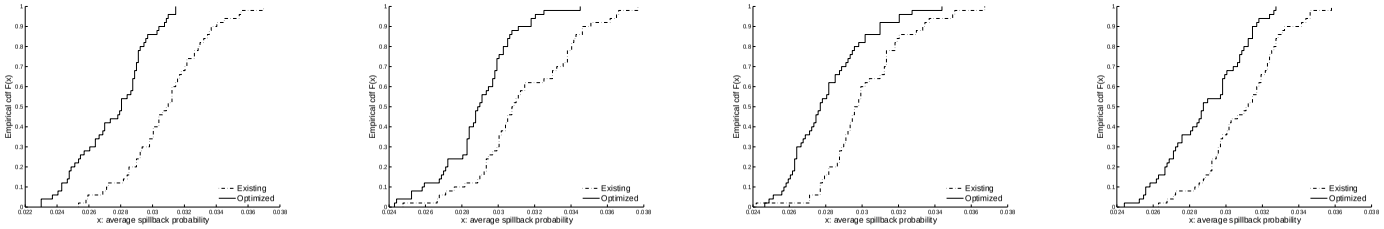


(g) average trip travel time for the third 15 mins

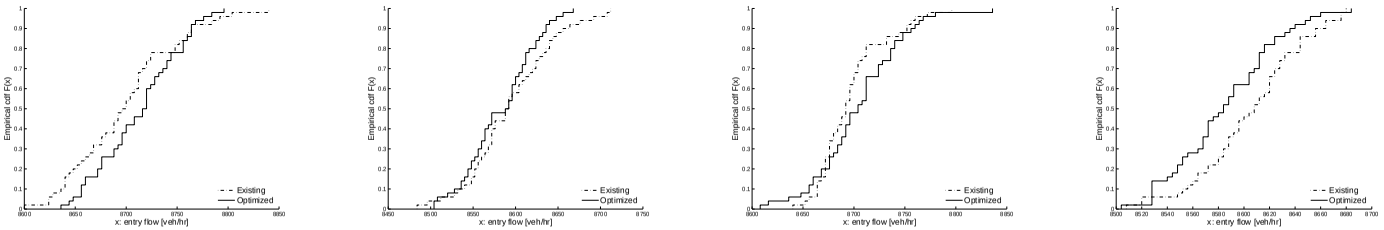


(h) average trip travel time for the fourth 15 mins

Figure D-1: Comparison of the average queue-length and average trip travel time of plan 1 and the existing signal plan.



(a) average spillback probability for the first 15 mins (b) average spillback probability for the second 15 mins (c) average spillback probability for the third 15 mins (d) average spillback probability for the fourth 15 mins



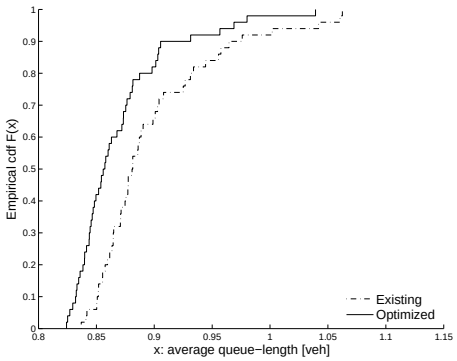
(e) entry flow for the first 15 mins (f) entry flow for the second 15 mins (g) entry flow for the third 15 mins (h) entry flow for the fourth 15 mins

Figure D-2: Comparison of the average spillback probability and entry flow of plan 1 and the existing signal plan.

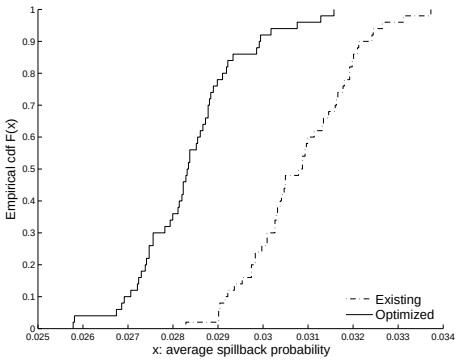
Figure D-2 shows the comparison of average spillback probability and entry flow for all time period. Figure D-2(a), Figure D-2(b), Figure D-2(c), and Figure D-2(d) show the comparison of the average spillback probability Figure D-2(e), Figure D-2(f), Figure D-2(g), and Figure D-2(h) show the comparison of entry flow. In all figures, the performance of the signal plan derived by SO is displayed in solid line, and the performance of the current plan is displayed in dashed line. The signal plan derived by SO yields smaller average queue length and average spillback probability for all time intervals. The signal plan derived by SO has smaller average trip travel time for the first and second time intervals, and larger entry flow for the first and third time intervals.

We then study the performance of the signal plan derived by SO and the existing signal plan for the whole simulation period aggregately. Figure D-3(a) shows the average queue length; Figure D-3(b) shows the average spillback probability; Figure D-3(c) shows the average travel time of unfinished trips; Figure D-3(d) shows the average travel time of all finished trips; Figure D-3(e) shows the number of unfinished trips, Figure D-3(f) shows the number of finished trips.

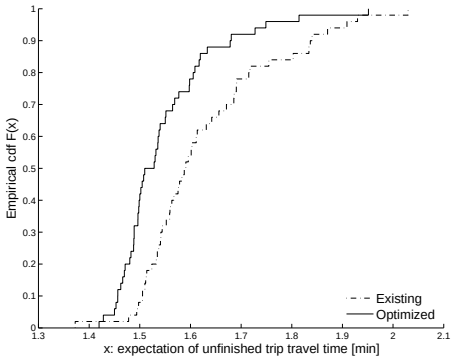
In Figure D-3, the derived signal plan yields smaller expected finished and unfinished trip travel time, smaller average queue length, and smaller average spillback probability. For number of finished and unfinished trips, the derived signal plan and the existing signal plan have similar performance.



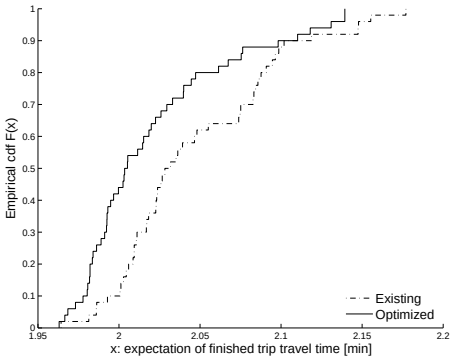
(a) average queue-length



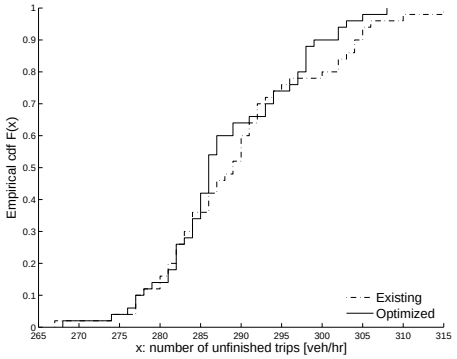
(b) average spillback probability



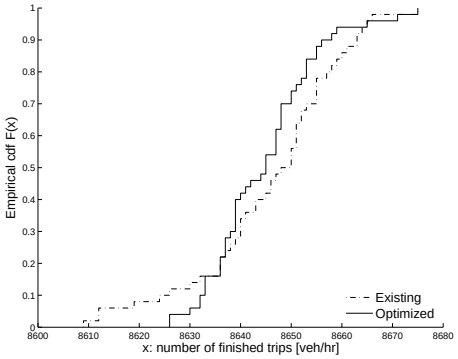
(c) expectation of unfinished trip travel time



(d) expectation of finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure D-3: Comparison of the performance of plan 1 and the existing signal plan for demand scenario 1.

	Existing plan				New plan			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	1.96	2.05	2.18	0.05	1.96	2.02	2.14	0.05
TP	8609	8645	8675	14.53	8626	8645	8675	10.56
TTun	1.37	1.62	2.03	0.13	1.42	1.55	1.95	0.10
TPun	267	290	315	10.15	268	288	308	8.72
QL	0.82	0.88	1.00	0.05	0.90	0.83	1.06	0.05
SP	0.0283	0.0308	0.0337	0.0012	0.0258	0.0284	0.0316	0.0012

Table D.1: Performance metrics statistics for plan 1 and existing signal plan.

Table D.1 shows the statistics of each performance measure over 50 replications. For each performance measure, we calculate the minimum value (min), mean value (mean), maximum value (max) and standard deviation (σ) for existing plan and new plan. TT represents average trip travel time of finished trips; TP is the number of finished trips; TTun is the average trip travel time for unfinished trips; TPun is the number of unfinished trips; QL is the average network queue-length; SP is the average spillback probability.

For each aggregated performance measure, the null hypothesis states that the performance measure of the signal plan proposed by SO method is equal to that of the existing signal plan, the alternative hypothesis states that the performance measure of the signal plan proposed by SO method is better (e.g.: smaller travel time for unfinished trips; smaller average travel time for finished trips; smaller average queue length; smaller average spillback probability; smaller number of unfinished trips and lager number of finished trips) than that of the existing signal plan. When running the 50 simulation replications to evaluate the performance of a given signal plan, we

use the same set of 50 replication seeds for each signal plan. The paired t-tests are carried out by pairing observations that have common seeds. Let $\overline{Y}$ denote the average paired difference between any two aggregated performance measures , let $\hat{s}$ denote its standard deviation, and let O denote the sample size. Then the paired t-statistic is given by (see, for instance, Hogg et al. (1977, p. 486)): $t = \sqrt{O}\ \overline{Y}/\hat{s}$.

