DYNAMIC ASSIGNMENT, SURVEILLANCE AND CONTROL FOR TRAFFIC NETWORK WITH UNCERTAINTIES

ZHONG REN XIN

Ph. D

THE HONG KONG POLYTECHNIC UNIVERSITY

2011

The Hong Kong Polytechnic University

Department of Civil and Structural Engineering

Dynamic Assignment, Surveillance and Control for Traffic Network with Uncertainties

ZHONG, Renxin

A Thesis Submitted in Partial Fulfilment

of the Requirements for the Degree of

Doctor of Philosophy

October, 2010

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis entitled "Dynamic Assignment, Surveillance and Control for Traffic Network with Uncertainties" is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Signature:

Name of student: Renxin ZHONG

Dedicated with love and gratitude to my family.

Abstract

This dissertation involves the development of three key components of advanced traffic management information systems (ATMIS), i.e. real-time traffic surveillance, dynamic traffic assignment with traffic volume (queue) control, and traffic management under demand and supply uncertainties.

The traffic volume (or queue) control scheme is widely used in traffic control practice and has been proven to be effective in managing congestion or gridlock. However, dynamic traffic assignment (DTA) considering the effects of traffic volume control schemes has been missing from literature. To fill this gap, this dissertation considers the analytical traffic volume (queue) control for traffic networks under two route choice behavior assumptions, i.e. dynamic user equilibrium (DUE) and dynamic system optimum (DSO). The traffic volume controls are related to the desired temporal traffic volumes on certain links, which can be set according to safety or environmental requirements. Both the DUE and DSO traffic assignment with traffic volume control are analyzed utilizing the optimal control theory. The existence of equilibrium to the DUE with traffic volume control is proven in this thesis. The DSO analysis highlights the differences between the dynamic externalities of the two vertical queue models, i.e. the whole link model and the deterministic queuing model. The results obtained from the DSO analysis are applied to investigate the traffic induced air pollution pricing.

For the surveillance part, this thesis concentrates on the development of a macroscopic traffic flow model to capture traffic dynamics on networks influenced by demand and supply uncertainties that are suitable for real-time traffic monitoring and control applications. To fulfill these objectives, a stochastic macroscopic dynamic traffic model, the stochastic cell transmission model (SCTM), which is based on the modified cell transmission model (MCTM) and the switching mode model (SMM), is proposed. The SCTM inherits the advantages of the MCTM and the SMM. However, there are several key differences between them, e.g. the MCTM and the SMM admit deterministic demand and stationary flow-density fundamental diagram while the SCTM accepts the random inflows (uncertain demand) as well as random parameters of the fundamental flow-density diagram (uncertain supply functions) with known means and variances of the freeway segment as exogenous inputs. Under the SCTM framework, the uncertain wavefronts are captured by probabilities of occurrence of operational modes which describe different congestion levels. The SCTM is calibrated and validated by several empirical studies. We also compare the performance of the SCTM with Monte Carlo Simulation of the MCTM (MCS-MCTM). The

results confirm that the SCTM outperforms the MCS-MCTM. We apply the SCTM to estimate the queues and delays at signalized intersections and compare the results with some well-known delay and queue estimation formulas, e.g., Webster, Beckmann, McNeil, and Akcelik. The comparison results show a good consistency between the SCTM and these formulas. In addition, the SCTM describes the temporal behavior of the queue and delay distributions at signalized junctions with stochastic supply functions and (non-stationary) arrivals.

In the traffic management part, optimal and robust decision making problems for managing uncertain network traffic are investigated. The proposed SCTM is applied to describe traffic dynamics on networks influenced by demand and supply uncertainties. The traffic management problems are formulated as stochastic dynamic programming problems. A closed form of optimal control law is derived in terms of a set of coupled generalized recursive Riccati equations. The robust decision making problem, which aims to act robustly with respect to the supply uncertainty and to attenuate the effect of demand uncertainty, can be recognized as an equivalent optimal decision making problem. Another implication of the proposed methodology is to make benefit from the inherent uncertainties, which is achieved by extending the conventional LQ optimal control theory to consider the indefinite terms of the state and input weighting matrices. The multiagent system (MAS) approach to access the traffic management for a general traffic network is discussed. The applications of the proposed methods to incident management are also highlighted.

In conclusion, this thesis contributes to the literature on dynamic traffic assignment, stochastic dynamic traffic modeling and management, and to support further analysis and development in this area.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor: Dr. Agachai Sumalee for his kind guidance, supervision and encouragement. I have gained a lot of knowledge from him that are applicable for my future career. The countless hours he spent with me on this research are gratefully appreciated. I would like to thank my co-supervisor, Prof. William H.K. Lam for the useful discussion in the early stage of the research and his criticisms on my presentations. I would like to thank the Hong Kong Polytechnic University for providing my PhD studentship.

I would also like to thank Prof. Terry Lee Friesz and Prof. Kann Özbay for serving in my dissertation committee, and for their valuable suggestions and advices. Their valuable suggestions and criticisms have greatly enhanced the quality of this research and dissertation. In particular, I would like to thank Prof. Friesz for his very interesting and constructive comments on some of my works as well as his encouragement. I would like to thank him also for sharing his experience in academic research during the Third International Symposium on Dynamic Traffic Assignment held at Takayama, Japan.

I am greatly indebted to all the professors who have passed on their skills, knowledge, and wisdom to me during my study. Among them, I would like to express my special thanks to Prof. Jie Huang for his guidance on my thesis research for a M.Phil degree in System and Control Theory, Prof. Frank Lewis for introducing me to the area of approximate dynamic programming, and Prof. Zhi Yang for his guidance on my thesis research during my undergraduate years at Sun Yat-Sen University. Apart from introducing me to the exciting area of system and control theory, they have enhanced the spectrum of my research activities in theoretical as well as practical projects.

Many past and present members in transportation group of the Hong Kong Polytechnic University deserve my thanks for their friendship and technical expertise, including Miss Tianlu Pan, Prof. Zhichun Li, Dr. Julio Ho, Dr. Karen Tam, Mr. Paramet Luathep, Mr. Biyu Chen, Miss Zhaokun Li. Particularly, I would like to thank Miss Tianlu Pan, Dr. Julio Ho and Prof. Zhichun Li for the fruitful discussions over the past two years. I would also like to thank them for always being there to support me through the tough times as well as to share the great moments, for filling my last two years with so many

unforgettable memories.

I would like to express my acknowledgement to my seniors in the Applied Control and Computing Laboratory of the Chinese University of Hong Kong, Dr. Tianshi Chen (currently Linköping University), Dr. Dabo Xu (currently University of New South Wales), Dr. Hongwei Zhang (currently University of Texas), Dr. Minghui Zhu (currently University of California), Dr. Xiaolin Hu (currently Tsinghua University) who have answered so many questions regarding my research and life and have given me so many helpful suggestions. I am also grateful to Dr. Zhijia Tan (currently Huazhong University of Science and Technology) and Miss Xiaolei Wang at the Hong Kong University of Science and Technology (HKUST), Miss Xin Pei at the University of Hong Kong (HKU), for the fruitful discussions on many other topics beyond this thesis. Special thanks are also due to Miss Bingxia Sun (at the HKBU), Mr. Lin Xiao, Mr. Xiaosu Ma and Miss Teng Yang (at the HKUST) and Miss Shenjun Yao (at the HKU) for their friendship.

Finally, my love and thanks go to my parents. Mere words, in any of the languages that I know, are inadequate to express my gratitude for the unconditional love and support they have given to me.

Renxin ZHONG

January 2011

Contents

Abstract i

List of Figures ix

1 Introduction and objectives 1

1.1 Need of the study . . . 1

1.2 Objectives . . . 2

1.3 Thesis organization and contributions . . . 7

2 Background and literature review 10

2.1 Background and the Intelligent Transportation Systems (ITS) . . . 10

2.1.1 Advanced traveler information system (ATIS) . . . 12

2.1.2 Advanced traffic management system (ATMS) . . . 13

2.2 Dynamic traffic assignment (DTA) . . . 15

2.2.1 A brief overview of the DTA . . . 15

2.2.2 Requirements for dynamic traffic models . . . 19

2.3 Macroscopic traffic flow models . . . 22

2.3.1 The Lighthill-Whitham-Richards (LWR) model . . . 23

2.3.2 The cell transmission model (CTM) . . . 24

2.3.3 Vertical (point) queue models . . . 25

2.4 Real-time traffic surveillance and control . . . 26

2.4.1 Real-time traffic surveillance . . . 26

2.4.2 Real-time traffic control . . . 28

2.4.3 DTA and dynamic traffic control . . . 31

3 Dynamic user equilibrium with side constraints for a traffic network with traffic volume (queue) control 32

3.1 Introduction . . . 33

3.2 Static side-constrained traffic assignment problem . . . 35

3.3 Dynamic user equilibrium with side constraints . . . 36
3.3.1 Preliminaries and dynamic user equilibrium . . . 36
3.3.2 Existence of the DUE with simultaneous departure-time-and-path-choice . . . 40
3.3.3 Problem formulation of dynamic user equilibrium with side constraint . . . 42
3.3.4 Existence of the DUE-SC with simultaneous departure-time-and-path-choice . . . 45
3.3.5 Necessary condition of the DUE-SC with simultaneous departure-time-and-path-choice . . . 45
3.3.6 Interpretation of DUE with side constraints . . . 53
3.4 Solution algorithm . . . 54
3.4.1 The fixed point problem and its optimal control formulation . . . 54
3.4.2 Solution algorithm for the DUE-SC . . . 56
3.5 Numerical examples . . . 58
3.5.1 A simple network case . . . 58
3.5.2 The Braess' Network . . . 63
3.6 Conclusions . . . 70

4 Dynamic marginal cost, access control, and pollution charge: a comparison of bottleneck and whole link models 75
4.1 Introduction . . . 76
4.2 Problem formulation of the DSO-AC and its solution—the WLM case . . . 79
4.2.1 Problem formulation of the DSO-AC . . . 79
4.2.2 Property of the DSO-AC . . . 82
4.3 Problem formulation of the DSO-AC and its solution—the DQM case . . . 87
4.4 DSO with access constraints, dynamic externality, and dynamic road pricing 89
4.4.1 Consistency between the results on the sensitivity value of the total system travel cost . . . 89
4.4.2 Difference between the dynamic externalities of the whole link model and the deterministic queuing model . . . 90
4.4.3 Marginal cost pricing, access pricing, and the access constraint . . . 93
4.5 Traffic induced air pollution pricing as a special case of the access pricing . 96
4.5.1 The traffic capacity and the environmental traffic capacity . . . 96
4.5.2 A brief review of the dynamic traffic pollution dispersion models . . 97

4.5.3 Environmental traffic capacity constraint and pollution charge . . . 104
4.6 Solution algorithm for the DSO-AC with the WLM as network loading model . . . 107
4.6.1 Reformulation of the optimal control problem and functional approximation . . . 107
4.6.2 Solution algorithm for the DSO-AC . . . 108
4.7 Numerical example . . . 109
4.8 Conclusions . . . 113

5 Stochastic cell transmission model (SCTM): a stochastic dynamic traffic model for freeway corridor traffic state surveillance 116
5.1 Introduction and motivation . . . 117
5.2 The MCTM and the SMM . . . 121
5.3 The stochastic cell transmission model . . . 124
5.3.1 The overall framework of the SCTM . . . 124
5.3.2 Formulation of demand and supply uncertainties . . . 126
5.3.3 Dynamic process of the SCTM and probabilistic conditions . . . 128
5.3.4 The SCTM as a class of stochastic bilinear system . . . 130
5.3.5 Mean and auto-correlation of stochastic traffic densities . . . 135
5.3.6 The one wavefront assumption, an interconnected SCTM approach to model a freeway corridor, and its implementation . . . 137
5.4 Numerical example . . . 138
5.5 An empirical study . . . 143
5.5.1 Test site description and model parameters calibration . . . 143
5.5.2 Test results against the supply uncertainty . . . 146
5.5.3 Test results against both demand and supply uncertainties . . . 154
5.6 Conclusion . . . 157

6 Stochastic cell transmission model for traffic networks with demand and supply uncertainties 161
6.1 Introduction . . . 161
6.2 Descriptions of a traffic network and the basic SCTM . . . 164
6.3 Uninterrupted facilities . . . 167
6.3.1 Freeway corridor . . . 167
6.3.2 On-/off- ramps, traffic merge and diverge . . . 170
6.4 Model of signalized junctions . . . 174

6.5 Numerical examples . . . 179
6.5.1 An application of the network SCTM as a stochastic dynamic traffic network model . . . 179
6.5.2 Queues and delays at a signalized junction . . . 186
6.6 Conclusions . . . 196

7 Traffic management under demand and supply uncertainties 199
7.1 Motivation and introduction . . . 199
7.2 The SCTM as a Markov switching state space model . . . 204
7.2.1 A model reduction of the SCTM for control and filtering . . . 206
7.2.2 A refinement of the control variables . . . 207
7.3 A stochastic optimal control framework for the SCTM . . . 209
7.3.1 The problem formulation and basic assumptions . . . 209
7.3.2 Definitions of operators . . . 211
7.3.3 Derivation of an optimal strategy . . . 212
7.4 A robust consideration . . . 218
7.4.1 An introduction to robust control . . . 218
7.4.2 A robust control formulation . . . 219
7.5 Multiple agent settings . . . 221
7.6 Conclusion . . . 223

8 Summary of the thesis and future research topics 228
8.1 Summary of thesis . . . 228
8.2 Future works . . . 234
8.2.1 On the DTA aspect . . . 234
8.2.2 On the traffic surveillance and control aspects . . . 236
8.2.3 An approximate dynamic programming (ADP) approach to overcome the curse of dimensionality of dynamic programming . . . 244
8.2.4 Multiagent reinforcement learning to coordinate the performance of agents . . . 247

Bibliography 249

List of Figures

1.1 The relationship between ITS and the study 3
1.2 The interconnection of different components of the study 4

2.1 Congestion delay, vehicle-miles traveled (VMT), and lane-miles of freeway in Los Angeles, normalized to 1982 levels (Source: Chen (2003)). 11
2.2 Flow conservation 21
2.3 A typical traffic control loop. Revised from Papageorgiou et al. (2003). . . . 29

3.1 Fundamental diagram of an urban street 42
3.2 Fixed point algorithm for DUE-SC 56
3.3 Solution algorithm for DUE-SC 58
3.4 Network connected with parallel links 59
3.5 Inflow profiles, link traffic volumes and travel costs of both links under the DUE with symmetric penalty 60
3.6 Inflow profiles, link traffic volumes and travel costs of both links under constant traffic volume control 61
3.7 Inflow profiles, link traffic volumes and travel costs of both links under time-varying traffic volume control 62
3.8 Inflow profiles, link traffic volumes and travel costs of both links under "step" traffic volume control 62
3.9 The Braess' network 63
3.10 Path departure rates and travel costs of the Braess' network under the DUE condition 64
3.11 Traffic volumes of the five links under the DUE condition 65
3.12 Path departure rates and travel costs of the Braess' network under constant traffic volume control 66
3.13 Traffic volumes of the five links and the additional travel cost under constant traffic volume control 66

3.14 Path departure rates and travel costs of the Braess' network under time-varying traffic volume control 67
3.15 Traffic volumes of the five links under time-varying traffic volume control . 68
3.16 Link traffic volumes, time-varying side constraints and additional travel costs of the Braess' network under the DUE-SC condition 68
3.17 Change in the convergence error with iteration for the Braess' network . . . 69

4.1 Dynamic marginal cost for the deterministic queuing model 91
4.2 Comparison of the dynamic marginal costs for the two models 94
4.3 Schematic illustration of flow and dispersion conditions in street canyons (Berkowicz et al., 2008) 99
4.4 Three flow regimes associated with different building-height-to-street-width ratios h/b (Oke, 1988) 100
4.5 optimal control formulation of the DSO-AC in a compact form 108
4.6 Network connected with parallel links 111
4.7 Inflow profiles, link traffic volumes and dynamic marginal costs of both links under the DSO condition 111
4.8 The externalities of the two links 111
4.9 Inflow profiles, link traffic volumes, dynamic marginal costs, and additional costs of both links under the DSO-AC condition 112
4.10 The externalities of the two links 112