Taking the average finished trip travel time as an example, we test the hypothesis that the average finished trip travel time from plan 1 is equal to the average finished trip travel time obtained from existing signal plan. The mean of the paired differences $\overline{Y}$ is around 0.0264 minutes. The standard deviation of the paired differences $\hat{s}$ is around 0.0397 minutes. The sample size O is 50. The critical value at the 2.5% significance level is $t_{0.025}(49) = 2.01$. The t values is 4.6998. Thus the null hypothesis is rejected. We show the paired t-test results for each performance measure in Table D.2. For average number of finished trips and unfinished trips, the t-values are 0.5693 and 1.0618 respectively, thus the null hypothesis is accepted. For all the other performance metrics, the t-values are larger than $t_{0.025}(49) = 2.01$, thus the null hypothesis is rejected. Plan 1 derived by SO leads to significant smaller average trip travel time for both finished and unfinished trips, shorter queue length and smaller average spillback probability. The number of finished and unfinished trips obtained from both plans are similar.

Plan 1 and existing signal plan leads to similar number of finished and unfinished

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	0.0264	0.0397	4.6998
TP	1.0400	12.9173	0.5693
TTun	0.0817	0.1180	4.8914
TPun	1.4200	9.4569	1.0618
QL	0.0224	0.0644	2.4565
SP	0.0024	0.0015	11.3563

Table D.2: Paired t-test for plan 1 and existing signal plan.

trips, however, travel time of finished and unfinished trips are significantly reduced, moreover, average queue length over all queues over time and the average spillback probability per queue over time are significant reduced. Plan 1 leads to better system performance while maintaining the same level of system throughout.

Plan 2 and plan 7 are the same as existing signal plan, the performance comparison of plan 4 and existing signal plan has been shown in Chapter 5.

Comparison of the performance of plan 3 and the existing signal plan

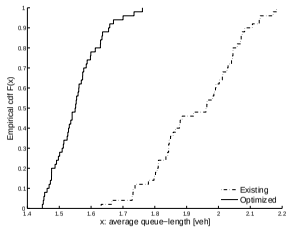
This section shows the comparison of the performance of the signal plan derived by SO and the existing signal plan under demand scenario 3.

Figure D-4 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure D-4(a), Figure D-4(b), Figure D-4(c), and Figure D-4(d) show the comparison of average queue-length of the adaptive signal settings and the existing signal plan for each 15 minutes; Figure D-4(e), Figure D-4(f), Figure D-4(g), and Figure D-4(h) show the the comparison of average trip travel time.

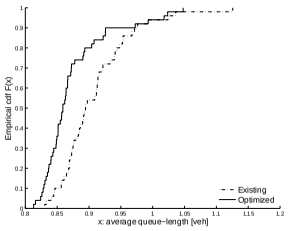
Figure D-5 shows the comparison of average spillback probability and entry flow for all time period. Figure D-5(a), Figure D-5(b), Figure D-5(c), and Figure D-5(d) show the comparison of the average spillback probability Figure D-5(e), Figure D-5(f), Figure D-5(g), and Figure D-5(h) show the comparison of entry flow. For all time intervals, the signal plan derived by SO yields smaller average queue-length, smaller average trip travel time, smaller average spillback probability, and larger entry flow.

We then study the performance of the signal plan derived by SO and the existing signal plan for the whole simulation period aggregately. Figure D-6(a) shows the average queue length; Figure D-6(b) shows the average spillback probability; Figure D-6(c) shows the average travel time of unfinished trips; Figure D-6(d) shows the average travel time of all finished trips; Figure D-6(e) shows the number of unfinished trips; Figure D-6(f) shows the number of finished trips. We use across-replication variability to represents the day-to-day variability. The proposed signal plan leads to steeper cdf curves in terms of number of finished trips and expectation of unfinished trip travel time, which means more stable performance.

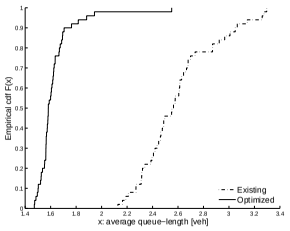
Table D.3 shows the statistics of each performance measure over 50 replications. The new plan derived by SO yields smaller average queue length, smaller spillback probability, smaller average finished trip travel time but larger expectation of unfinished trip travel time. The proposed signal plan leads to larger number of finished trips which means the system throughput is increased, more travelers could pass the net-



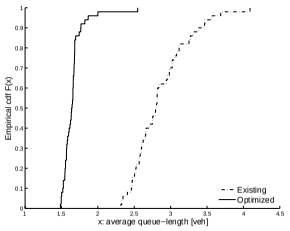
(a) average queue-length for the first 15 mins



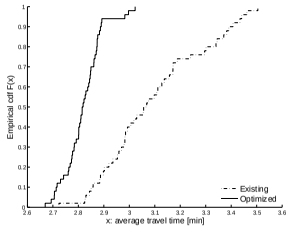
(b) average queue-length for the second 15 mins



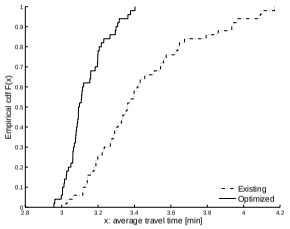
(c) average queue-length for the third 15 mins



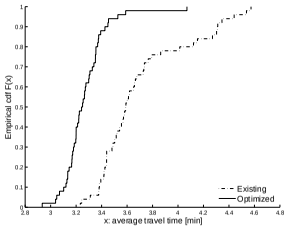
(d) average queue-length for the fourth 15 mins



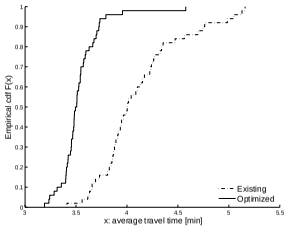
(e) average trip travel time for the first 15 mins



(f) average trip travel time for the second 15 mins

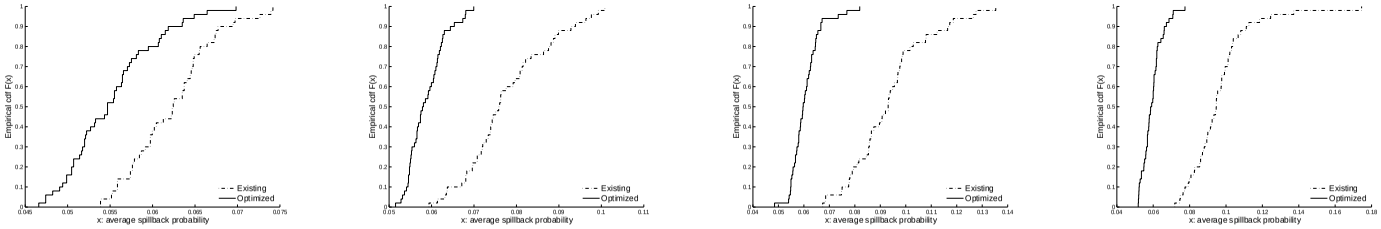


(g) average trip travel time for the third 15 mins



(h) average trip travel time for the fourth 15 mins

Figure D-4: Comparison of the average queue-length and average trip travel time of plan 3 and the existing signal plan.

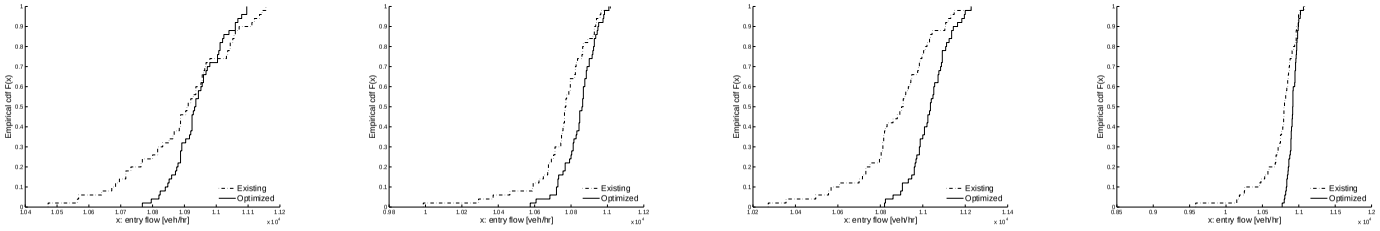


(a) average spillback probability for the first 15 mins

(b) average spillback probability for the second 15 mins

(c) average spillback probability for the third 15 mins

(d) average spillback probability for the fourth 15 mins



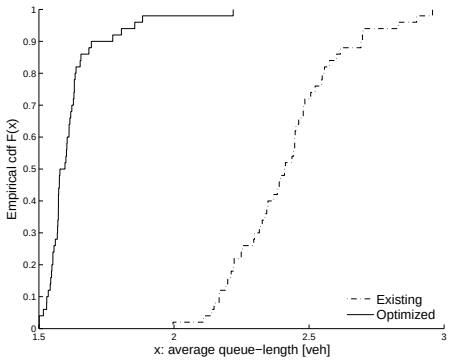
(e) entry flow for the first 15 mins

(f) entry flow for the second 15 mins

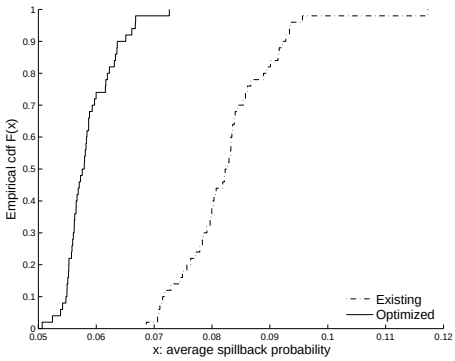
(g) entry flow for the third 15 mins

(h) entry flow for the fourth 15 mins

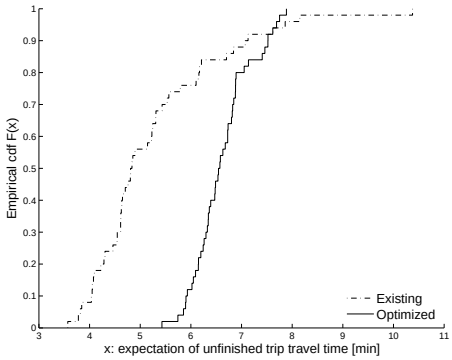
Figure D-5: Comparison of the average spillback probability and entry flow of plan 3 and the existing signal plan.