5.1 A fundamental flow-density diagram of traffic flow 119
5.2 A trapezoidal fundamental diagram for the modified cell transmission model 122
5.3 Five traffic operational modes for a freeway segment with p cells 123
5.4 The overall framework of the stochastic cell transmission model 125
5.5 The flow chart for implementation of the SCTM for a freeway segment . . . 139
5.6 Freeway segment consisting of 4 cells 139
5.7 The nominal fundamental diagrams 140
5.8 Traffic density generated by the SCTM for the test case 141
5.9 Traffic density generated by the Monte Carlo Simulation of MCTM 142
5.10 Propagation of SD of the traffic density 142
5.11 Map of the test site (Source: Google map) 144
5.12 A section of I210-W divided into 4 cells and its detector configuration . . . 144
5.13 The fundamental diagrams of the four cells calibrated from the traffic flow data collected on April 22, 2008 145

5.14 Measured densities and the MCTM's estimated densities for a segment of I-210W on April 22, 2008 . . . 147
5.15 Measured densities, simulated mean densities obtained by the MCS of M-CTM for a segment of I-210W on April 22, 2008 . . . 148
5.16 Measured densities, simulated mean densities, and the 68 percent confidence interval obtained by the MCS of MCTM for a segment of I-210W on April 22, 2008 . . . 148
5.17 Measured densities and the SCTM's estimated mean densities for a segment of I-210W on April 22, 2008 . . . 149
5.18 Measured densities, the SCTM's estimated mean densities and the 68 percent confidence interval for a segment of I-210W on April 22, 2008 . . . 150
5.19 Probability distributions of different modes in the "simplified" SCTM approach over time . . . 150
5.20 Measured densities and estimated mean densities by the interconnected SCTM for a segment of I-210W on April 22, 2008 . . . 152
5.21 Measured densities, estimated mean densities and the 68 percent confidence interval by the interconnected SCTM for a segment of I-210W on April 22, 2008 . . . 152
5.22 Probability distributions of different modes in the interconnected SCTM approach over time . . . 153
5.23 The measured "missing" flow q_m and its estimated value . . . 153
5.24 The fundamental diagrams of the four cells calibrated from the historical data over the selected days . . . 155
5.25 A demonstration of the demand uncertainty . . . 155
5.26 The estimated mean densities against the historical mean densities . . . 156
5.27 The estimated mean densities and the 68 percent confidence against the historical data over the selected days . . . 156

6.1 Five traffic operational modes for a freeway segment with 2 cells . . . 165
6.2 A block diagram of the basic SCTM subsystem . . . 165
6.3 An interconnected SCTM approach to model a freeway corridor: (a) a short segment as one SCTM subsystem, segment variables, and segment inputs; (b) a freeway corridor as interconnected SCTM subsystems. . . . 168
6.4 The interconnected SCTM subsystems approach as paired up two neighboring cells . . . 169

6.5 A link-node model of a freeway segment . . . 171
6.6 The model of a signalized cell . . . 174
6.7 A signalized merge . . . 174
6.8 A typical signalized junction . . . 175
6.9 A link-node representation of the junction in green and red phases for one direction . . . 175
6.10 A typical signalized junction represented by four SCTM subsystems . . . 177
6.11 Three phases of Subsystem 1 and two signalized movements . . . 177
6.12 The SCTM subsystems representation under permitted signal . . . 177
6.13 Specification of the test network . . . 180
6.14 Nominal fundamental diagram . . . 181
6.15 Traffic densities of links R and S, and the 68% confidence interval . . . 181
6.16 Probabilities of occurrence of the five modes of links R and S . . . 182
6.17 Traffic densities of links V and W, and the 68% confidence interval . . . 182
6.18 Probabilities of occurrence of the five modes of links V and W . . . 183
6.19 Traffic densities of links X and Y, and the 68% confidence interval . . . 184
6.20 Probabilities of occurrence of the five modes of links X and Y . . . 185
6.21 Traffic densities of links T and M, and the 68% confidence interval . . . 185
6.22 Probabilities of occurrence of the five modes of links T and M . . . 186
6.23 Delay and queue evaluation scenario . . . 188
6.24 Nominal fundamental diagram of the segment . . . 189
6.25 Comparison of delay estimations for different degrees of saturation . . . 191
6.26 Concept of the stochastic time-dependent delay model . . . 191
6.27 Comparison of the average queue lengths obtained by the SCTM with those of Akcelik's formula . . . 192
6.28 An illustration of choice of the "*sampling*" process . . . 193
6.29 The PMF and the corresponding CMF with respect to exit time index for entry time index k . . . 194
6.30 The PMF with respect to exit time index for entry time index l . . . 194
6.31 Estimated queue length and its 68% confidence interval . . . 195
6.32 Comparison of delay estimations for different degrees of saturation (non-stationary inflows . . . 196

7.1 Different on-ramp metering control structures for different congestion modes of the SMM. Source: Sun (2005) . . . 208

7.2 A unified on-ramp metering control structure for the SCTM 209
7.3 A network with a hierarchy of agents and their responsibilities 222

8.1 A typical fundamental diagram of a freeway lane (Source: Hegyi (2004)) . . 240
8.2 A typical fundamental diagram of a freeway lane with speed limit control . 241
8.3 The traffic diversion control in case of incident (Source: Liu et al. (2009)) . 242
8.4 Sequence of the incident detection and traffic management 243
8.5 An illustration of the five fundamental dimensions of a stochastic (dynamic) optimization and the basic idea of implementation of ADP to incident management 245

Chapter 1

Introduction and objectives

1.1 Need of the study

Traffic congestion and environmental issues associated with vehicle use have been recognized as serious problems for their negative effects on productivity, health and living conditions. Harmful side effects of such traffic congestion include reduction of safety, deterioration of air quality, wasteful fuel consumption, and reduction of productivity. For example, in USA, 1992, traffic congestion accounted for 100 billion loss in national productivity (Strategic Plan, U.S. Department of Transportation, 1992). The situation is worsening with continuously increasing traffic volume worldwide. According to the 2009 Urban Mobility Report (Schrank and Lomax, 2009), congestion caused urban Americans to travel 4.2 billion hours more and to purchase an extra 2.8 billion gallons of fuel for a congestion cost of 87.2 billion USD[1]–an increase of more than 60% over the previous decade, and an increase of more than 420% since 1982. Research has also indicated that vehicles are responsible for at least 50 percent of the air pollution in urban areas. Only about 20 % of the town residents enjoy good enough air quality according to the estimation of the World Health Organization (WHO) in terms of the measured levels of emissions. According to the white paper of European Commission (2001), transport is responsible for 28% of carbon dioxide emissions in Europe, of which road transport accounts for about 84%.

Traffic congestion can be classified as recurrent and non-recurrent. Recurrent congestion is caused by the peak hour traffic demand exceeding the available roadway capacity.

[1]It is also reported that there was a decrease of 40 million hours and a decrease of 40 million gallons, but an increase of over 100 million USD from 2006 due to an increase in the cost of fuel and truck delay. Small traffic volume declines brought on by increases in fuel prices over the last half of 2007 caused a small reduction in congestion from 2006 to 2007.

Non-recurrent congestion is largely produced by traffic incidents and managements. Incidents vary widely in severity, from vehicles stranding on the roadway shoulder with a flat tire to the closing of an entire highway section caused by vehicle crashes or hazardous materials. Studies have shown that 60 percent of the urban freeway delay may be caused by freeway incidents (Lindley, 1987). Research suggests that this may increase to approximately 70 % by 2005 (Özbay and Kachroo, 1999). It is reported that there is a symbiotic relationship between congestion and traffic incidents. Congested traffic condition is one of the main reasons for traffic accidents. Incidents on freeways interrupt traffic flows unexpectedly. They can be the major cause of "unusual" bottlenecks and secondary accidents. Those accidents cause more congestion, which in turn, causes more accidents (Özbay and Kachroo, 1999). For instance, the incident on 9^{th} May 2005 involved a heavy rainstorm and incidents on three roads[2] in Kowloon area in Hong Kong causing an extended and wide traffic congestion problem (Cheng et al., 2005). It has been suggested that the risk of secondary accidents can be significantly reduced by early detection and warning. For this reason, real-time freeway incident detection and characterization is an important function for freeway traffic management. Many of delays are caused by the capacity reduction due to the lane blockage during the incident response and clearance. In most scenarios, if proper traffic control and rerouting strategies can be implemented in time, travelers can go through the congested segments efficiently and/or circumvent the congested segments by traffic diversion control.

Hand in hand with the advances in computer science and technology, telecommunication, and control system engineering, Intelligent Transportation System (ITS)[3] has become an effective tool to alleviate traffic congestion, improve safety and efficiency, and reduce vehicle emissions for urban traffic networks. Therefore, this thesis concentrates on three aspects which aim to support ITS development. We depict the structure of an ITS and the relationship between an ITS and this dissertation in a block diagram as demonstrated in Figure 1.1. Detail objectives of this dissertation and their interconnection are discussed in the forthcoming section.

1.2 Objectives

Figure 1.2 depicts a block diagram which illustrates the overall framework and the interconnection of different components of this dissertation. Traffic networks are exposed to

[2]They were a fallen tree across Waterloo Road, loose scaffolding at Argyle Street, and fallen scaffolding at Prince Edward Road East.

[3]We will introduce ITS in detail in Chapter 2.

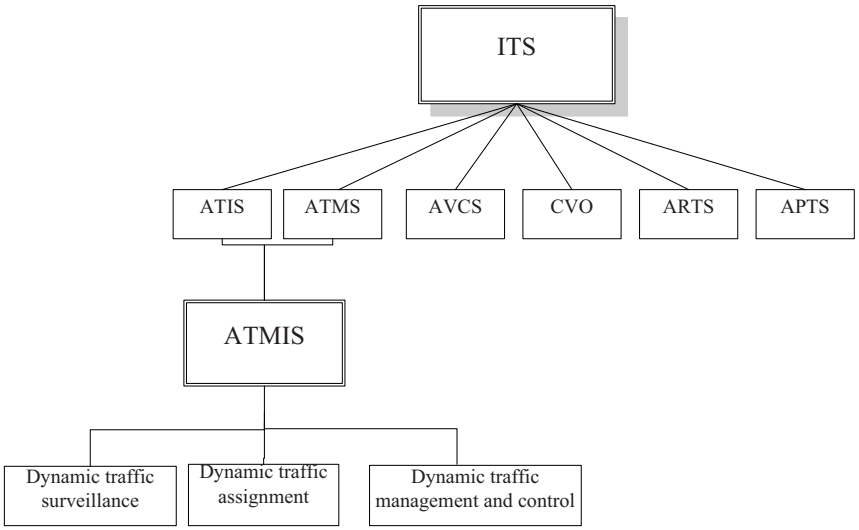


Figure 1.1: The relationship between ITS and the study

both demand and supply uncertainties by their very nature. A traffic network is usually equipped with measurement devices such as ultrasonic, microwave, and laser sensors, processed video, and tube-type vehicle counters. These sensors are installed at fixed locations and measure different combinations of speed, vehicle count, and occupancy. Travel time can be measured using probe vehicles equipped with GPS devices or automatic vehicle identification (AVI) data. All the detected data is transmitted to a central data collection point, e.g. the traffic management center (TMC), and is displayed by a human-machine interface. On the other hand, the traffic network is also equipped with control actuators, e.g. traffic signals. The data obtained from measurement devices is used to feed the surveillance tools so as to monitor the state of a traffic network and support traffic control strategies design. Therefore, the major problems are how to make use of the detected data and how to perform control design so as to better manage the road network and make efficient use of the existing capacity, which are also typical objectives of an ITS.

As suggested by Figure 1.1, this dissertation concentrates on three key components of an ITS, i.e. dynamic traffic assignment with traffic volume (or queue) control, real-time traffic surveillance and decision making for traffic management/control under demand and supply uncertainties. The above three functions are further grouped into two subsystems: the DTA subsystem and the dynamic traffic surveillance and control subsystem. The DTA component, as a short-term planning strategy and a traffic state prediction method, describes the idea performance of a traffic network under different user behavior assumptions, e.g. dynamic user equilibrium (DUE), dynamic user optimum (DUO), and dynamic system optimum (DSO)[4]. Due to many unsolved technical issues regarding the application

[4]These user behavior assumptions and the corresponding equilibriums will be reviewed in Chapter 2.

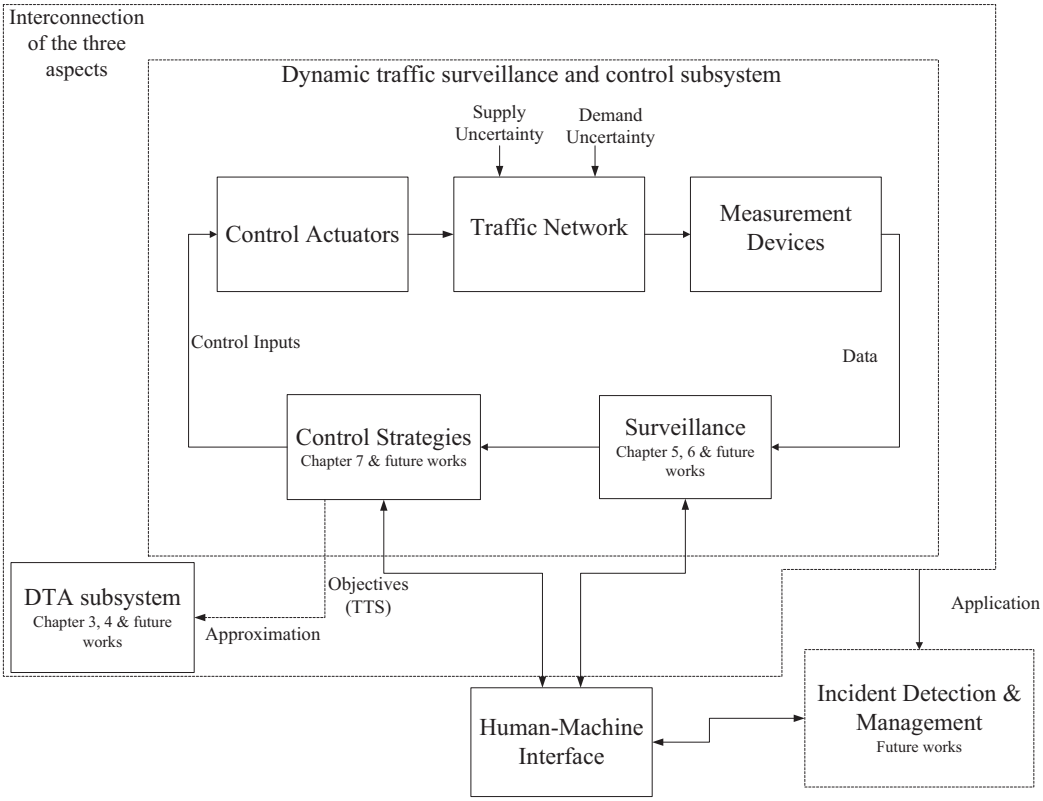


Figure 1.2: The interconnection of different components of the study

of DTA (especially regarding to the real time and large network applications) the dynamic traffic surveillance and control subsystem aims at approximating the equilibrium(s) obtained from DTA rather than achieving the exact DTA equilibrium(s). For instance, if the objective of traffic control strategy design is to minimize the total travel time spent (TTS) in the network, the traffic control strategy design aims to approximate a DSO equilibrium of the network. Within the dynamic traffic surveillance and control subsystem, the data detected by measurement devices feeds the surveillance tool to generate traffic states of the underlying network. Traffic information generated by the surveillance tool is then transmitted to support the design of control strategy and to feed necessary information to the controller. The controller generates different control inputs corresponding to different objectives prescribed. The control actuators adjust the traffic control strategies in terms of control inputs, e.g. the green time of a signal, which in turn affects the network states. The proposed methodologies can also be applied to detect incident and to support incident management, which will be highlighted in the future work.

Traffic volume (queue) control or access control, which restricts the traffic to access the congested (or controlled) locations, is essential for alleviating traffic congestion, especially traffic congestion under incident scenarios. The traffic volume (queue) control scheme is

widely used in traffic control practice and has been proven to be effective in preventing the traffic network from spillback effect of congestion or gridlock. Controlling the traffic volumes on critical infrastructures is also an easy and efficient approach to increase the safety of these facilities. The first part of this dissertation aims to investigate two important yet underdeveloped areas: the dynamic user equilibrium traffic assignment with traffic volume (queue) control and the dynamic system optimal traffic assignment with access control. It is important for us to look into the optimality state(s) of a traffic network under traffic volume (queue) control strategies. These results provide us some guidelines on queue control design, which in turn yields better implementations of the queue control strategies. The results also give us some guidelines to design the rerouting strategies, wherein the motorists can go through the congested segments (controlled segments or segments with incidents) efficiently and/or circumvent the congested segments by traffic diversion control. To be precise, we aim to address the following problems in this part:

1. Parallel to the static traffic assignment counterparts, we try to formulate the dynamic user equilibrium (DUE) and system optimal (DSO) traffic assignment problems with side constraints wherein the side constraints are regarded as link traffic controls.

2. We try to provide a rigorous mathematical analysis of dynamic traffic assignment with traffic volume control, e.g. necessary conditions, existence of equilibriums.

3. By restricting the link traffic volumes to be equal to or less than the link storage capacities, we try to capture the spillback effect of traffic congestion while avoiding the drawbacks of physical queue models, e.g. discontinuous and non-differentiable traffic dynamics and travel cost functions.

4. To satisfy some requirements of the DTA problems (e.g. first-in-first-out (FIFO) condition), the link travel time function is always assumed to be separable (i.e. depends on the link traffic volume only) in DTA literature. This assumption may not be reasonable if there is spillback. We try to make this separable link travel time function meaningful by controlling the traffic to access the saturated links such that no spillback would happen.

5. We try to develop a benchmark for evaluating various transport policy measures under different user behavior assumptions, e.g. optimal design of queue balancing for ramp metering and traffic signal control, as well as optimal design of the rerouting strategies.

6. To furnish the preceding objective, we attempt to perform the analysis of DSO with access control to provide a bound on the best performance of a traffic network.

7. Furthermore, we attempt to analyze the dynamic marginal cost, dynamic externality and the optimality condition under which the user surplus is maximized.

8. By the DSO analysis results, we attempt to internalize the external cost caused by traffic congestion and traffic induced air pollution by the access pricing framework. To be more specific, we will address the following subproblems:

   (a) Which pricing scheme should be imposed on each link, the access control based or the environmental traffic capacity based pricing scheme?

   (b) Which is the dominant pricing scheme at a specific time instant?

   (c) How to determine the boundaries under which a traveler on a link should pay either an access toll or an extra pollution charge?

The second part of this dissertation attempts to address the traffic state surveillance and control problems under stochastic environments. We emphasize the robustness in this part. To this end, this study first aims to develop a real time traffic surveillance tool which enables us to conduct robust real-time traffic state surveillance. When macroscopic traffic flow models are applied to a specific freeway segment or a traffic network, appropriate model parameters, such as free-flow speed, critical density etc., are needed. However, these parameters are usually not precisely known before hand. They may be different from site to site and even different within the same site for different time periods and adverse weather conditions, etc. Therefore, before a macroscopic traffic model, such as the Cell Transmission Model (CTM), can be applied to a specific site, a tedious model calibration procedure has to be conducted off-line based on available traffic measurement data to identify the corresponding values of the model parameters (regarded as supply functions). Errors and uncertainties (or variations/variances) cannot be avoided in such calibration and thus introduce uncertainties to the supply functions. For the demand side, day-to-day fluctuation in travel demand and travelers' responses to their information about the traffic network are considered as the leading causes of the demand uncertainty. The demand uncertainty is always regarded as recurrent uncertainty or disturbance to traffic flow models. It is hoped that the proposed traffic surveillance tool can act robust to these demand and supply uncertainties. Conventionally, variance is utilized to measure the risk under uncertainties. It is preferable that the proposed traffic surveillance tool can release the variance of traffic flow under demand and supply uncertainties to provide some

risk measure for traffic state surveillance. To be more specific, we attempt to address the following problems in this part:

1. We try to propose a stochastic macroscopic traffic flow model for traffic state modeling under demand and supply uncertainties by extending the well-known CTM.
2. We aim to calculate the mean and variance of the stochastic traffic flow analytically.
3. We attempt to capture the stochastic dynamic flow propagation.
4. We would like to capture the stochastic queues and delays at signalized junctions.
5. We intend to propose a potential way to avoid discontinuity and non-differentiability of the CTM, which can be viewed as making benefit from the uncertainties.

Despite the promising progress from DTA and integrated control models, development of efficient integrated optimal control strategies for both urban arterials and freeways remains challenging, especially regarding the following issues:

1. How to conduct optimal control design in presence of demand and supply uncertainties?
2. To pursue the robustness:
   (a) How to be aware of supply uncertainty?
   (b) How to attenuate the effect of disturbances which are regarded as demand uncertainty?
3. How to make benefit from uncertainties?

In the control (or traffic management) part of this dissertation, we aim to tackle these issues in the traffic management (or control) part.

1.3 Thesis organization and contributions

As previously explained, the presentation of the dissertation is roughly divided into two parts. After a brief introduction to the background materials and literature review, we present two dynamic traffic assignment models with traffic volume (or queue) control. Then we move to the second part which concentrates on stochastic traffic state modeling and control under demand and supply uncertainties. The reminder of this thesis is organized as follows.

Chapter 2 carries out a brief literature review on several related topics and provides a background information on ITS.

Chapter 3 formulates and analyzes the existence of equilibrium as well as the necessary condition to the dynamic user equilibrium with traffic volume (queue) control. Similar to the static counterpart, this problem is formulated as a dynamic user equilibrium with side constraints (DUE-SC) wherein the side constraints represent the restrictions on the link traffic volumes. The DUE-SC is formulated as an infinite-dimensional variational inequality (VI) problem. Based on this VI formulation, we discuss the existence of equilibrium to the DUE-SC. To analyze the necessary condition of the DUE-SC, the DUE-SC problem is further converted into an equivalent constrained optimal control problem for dynamic systems with state dependent time shifts and constraints on controls and states. We then derive the necessary condition by the Minimum Principle. An optimal control based numerical algorithm is developed to solve the DUE-SC problem.

Chapter 4 formulates the dynamic system optimal with access control (DSO-AC) for two vertical queue models: the whole link model and the deterministic queueing model. The access control constraints represent the restrictions on the traffic volumes and/or environmental constraints (e.g. vehicle emission). The dynamic externalities obtained from these two models are compared. Different structures of the dynamic externalities yield different tolling structures for these two vertical models to achieve DSO. The DSO-AC analysis also reveals the variety of economic effect of a certain amount of road capacity with respect to its spatial and temporal allocation, e.g. decide which links can be used and how to use their available capacities as "holding" capacities for queues. As another application of the dynamic system optimal with access control, an access pricing scheme (networks operate in a competitive market for subscribers, and yet have a monopoly position for providing access to these subscribers) is proposed to internalize the externalities caused by the traffic congestion under traffic volume control and the traffic induced pollution. A set of boundary conditions for the access pricing, i.e. whether we should charge the travelers based on the traffic volume control or environmental capacity restriction, is derived.

Chapter 5 proposes a stochastic traffic model for traffic state surveillance. The proposed dynamic traffic flow model, named as the stochastic cell transmission model (SCTM), which extends the CTM to consider stochastic supply functions as well as the stochastic travel demand for a freeway corridor. In the SCTM, the supply uncertainties are governed by the random parameters of the triangular fundamental flow-density diagrams, e.g., free-flow speed, jam-density, and backward wave speed, etc. The stochastic demand is also modeled as stochastic exogenous input to the SCTM. The model is calibrated and

validated by an empirical study. Numerical simulation and empirical study results are quite satisfactory and promising for us to conclude that SCTM is suitable for real-time traffic monitoring and control applications.

Chapter 6 extends the SCTM to capture the traffic dynamics on traffic networks consisting of freeways and arterials under demand and supply uncertainties, and to measure the queues at signalized junctions by assuming any temporal distribution of arrivals for a better implementation of traffic control strategies. By applying the proposed signalized SCTM and the proposed stochastic dynamic travel time estimation method to a signalized junction, the time average delays with respect to different levels of saturation are obtained for both stationary and non-stationary arrivals. The results are then compared with the delay estimations obtained from the traditional methods, e.g., Webster's, McNeill's and Akcelik's formulas. A comparison of the queue length is also carried out. The results show good consistency between the SCTM and these formulas. In addition, the SCTM describes the temporal behavior of the queue and delay distributions at signalized junctions with stochastic supply functions and (non-stationary) arrivals.

Chapter 7 derives optimal and robust decision laws for traffic management under demand and supply uncertainties. Stochastic dynamic programming is applied to design these traffic management policies based on the stochastic traffic flow obtained from the SCTM. For the control purpose, the SCTM is reformulated as a class of discrete time stochastic bilinear systems with Markov switching. Optimal traffic management policy for a freeway segment is derived based on this reformulation. A closed form of optimal control law is derived in terms of a set of coupled generalized recursive Riccati equations. The optimal control may be fragile with respect to the model miss-specifications. Furthermore, traffic manager would prefer a policy that would be robust for the uncertainties. We further pursue a robust (optimal) decision making law which is aimed to act robust with respect to the parameter miss-specifications in the traffic flow model, and to attenuate the effect of disturbance (which are regarded as travel demand). By extending the conventional LQ control theory, the proposed methodology can address the problem of making benefit from the inherent uncertainties, e.g. risk adjustment. Finally, we list some practical issues in traffic management that can be addressed by extending the current framework.

Chapter 8 gives a summary of this thesis. Some topics for the future research are also highlighted in this chapter.

This thesis was typeset using LaTeX.

Chapter 2

Background and literature review

2.1 Background and the Intelligent Transportation Systems (ITS)

Figure 2.1 shows the trend in delay, vehicle-miles traveled (VMT), and lane-miles of freeway in the Los Angeles region between 1982 and 2000 (Chen, 2003). As shown in the figure, the number of lane-miles increased about 30%, while vehicle-miles traveled increased by 70% with delay increased by 270%. Between 1982 and 2000, the average annual delay per peak road traveler grew from 16 hours to 62 hours, a 288% increase. During the same period, the total number of miles of freeway grew by only 35%. From the above statistics, it is clear that the explosive growth in traffic volume and travel demand cannot be handled solely by building and expanding highway facilities because of the prohibitively high costs, as well as social, political, and environmental issues resulting from urban and suburban infrastructure construction. The idea of using advanced technology to better manage the road network infrastructure and make efficient use of the existing capacity becomes more and more popular, which in turn motivates the development of Intelligent Transportation Systems (ITS). Recent advances in telecommunications, electronics, computing, networking, and control technologies have made it possible to build intelligent transportation systems.

It is believed that the ITS is an effective tool to alleviate traffic congestion, which may lead to more efficient travel demand and transportation network management, improve safety and efficiency, and hence reduce vehicle emissions. Current research and flied practices on ITS have been seeking to apply well established technologies in the areas of telecommunications, electronics, computing, networking, and control to vehicles, roadway networks and operational plans, to make it possible for vehicles and infrastructures to

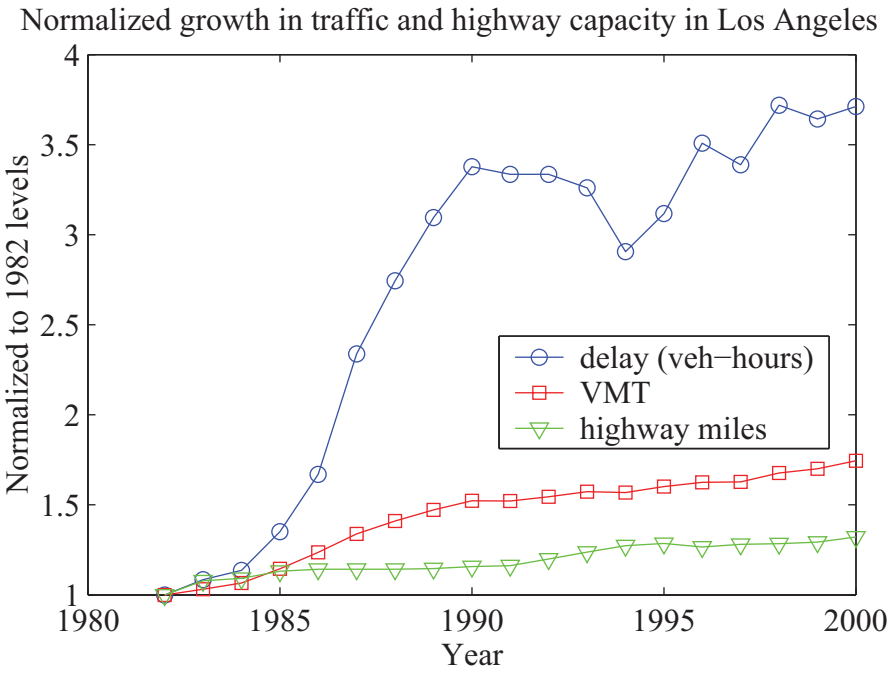


Figure 2.1: Congestion delay, vehicle-miles traveled (VMT), and lane-miles of freeway in Los Angeles, normalized to 1982 levels (Source: Chen (2003)).

exchange vast amounts of data back and forth and finally increase the utility of the entire transportation system. The development of ITS has brought many changes to traffic planning, control, and management in recent years. In particular, advanced sensing and surveillance technologies have made real-time traffic data available from various sources, the GPS equipped vehicles, smart phones, infrastructure based traffic sensors (e.g. the Freeway performance measurement system (PeMS) and Tools for Operational Planning (TOPL) in California, USA) and electronic toll collection tags (e.g. Autotoll in Hong Kong). These data sources provide rich information to better understand the congestion phenomena, and can be used to plan and manage transportation networks efficiently.

As illustrated in Figure 1.1, there are six major areas that the ITS focus on: Advanced Traveler Information System (ATIS), Advanced Traffic Management System (ATMS), Advanced Vehicle Control System (AVCS), Commercial Vehicle Operations (CVO), Advanced Public Transportation System (APTS), and Advanced Rural Transportation System (ARTS). As explained in the previous chapter, in this thesis, we will mainly focus on three key functions of ATIS and ATMS, which will be introduced in detail later, thus we only give a brief description of the other four areas as follows (Peeta, 1994):

> The AVCS is part of the "Smart Highway" initiative which aims at developing an automated vehicle guidance system to better utilization of highway space and safety.

The AVCS tries to maximize the usage of highway space by dramatically reducing reaction times of divers and headways. The AVCS could be more precise than human divers in controlling vehicle movements, e.g. lane changes, maintaining safe following distances between vehicles, and thus require less space and improve safety. The AVCS aims to comprehensively manage accident risk factors, e.g. lane keeping and collision avoidance by lateral and longitudinal control. Different kinds of actuators and sensors are utilized to manage human factors including driver fatigue.

The CVO intends to improve productivity, safety, and management of commercial vehicle operations. Research on the CVO is close related to logistics.

The APTS utilizes technologies such as advanced navigation (e.g. GPS and automatic vehicle location (AVL) system) computer and communication technologies to public transportation system operations to improve the efficiency and effectiveness of public transportation operations, vehicle maintenance, and administration, thereby attract travelers to transit, and ride-sharing modes. As a result, traffic congestion and air pollution caused by private vehicles can be reduced.

Finally, the ARTS involves the application of advanced technologies for travel in rural areas, which mainly focuses on freeway safety.