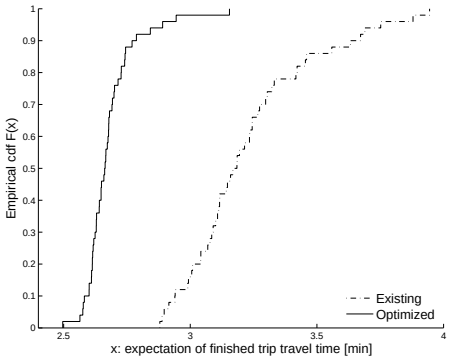
(a) average queue-length



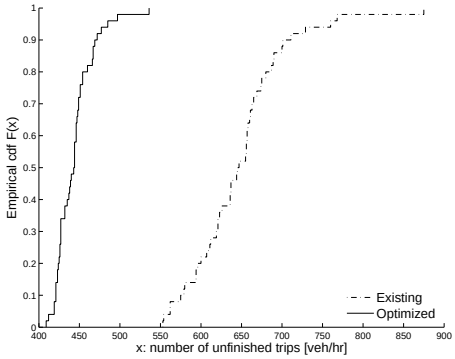
(b) average spillback probability



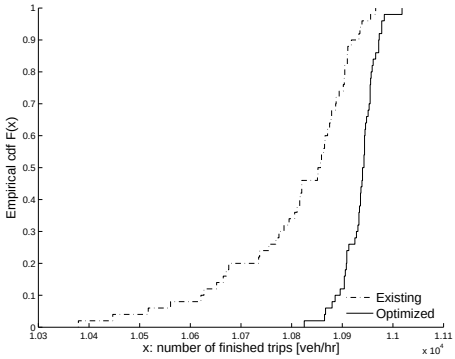
(c) expectation of unfinished trip travel time



(d) expectation of finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure D-6: Comparison of the performance of plan 3 and the existing signal plan for demand scenario 3.

	Existing plan				New plan			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	2.88	3.23	3.94	0.25	2.50	2.68	3.15	0.11
TP	10379	10805	10966	132.34	10825	10936	11018	34.07
TTun	3.57	5.26	10.38	1.33	5.43	6.62	7.89	0.56
TPun	553	647	875	58.92	409	444	536	23.33
QL	2.00	2.41	2.96	0.2	1.50	1.62	2.22	0.12
SP	0.0686	0.0826	0.1173	0.0084	0.0507	0.0586	0.0726	0.0042

Table D.3: Performance metrics statistics for plan 3 and existing signal plan.

work with a reduced average trip travel time. Comparing to the existing signal plan, the proposed signal plan also reduces the number of unfinished trips which includes the travelers that are blocked in the network. Improving total system throughput (number of finished trips) while reducing number of cars being blocked is not a simple task, thus the average travel time for unfinished trips is increased. For those small amount of travelers, they are suffering longer travel time, but for majority of the road network users, the travel time are reduced significantly, furthermore, more traveller are served.

Table D.4 shows the paired t-test for each performance measure. Besides average unfinished trip travel time, plan 3 yields significant better performance than existing signal plan for all other performance metrics. Note that the standard deviation of the paired differences $\hat{s}$ for number of finished trips is large, that is because of the long tail in the cdf curve of the number of finished trips obtained by existing signal plan.

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	0.5492	0.2528	15.3637
TP	130.9400	130.1554	7.1137
TTun	-1.3549	1.2604	-7.6012
TPun	203.6200	65.1601	22.0965
QL	0.7963	0.2345	24.0166
SP	0.0240	0.0094	18.0012

Table D.4: Paired t-test for plan 3 and existing signal plan.

Comparison of the performance of plan 5 and the existing signal plan

This section shows the comparison of the performance of the signal plan derived by SO and the existing signal plan under demand scenario 5.

Figure D-7 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure D-7(a), Figure D-7(b), Figure D-7(c), and Figure D-7(d) show the comparison of average queue-length of the adaptive signal settings and the existing signal plan for each 15 minutes; Figure D-7(e), Figure D-7(f), Figure D-7(g), and Figure D-7(h) show the the comparison of average trip travel time.

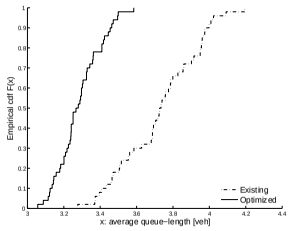
Figure D-8 shows the comparison of average spillback probability and entry flow for all time period. Figure D-8(a), Figure D-8(b), Figure D-8(c), and Figure D-8(d) show the comparison of the average spillback probability Figure D-8(e), Figure D-8(f), Figure D-8(g), and Figure D-8(h) show the comparison of entry flow. For all performance measures and all time intervals, the signal plan derived by SO yields smaller average queue-length, smaller average trip travel time, smaller average spillback prob-

ability, and larger entry flow. Furthermore, proposed signal plan leads to smaller cross replication variability for average trip travel time, average spillback probability and entry flow comparing to existing signal plan.

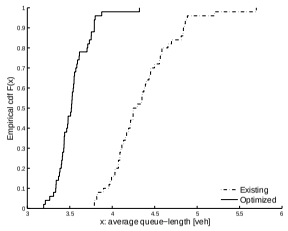
This section shows the comparison of the performance of the signal plan derived by SO and the existing signal plan under demand scenario 5.

Figure D-9(a) shows the average queue length; Figure D-9(b) shows the average spillback probability; Figure D-9(c) shows the average travel time of unfinished trips; Figure D-9(d) shows the average travel time of all finished trips; Figure D-9(e) shows the number of unfinished trips; Figure D-9(f) shows the number of finished trips. Besides average queue length and expectation of finished trip travel time, proposed signal plan leads to smaller across-replication variability for all other performance metrics and offers more stable service.

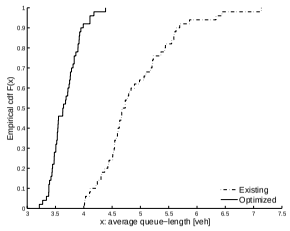
Table D.5 shows the statistics of each performance measure over 50 replications. The new plan derived by SO yields smaller expectation of finished and unfinished trip travel time, larger number of finished trips, smaller number of unfinished trips, smaller average queue-length and smaller average spillback probability. Table D.6 shows the paired t-test for each performance measure. Plan 5 yields significant better performance than existing signal plan in terms of all performance metrics. Plan 5 systematically increases the number of finished trip without imposing extra travel time, and reduces the number of vehicles being blocked in the network.



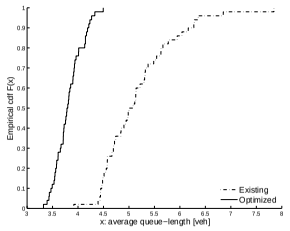
(a) average queue-length for the first 15 mins



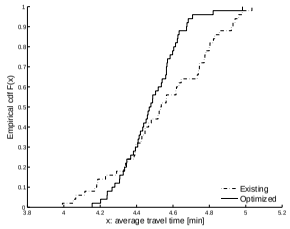
(b) average queue-length for the second 15 mins



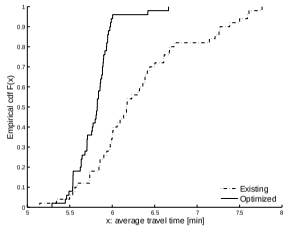
(c) average queue-length for the third 15 mins



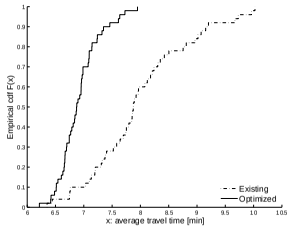
(d) average queue-length for the fourth 15 mins



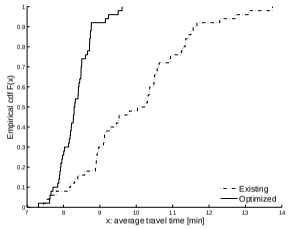
(e) average trip travel time for the first 15 mins



(f) average trip travel time for the second 15 mins

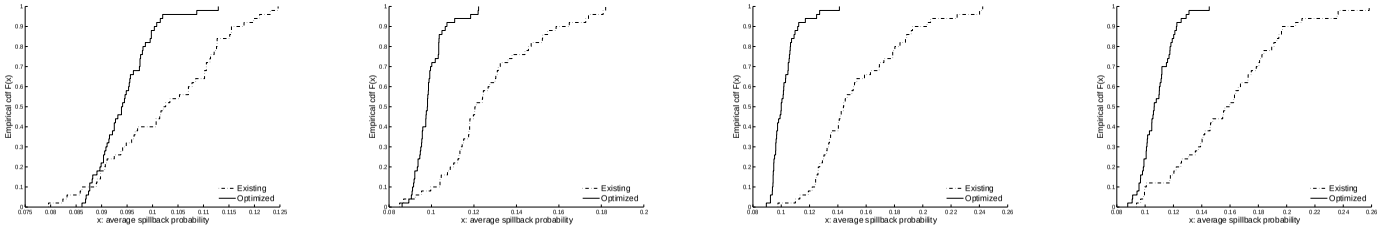


(g) average trip travel time for the third 15 mins



(h) average trip travel time for the fourth 15 mins

Figure D-7: Comparison of the average queue-length and average trip travel time of plan 5 and the existing signal plan.