2.1.1 Advanced traveler information system (ATIS)

The ATIS plays an important role in the ITS. The basic objectives of the ATIS are to acquire, analyze, communicate, and present information to users to enhance personal mobility and hence the efficiency of travel, safety and the productivity of transportation, and reduce air pollution. The ATIS provides pre-trip and/or en-route travel information concerning traffic conditions such as traffic flow, travel time, and speed, and route guidance through various information media such as TV, radio, Internet, variable message signs (VMS), smart phones (IPhone, Nexus One, etc.), and in-vehicle (GPS) navigation systems, etc. The information is broadcasted in order to support travelers' decision making which in turn influences their travel choices and consequently reduce the (total/individual) travel time and improves efficiency of the traffic network. The development of an ATIS can be divided into three stages (Peeta, 1994; Zhou, 2002):

1. The information stage: In this stage, information is provided to the travelers to help their pre-trip planning and en route decision making.

2. The advisory stage: In the advisory stage, more real-time information, such as link travel times, traffic incident information, weather information, are collected and broadcasted by the ATIS to travelers. The in-vehicle information systems analyze the received information and promote advices to travelers.

3. The coordination stage: In this stage, the ATIS integrates the feedback information from in-vehicle information systems and other sources to predict traffic conditions. Under some circumstances, optimization of traffic flows over the entire network is possible. This enables some advanced functions such as coordinated routing, coordinated ramp metering, coordinated traffic signal control, and transit dispatching. In this stage, ATIS and ATMS will merge into an integrated system (ATMIS) to optimize the performance of traffic networks.

However, one should be aware that providing traffic information to travelers may not necessarily improve the traffic condition. Traffic information affects travelers' behavior, which in turn changes the traffic condition. A critical requirement and challenge for the successful deployment of ATIS strategies is to explicitly encounter drivers' behavior and create consistent predictions that are valid when users modify their behavior based on the information broadcasted to them. Information strategies developed based on the approaches that are behaviorally restrictive and limited in their ability to incorporate drivers' response behavior can result in misleading control strategies, and potentially deteriorate network performance (Paz, 2007; Paz and Peeta, 2009a,b).

2.1.2 Advanced traffic management system (ATMS)

In the meantime, another major component of the ITS, the ATMS also plays an important role in collecting data from a variety of sources, such as loop detectors, probe vehicles, video cameras, and other communication systems. The ATMS aims at managing and adjusting the traffic control systems in the network to respond to dynamic traffic conditions through real-time measurement, and communication to alleviate congestion and promote efficiency in utilizing the traffic network. The ATMS aids in providing real-time route guidance to travelers (drivers) through the ATIS. The ATMS is helpful in optimizing urban traffic signals, ramp-metering control, variable speed limit (VSL) control, etc. The ATMS is divided into three stages, i.e. the near term, middle term and longer term developments (Peeta, 1994; Zhou, 2002):

The near term development which aims at conducting basic research in five primary areas:

1. Traffic surveillance and monitoring: As the sources of traffic information to the ATMS, traffic surveillance techniques and (feedback) information from travelers (e.g. feedback information from ATIS) are needed to be developed and integrated within the ATMS.

2. Traffic control: Traffic control aims to regulate the traffic flows on urban transportation networks to improve their efficiency. We briefly review traffic surveillance and control in Section 2.4.

3. Traffic management: To develop models, algorithms and strategies to optimize the performance of traffic networks. Several traffic management strategies are discussed in this thesis, e.g. traffic volume (queue) control, dynamic road pricing, hybrid of ramp metering control and dynamic road pricing.

4. Human factors: As explained in the previous section, the divers' behavior is a critical issue for ITS design. From an ATMS viewpoint, traveler behavior, e.g. departure time choice, route choice, mode choice, en-route switching choice, risk taking attitude, and compliance with the guidance, are critical for developing traffic management techniques. There is a need to understand how travelers respond to the information provided by the ATMIS. Some advances on behavior-consistent information-based network traffic control strategies have been reported by Paz (2007); Paz and Peeta (2009a,b).

5. Integrated systems: Research on this area mainly concentrates on how to design and integrate different components of the system with respect to different objectives and functional requirements.

The middle term development: In this stage, the ATMS emphasizes the implementation of the technologies developed in the near term development. Meanwhile, off-line updates will be conducted to refine the performance of the technologies developed in the previous stage based on field experiments.

The longer term development: The longer term consists of integrated, interactive, and adaptive systems.

Since the ATIS and ATMS are interrelated, they are also known as advanced traffic management information systems (ATMIS) as depicted in Figure 1.1. In terms of the ATIS and ATMS, we can monitor and manage the transportation system more effectively and efficiently in a systematic manner. Dynamic traffic surveillance, assignment, and management (or control) of network traffic are three key components of the ATMIS. Both

the ATIS and ATMS rely on real-time and/or predictive traffic states, especially traffic density/volume/occupancy, travel time and speed. As a matter of fact, we try to develop a systematic tool that enables us to conduct traffic surveillance, to be more specific, the tool enables us to estimate/predict some basic traffic states, such as traffic density, under supply and demand uncertainties. Many traffic control strategies, e.g. urban traffic signal control and ramp metering control, could cause side effects if the queues induced by these control mechanicians were not properly addressed (Papageorgiou and Kotsialos, 2002; Papageorgiou et al., 2003; Varaiya, 2008). Therefore, queue control is a very important and has been proven to be efficient to prevent spillback effects of traffic congestion, either normal recurrent congestion or non-recurrent, e.g. incident induced, congestion. However, this important topic remains underdeveloped in DTA. In this study, we will formulate the dynamic user equilibrium (DUE) and dynamic system optimal (DSO) traffic assignment problems with queue control, which reveal the effect of queue control under the framework of DTA. As mentioned in the DTA research literature that the DUE and DSO are too idea to apply. Many technical issues regarding to the application of DTA results, such as existence, convergence, uniqueness, and stability of equilibrium(s) of DTA under general assumptions, estimation of real-time OD matrices, the driver behavior-consistent route guidance, etc., are yet not clear. Usually, only approximation of DUE or DSO can be achieved in practice by open-loop optimal control based approaches (see e.g. Kotsialos et al. (2002); Kosmatopoulos et al. (2006b); Papageorgiou et al. (2003) and the references therein), closed-loop feedback control based approaches (see Papageorgiou and Kotsialos (2002); Papageorgiou et al. (2003) for a review), and some other intelligent learning framework based approaches, (see e.g. Paz and Peeta (2008), Paz and Peeta (2009a), and Paz and Peeta (2009b)). In this study, we try to develop optimal and robust control based traffic control strategies which enable us to conduct traffic management under both supply and demand uncertainties. We provide a brief literature review on these three aspects of the ATMIS in this chapter.

2.2 Dynamic traffic assignment (DTA)

2.2.1 A brief overview of the DTA

As the last step of the well known four-step transportation planing procedure (trip generation, trip distribution, mode choice and traffic assignment), traffic assignment involves allocating a set of origin-destination (OD) demands onto a traffic network, usually under certain constraints that reflect some appropriate behavior of the travelers. If the traffic

dynamics is not taken into account, *i.e.* the traffic is assumed to be operating in its steady state, the corresponding traffic assignment problem reduces to a static traffic assignment problem which was originally developed for long term transportation planning.

Dynamic traffic assignment (DTA), as one of the key functions of the ATMIS architecture, has been recognized as a key component for network planning and transport policy evaluation as well as for real-time traffic operation and management for the last three decades. The readers can refer to two recent excellent review papers, i.e. Peeta and Ziliaskopoulos (2001) and Szeto and Lo (2006) for a comprehensive review of the history and development of DTA. To our purpose, a brief overview of the DTA is given in the section. For the sake of clarity, some general background of DTA, such as its relationship and implication with different traffic flow models, will also be briefly reviewed. With respect to the ITS, the DTA system provides the ability to model the time-varying dynamics of traffic flows in a network, and serves mainly the following two functions:

1. The first one is the descriptive capability which estimates the current network traffic states and predicts future network traffic states over time (mainly short-term traffic state prediction), in terms of time-varying network traffic flows/densities, travel times, and other performance characteristics on various components of the traffic network[1]. These estimated/predicted states are used in the on-line generation and real-time evaluation of a wide range of ATMS measures and ATIS messages.

2. The second one is the normative capability. This function aims to provide real-time route guidance information to travelers to achieve some system-wide objectives by taking into account the individual welfare of travelers and/or the system-wide social welfare. The most common way of seeking this capability is to search for path/route and/or departure time assignment. The normative capability involves a traffic controller or a traffic management center, with historical and real-time information on origin-destination (OD) matrices (demand information) and network supply conditions (such as calibrated flow-density fundamental diagram (or link performance function), incidents, weather conditions, *etc.*) for a given day (or a time interval for which the ATMIS is applied), that seeks to guide the travelers from their origins (or current positions) to their destinations so as to achieve certain

[1]DTA is not the only way for traffic state prediction. Another important approach for traffic prediction is mainly based on the use of statistical techniques, such as smoothing, auto-regressive moving average (ARMA), Kalman Filters, non-parametric regression, and neural networks (see e.g. Vlahogianni et al. (2004) for an overview). Compared with the approaches based on DTA, these approaches could offer some computational advantages. However, they do not make use of any behavioral rule.

system-wide objectives (e.g. minimize the total travel time), subject to individual routing constraints. Two different control architectures can be used for DTA: the decentralized and the centralized architectures.

Depending on how the path departure rates are calculated, DTA can be implemented in two ways: dynamic equilibrium assignment (wherein path flows are determined by the dynamic equilibrium conditions) and dynamic en-route assignment (wherein path flows are assigned according to a stochastic route choice model and the drivers can divert to other routes with less cost).

Three route choice behavior assumptions are often used in dynamic traffic assignment. The first states that each driver chooses a route/path at a specified departure time that minimizes the travel cost of all vehicles traveling on the network during a certain time period. The resulting traffic assignment problem is the so-called dynamic system optimal assignment (DSO). The second behavior rule states that each driver chooses a route and a departure time such that his/her own travel cost is minimized. If all the drivers behave this way, the transportation system may reach a dynamic equilibrium status under which no drivers could get better off by switching to another route and/or departure time. Traffic assignment based on this assumption is known as predictive dynamic user optimal (PDUO) assignment (also known as dynamic user equilibrium (DUE)). The third rule assumes that all drivers seek to minimize their own travel time by continuously updating their route choices according to the prevailing traffic conditions of the roadway network. This leads to the reactive dynamic user-optimal (RDUO) assignment.

In terms of their methodologies, there are two distinctive approaches to access DTA: one relies on analytical formulations and solutions algorithms, while the other is simulation based model. We concentrate the analytical formulations in this thesis, the readers can refer to Balakrishna (2006); Wen et al. (2006); Wen (2008) for overviews of the simulation based DTA models. There are mainly several approaches to the analytical dynamic traffic assignment problem[2]:

1. Mathematical programming (MP) approach: Dynamic traffic assignment problems were first formulated as mathematical programming problems by Merchant and Nemhauser (1978a,b) to study the DSO. Carey (1987) extended Merchant and Nemhauser's formulation to more general cases by utilizing convex nonlinear programming such that the problem can then be solved using standard NLP solvers. Drissi-Kaitouni and Hameda-Benchekroun (1992) first formulated the DUE problem

[2]See Boyce et al. (2001) for an overview.

as an NLP program over the static temporal (or time) expanded network (STEN). Smith (1993) proved that DUE has equivalent fixed-point, minimization, and VI formulations. He also proved that, under some assumptions, the path travel time is continuous with respect to path inflow, and hence showed the existence of DUE. Ziliaskopoulos (2000) proposed a linear programming model for DSO with single destination.

2. Optimal control approach: This approach was first proposed by Friesz et al. (1989) to establish the instantaneous DUE model with multiple origins but only single destination. This approach was followed by many researchers. However, at the early stage of the optimal control approach, the proposed models treated the exit flow as a function of the inflow, which introduces theoretical difficulty to the models unless the function is linear. Later, a new whole link model (also known as the link delay model) was introduced by Friesz et al. (2001). This model fully utilizes the consistency between link travel time and cumulative inflow-outflow curves. The outflow is defined as a function of the inflow with a time-shift that is equal to the link travel time. Based on this model, the DUE problem was reformulated. Chow (2007a, 2009a) applied this framework to study the DSO problem for a general traffic network.

3. Variational inequality (VI) approach: The approach was first introduced to investigate the DUE problem by Friesz et al. (1993). Later, link-based VI formulation was introduced by Ran and Boyce (1996a,b). To capture the effect of physical queue, Lo et al. (2001) investigated the DUE problem using the extended cell transmission model (CTM) as network loading model based on the VI formulation. Due to the powerful capability of VI to model various traffic dynamics and its rich literature in mathematical society, the VI approach has been widely employed and become a principal analytical method in the DTA literature, see, e.g. Friesz et al. (2001); Peeta and Ziliaskopoulos (2001); Szeto and Lo (2006).

4. Differential VI (DVI) approach: The DUE problem was claimed to be equivalent to a DVI problem in Friesz and Mookherjee (2006). Recently, Friesz (2010); Friesz et al. (2011) proved this equivalence.

5. Quasi-VI (QVI) approach: It is claimed that most of the discrete-time DUE models in terms of VI formulation proposed so far are actually QVI problems (see e.g. Ban et al. (2008) and the reference therein).

6. Game theory based approach: Wie (1993, 1995) and Wie and Tobin (1998) introduced differential game theory into investigate the DUE. The author was trying to establish the relationship between the Nash equilibrium and the dynamic user equilibrium. To be more specific, to classify the DUE as a particular Nash game, i.e. (differential) Nash non-cooperative game.

The analytical approaches are mathematically rigorous and attractive, however, they suffer from many limitations such as oversimplified assumptions to capture the "real" dynamics, etc. (Peeta and Ziliaskopoulos, 2001). For many practical applications, a model need to capture the stochastic characteristics of traffic dynamics, such as traffic flows, travel times, queues and spill-backs. Those capabilities are beyond existing analytical models, and thus simulation based model are more preferable for practical implementations. The simulation-based approach has the merit of closely approximating the travel behavior of individual drivers and easily incorporating traffic control measures. This approach has been used widely in formulating DTA problems, see e.g. Wen (2008).

As mentioned in Friesz et al. (2008), DTA models tend to be comprised of four essential sub-models:

1. a model of path delay,
2. flow dynamics,
3. flow propagation constraints,
4. a route/departure-time choice model.

With respect to the above prospects, we give an overview on some common requirements for a proper DTA problem.

2.2.2 Requirements for dynamic traffic models

Dynamic traffic flow models for DTA are used to describe the link traffic dynamics, travel time of the link, and flow propagation. Link travel time is the time consumption incurred by vehicles entering the link according to the time of entry and the cumulative traffic volume on the link at the same time. Dynamic traffic flow models describe time-varying traffic flow characteristics on the links of a specific network which can be viewed as a major difference from static traffic assignment, wherein the traffic network is assumed to be running in a steady state all the time. Compared with static traffic assignment, flow propagation as described by traffic flow models, is another important feature of DTA. For

the purposes of DTA such as route choice, we need to enforce some restrictions to describe flow propagation and travel time in an appropriate manner. These requirements are:

Theoretically: non-negativity, flow conservation, flow propagation, the first in/first out (FIFO) principle, causality, continuity.

Computational tractability is also an important requirement for application to large networks (Mun, 2007).

In this section, we provide a brief review on these properties as follows. For the sake of clarity, we first define the FIFO principle.

Definition 2.1. First-In-First-Out (FIFO) Principle[3]: Roughly speaking, the FIFO principle requires that vehicles entering a link first must also exit from the link first. Mathematically, this condition takes the form,

$$t' > t'' \Rightarrow t' + \tau(t') > t'' + \tau(t''), \tag{2.1}$$

where $\tau(t)$ is the link travel time for vehicles entering the link at time t. ∎

Readers can refer to Nie and Zhang (2005) for an overview on the role of FIFO principle in modeling network traffic dynamics. As explained in Mun (2007), in the real world the FIFO principle can not be always fulfilled and its violation is permitted through overtaking in microscopic models because vehicles can be given different characteristics even though they entered a link at the same time. Different from microscopic models, in macroscopic (or aggregate) models vehicles are considered to take the same travel time to traverse a link if they enter it at the same time since macroscopic (or aggregate) models describe the average (statistical) behavior of vehicles. The FIFO principle is therefore should be enforced in dynamic assignment models for the issue of equity (Mun, 2007). If the FIFO principle is enforced for a traffic flow model, we have the following proposition.

Proposition 2.1. (Chow, 2007b) If a traffic model satisfies the FIFO principle and the link travel time function $\tau(\cdot)$ is differentiable, then the following condition will be satisfied

$$1 + \frac{d\tau(t)}{dt} \geq 0, \tag{2.2}$$

for all times of entry t to the link. ∎

[3]As static assignment models do not consider traffic dynamics but are focused on network flows in steady state, the time consumption to traverse a link is assumed to be the same for all the vehicles that are assigned to the link irrespective of their origin, destination, and entry time. Therefore, the FIFO principle is automatically fulfilled in static assignment framework.

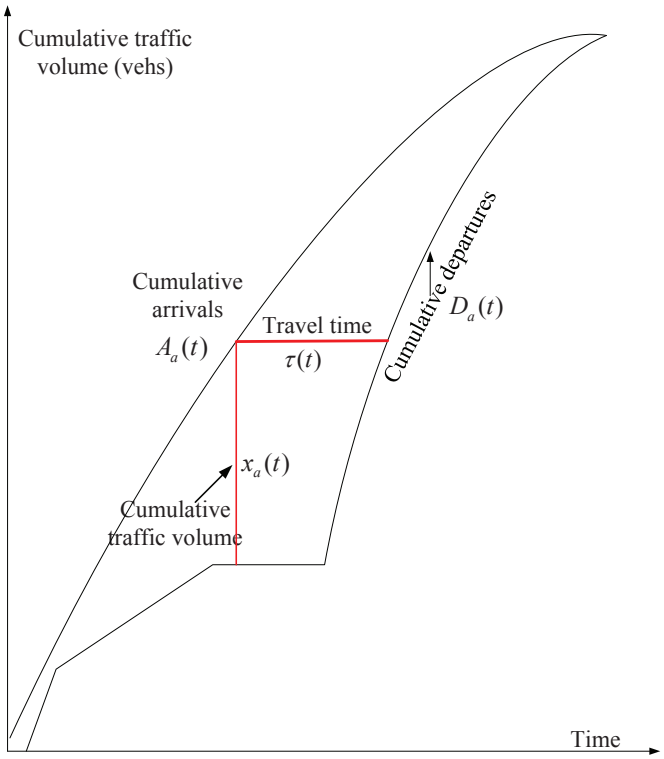


Figure 2.2: Flow conservation

Non-negativity: By definition, the inflow rate $h_a(t)$, the cumulative link traffic volume $x_a(t)$ and link outflow rate $g_a(t)$, are required to be non-negative, i.e.

$$h_a(t) \geq 0,\ x_a(t) \geq 0 \text{ and } g_a(t) \geq 0,\ \forall t. \tag{2.3}$$

It can be shown that, if FIFO principle is enforced, the non-negativity of $x_a(t)$ and $g_a(t)$ can be guaranteed by $h_a(t) \geq 0$ automatically.

Flow Conservation: The conservation of traffic flow is one of the most important requirements that dynamic traffic flow models should possess. The conservation is enforced to prevent the situations that travelers enter the network vanish before reaching the destination during the planning horizon, or the total outflow exceeds the total inflow to a link at any time instance. As depicted in Figure 2.2, the flow conservation for a link a can be expressed as follows:

$$A_a(t) = D_a(t) + x_a(t), \tag{2.4}$$

where $A_a(t)$ denotes the cumulative arrivals up to time t and $D_a[t]$ is the cumulative departures up to time t, respectively. This equation states that cumulative traffic volume on link a at time t is equal to the difference between cumulative arrivals and cumulative departures up to time t from the initial time t_0 with initial condition $x_a(t_0) = 0$. For the

case that link a is not empty at the initial moment, we can revise (2.4) as

$$A_a(t) + x_a(t_0) = D_a(t) + x_a(t). \tag{2.5}$$

Flow propagation: In DTA, flow should propagate through a link in a consistent manner with the speed of vehicles under FIFO principle. The minimum time for a vehicle to traverse a link should not be shorter than its free-flow travel time. The flow propagation can be expressed according to the flow conservation equation.

$$A_a(t) = D_a[t + \tau(t)], \tag{2.6}$$

Differentiating this with respect to entry time t gives:

$$h_a(t) = g_a\left(t + \tau(t)\right)\left(1 + \dot{\tau}(t)\right). \tag{2.7}$$

(2.7) is referred to as a time-flow consistency equation because it ensures that inflow, outflow and travel time are consistent with each other under FIFO principle.

As we can see from the flow propagation equation (2.7) that if the FIFO condition is enforced, i.e. $1 + \dot{\tau}(t) \geq 0$, non-negative inflow rate yields non-negative outflow rate. The consequence of FIFO violation can be seen in equation (2.7) where the outflow rate will be negative whenever $1 + \dot{\tau}(t)$ is negative. Theoretically, the FIFO discipline is a necessary and sufficient condition to ensure non-negativity of traffic and consistency between traffic flows and corresponding travel times (Daganzo, 1995b; Astarita, 1996; Carey, 2004a).

Causality: The travel behavior of vehicles is affected by the vehicles already on the link at the time of entry, but not by any future entering vehicles. Carey (2004b) referred to this as "strict causality" and also introduces the term "partial causality" to describe travel times affected by vehicles ahead as well as behind.

Computational efficiency should also be considered in addition to the aforementioned requirements. This is because substantial computational effort is inherently required to achieve dynamic equilibrium, especially when the size of the networks and the number of origin-destination pairs are large.

2.3 Macroscopic traffic flow models

The macroscopic dynamic traffic flow models ignore the behavior of the individual driver and attempt to replicate the aggregate response of a large number of vehicles. These models represent traffic as a compressible fluid, in terms of traffic flow, density, and speed. Almost all the analytical DTA approaches, nearly all model-based on-ramp metering control designs, and practical traffic engineering have applied macroscopic traffic flow models.

As dynamic traffic flow models are essential for dynamic traffic surveillance, assignment, and control, we review some fundamental quantities and the evolution of macroscopic traffic flow models related to this thesis. A comparative study of some macroscopic link models used in DTA can be found in Nie and Zhang (2005).

According to the HCM (2000), the following quantities are defined.

Speed $v(y,t)$ is defined as a rate of motion expressed as distance per unit of time, where y,t represents position (measured in the direction of traffic flow) and time, respectively. Depending on how it is measured, $v(y,t)$ is referred to as either space mean speed or time mean speed (HCM, 2000). The other speed concept used in dynamic traffic flow models is the so-called free-flow speed, which is defined as the **average** speed of traffic measured under light conditions so that vehicles can move freely at their desired speed.

Flow $f(y,t)$ is defined as the total number of vehicles that pass by the measure point y during a given time interval including t, divided by the length of the time interval.

Density $\rho(y,t)$ is the number of vehicles occupying a (unit) length of roadway around the measure point y at time t. The term can be obtained by the flow-density relationship, i.e.

$$\rho(y,t) = \frac{f(y,t)}{v(y,t)}. \tag{2.8}$$

The flow-density relationship of a freeway segment is often referred to as the fundamental diagram of the segment. Depending on the speed-density relationship, the fundamental diagram can have different shapes. The Greenshields' quadratic fundamental diagram and the Newell's triangular fundamental diagram are two most common used fundamental diagrams.

2.3.1 The Lighthill-Whitham-Richards (LWR) model

Among the macroscopic traffic flow models, Lighthill-Whitham-Richards (LWR) model would be the most popular and most-cited one. In terms of fluid dynamics, the traffic dynamics modeled by the LWR model is governed by the following two equations.

$$\begin{aligned} \frac{\partial \rho(y,t)}{\partial t} + \frac{\partial f(y,t)}{\partial y} &= 0, \\ f(y,t) &= F\left(\rho(y,t)\right). \end{aligned} \tag{2.9}$$

The first equation of (2.9) is the principle of conservation of vehicles, which is followed from fluid mechanics. The second equation of (2.9) is a flow-density relationship which is

also known as the "fundamental diagram". As a "fluid-dynamic" traffic flow model, since the LWR model does not contain a second-order derivative (such as a diffusion term), it is classified into the category of first-order model. By substituting the second equation of (2.9) into the first equation we have that

$$\frac{\partial \rho(y,t)}{\partial t} + \frac{\partial F\left(\rho(y,t)\right)}{\partial \rho(y,t)} \frac{\partial \rho(y,t)}{\partial y} = 0. \tag{2.10}$$

Detailed discussions on the LWR model can be found in Haberman (1977); Gomes (2004); Schönhof and Helbing (2007). The LWR model is capable of reproducing many important phenomena of freeway traffic. For instance, it captures the main difference between free-flow and congested traffic, which is that they propagate small disturbances in opposite directions and at different speeds. The LWR model also explains the formation and dissipation of queues upstream of a bottleneck, the dynamics of deceleration shock waves, and the absence of naturally forming acceleration shock waves (Gomes, 2004). However, in the meantime, the model is criticized for predicting some unrealistic traffic behavior. Criticisms of the LWR model can be summarized as (Gomes, 2004; Schönhof and Helbing, 2007):

1. It would not be able to describe unstable flow;
2. It would not describe spontaneous breakdowns of traffic flow;
3. It cannot incorporate any abrupt losses in capacity due to congestion (or capacity drop, two-capacity phenomenon);
4. The hypothesis of a steady (or static) flow-density fundamental diagram fails in the congested regime;
5. Field measurements suggest that flow-density fundamental diagram seem to be different with respect to time, depending on whether the traffic stream is decelerating or accelerating
6. The model does not consider the distribution of driver behaviors and desired speeds.

2.3.2 The cell transmission model (CTM)

The cell transmission model (CTM) was proposed by Daganzo (1994, 1995a). The CTM discretize the LWR model in both time and space. The CTM defines the flow propagation based on the intuitive concepts of sending and receiving flows. The model is shown to be computationally efficient and easy to analyze yet capture many important traffic

phenomena, such as queue build-up and dissipation, backward propagation of congestion waves, *etc.* A freeway segment is divided into homogeneous, consecutively numbered cells of length l_i, where i is a cell index. Time is discretized into uniform intervals of duration Δt, such that $v_{f,i}\Delta t \leq \min_i l_i$ in order to enforce a numerical stability condition and flow conservation, where $v_{f,i}$ is the free-flow speed of cell i. The number of vehicles on cell i at time $k\Delta t$ is taken as the state variable for the cell. The dynamics is evaluated according to the following flow conservation equation in conjunction with the fundamental diagram.

$$\begin{aligned} n_i(k+1) &= n_i(k) + q_{i,in}(k) - q_{i,out}(k),\ q_{i,out}(k) = q_{i+1,in}(k), \\ q_{i,in}(k) &= \min\left\{n_{i-1}(k), Q_i, \frac{w_{c,i}}{v_{f,i}}\left(N_i - n_i(k)\right)\right\}, \end{aligned} \tag{2.11}$$

where $v_{f,i}$ and $w_{c,i}$ are the slopes of the free-flow and congested portions of the triangular fundamental diagram of cell i (the free-flow speed and congestion backward wave speed). N_i is the maximum number of vehicles that can be accommodated by cell i (related to jam density). Q_i is the maximum number of vehicles that can move from cell $i-1$ to cell i during one time interval (related to capacity). $q_{i,in}(k)$ is the number of vehicles that actually move from cell $i-1$ to i during time step k. The amount of $q_{i,in}(k)$ can be computed with (2.11) or obtained by defining the sending and receiving functions: $S_{i-1}(k) = \min\{n_{i-1}(k), Q_i\}$ the maximum flow supplied by cell $i-1$ under the free-flow condition, over the interval $[k,\ k{+}1)$, and $R_i(k) = \min\left\{Q_i, \frac{w_{c,i}}{v_{f,i}}\left(N_i - n_i(k)\right)\right\}$ the maximum flow received by cell i under the congested condition over the same time interval. The definitions of sending and receiving functions are useful when the model is extended to tackle general network topologies (Daganzo, 1995a). The CTM was further extended to track the path flows for the purpose of dynamic traffic assignment by Lo (1999a); Lo et al. (2001). As a discrete version of the LWR model, the CTM suffers from most of the drawbacks of the LWR model.

2.3.3 Vertical (point) queue models

In this section, we review two vertical queue models[4], i.e. the deterministic queueing model and the whole link (or link delay) model.

2.3.3.1 The deterministic queueing model (DQM)

The deterministic queuing model, which is also known as the bottleneck model, was first proposed by Vickrey (1969). The DQM is described by the following hybrid system who

[4]Vertical queue model is also referred to as point queue model.

has a non-differentiable point.

$$\frac{dx_a(t)}{dt} = \begin{cases} h_a(t-\tau) - Q_a, & x_a(t) > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{2.12}$$

In the DQM, the travel time is flow-invariant, i.e. τ, if it is in free-flow condition. Under congested condition, a deterministic queue forms at its downstream end and being discharged with a maximum service rate Q_a, i.e. the link outflow is equal to the capacity. Under this circumstances, all travelers arriving before the queue dissipates will incur travel delay. To simplify the analysis, the free-flow time is commonly ignored in the DTA literature, i.e.

$$\frac{dx_a(t)}{dt} = \begin{cases} h_a(t) - Q_a, & x_a(t) > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

The DQM may be the simplest dynamic traffic flow model satisfying all the requirements for DTA models. It has been adopted by many authors in analyzing dynamic network traffic and various control policies. However, the DQM has also been criticized for over-simplifying real traffic behavior (Nie and Zhang, 2005; Mun, 2007). For example, the DQM does not give any delay until the link has been over-saturated. In addition, the non-differentiability in the state equation also causes analytical and computational difficulties.

2.3.3.2 The whole link (or link delay) model (WLM)

The whole link model is another vertical queue model, which aims to avoid the non-differentiability of the DQM. As we will discuss the WLM in detail in the chapters related to DTA, we refer the readers to the comparative studies of these two vertical queue models by Nie and Zhang (2005); Mun (2007).

2.4 Real-time traffic surveillance and control

Real-time traffic control is always in conjunction with real-time traffic surveillance as real-time traffic states are required in order to implement traffic control strategies. The importance of real-time traffic state surveillance within a traffic control loop had been recognized already in the 1970s (Papageorgiou, 1983).

2.4.1 Real-time traffic surveillance

Data of the ITS is usually generated by traffic surveillance. In this sense, traffic surveillance forms the basis for the formation of information for an ITS. Comprehensive reviews

on traffic surveillance and freeway performance measures can be found in Wang and Papageorgiou (2006b); Chen (2003); Kurzhanskiy (2007). In this section, we only provide some aspects of the traffic surveillance related to this thesis.

Freeway networks are usually equipped with a number of measurement devices (such as inductive loops, video cameras, radar detectors, *etc.*) which detect and deliver the real-time traffic information (Wang and Papageorgiou, 2006b; Chen, 2003; Kurzhanskiy, 2007). As it is mentioned in Wang and Papageorgiou (2005), due to significant space inhomogeneities of traffic flow the spatial resolution of the corresponding real-time traffic measurements may not be sufficient for direct use or further exploitation of the measurements. On the other hand, loop detector data sets are often incomplete or contain bad samples. For instance, Chen (2003); Muñoz et al. (2003) pointed out that approximately 30 percent of the possible loop samples in California's District 7, which contains over 30 freeways, were missing, on average, over the period from March 2002 to February 2003. However, traffic control strategies, especially the optimal control based strategies, e.g. the advanced motorway optimal control (AMOC) strategy , require accurate traffic state information in order to effectively regulate the traffic flows on the freeways. Therefore, besides the measurement devices installed, appropriate traffic flow models and/or traffic state estimators are required to produce good traffic state information to support traffic control strategies.

Traffic state estimation for a freeway network refers to estimating traffic conditions (such as volume, density, levels of service, wave-fronts, etc.) of the network at the current time instant based on the available real-time traffic measurements, to be more precise, based on a limited amount of available measurement data from traffic detectors. Usually, the number of traffic variables to be estimated is much larger than the number of traffic variables which can be directly measured. For example, we cannot install as many detectors as we want due to many reasons such as cost, maintenance, privacy issue, and geometric constraints. Even though we can install as many detectors as we want, we still suffer from the incomplete and bad traffic data. These constraints emerge the essential contribution of the freeway traffic state estimation (Chen, 2003; Wang and Papageorgiou, 2005, 2006b).