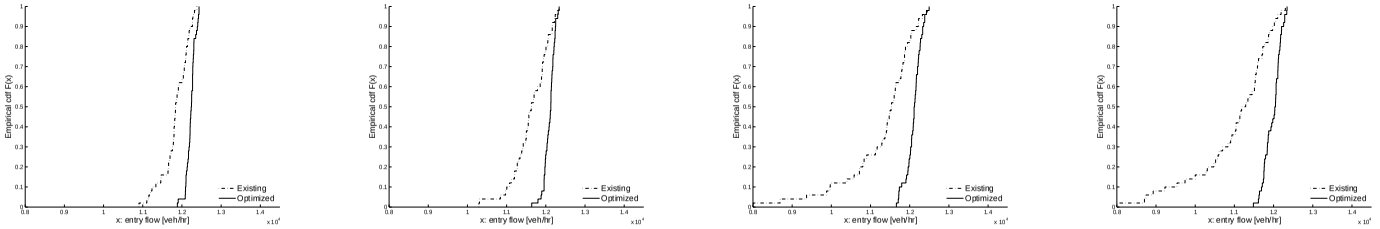


(a) average spillback probability for the first 15 mins

(b) average spillback probability for the second 15 mins

(c) average spillback probability for the third 15 mins

(d) average spillback probability for the fourth 15 mins



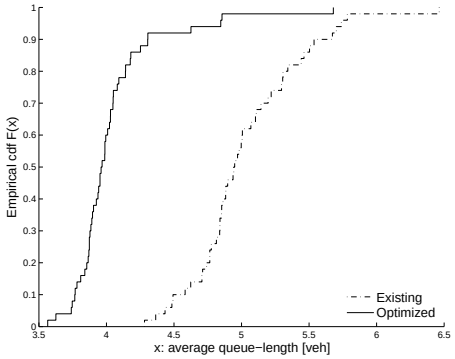
(e) entry flow for the first 15 mins

(f) entry flow for the second 15 mins

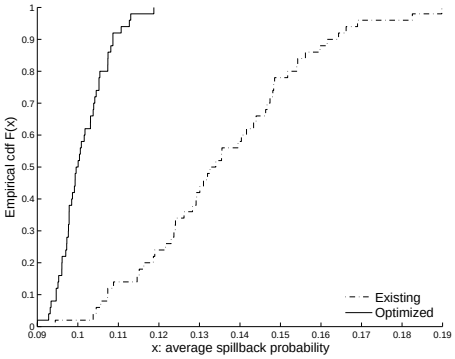
(g) entry flow for the third 15 mins

(h) entry flow for the fourth 15 mins

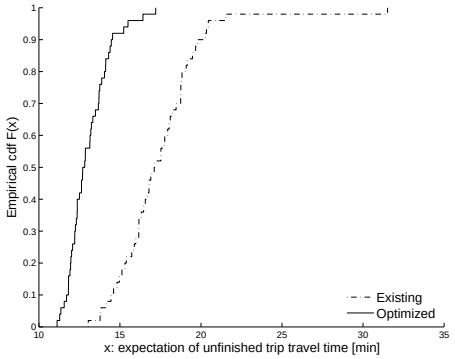
Figure D-8: Comparison of the average spillback probability and entry flow of plan 5 and the existing signal plan.



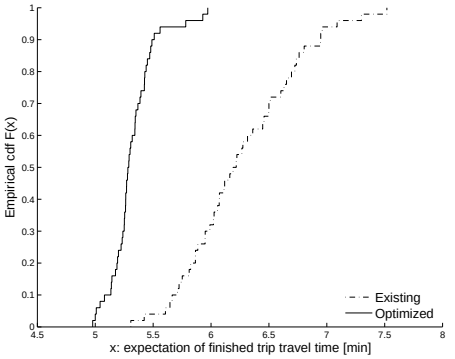
(a) average queue-length



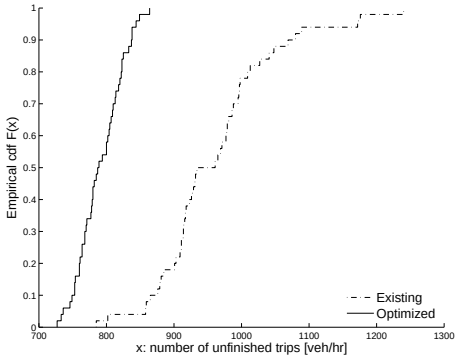
(b) average spillback probability



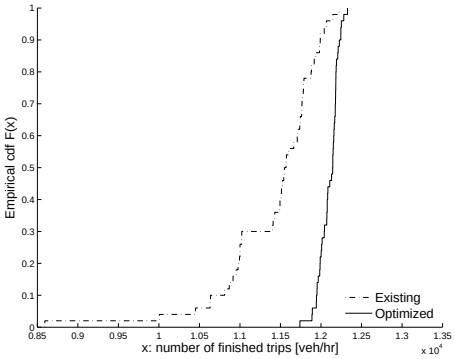
(c) expectation of unfinished trip travel time



(d) expectation of finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure D-9: Comparison of the performance of plan 5 and the existing signal plan for demand scenario 5.

	Existing plan				New plan			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	5.31	6.27	7.52	0.49	4.98	5.32	5.97	0.20
TP	8593	11415	12225	632.07	11740	12102	12327	117.76
TTun	13.05	17.46	31.51	2.82	11.12	13.06	17.21	1.29
TPun	785	961	1240	88.71	727	791	864	33.12
QL	4.28	5.02	6.46	0.41	3.57	4.04	5.68	0.34
SP	0.0944	0.1356	0.1898	0.0209	0.0900	0.1012	0.1188	0.0059

Table D.5: Performance metrics statistics for plan 5 and existing signal plan.

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	0.9443	0.5131	13.0132
TP	686.3400	645.2933	7.5209
TTun	4.3999	3.0395	10.2358
TPun	170.3600	91.2737	13.1980
QL	0.9907	0.4995	14.0253
SP	0.0344	0.0219	11.0871

Table D.6: Paired t-test for plan 5 and existing signal plan.

Comparison of the performance of plan 6 and the existing signal plan

This section shows the comparison of the performance of the signal plan derived by SO and the existing signal plan under demand scenario 6.

Figure D-10 shows the comparison of average queue-length and average trip travel time from the first time interval until the fourth time interval. Figure D-10(a), Figure D-10(b), Figure D-10(c), and Figure D-10(d) show the comparison of average queue-length of the adaptive signal settings and the existing signal plan for each 15 minutes; Figure D-10(e), Figure D-10(f), Figure D-10(g), and Figure D-10(h) show the the comparison of average trip travel time.

Figure D-11 shows the comparison of average spillback probability and entry flow

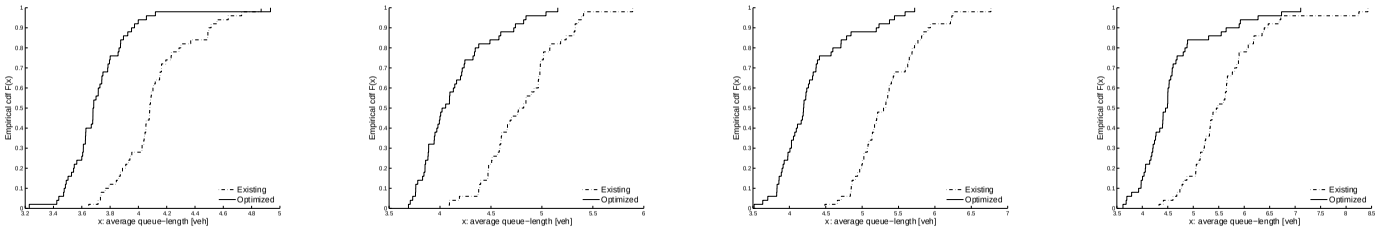
	Existing plan				New plan			
	Min	Mean	Max	σ	Min	Mean	Max	σ
TT	6.74	8.09	9.47	0.60	5.53	6.90	7.76	0.47
TP	8370	11342	12248	719.72	10490	12071	12420	351.35
TTun	18.27	22.19	35.89	2.87	14.01	18.09	25.83	1.70
TPun	896	1015	1249	85.35	760	905	1213	97.97
QL	4.36	4.92	6.14	0.33	3.57	4.04	5.68	0.34
SP	0.1153	0.1549	0.1980	0.0180	0.1291	0.1525	0.2050	0.0162

Table D.7: Performance metrics statistics for plan 6 and existing signal plan.

for all time period. Figure D-11(a), Figure D-11(b), Figure D-11(c), and Figure D-11(d) show the comparison of the average spillback probability Figure D-11(e), Figure D-11(f), Figure D-11(g), and Figure D-11(h) show the comparison of entry flow. For the first and second time intervals, the signal plan derived by SO yields smaller average queue-length, smaller average trip travel time, larger network throughput but larger average spillback probability. For the last two time intervals, plan 6 leads to better performance in terms of all performance metrics.

Figure D-12(a) shows the average queue length; Figure D-12(b) shows the average spillback probability; Figure D-12(c) shows the average travel time of unfinished trips; Figure D-12(d) shows the average travel time of all finished trips; Figure D-12(e) shows the number of unfinished trips; Figure D-12(f) shows the number of finished trips.

Table D.7 shows the statistics of each performance measure over 50 replications. The new plan derived by SO yields smaller average queue-length, similar spillback probability, smaller expectation of finished and unfinished trip travel time, smaller number of unfinished trips, and larger number of finished trips.

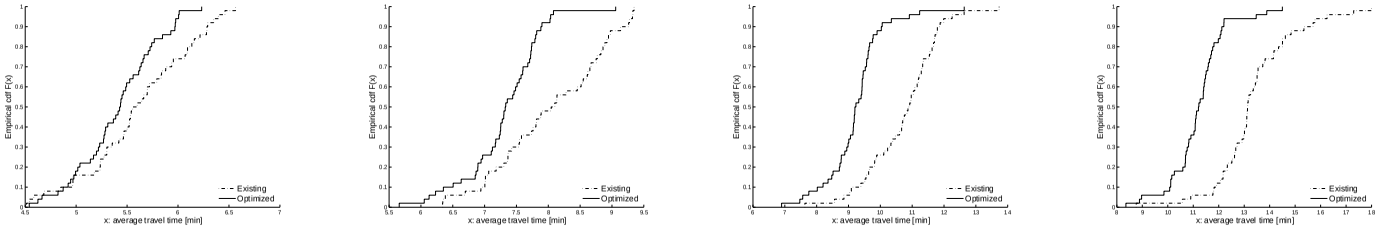


(a) average queue-length for the first 15 mins

(b) average queue-length for the second 15 mins

(c) average queue-length for the third 15 mins

(d) average queue-length for the fourth 15 mins



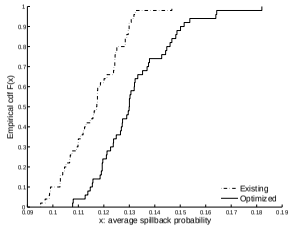
(e) average trip travel time for the first 15 mins

(f) average trip travel time for the second 15 mins

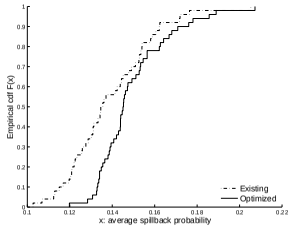
(g) average trip travel time for the third 15 mins

(h) average trip travel time for the fourth 15 mins

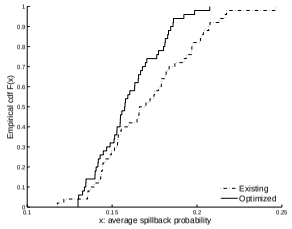
Figure D-10: Comparison of the average queue-length and average trip travel time of plan 6 and the existing signal plan.



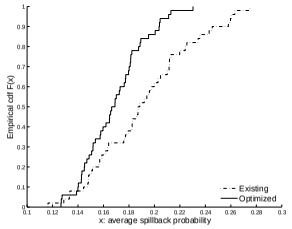
(a) average spillback probability for the first 15 mins



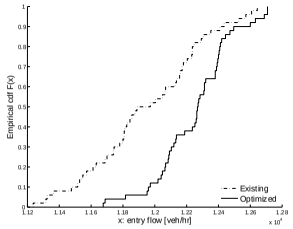
(b) average spillback probability for the second 15 mins



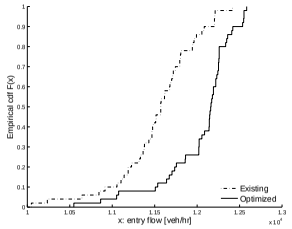
(c) average spillback probability for the third 15 mins



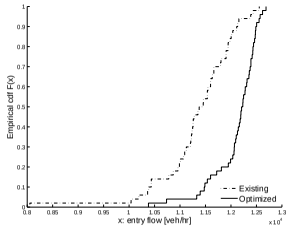
(d) average spillback probability for the fourth 15 mins



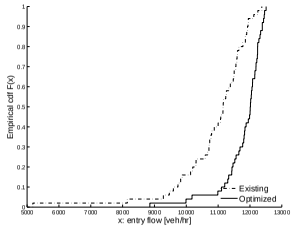
(e) entry flow for the first 15 mins



(f) entry flow for the second 15 mins

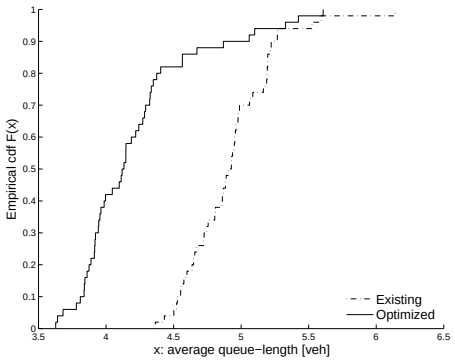


(g) entry flow for the third 15 mins

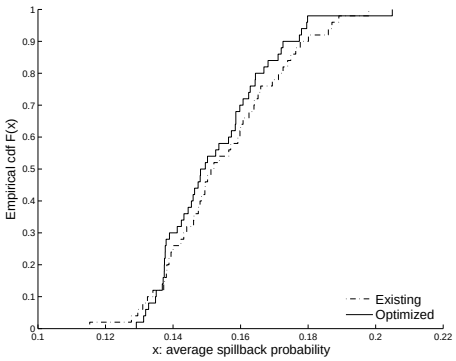


(h) entry flow for the fourth 15 mins

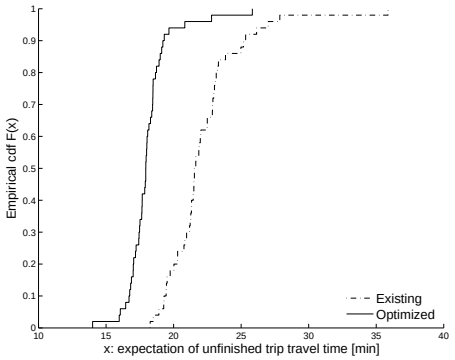
Figure D-11: Comparison of the average spillback probability and entry flow of plan 6 and the existing signal plan.



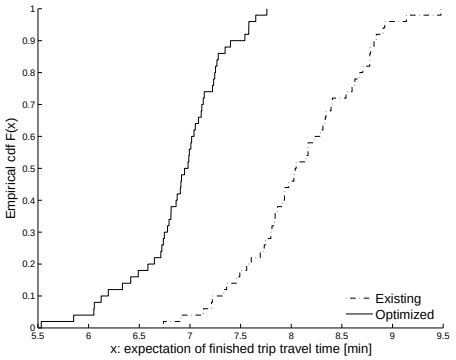
(a) average queue-length



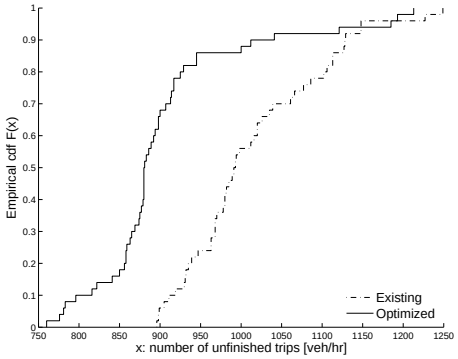
(b) average spillback probability



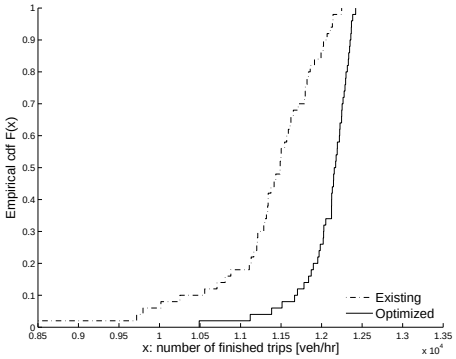
(c) expectation of unfinished trip travel time



(d) expectation of finished trip travel time



(e) number of unfinished trips



(f) number of finished trips

Figure D-12: Comparison of the performance of the plan 6 and the exiting signal plan for demand scenario 6.

	$\overline{Y}$	$\hat{s}$	t-statistic
TT	1.1950	0.5857	14.4271
TP	728.7800	816.4712	6.3116
TTun	4.0924	3.6788	7.8661
TPun	110.0800	134.0330	5.8074
QL	0.9907	0.4995	14.0253
SP	0.0023	0.0267	0.6130

Table D.8: Paired t-test for plan 6 and existing signal plan.

Table D.8 shows the paired t-test for each performance measure. Besides average spillback probability, plan 6 leads to significant better performance than existing signal plan for all other performance metrics. Plan 6 and existing signal plan have similar performance in terms of average spillback probability. Under demand scenario 6, the demand level is 20% higher than the normal morning peak demand, in which spillback could easily happen. In this case, the proposed signal plan does not outperform existing signal plan in terms of spillback probability might due to the high level of demand.

Appendix E

Look-up tables

We show the look-up table for each signal plan in Table E.1, Table E.2, Table E.3, Table E.4, Table E.5, and Table E.6. In each table, PM_j represents the vector of total average link travel time for demand scenario j, which contains 300 simulation replications. In each table, minimum, maximum, mean and standard deviation of the total average link travel times for traffic condition j are indicated by $minPM_j$, $maxPM_j$, $\overline{PM_j}$ and σ_j, TT intervals shows the lower and upper bounds of total average link travel time specified for each demand scenario. Taking Table E.1 as an example, from demand scenario 1 to demand scenario 7, the values of $\overline{PM_j}$ and σ_j increase, which indicates the growing across-replication variability. We use the across-replication variability to indicates day-to-day travel time variability, as demand increases, traffic dynamics are more complicated, thus the variability of the performance measure increases. For highly congested network, reducing travel time variability or focusing on eliminating

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	29.70	31.00	35.26	1.21	(0, 34)
scenario 2	32.24	39.94	73.10	5.45	[34,61)
scenario 3	63.98	98.17	127.46	15.97	[61,126)
scenario 4	97.68	148.43	207.94	21.27	[126,172)
scenario 5	102.13	182.39	250.47	26.30	[172, 207)
scenario 6	131.58	211.68	301.75	33.15	[207,240)
scenario 7	163.43	257.20	363.97	34.52	[240, inf)

Table E.1: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 1.

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	28.56	30.42	34.71	1.25	(0, 34)
scenario 2	31.76	37.64	46.92	2.22	[34,45)
scenario 3	46.10	56.37	70.80	4.83	[45,71)
scenario 4	68.68	85.93	116.48	7.80	[71,104)
scenario 5	102.12	126.11	160.03	13.46	[104, 135)
scenario 6	128.17	160.42	210.51	17.81	[135, 164)
scenario 7	150.68	186.12	271.47	18.34	[164, inf)

Table E.2: Average total link travel time statistic and link travel time interval classification corresponding to different demand scenarios under plan 2 & plan 7 (existing signal plan plan).

the long tail of the cdf curve might improve system performance.

From Figure E-1 to Figure E-6, in each figure seven cdf curves of the performance measure corresponding to each level of demand are displayed. X-axis represents the average total link travel time over all links of interest. From the left to the right, each curve represents a scenario from lowest demand (scenario 1) to highest demand (scenario 7). The vertical lines classify travel time boundaries from b_1 to b_6. It is more clear that from the left to the right, the variability across-replications become larger as demand increases in most of the cases.