Besides the traffic state estimation, other advanced real-time traffic surveillance tasks including short-term traffic state prediction, travel time estimation/prediction, queue estimation/prediction are of interest to ITS, which are useful to support the functions of ITS. The short-term traffic state prediction attempts to forecast the traffic condition for a freeway network over a future time horizon given the short-term historical and real-time

traffic information. Most of the existing methods on short-term traffic state forecasting focus on incorporating temporal and spatial traffic characteristics into the forecasting process, e.g. using k-nearest neighbor algorithm (KNN) (Tam and Lam, 2008). A review on the short-term traffic state prediction can be found in Vlahogianni et al. (2004). The travel time estimation (prediction) refers to estimating (forecasting) the travel time along any specified route inside a traffic network. Travel time estimation and prediction have attracted much attention of researchers and traffic engineers, due to its fundamental significance for DTA, traffic control, reliability analysis of traffic network, and ITS. Reviews on the travel time estimation and prediction methods can be found in van Lint (2004); Tu (2008). The queue length (sometimes we would further need information on its tail and head) estimation/prediction aims at estimating (forecasting) the length and locations of any queue tail and head along any specified route. The queue information is essential for traffic signal control design, e.g. the traffic (or queue) response signal control, the store-and-forward control. The information on traffic queue is also important for incident management, e.g. traffic control centers wish to issue queue length information or queue tail warnings to avoid secondary accidents and to design incident response strategies. Queue length estimation methods are briefly reviewed in Liu et al. (2009).

2.4.2 Real-time traffic control

Some excellent review articles on real-time traffic control are available in literature, e.g. Papageorgiou and Kotsialos (2002); Papageorgiou et al. (2003). To this end, we just give some basic idea about real-time traffic control in this section.

In this thesis, we concentrate on a typical traffic control loop in conjunction with a traffic surveillance tool as depicted in Figure 2.3. Demand and supply uncertainties are considered as external quantities which feed the traffic network. Conventionally, the supply uncertainty is taken as noise and/or parameter uncertainty of a traffic flow model/estimator, while the demand uncertainty is always regarded as disturbance. The control inputs are directly related to corresponding control actuators, such as traffic signals, variable message signs, ramp metering etc. The control inputs may be selected from an admissible control region subject to technical, physical, and operational constraints, e.g. ramp metering rate, cycle time length, etc. The kernel of the control loop is the control strategy, whose task is to produce the control inputs, based on available measurements/estimations/predictions from the traffic surveillance tool. The control strategies generate the control inputs to render the closed-loop system to achieve the pre-specified objectives despite the influence of various disturbances and uncertainties. We have a

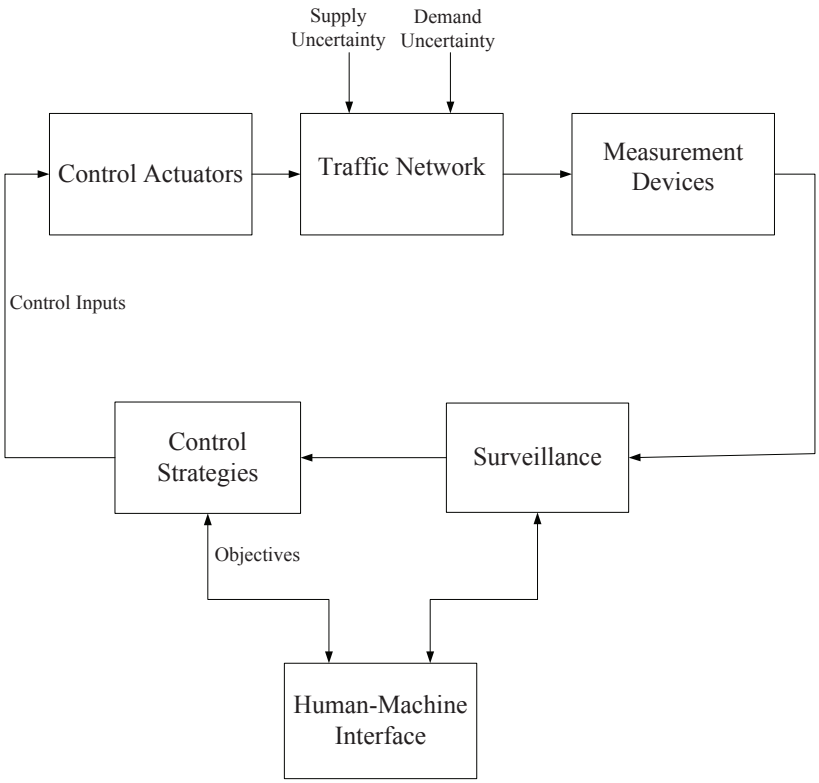


Figure 2.3: A typical traffic control loop. Revised from Papageorgiou et al. (2003).

human-machine interface as the control objective is to be specified by a human operator (or the traffic control center) and the control output is monitored by human. The relevance and efficiency of the control strategy largely determines the efficiency of the overall control system. Therefore, whenever possible, control strategies should be designed with care, via application of powerful and systematic methods of optimization and automatic control, rather than via questionable heuristics (Papageorgiou et al., 2003). Without considering the uncertainties and disturbances, a considerable amount of literature has been published on both the open-loop and closed-loop based control strategies for traffic network operation. Traffic control strategies can be classified into two categories: control strategies for urban roads, and control strategies for freeways.

Within an urban traffic network, traffic signals at the intersections are major control measurements. Control strategies for urban road traffic may be classified as Papageorgiou et al. (2003):

> The earliest traffic signal control may be the fixed-time strategies. A fixed-time scheme may not be "fixed" for a whole day but for a given time of day, e.g. the morning peak. The fixed-time strategies are derived off-line by using appropriate optimization methods based on historical demands and turning rates for each stream of the intersection (which are usually taken as constants). Some well known softwares,

such as MAXBAND, TRANSYT, etc., are based on the fixed-time traffic signal control strategies. By their very nature, fixed-time traffic signal control strategies are applicable to under-saturated traffic conditions only.

When more and more real-time measurements on the traffic states are available in accompany with the development of ITS, traffic-responsive strategies were proposed. These strategies make use of these real-time measurements to calculate in real time the suitable signal settings. Traffic response strategies are also extended to freeway traffic control by researchers due to their feedback structure (Papamichail and Papageorgiou, 2008). Due to the fact that the queues at the signalized junctions are stochastic and very difficult to estimate/predict, most of the existing signal control strategies are only applicable to under-saturated[5] traffic conditions. Very few strategies are suitable also for over-saturated conditions with partially increasing queues that reach the upstream intersections[6] (Papageorgiou et al., 2003).

Isolated traffic signal strategies aim to handle single intersections while coordinated traffic signal strategies aim to tackle an urban zone or even a whole network comprising many intersections.

Another category of control strategies is the freeway traffic control. The control measures that are typically employed in freeway networks are Papageorgiou et al. (2003):

Ramp metering, is implemented by installing traffic signals at on-ramps or freeway interchanges. The ramp metering rate is designed to regulate the traffic flows to access the freeways Papageorgiou and Kotsialos (2002); Gomes and Horowitz (2006); Gomes et al. (2008).

Link control, mainly includes lane control, variable speed limits, congestion warning, tidal flow, keep-lane instructions, etc.

Driver information and guidance systems, which are functions of ITS. The information and guidance can be broadcasted either by use of roadside variable message signs or via ATIS.

As traffic networks are exposed to supply and demand uncertainties and disturbances, we have to take them into account. Yet, when take into account these uncertainties and disturbances, the existing traffic control strategies either perform not good or even become

[5] Queues are only created during the red phases and are dissolved during the green phases.

[6] One difficulty is we are unable to estimate the queues under these circumstances.

unstable as the network is influenced by the demand and supply uncertainties. We will discuss this issue in detail in Chapter 7, thus we omit them here for brevity.

2.4.3 DTA and dynamic traffic control

DTA and dynamic traffic control are two key functions of the ATMIS. They ought to be interrelated. Research on these two topics seem to be parallel developed, although dynamic traffic control aims to approximate the DTA. However, no systematic method has been proposed in literature. Also there is no literature tells how close the designed control law can converge to the DTA. Some early works, e.g. Wang et al. (2003), have been done to combine DTA with dynamic traffic control by using the real-time prediction of travel time. Model predictive control based approach has also been proposed to integrate the dynamic route guidance and dynamic traffic control, see e.g. Hegyi (2004), Karimi et al. (2004). Some disadvantages of these strategies are: too complex, too computational expensive, precise model and precise disturbance prediction are necessary required, lack of stability proof, and optimality is not guaranteed. Recently, in Paz (2007), Paz and Peeta (2008), Paz and Peeta (2009a), and Paz and Peeta (2009b), the authors try to use intelligent learning mechanism, the fuzzy based learning approach to be more specified, to approximate the DTA by giving the travelers real-time route guidance. The proposed method is designing a fuzzy based system to approximate the equilibrium(s) of DTA. The guidance to a specific traveler is somehow the value of the corresponding defuzzification function.

To tell how accurate the dynamic traffic control is approximating the equilibrium(s) of DTA is not easy. We first need to address the uniqueness of the equilibrium of DTA. If a DTA model, say DUE, admits multiple equilibriums, we need to first identify the stability and its region of attraction for each equilibrium of the DUE, then answer the accuracy we can achieve. All these jobs are not easy, and research works on the stability and uniqueness analysis of the equilibriums of DTA by different researchers tell very different stories. For example, the works by Mounce and Smith, see e.g. Mounce (2006), Mounce (2007) and Mounce and Smith (2007), proposed that the uniqueness and stability of DUE can be guaranteed only for network with single bottleneck per route case under various assumptions, while the works by Peeta and Yang (2003) and Iryo (2008) proposed that the DUE and DSO are unique and stable for general network under certain assumptions. These topics are still open for research. Many issues are needed to be addressed before the application of DTA results. We quote a Chinese proverb to end the chapter: “The journey of a thousand miles begins with a single step”.

Chapter 3

Dynamic user equilibrium with side constraints for a traffic network with traffic volume (queue) control

This chapter investigates a traffic volume control scheme for a dynamic traffic network model which aims to ensure that traffic volumes on specified links do not exceed preferred levels. The problem is formulated as a dynamic user equilibrium problem with side constraints (DUE-SC) in which the side constraints represent the restrictions on the traffic volumes. Travelers choose their departure times and routes to minimize their generalized travel costs, which include early/late arrival penalties. An infinite-dimensional variational inequality (VI) is formulated to model the DUE-SC. Based on this VI formulation, we establish an existence result for the DUE-SC by showing that the VI admits at least one solution. To analyze the necessary condition for the DUE-SC, we restate the VI as an equivalent optimal control problem. The Lagrange multipliers associated with the side constraints as derived from the optimality condition of the DUE-SC provide the traffic volume control scheme. The control scheme can be interpreted as additional travel delays (either tolls or access delays) imposed upon drivers for using the controlled links. This additional delay term derived from the Lagrange multiplier is compared with its counterpart in a static user equilibrium assignment model. If the side constraint is chosen as the storage capacity of a link, the additional delay can be viewed as the effort needed to prevent the link from spillback. Under this circumstance, it is found that the flow is incompressible when the link traffic volume is equal to its storage capacity. An algorithm

based on Euler's discretization scheme and nonlinear programming is proposed to solve the DUE-SC. Numerical examples are presented to illustrate the mechanism of the proposed traffic volume control scheme.

3.1 Introduction

Various forms of traffic control scheme have been introduced to maintain or control the level of service of a traffic network, such as traffic signals, ramp metering, and traffic volume (or queue) control. These schemes mainly mitigate congestion or maximize capacity. However, there is a need in some cases to ensure the safety or desired environmental condition of an urban area or road section that may require a restriction on traffic volume. This chapter focuses on that type of traffic volume control problem, wherein one aims to derive an appropriate control scheme from a dynamic traffic network model.

Similar studies have been conducted in the context of side constrained static traffic assignment problem (SC-TAP) or capacitated static traffic assignment problem (Larsson and Patriksson, 1995, 1999; Larsson et al., 2004; Yang and Huang, 2005). Typically, the side constraints impose restrictions on traffic volumes or maximum delays of certain links. There are two types of side constraint in traffic assignment models: prescriptive (hard) and descriptive (soft) side constraints. Prescriptive side constraints typically arise from traffic management and control policies (e.g., link capacity constraint, traffic signal, access control, and traffic volume control, etc.). With this type of constraint, the SC-TAP can be used to model and evaluate necessary optimal control parameters (implemented as access delay times or tolls). The validity of such controls depends on the assumption that their effects are transferable to travelers' perceptions as additional disutilities of travel. Descriptive side constraints, in contrast, can be introduced to better represent the physical restriction of highways or junctions (e.g., joint capacities of roundabouts), which may not be fully represented by the standard flow-delay relationship (e.g., the BPR function). In this chapter, we concentrate on the prescriptive side constraints for the purposes of access control and traffic volume control.

Under the SC-TAP, Larsson and Patriksson (1995, 1999) show that the Lagrange multiplier as derived from the optimality condition of the SC-TAP can be used as the control parameter (equivalent delay or adjusted travel cost function) on that link. However, the basic assumption of the static model is that the traffic system operates at its steady state with constant demand and link traffic volume. This may not be suitable or plausible for the analysis of urban traffic management, which involves temporal flow and congestion.

We thus aim to extend the SC-TAP to the dynamic case and derive the dynamic control parameter similar to the Lagrange multiplier of the SC-TAP, which can be adopted in the dynamic traffic volume control scheme.

The dynamic user equilibrium (DUE) condition is a type of route/departure time choice principle for dynamic traffic assignment (DTA) problems. Under the DUE, each traveler chooses a route and/or departure time to minimize his generalized travel time/cost (including early/late arrival penalties) (Carey, 2008). Several approaches have been proposed to formulate and solve the DTA problem, e.g., Carey (2008); Friesz et al. (1993, 2001); Friesz and Mookherjee (2006); Friesz et al. (2008); Wie et al. (2002); Huang and Lam (2002); Lindsey (2004), and Chow (2009a). The dynamical framework allows a representation of temporal travel demand and traffic condition in which a time-varying traffic control scheme can also be derived from the model.

The need for the dynamic traffic volume control scheme can be found in several practical cases. For example, in Hong Kong there are several cross harbor tunnels, with the central tunnel the most congested due to its lower toll level. It is thus necessary to control the number of vehicles inside the tunnel to maintain sufficient reserve capacity/space for handling any possible incident (e.g., car accident or disruption due to disaster). We provide a sound theoretical analysis of the optimal design of dynamic traffic volume control scheme by exposing the DUE-SC condition and the additional travel cost induced by side constraints. The introduction of the DUE-SC also provides a potential approach to determining the DUE condition under a pre-specified level of service (LOS) for a traffic network (Yang et al., 2000).

In this chapter, the DUE-SC is formulated as an infinite-dimensional variational inequality following Friesz et al. (1993); Friesz (2010). We adopt the whole-link linear travel time model as proposed by Friesz et al. (1993) that guarantees the first-in-first-out (FIFO) discipline of dynamic flow. The existence of equilibrium to the DUE-SC is verified by showing the VIP admits at least one solution. The Pontryagin minimum principle is then applied to analyze the necessary condition of the DUE-SC for a general traffic network by restating the VI as an equivalent optimal control problem. We then interpret the physical meaning of the Lagrange multipliers, as in the static case. Based on the VI formulation, a fixed point algorithm is developed for the DUE-SC. An algorithm based on Euler's discretization scheme and nonlinear programming is then proposed to solve the DUE-SC. Numerical tests are conducted to illustrate the applicability and mechanism of the proposed traffic volume control scheme.