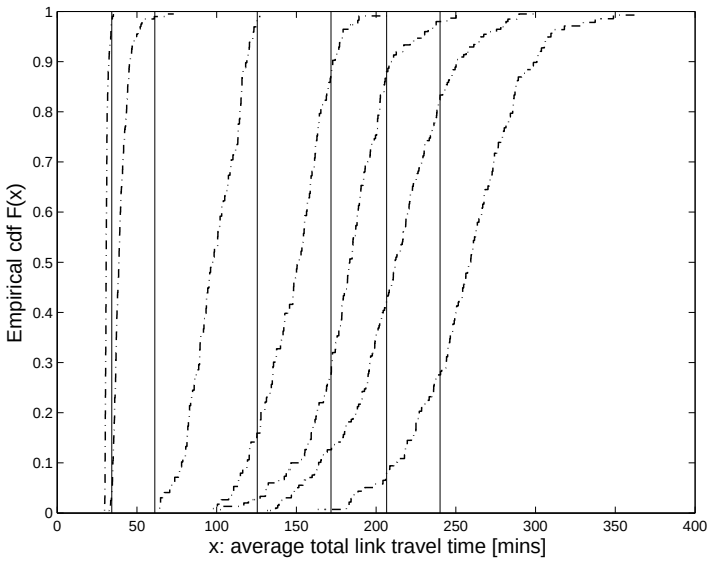


Figure E-1: Average total link travel time cdfs according to different demand levels based on signal plan 1.

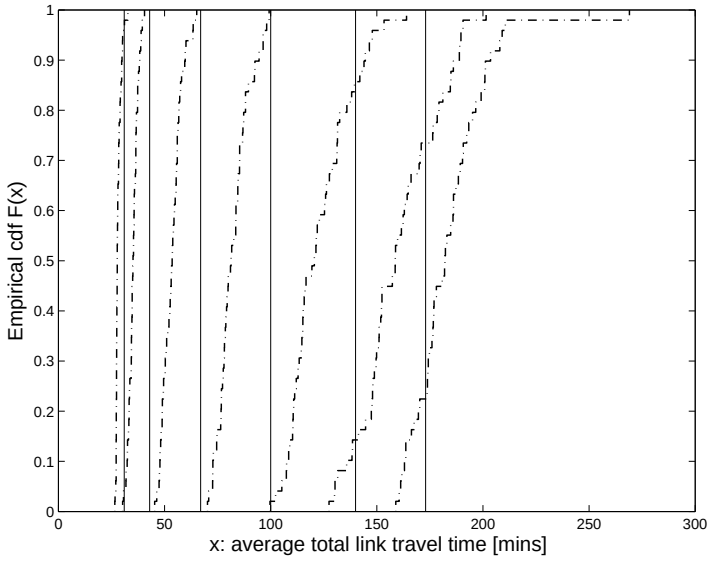


Figure E-2: Average total link travel time cdfs according to different demand levels based on plan 2 & plan 7 (existing signal plan plan).

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	40.77	42.94	45.14	0.83	(0, 44)
scenario 2	42.68	44.50	47.12	0.78	[44,47)
scenario 3	46.47	49.66	64.77	3.44	[47,60)
scenario 4	59.94	86.75	163.67	15.52	[60,104)
scenario 5	97.96	127.57	211.81	17.37	[104, 136)
scenario 6	123.20	157.37	199.50	18.25	[136,171)
scenario 7	159.51	189.34	239.15	14.40	[171, inf)

Table E.3: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 3.

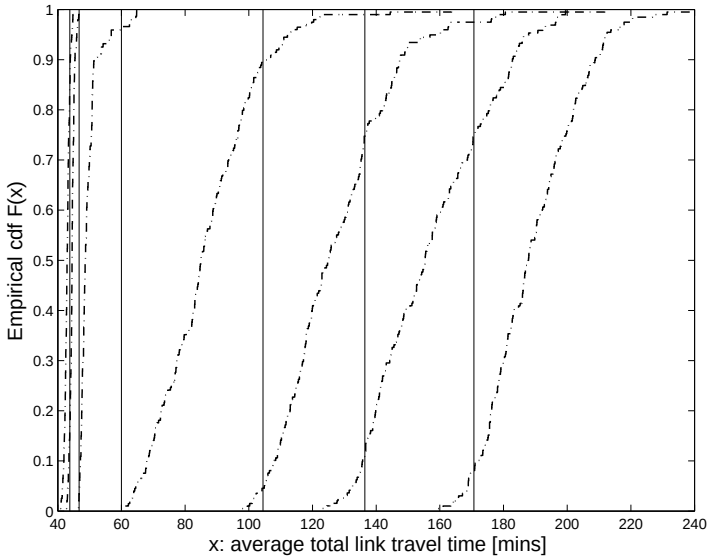


Figure E-3: Average total link travel time cdfs according to different demand levels based on signal plan 3.

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	36.09	38.50	41.82	1.14	(0, 40)
scenario 2	37.08	40.98	44.82	1.27	[40,43)
scenario 3	40.85	45.76	53.19	1.96	[43,53)
scenario 4	53.54	64.95	85.64	5.49	[53,81)
scenario 5	72.83	93.74	108.93	6.60	[81, 109)
scenario 6	109.86	125.91	154.54	10.41	[109,140)
scenario 7	136.88	150.85	166.87	6.97	[140, inf)

Table E.4: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 4.

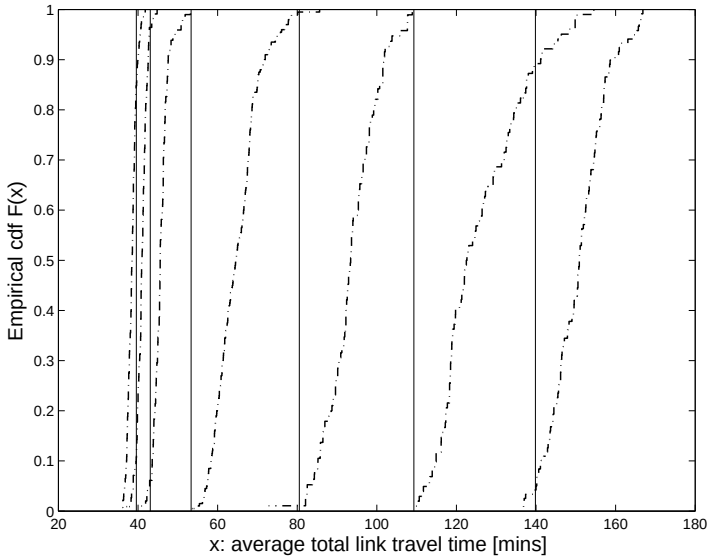


Figure E-4: Average total link travel time cdfs according to different demand levels based on signal plan 4.

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	31.84	39.90	41.91	1.13	(0, 41)
scenario 2	38.33	42.90	46.20	1.00	[41,46)
scenario 3	46.10	50.09	62.06	2.79	[46,60)
scenario 4	60.45	75.04	123.36	7.46	[60,89)
scenario 5	89.31	109.65	172.22	12.32	[89, 124)
scenario 6	127.96	145.00	196.05	16.22	[124,159)
scenario 7	146.53	169.25	194.77	10.28	[159, inf)

Table E.5: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 5.

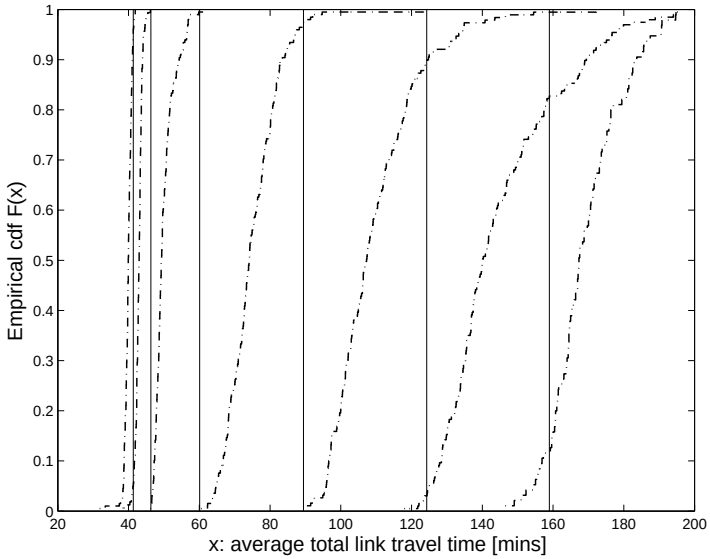


Figure E-5: Average total link travel time cdfs according to different demand levels based on signal plan 5

	$minPM_j$	$\overline{PM_j}$	$maxPM_j$	σ_j	TT interval
scenario 1	89.28	92.45	95.23	1.09	(0, 94)
scenario 2	92.57	96.19	98.94	1.16	[94,98)
scenario 3	96.35	100.85	107.46	1.88	[98,106)
scenario 4	106.16	114.27	151.30	6.67	[106,123)
scenario 5	116.91	134.11	173.47	10.13	[123, 143)
scenario 6	142.06	157.97	197.23	10.94	[143,171)
scenario 7	168.41	182.85	208.21	6.98	[171, inf)

Table E.6: Average total link travel time statistic and link travel time interval classification according to different demand scenario under signal plan 6.

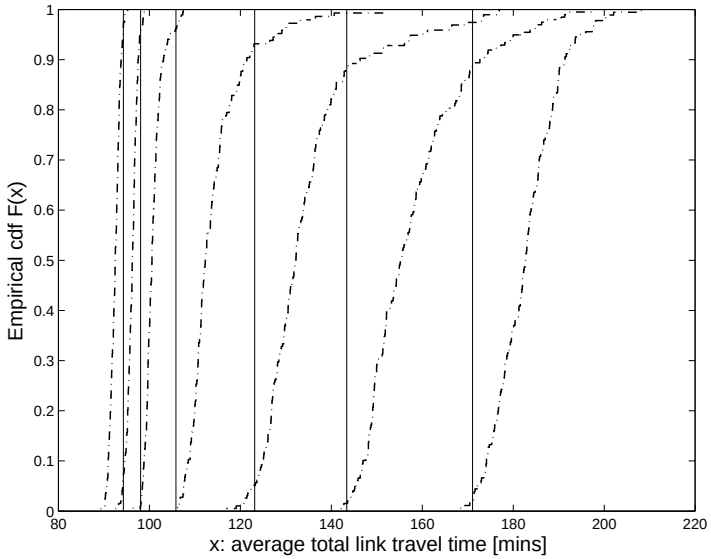


Figure E-6: Average total link travel time cdfs according to different demand levels based on signal plan 6

Figure E-1 shows the cdfs of the performance measure obtained from signal plan 1 for each demand scenario. Plan 1 is derived under the demand scenario with the lowest demand. Compared to other signal plans designed for demand scenario with higher demand such as plan 4 and 5, average link travel time obtained from plan 1 significantly increases under demand scenario 4, 5, 6 and 7.