The remainder of this chapter is organized as follows. In Section 3.2, the SC-TAP is

briefly reviewed. The DUE-SC is formulated and analyzed in Section 3.3. The solution algorithm for solving the DUE-SC problem is proposed in Section 3.4. Numerical examples are then provided to demonstrate the proposed method in Section 3.5. Conclusions and directions for future work are highlighted in Section 3.6. Companion materials are given in the Appendix.

3.2 Static side-constrained traffic assignment problem

Before addressing the DUE-SC, it is necessary to review the formulation and properties of the SC-TAP. An urban road transportation system can be described as a strongly connected directed network $G(N, A)$ where N and A denote the sets of nodes and links, respectively. We assume that the network has n links in total. Let C_a be the capacity of link $a \in A$. Let W be the set of OD pairs and Q_w be the number of trips made between OD pair $w \in W$. A path is defined as a connected sequence of links, P_w denotes the set of all non-cyclic paths connecting OD pair $w \in W$ and P denotes the set of all non-cyclic paths. f_p denotes the flow on path $p \in P$ and υ_a represents the traffic volume on link $a \in A$. We presume that the SC-TAP has at least one feasible solution, which means that link capacities are sufficiently large to allow all travel demands to traverse the network for at least one assignment. The SC-TAP (Larsson and Patriksson, 1995) is formulated as

$$\min Z(\upsilon) = \sum_{a \in A} \int_0^{\upsilon_a} t_a(\omega) d\omega, \tag{3.1}$$

subject to

$$\sum_{p \in P_w} f_p = Q_w, \ w \in W, \tag{3.2}$$

$$\sum_{p \in P} f_p \delta_a^p = \upsilon_a, \ a \in A, \tag{3.3}$$

$$\upsilon_a \leq C_a, \ a \in A, \tag{3.4}$$

$$f_p \geq 0, \ p \in P, \tag{3.5}$$

where $\upsilon = (\upsilon_a : \forall a \in A)$, (3.4) defines the capacity constraints, $t_a(\upsilon_a)$ is the link travel time on link $a \in A$ when the traffic volume on link a is υ_a, and

$$\delta_a^p = \begin{cases} 1, \text{ if } a \in p, \\ 0, \text{ otherwise}, \end{cases}$$

is the Kronecker Delta function. We have the following Lemma for the SC-TAP.

Lemma 3.1. (Larsson and Patriksson, 1995) Let μ_w and λ_a be the Lagrange multipliers associated with constraints (3.2) and (3.4), respectively. The first-order optimality condition for UE with side constraints is given as follows:

$$\begin{aligned}
f_p(c_p - \mu_w) &= 0,\ w \in W, && (3.6)\\
c_p - \mu_w &\geq 0,\ p \in P_w,\ w \in W, && (3.7)\\
\sum_{p \in P_w} f_p - Q_w &= 0,\ w \in W, && (3.8)\\
f_p &\geq 0,\ p \in P, && (3.9)\\
\upsilon_a &\leq C_a,\ a \in A, && (3.10)\\
\lambda_a(\upsilon_a - C_a) &= 0,\ a \in A, && (3.11)\\
\lambda_a &\geq 0, a \in A, && (3.12)
\end{aligned}$$

where c_p, the generalized link travel time, is given by

$$\begin{aligned}
c_p &= \sum_{a \in A} \hat{t}_a(\upsilon_a)\delta_a^p,\ p \in P, && (3.13)\\
\hat{t}_a(\upsilon_a) &= t_a(\upsilon_a) + \lambda_a,\ a \in A. && (3.14)
\end{aligned}$$

The generalized link travel time comprises two components: $t_a(\upsilon_a)$ is the normal travel time on link $a \in A$, and λ_a is the Lagrange multiplier associated with constraint (3.4). $\lambda_a \geq 0$ only if $\upsilon_a = C_a$, and $\lambda_a = 0$ if $\upsilon_a < C_a$. We can regard this multiplier, λ_a, as an additional time or cost penalty (besides the normal travel time) that users traveling on this saturated link are willing to wait or pay for using the link. ∎

3.3 Dynamic user equilibrium with side constraints

3.3.1 Preliminaries and dynamic user equilibrium

The dynamic user equilibrium formulation adopted here can be traced back to Friesz et al. (1993), and is also discussed in detail by Friesz et al. (2001). According to Friesz and Mookherjee (2006); Friesz (2010), DUE models tend to be comprised of four essential submodels: (i) a model of path delay; (ii) flow dynamics; (iii) flow propagation constraints; and (iv) a route/departure-time choice model. We will consider a finite time planning horizon $T > 0$ and regard time $t \in [0, T] \subset R_+^1$ as a continuous variable. Let P denote the set of all paths and $|P|$ denote the number of paths in a network. An arbitrary path $p \in P$ of the network of interest is defined by a sequence of the links used by that path which is denoted by $p \doteq \{a_1, a_2, \cdots, a_{m(p)}\}$, where $m(p)$ is the number of links used by path p.

We denote the vector of path flows as $\mathbf{h} = (h_p : p \in P)$ and assume that $\mathbf{h} \in \mathcal{L}_+ = \left(L_+^2[0,T]\right)^{|P|}$, where $\mathcal{L}_+$ denotes the nonnegative cone of the $|P|$-fold product of the Hilbert space $L_+^2[0,T]$ of square-integrable functions on $[0,T]$. As argued in Friesz et al. (2011), the most essential component of a dynamic user equilibrium model is the path delay operator which is denoted as $D_p(t,\mathbf{h}), \forall p \in P$. The path delay operator $D_p(t,\mathbf{h})$ provides the travel time needed to traverse path p for per unit of flow departing from the origin at time t. When departure time choice is enabled, the schedule delay cost function (or early/late arrival penalty) $\kappa(\chi)$ is employed, whereby χ is defined as the difference between actual and preferred arrival time denoted by t^* and $t^* < T$: $\chi = t + D_p(t,\mathbf{h}) - t^*$. To this end, we define effective path delay operator $\Psi_p(t,\mathbf{h})$ by adding the schedule delay to $D_p(t,\mathbf{h})$, i.e.

$$\Psi_p(t,\mathbf{h}) = D_p(t,\mathbf{h}) + \kappa(\chi). \tag{3.15}$$

Let Q_w denote the fixed total travel demand for origin-destination (OD) pair $w \in W$ and P_w denote the set of paths connecting OD pair w. We need to enforce flow conservation, i.e.

$$\sum_{p \in P_w} \int_0^T h_p(t)dt = Q_w, \quad \forall w \in W, \tag{3.16}$$

where (3.16) is comprised of Lebesgue integrals. Thus, we define the feasible region for the DUE problem as

$$\check{\Lambda} = \left\{ \mathbf{h} : \mathbf{h} \in \mathcal{L}_+, \ \sum_{p \in P_w} \int_0^T h_p(t)dt = Q_w, \ \forall w \in W \right\}, \tag{3.17}$$

Definition 3.1. Dynamic user equilibrium (Friesz, 2010). For any $\mathbf{h}^* \in \check{\Lambda}$ and any nonnegative vector $\varrho = (\varrho_w : w \in W) \in R_+^{|W|}$, the pair $(\mathbf{h}^*, \varrho)$ is a simultaneous departure-time-and-path-choice dynamic user equilibrium if and only if the following two conditions are satisfied for all $p \in P_w$ and for all $w \in W$:

$$\begin{aligned} h_p^*(t) > 0 \quad &\Rightarrow \quad \Psi_p(t,\mathbf{h}^*) = \varrho_w, \\ \Psi_p(t,\mathbf{h}^*) > \varrho_w \quad &\Rightarrow \quad h_p^*(t) = 0, \end{aligned} \tag{3.18}$$

where ϱ_w is the smallest travel time for the OD pair w, given by

$$\varrho_w = \min\{\varrho_p : p \in P_w\} \geq 0, \text{ and } \varrho_p = \operatorname{ess\,inf}\{\Psi_p(t,\mathbf{h}) : t \in [0 \ \ T]\} \geq 0, \tag{3.19}$$

where ess inf is essential infimum operator which defines the largest essential lower bound for a given function f in which all $\inf f \ \leq \ \operatorname{ess\,inf} \ f$. ∎

As shown by Friesz et al. (1993), any solution of the following variational inequality is a solution of the DUE problem with simultaneous departure-time-and-path-choice: find $\mathbf{h}^* \in \check{\Lambda}$ such that

$$\langle \Psi(t, \mathbf{h}^*), (\mathbf{h} - \mathbf{h}^*) \rangle = \sum_{p \in P} \int_0^T \Psi_p(t, \mathbf{h}^*) \left(h_p(t) - h_p^*(t) \right) dt \geq 0, \tag{3.20}$$

for all $\mathbf{h} \in \check{\Lambda}$, where $\Psi(t, \mathbf{h}) = (\Psi_p(t, \mathbf{h}) : p \in P)$ is the effective network delay operator.

In the above formulation of DUE problem, only two of the four essential sub-models of DUE are considered, i.e. a path delay model and a route/departure-time choice model. Since the path delay operator does not explicitly depend on the detailed analytical traffic dynamics, i.e. flow dynamics and proper flow propagation constraints, this formulation is reasonable

> if the delays are assumed to be exogenous (Friesz, 2010), i.e. the path delay operators are known in advance or represented by a simulation model, or
>
> if one uses the link traffic dynamics to eliminate the vector of link traffic volumes and recognizes that link exit flows are completely determined by path flows (i.e. $h_p(t),\ \forall p \in P$) and link traffic volumes according to the flow propagation constraints (Friesz et al., 1993; Friesz and Mookherjee, 2006).

To make the formulation self-contained and to investigate more analytical properties of the DUE problem, we further introduce the other two sub-models, i.e. (ii)-(iii), which are also known as network loading models. For a path $p \doteq \{a_1, a_2, \cdots, a_{m(p)}\}$, the dynamics of link a_i is assumed to be described by the following equations:

$$\frac{dx_{a_i}^p(t)}{dt} = g_{a_{i-1}}^p(t) - g_{a_i}^p(t),\ \forall p \in P,\ i \in [1,\ m(p)], \tag{3.21}$$

$$x_{a_i}(t) = \sum_{p \in P} x_{a_i}^p(t) \delta_{a_i}^p,\ \forall a_i \in A, \tag{3.22}$$

where $x_{a_i}^p(t)$ denotes the traffic volume on path p traversing link a_i at time t, $g_{a_i}^p(t)$ is the flow exiting link a_i and $g_{a_{i-1}}^p(t)$ is the flow entering link a_i of path $p \in P$ at time t. The total traffic volume $x_{a_i}(t)$ of link a_i at time t is defined by (3.22). For $i = 1$, that is the link connecting the origin of path p, we have

$$\frac{dx_{a_1}^p(t)}{dt} = h_p(t) - g_{a_1}^p(t),\ \forall p \in P, \tag{3.23}$$

where $g_{a_0}^p(t) = h_p(t)$, i.e. the departure rate from the origin of path p at time t. We assume that $\mathbf{x}_a^P = (x_{a_i}^p : p \in [1, |P|],\ i = [1, m(p)]) \in \left(\mathcal{H}^1[0, T]\right)^{n_1}$, where $n_1 = \sum_{p=1}^{|P|} m(p)$

and $\mathcal{H}^1[0,T]$ denotes the Sobolev space[1]. Under this assumption, we can represent the vector of link traffic volumes defined by (3.22), i.e. $\mathbf{x} = (x_{a_i} : a_i \in A)$, as a function of time and path flows, i.e. $\mathbf{x}(t,\mathbf{h}) : R^1_+ \times \mathcal{L}_+ \to \left(\mathcal{H}^1[0,T]\right)^n$, where $n = |A|$ is the number of links in the network. Under certain conditions, e.g. strong FIFO (Theorem 3.3 of Zhu and Marcotte (2000)), we can represent the link flow operator as $\mathbf{x}(\mathbf{h}) = \mathbf{x}(t,\mathbf{h})$ which is weakly continuous on $\check{\Lambda}$. We assume also that $\mathbf{g} = (g^p_{a_i} : p \in [1,|P|],\ i = [1,m(p)]) \in \left(L^2_+[0,T]\right)^{n_1}$.

In addition, a link delay function (or link travel time function) $D_{a_i}(x_{a_i}(t))$, which defines the link travel time as a function of the link traffic volume at the entry time to the link, is required to characterize travel time required to traverse a link, say link a_i:

$$\begin{aligned} \tau^p_{a_1}(t) &= t + D_{a_1}(x_{a_1}(t)),\ \forall p \in P, \\ \tau^p_{a_i}(t) &= \tau^p_{a_{i-1}}(t) + D_{a_i}\left(x_{a_i}(\tau^p_{a_{i-1}}(t))\right),\ \forall p \in P,\ i \in [2,m(p)], \end{aligned} \tag{3.24}$$

where $\tau^p_{a_i}(t)$ is known as the exit time for vehicles entering link a_i at time t. As noticed the link delay function depends on its link traffic volume only, hence it is known as separable link delay model. To proceed the analysis, we assume that all link delay functions are differentiable with respect to their own arguments. By differentiating (3.24) and using the link flow conservation, we can obtain the following proper flow propagation constraints (Friesz et al., 2001):

$$g^p_{a_1}\left(t + D_{a_1}(x_{a_1}(t))\right)\left(1 + D'_{a_1}(x_{a_1}(t))\,\dot{x}_{a_1}(t)\right) = h_p(t),\ \forall p \in P, \tag{3.25}$$

$$g^p_{a_i}\left(t + D_{a_i}(x_{a_i}(t))\right)\left(1 + D'_{a_i}(x_{a_i}(t))\,\dot{x}_{a_i}(t)\right) = g^p_{a_{i-1}}(t),\ \forall p \in P,\ i \in [2,m(p)], \tag{3.26}$$

the superscript "$'$" denotes differentiation with respect to the associated function argument, and the superscript "$\cdot$" denotes differentiation with respect to time. It is clear that we need

$$1 + D'_{a_i}(x_{a_i}(t))\,\dot{x}_{a_i}(t) \geq 0 \tag{3.27}$$

to ensure the non-negativeness of traffic flows, i.e. given $h_p(t) \geq 0,\ \forall p \in P$, we have $g^p_{a_i}(t) \geq 0,\ \forall p \in P,\ \forall a_i \in A$, and $x_{a_i} \geq 0,\ \forall a_i \in A$. The condition (3.27) ensures the FIFO queue discipline for the DTA problems. Without loss of generality, we assume zero initial conditions, that is $x_{a_i}(0) = 0,\ \forall a_i \in A$, $h_p(0) = 0, \forall p \in P$. When $D_{a_i}(\cdot)$ is linear, Friesz et al. (1993) first showed that the model satisfies the FIFO for all continuous inflow profiles and later Zhu and Marcotte (2000) showed the strong FIFO property for

[1]Suppose that $u \in L^p(\Omega)$, $1 \leq p < \infty$ and there exist weak derivatives $\partial^\alpha u$ for any α with $|\alpha| < \ell$, $\ell \in Z_+$ (all derivatives up to order ℓ), such that $\partial^\alpha u \in L^p(\Omega), |\alpha| < \ell$, Then we say that $u \in W^{\ell,p}(\Omega)$. We denote $\mathcal{H}^1[0,T] = W^{1,2}[0,T]$ as the Sobolev space, see e.g. Minoux (1986); Leoni (2009).

all continuous and bounded inflow profiles. Xu et al. (1999) obtained fairly weakly FIFO conditions for nonlinear $D_{a_i}(\cdot)$. In this chapter, the whole-link linear travel time model proposed by Friesz et al. (1993) is adopted. The exit time of a vehicle entering the link at time t can be calculated as:

$$\tau_{a_i}(t) = t + b_{a_i} + \frac{x_{a_i}(t)}{R_{a_i}}, \tag{3.28}$$

where b_{a_i} is a flow-invariant travel time (free-flow travel time) of link a_i. A suitable physical interpretation for the parameter R_{a_i} is the maximum feasible constant outflow from the link (Carey and McCartney, 2002) and thus we state it here as the service rate of link a_i. To calculate the path delays in terms of link delays, the nested path delay operators were proposed by Friesz et al. (1993) which are defined as:

$$D_p(t, \mathbf{x}) := \sum_{i=1}^{m(p)} \delta_{a_i}^p \Phi_{a_i}(t, \mathbf{x}), \forall p \in P, \tag{3.29}$$

where

$$\begin{aligned}
\Phi_{a_1}(t, \mathbf{x}) &= D_{a_1}\left(x_{a_1}(t)\right), \\
\Phi_{a_i}(t, \mathbf{x}) &= D_{a_i}\left(x_{a_i}(t + \Phi_{a_1} + \cdots + \Phi_{a_{i-1}})\right) = D_{a_i}\left(x_{ai}\left(t + \sum_{j=1}^{i-1} \Phi_{a_j}\right)\right), \ \forall i \in [2, m(p)].
\end{aligned}$$

The value of path delay operator $D_p(t, \mathbf{h})$ is calculated by identifying $D_p(t, \mathbf{h})$ to the nested path delay operator $D_p(t, \mathbf{x})$, i.e. $D_p(t, \mathbf{h}) = D_p(t, \mathbf{x})$.