On the contrary, Figure E-6 shows the cdfs of the performance measure obtained from signal plan 6 for each demand scenario. Plan 6 is derived under demand scenario 6. Compared to other signal plans designed for demand scenario with lower demand, average link travel time obtained from plan 6 is much larger under demand scenario 1, 2, and 3.

This proves that signal plans designed for light traffic is not suitable for heavy traffic. Similarly, signal plan designed for heavy traffic might have poor performance under light traffic. Selecting the most appropriate signal plan under different traffic conditions helps to reduce travel time and enhance system throughput.

Bibliography

Abdel-Aty, A., K. R. and Jovanis, P. (1996). Investigation of effect of travel time variability on route choice using repeated measurement stated preference data. 1493:39–45.

Aboudolas, K., Papageorgiou, M., Kouvelas, A., and Kosmatopoulos, E. (2010). A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks. *Transportation Research Part C: Emerging Technologies*, 18(5):680 – 694.

Abu-Lebdeh, G. and Benekohal, R. (1997). Development of traffic control and queue management procedures for oversaturated arterials. *Transportation Research Record*, 1603:119–127.

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, Hampoton, VA, USA.

Barceló, J. (2010). *Fundamentals of traffic simulation*, volume 145 of *International Series in Operations Research and Management Science*. Springer, New York, USA.

Barton, R. R. and Meckesheimer, M. (2006). Metamodel-based simulation optimization. In Henderson, S. G. and Nelson, B. L., editors, *Handbooks in operations research and management science: Simulation*, volume 13, chapter 18, pages 535–574. Elsevier, Amsterdam.

Bates, J., Polak, J., Jones, P., and Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, 37(2):191–229.

Batley, R. and Ibáñez, N. (2009). Demand effects of travel time reliability. In *International Choice Modelling Conference*.

Ben-Akiva, M., Cuneo, D., Hasan, M., Jha, M., and Yang, Q. (2003). Evaluation of freeway control using a microscopic simulation laboratory. *Transportation Research Part C*, 11(1):29–50.

Bertsimas, D. and Nakazato, D. (1995). The distributional Little's law and its applications. *Operations Research*, 43(2):298–310.

Black, I. G. and Towriss, J. G. (1997). *Demand Effects of Travel Time Reliability*. Centre for Transport Studies, Cranfield Institute of Technology.

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

Branke, J., Goldate, P., and Prothmann, H. (2007). Actuated traffic signal optimization using evolutionary algorithms. In *Proceedings of the 6th European Congress and Exhibition on Intelligent Transport Systems and Services*.

Brumelle, S. L. (1972). A generalization of L= λW to moments of queue length and waiting times. *Operations Research*, 20(6):1127–1136.

Bullock, D., Johnson, B., Wells, R. B., Kyte, M., and Li, Z. (2004). Hardware-in-the-loop simulation. *Transportation Research Part C*, 12(1):73 – 89.

Carrion, C. and Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, 46(4):720–741.

Charle, W., Viti, F., and Tampère, C. (2010). Estimating route travel time variability from link data by means of clustering. In *Proceedings of the 12th World Conference on Transport Research WCTR*.

Chen, A., Yang, H., Lo, H. K., and Tang, W. H. (1999a). A capacity related reliability for transportation networks. *Journal of Advanced Transportation*, 33(2):183–200.

Chen, B. Y., Lam, W. H., Sumalee, A., and Li, Z.-l. (2012a). Reliable shortest path finding in stochastic networks with spatial correlated link travel times. *International Journal of Geographical Information Science*, 26(2):365–386.

Chen, C., Skabardonis, A., and Varaiya, P. (2003). Travel time reliability as a measure of service. *Transportation Research Record: Journal of the Transportation Research Board*, 1855(1):74–79.

Chen, C.-H., Lin, J., Yücesan, E., and Chick, S. E. (2000). Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270.

Chen, C.-H., Wu, S. D., and Dai, L. (1999b). Ordinal comparison of heuristic algorithms using stochastic optimization. *Robotics and Automation, IEEE Transactions on*, 15(1):44–56.

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

Clark, S. and Watling, D. (2005). Modelling network travel time reliability under stochastic demand. *Transportation Research Part B: Methodological*, 39:119–140.

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., Scheinberg, K., and Vicente, L. N. (2009). Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points. *SIAM Journal on Optimization*, 20(1):387–415.

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

D'Este, G. and Taylor, M. A. (2003). Network vulnerability: an approach to reliability analysis at the level of national strategic transport networks. In *Network Reliability of Transport. Proceedings of the 1st International Symposium on Transportation Network Reliability (INSTR)*.

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.

EuropeanComission (2001). White paper european transport policy for 2010: time to decide. `http://ec.europa.eu/transport/themes/strategies/2001_white_paper_en.htm`. Accessed: 2014-07-25.

EuropeanComission (2011). White paper roadmap to a single european transport area - towards a competitive and resource efficient transport system. `http://ec.europa.eu/transport/themes/strategies/2011_white_paper_en.htm`. Accessed: 2014-07-25.

Flötteröd, G. and Osorio, C. (2013). Approximation of time-dependent multi-dimensional queue-length distributions. In *Proceedings of the Triennial Symposium on Transportation Analysis (TRISTAN)*.

Fu, L. and Hellinga, B. (2000). Delay variability at signalized intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 1710(1):215–221.

Fu, L. and Rilett, L. R. (1998). Expected shortest paths in dynamic and stochastic traffic networks. *Transportation Research Part B: Methodological*, 32(7):499–516.

Fu, M. C., Chen, C.-H., and Shi, L. (2008). Some topics for simulation optimization. In *Proceedings of the 40th Conference on Winter Simulation*, pages 27–38. Winter Simulation Conference.

Fu, M. C., Glover, F. W., and April, J. (2005). Simulation optimization: a review, new developments, and applications. In Kuhl, M. E., Steiger, N. M., Armstrong, F. B., and Joines, J. A., editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 83–95, Piscataway, New Jersey, USA.

Gartner, N. H. (1983). Opac: A demand-responsive strategy for traffic signal control. *Transportation Research Record*, (906).

Gettman, D., Fok, E., Curtis, E., Ormand, K. K. D., Mayer, M., and Flanigan, E. (2013). Measures of effectiveness and validation guidance for adaptive signal control technologies, FHWA-HOP-13-031. Technical report, U.S. Department of Transportation.

Gradshteyn, I. and Ryzhik, I. (2007). *Table of Integrals, Series, and Products*. Academic Press.

Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. (1998). *Fundamentals of queueing theory*. Wiley-Interscience.

Hachicha, W., Ammeri, A., Masmoudi, F., and Chachoub, H. (2010). A comprehensive literature classification of simulation optimisation methods. In *Proceedings of the International Conference on Multiple Objective Programming and Goal Programming MOPGP10*, Sousse, Tunisia.

Hagemann, G., Michaels, J., Minnice, P., Pace, D., Radin, S., Spiro, A., and West, R. (2010). Its technology adoption and observed market trends from its deployment tracking. Technical report.

Hajbabaie, A., Medina, J. C., Benekohal, R. F., and NEXTRANS, C. (2011). Traffic signal coordination and queue management in oversaturated intersections. Technical Report No. 047IY02.

Haji, R. and Newell, G. F. (1971). A relation between stationary queue and waiting time distributions. *Journal of Applied Probability*, pages 617–620.

Hale, D. (2005). Traffic network study tool TRANSYT-7F. Technical report, McTrans Center in the University of Florida, Gainesville, Florida.

He, R. R., Liu, H. X., Kornhauser, A. L., and Ran, B. (2002). Temporal and spatial variability of travel time. *Center for Traffic Simulation Studies. Paper UCI-ITS-TS-02*, 14.

Head, K. L., Mirchandani, P. B., and Sheppard, D. (1992). Hierarchical framework for real-time traffic control. *Transportation Research Record*, 1360:82–88.

Heffes, H. (1982). Moment formula for a class of mixed multi-job-type queueing networks. *Bell Syst. Tech. J*, 61(5):709–745.

Hogg, R. V., Tanis, E. A., and Rao, M. J. M. (1977). *Probability and statistical inference*, volume 993. Macmillan New York.

Hollander, Y. (2006). Direct versus indirect models for the effects of unreliability. *Transportation Research Part A: Policy and Practice*, 40(9):699–711.

Hu, Y. (2014). *The Impact of Pedestrian Activities in Adaptive Traffic Signal Control System Operations*. PhD thesis, University of Pittsburgh.

Hunt, P., Robertson, D., Bretherton, R., and Royle, M. (1982). The SCOOT on-line traffic signal optimisation technique. *Traffic Engineering & Control*, 23(4).

Hutton, J. M., Bokenkroger, C. D., and Meyer, M. M. (2010). Evaluation of an adaptive traffic signal system: Route 291 in Lee's Summit, Missouri. Technical Report OR 10-020.

Jackson, W. B. and Jucker, J. V. (1982). An empirical study of travel time variability and travel choice behavior. *Transportation Science*, 16(4):460–475.

Joshi, S., Rathi, A., and Tew, J. (1995). An improved response surface methodology algorithm with an application to traffic signal optimization for urban networks. In Alexopoulos, C., Kang, K., Lilegdon, W. R., and Goldsman, D., editors, *Proceedings of the 1995 Winter Simulation Conference*, pages 1104–1109.

Kamarajugadda, A. and Park, B. (2003). Stochastic traffic signal timing optimization. Technical report, Dept. of Civil Engineering, Center for Transportation Studies, Univ. of Virginia, Charlottesville, VA, USA.

Keilson, J. and Servi, L. (1988). A distributional form of Little's law. *Operations Research Letters*, 7(5):223–227.

Li, J.-Q. (2011). Discretization modeling, integer programming formulations and dynamic programming algorithms for robust traffic signal timing. *Transportation Research Part C: Emerging Technologies*, 19(4):708–719.

Li, P., Abbas, M., Pasupathy, R., and Head, L. (2010a). Simulation-based optimization of maximum green setting under retrospective approximation framework. *Transportation Research Record*, 2192:1–10.

Li, Z., Hensher, D. A., and Rose, J. M. (2010b). Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. *Transportation Research Part E*, 46(3):384–403.

Lin, S. (2011). *Efficient model predictive control for large-scale urban traffic networks*. PhD thesis, Delft University of Technology.

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

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

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

Liu, Y. and Chang, G.-L. (2011). An arterial signal optimization model for intersections experiencing queue spillback and lane blockage. *Transportation research part C: emerging technologies*, 19(1):130–144.

Lo, H. K., Chang, E., and Chan, Y. C. (2001). Dynamic network traffic control. *Transportation Research Part A: Policy and Practice*, 35(8):721–744.

Luyanda, F., Gettman, D., Head, L., Shelby, S., Bullock, D., and Mirchandani, P. (2003). Acs-lite algorithmic architecture: applying adaptive control system technology to closed-loop traffic signal control systems. *Transportation Research Record*, 1856(1):175–184.

Marshall, K. T. and Wolff, R. W. (1971). Customer average and time average queue lengths and waiting times. *Journal of Applied Probability*, pages 535–542.

Martin, P. T. (2007). Applications and benefits of adaptive traffic control systems in oversaturated conditions. Presentation at the 2007 TRB Workshop on Operating Traffic Signal Systems in Oversaturated Conditions, Washington DC. http://www.signalsystems.org.vt.edu/documents/Jan2007AnnualMeeting/Presentations/ATCS_Oversaturation_TRB.pdf. Accessed: 2014-07-25.

Mathworks, Inc. (2011). *Optimization Toolbox Version 6. User's Guide Matlab*. Natick, MA, USA.

McKenna, J. (1989). A generalization of Little's law to moments of queue lengths and waiting times in closed, product-form queueing networks. *Journal of Applied Probability*, pages 121–133.

Meier, P. (2007). Simulation of finite capacity queueing networks. Technical report, Transport and Mobility Laboratory, ENAC, Ecole Polytechnique Fédérale de Lausanne (EPFL).

Michalopoulos, P. G. and Stephanopoulos, G. (1977). Oversaturated signal systems with queue length constraintsii: Systems of intersections. *Transportation Research*, 11(6):423–428.

Mirchandani, P. and Soroush, H. (1987). Generalized traffic equilibrium with probabilistic travel times and perceptions. *Transportation Science*, 21(3):133–152.

Ng, M., Kockelman, K. M., and Waller, S. T. (2011). A review of the correlation coefficient as a dependence modeling tool. In *Proceedings of the Transportation Research Board (TRB) Conference*, Washington DC, USA.

Noland, R. B. and Polak, J. W. (2002). Travel time variability: a review of theoretical and empirical issues. *Transport Reviews*, 22(1):39–54.

NYCDOT (2014). Bridges. http://www.nyc.gov/html/dot/html/infrastructure/bridges.shtml. Accessed: 2014-07-25.

OECD (2010). Improving reliability on surface transport networks. Technical report, Organisation for Economic Co-operation and Development.

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

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

Osorio, C. and Bierlaire, M. (2009b). A surrogate model for traffic optimization of congested networks: an analytic queueing network approach. Technical Report 090825, Transport and Mobility Laboratory, ENAC, Ecole Polytechnique Fédérale de Lausanne. Available at: http://web.mit.edu/osorioc/www/papers/osorBier09TechRepQgTraf.pdf.

Osorio, C. and Bierlaire, M. (2013). A simulation-based optimization framework for urban transportation problems. *Operations Research*, 61(6):1333–1345.

Osorio, C. and Chong, L. (2012). Large-scale simulation-based traffic signal control. In *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA. Available at: http://web.mit.edu/osorioc/www/papers/osoChoLgeScaleSO.pdf.

Osorio, C. and Chong, L. (2013). A computationally efficient simulation-based optimization algorithm for large-scale urban transportation problems. *Transportation Science*. Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoChoLgeScaleSO.pdf.

Osorio, C. and Flötteröd, G. (2012). Capturing dependency among link boundaries in a stochastic network loading model. In *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA. Submitted to Transportation Science (minor revision status), available at: http://web.mit.edu/osorioc/www/papers/osoFlo13.pdf.

Osorio, C., Flötteröd, G., and Wang, C. (2013). Efficient calibration of stochastic traffic simulation models. In *Proceedings of the Symposium of the European Association for Research in Transportation (hEART)*.

Osorio, C. and Nanduri, K. (2012). Energy-efficient traffic management: a microscopic simulation-based approach. In *International Symposium on Dynamic Traffic Assignment (DTA)*, Martha's Vineyard, USA. Submitted to Transportation Science. Available at http://web.mit.edu/osorioc/www/papers/osoNanEnergySO.pdf.

Osorio, C. and Nanduri, K. (2013). Energy-efficient urban traffic management: a microscopic simulation-based approach. *Transportation Science*. Forthcoming. Available at: http://web.mit.edu/osorioc/www/papers/osoNanEnergySO.pdf.

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

Park, B. B. and Kamarajugadda, A. (2007). Development and evaluation of a stochastic traffic signal optimization method. *International Journal of Sustainable Transportation*, 1(3):193–207.

Peterson, M. D., Bertsimas, D. J., and Odoni, A. R. (1995). Decomposition algorithms for analyzing transient phenomena in multiclass queueing networks in air transportation. *Operations Research*, 43(6):995–1011.

Quinn, D. (1992). A review of queue management strategies. *Traffic Engineering+ Control*, 33(11):600–5.

Rakha, H., El-Shawarby, I., Arafeh, M., and Dion, F. (2006). Estimating path travel-time reliability. In *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*, pages 236–241. IEEE.

Robert L. Gordon, P.E., W. T. P. (2005). Traffic control systems handbook. Technical report, Federal Highway Administration Report FHWA-HOP-06-006.

Santos, B. F., Antunes, A. P., and Miller, E. J. (2010). Interurban road network planning model with accessibility and robustness objectives. *Transportation planning and technology*, 33(3):297–313.

Serafini, D. B. (1998). *A framework for managing models in nonlinear optimization of computationally expensive functions*. PhD thesis, Rice University.

Sims, A. and Dobinson, K. (1979). SCATS the sydney coordinated adaptive traffic system: philosophy and benefits. In *International Symposium on Traffic Control Systems, 1979, Berkeley, California, USA*, volume 2.

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.

Stevanovic, A. (2010). *Adaptive traffic control systems: Domestic and foreign state of practice*. Number Project 20-5: Topic 40-03. Transportation research board.

Stevanovic, A., Stevanovic, J., Zhang, K., and Batterman, S. (2009). Optimizing traffic control to reduce fuel consumption and vehicular emissions. *Transportation Research Record*, 2128:105–113.

Stevanovic, J., Stevanovic, A., Martin, P. T., and Bauer, T. (2008). Stochastic optimization of traffic control and transit priority settings in VISSIM. *Transportation Research Part C*, 16(3):332 – 349.

Texas Transportation Institute (2012). 2012 Urban mobility report. Technical report, Texas Transportation Institute (TTI), Texas A&M University System.

Tijms, H. C. (2003). *A First Course in Stochastic Models*. Wiley, Chichester, West Sussex, England.

Transport for London (2010). Traffic modelling guidelines. version 3.0. Technical report, Transport for London (TfL).

TRB (2000). *Highway capacity manual*. Transportation Research Board, National Research Council, Washington, D.C., USA.

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

TSS (2013). *AIMSUN 7 Dynamic Simulators User's Manual*. Transport Simulation Systems.

van Lint, J. W. and van Zuylen, H. J. (2005). Monitoring and predicting freeway travel time reliability: Using width and skew of day-to-day travel time distribution. *Transportation Research Record: Journal of the Transportation Research Board*, 1917(1):54–62.

Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*.

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.

Wakabayashi, H. and Iida, Y. (1991). Evaluation of reliability of road network for better performance, advanced management and future network design. In *Applications of Advanced Technologies in Transportation Engineering*, pages 121–125. ASCE.

Webster, F. V. (1958). Traffic signal settings. Technical Report 39, Road Research Laboratory.

Westgate, B. S. (2013). *Vehicle travel time distribution estimation and map-matching via Markov chain monte carlo Methods*. PhD thesis, Cornell University.

Whitt, W. (2012). Extending the FCLT version of L= λW. *Operations Research Letters*, 40(4):230–234.

Wolff, R. W. and Yao, Y.-C. (2013). Little's law when the average waiting time is infinite. *Queueing Systems*, pages 1–15.

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*, 36(4):291–312.

Xing, T. and Zhou, X. (2011). Finding the most reliable path with and without link travel time correlation: A Lagrangian substitution based approach. *Transportation Research Part B: Methodological*, 45(10):1660–1679.

Yin, Y. (2008). Robust optimal traffic signal timing. *Transportation Research Part B*, 42(10):911–924.

Yun, I. and Park, B. (2006). Application of stochastic optimization method for an urban corridor. In *Proceedings of the Winter Simulation Conference*, pages 1493–1499.

Zhang, L., Yin, Y., and Lou, Y. (2010). Robust signal timing for arterials under day-to-day demand variations. *Transportation Research Record*, 2192:156–166.

Zheng, F. and Van Zuylen, H. (2013). Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies*, 31:145–157.