3.3.2 Existence of the DUE with simultaneous departure-time-and-path-choice

The existence of equilibrium to the DUE without departure time choice is proven by Zhu and Marcotte (2000), wherein $\Psi_p(t, \mathbf{h}) = D_p(t, \mathbf{h})$, i.e. no early/late penalty is considered, and the feasible region is defined as:

$$\hat{\Lambda} = \left\{ \mathbf{h} : \mathbf{h} \in \mathcal{L}_+, \ \sum_{p \in P_w} h_p(t) = Q_w(t), \ h_p(t) \leq B_p, \ \forall p \in P_w, \ \forall w \in W, \ \forall t \in [0, T] \right\},$$

where B_p is the upper bound of the path flow. The existence of the DUE without departure time choice is proven by showing that the link traffic dynamics, link delay sub-model and path delay operators under rather mild regularity conditions are continuous under the assumption that departure rates are bounded from above. The existence result has not been shown for the DUE with simultaneous departure-time-and-path-choice, i.e. the variational inequality defined by (3.17) and (3.20) admits a least one solution. To prove the

existence of the DUE-SC, we will first prove the existence of the DUE with simultaneous route/departure time choice. In the discrete time framework, Wie et al. (2002) show the existence result for the DUE with simultaneous departure-time-and-path-choice and elastic demand under certain assumptions, e.g. the path delay operator $\Psi_p(t, \mathbf{h})$ is continuous and each integral constraint comprising (3.17) is a Riemann integral. As we consider the integrals in (3.17) are comprised of Lebesgue integrals and the flow conservation constraints themselves are interpreted as valid almost everywhere, the finiteness of each Q_w is not enough to assure bounded path flows, or the path flows may not be well defined. In this case, we need to assume that the path flows are bounded as done by Zhu and Marcotte (2000), i.e. the feasible region is further restricted as

$$\Lambda = \left\{ \mathbf{h} : \mathbf{h} \in \mathcal{L}_+, \sum_{p \in P_w} \int_0^T h_p(t)dt = Q_w, h_p(t) \leq B_p, \forall p \in P_w, \forall w \in W, \forall t \in [0, T] \right\},$$

where B_p is the upper bound of path flow $h_p(t)$. In the continuous time framework, we state the following lemma for the proof of the existence result for the DUE with simultaneous departure-time-and-path-choice:

Lemma 3.2. Assume that there is a finite instant T such that

(i) all path flow departure rates $h_p(t)$ are well defined over Λ;

(ii) all links satisfy the strong FIFO condition with a uniform constant over $[0, T]$;

(iii) the link travel time functions are nonnegative, nondecreasing, differentiable and Lipschitz continuous over $[0, T]$;

(iv) the early/late penalty function $\kappa : R \to R$ is continuous.

Then the effective network delay operator $\Psi(\mathbf{h})$ is weakly continuous over Λ. ∎

Proof of Lemma 3.2 If each integral constraint comprising (3.17) is assumed to be a Lebesgue integral and $h_p(t)$ is finite, condition (i) holds. Because we apply the linear link travel time function and link dynamics of Zhu and Marcotte (2000), conditions (ii) and (iii) are satisfied. Since all links satisfy the strong FIFO condition, a similar condition holds for all paths, i.e. $D_p(t_j, \mathbf{h}) > D_p(t_i, \mathbf{h})$ whenever $t_j > t_i$. From Theorem 2.1 of Zhu and Marcotte (2000), $D_p(t_i, \mathbf{h})$ is measurable and square integrable. Moreover, we can represent the path delay operator as $D_p(t_i, \mathbf{h}) = D_p(\mathbf{h})$. From Theorem 4.1 of Zhu and Marcotte (2000), $D(\mathbf{h}) = (D_p(\mathbf{h}) : \forall p \in P)$ is weakly continuous over Λ. By definition, i.e. (4.16), $\Psi_p(\mathbf{h}) = D_p(\mathbf{h}) + \kappa(D_p(\mathbf{h}))$. Similar to the proof of Theorem 4.1 of Zhu and Marcotte (2000) we have $\kappa(D_p(\mathbf{h}))$ is weakly continuous over Λ as $D_p(\cdot)$ is weakly

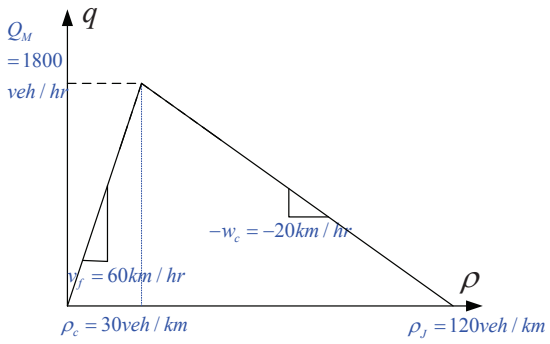


Figure 3.1: Fundamental diagram of an urban street

continuous over Λ and $\kappa(\cdot)$ is continuous over R. Thus $\Psi_p(\mathbf{h})$ is weakly continuous. The effective network delay operator $\Psi(\mathbf{h})$ is weakly continuous over Λ as each of its components is weakly continuous. ■

Proposition 3.1. Suppose all the conditions in Lemma 3.2 hold. Then the solution set of the infinite dimensional variational inequality (3.17) and (3.20) is nonempty. ∎

Proof of Proposition 3.1 With Lemma 3.2, the existence of the DUE with simultaneous departure-time-and-path-choice can be shown by following the proof of Theorem 4.2 of Zhu and Marcotte (2000) for the feasible set Λ is closed, convex, bounded and $\Psi(\mathbf{h})$ is weakly continuous. ■

In fact, condition (iv) in Lemma 3.2 is consistent with the existence result of the DUE with simultaneous departure-time-and-path-choice for bottleneck model proposed by Lindsey (2004). Lemma 3.2 proves the existence of equilibrium for the DUE with simultaneous departure-time-and-path-choice problem. This result will be useful for us to prove the existence result for dynamic user equilibrium with side constraints.

3.3.3 Problem formulation of dynamic user equilibrium with side constraint

To begin with, we first look into an illustrative example of traffic volume control and define the side constraints for dynamic user equilibrium traffic assignment.

An illustrative example of access control by controlling link traffic volume. Consider a two-lanes urban street with a length of 100 meters that admits a fundamental diagram as depicted in Figure 3.1. The free-flow travel time of the street is then 0.1 minutes. The service rate is 60 veh/min. If the linear travel time function is adopted, the travel time of this street is $\theta(t) = 0.1 + \frac{x(t)}{60}$ minutes. If vertical queue models are adopted to model the traffic dynamics and the link is assumed to be capable of accommodating

arbitrary large number of vehicles, the congestion effect would be captured by the above linear travel time function. However, there are physical storage capacities for links to accommodate queuing vehicles. In this example, the street would be completely blocked if the traffic volume on it equals to 24 vehicles, which may render the travel time of the link be larger than 1 minute. However, it is unlikely for us to capture the spillback effect of congestion by applying the linear travel time function in conjunction with vertical queue models (since the link can accommodate 24 vehicles only, which implies the maximal travel time for this link is 0.35 min).

In DTA literature, the side constraint is usually selected as the inflow rate to a link is less than or equal to its service rate. By such kind of restriction, one can eliminate the traffic congestion (or queues) (Carey and Ge, 2003). On the other hand, the effort needed to prevent the congestion happening is captured by the Lagrange multiplier associated with the side constraint, which can be viewed as an additional travel cost imposed on the travelers. The travelers are then charged by this additional travel cost for using the saturated bottleneck (Yang and Meng, 1998). However, this is not practical due to fact that congestion is somehow inevitable. A practical method that is widely applied and has been validated by traffic control engineering field applications is to control the link traffic volumes to capture the potential spillback effects of traffic on ramps and queues in store-and-forward based methods (see e.g. Papageorgiou et al. (2003); Aboudolas et al. (2009) and the references therein). The traffic volume control can capture the potential spillback effect in the sense that the effort needed to prevent the spillback happening can be simulated by restricting the link traffic volume be less than or equal to its jam traffic volume (i.e. storage capacity), e.g. $x_a(t) \leq 24$ vehicles in the previous example. Therefore, we consider the traffic volume control as a kind of side constraints in this chapter. Another possible way to capture the spillback effect by the WLM is to apply a special piecewise linear exit-flow function and explicitly restrict the exit-flow according to downstream conditions, which results in a relaxed cell transmission model (CTM) when the restricted WLM is discretized (Nie, in press). The other way to capture the spillback effect is to apply the physical queue models (Lo and Szeto, 2002; Gentile et al., 2005; Nie and Zhang, 2010), e.g. the CTM. A critical drawback of these two approaches is the physical queue models introduce discontinuous and non-differentiable traffic dynamics and travel cost functions, e.g. path delay operator. This prevents us from solving the DTA problems analytically. The non-monotone and non-differentiability properties also lead to difficulties in finding numerical solutions, because the convergence of existing algorithms rely on either monotonicity and/or differentiability (Szeto and Lo, 2006).

Recent studies have revealed that these macroscopic link models (including vertical queue models, e.g. WLM and bottleneck model; and physical queue models, e.g. CTM) would produce almost the same traffic assignment result unless there is a spillback (shock-wave) (Nie and Zhang, 2005; Mun, 2007; Nie and Zhang, 2010). By this result, another way to capture the spillback effect is to adopt vertical queue models, to be precise the WLM in this chapter, meanwhile to restrict the link traffic volumes to be less than or equal to the link storage capacities on the network such that no spillback would happen. The spillback effect is captured in terms of the effect needed to prevent the network from spillback. By doing this, we can avoid the drawbacks of physical queue models while capturing the spillback effect. Similar to the bottleneck case, we can regard this effect as additional travel cost imposed on travelers for using the controlled links. The travel time function adopted in DTA literature, i.e. $D_a(x_a(t))$, would be more meaningful under this circumstance since if there is spillback the link travel time function would not be separable, e.g. depends on the downstream traffic conditions like the case of physical queue model (Szeto and Lo, 2006). Practically, the dynamic traffic volume control can also be used to alleviate congestion spillback, avoid gridlock, and increase the safety of some critical facilities.

From the previous example and the analysis above, the side constraint considered here is the link traffic volume control or restriction on link traffic volume, i.e.

$$x_{a_i}(t) \leq C_{a_i}(t), \quad \forall a_i \in A, \ \forall t \in [0,T]. \tag{3.30}$$

(3.30) defines the side constraints imposed on the links of the network. Let us define the following vector-valued function for the side constraints as $\mathbf{C} = (C_{a_i} : a_i \in A)$, where $\mathbf{C} : [0,T] \to \Sigma \subset R_+^n$, with Σ being a prescribed compact subset of R_+^n with known bounds. We assume that $\mathbf{C} \in \left(W^{1,\infty}[0,T]\right)^n$, where $W^{1,\infty}[0,T]$ denotes the space of Lipschitz functions on $[0,T]$ (Leoni, 2009). We rewrite the side constraints in a compact form as

$$\mathbf{x}(t) \leq \mathbf{C}(t), \ \forall t \in [0,T]. \tag{3.31}$$

By construction, the set $\Sigma_x \triangleq \{\mathbf{x}(t) : 0 \leq \mathbf{x}(t) \leq \mathbf{C}(t)\}$ is compact.

The DUE-SC with simultaneous departure-time-and-path-choice can be formulated as: find $\mathbf{h}^* \in \bar{\Lambda}$ such that

$$\langle \Psi(t,\mathbf{h}^*), (\mathbf{h}-\mathbf{h}^*) \rangle = \sum_{p \in P} \int_0^T \Psi_p(t,\mathbf{h}^*) \left(h_p(t) - h_p^*(t)\right) dt \geq 0, \tag{3.32}$$

for all $\mathbf{h} \in \bar{\Lambda}$, where

$$\bar{\Lambda} = \Lambda \bigcap \tilde{\Lambda}, \tag{3.33}$$

and $\tilde{\Lambda} = \{\mathbf{h} : \mathbf{h} \geq 0,\ \mathbf{x}(t) \leq \mathbf{C}(t)\}$.

3.3.4 Existence of the DUE-SC with simultaneous departure-time-and-path-choice

Following the standard assumption of the SC-TAP, we need to presume that the feasible set $\bar{\Lambda}$ of the DUE-SC with simultaneous departure-time-and-path-choice is not empty. For the DUE-SC with simultaneous departure-time-and-path-choice, we have the following proposition:

Proposition 3.2. Suppose all the conditions in Lemma 3.2 hold. Then the infinite dimensional variational inequality (3.32)-(3.33) admits a least one solution. ∎

Proof of Proposition 3.2 For the DUE-SC, as the side constraints are introduced, the feasible region Λ of the original DUE problem is further restricted to $\bar{\Lambda}$. To show the variational inequality (3.32)-(3.33) admits at least one solution, we need to show that the new feasible region $\bar{\Lambda}$ restricted by the side constraints is bounded, convex and closed. As shown by Zhu and Marcotte (2000), the flow operator $\mathbf{x}(\mathbf{h})$ as a function of the path flow $\mathbf{h}$ is weakly continuous given that conditions (i)-(iii) of Lemma 3.2 hold. By following the weak continuity of $\mathbf{x}(\mathbf{h})$ and the fact that the set Σ_x is constructed to be compact, the set $\tilde{\Lambda}$ is closed (Noiri, 1974; Rose, 1984). Therefore, $\bar{\Lambda}$ is a closed set. In conjunction with the assumption that $\mathbf{h} \geq 0$ and is bounded from above, the set $\bar{\Lambda}$ is bounded and convex. If all the conditions in Lemma 3.2 hold, the weak continuity of the effective network delay operator $\Psi(\mathbf{h})$ can be shown. Thus, the variational inequality defined by (3.32) and (3.33) admits at least one solution. The existence of equilibrium to the DUE-SC with simultaneous departure-time-and-path-choice is established. ■

3.3.5 Necessary condition of the DUE-SC with simultaneous departure-time-and-path-choice

In the previous sections, we have studied the problem of existence of equilibriums to the DUE and the DUE-SC with simultaneous departure-time-and-path-choice, wherein the delay operators are exogenous. To analyze the necessary conditions for the DUE and DUE-SC problems, we consider the circumstance where the delay operators are endogenous. An equivalent form that facilitates derivation of necessary conditions for variational inequality (3.32)-(3.33) is the following differential variational inequality (DVI) (Friesz et al., 2001;

Friesz, 2010): find $(\mathbf{x}^*, \mathbf{h}^*, \mathbf{g}^*) \in \Gamma$ such that

$$\langle \Psi(t, \mathbf{x}^*), (\mathbf{h} - \mathbf{h}^*) \rangle = \sum_{p \in P} \int_0^T \Psi_p(t, \mathbf{x}^*) \left(h_p(t) - h_p^*(t) \right) dt \geq 0, \tag{3.34}$$

for all $(\mathbf{x}, \mathbf{h}, \mathbf{g}) \in \Gamma$, where Γ is the admissible set

$$\Upsilon = \{(\mathbf{x}, \mathbf{h}, \mathbf{g}) \geq 0 : (3.23),\ (3.21),\ (3.16),\ (3.25),\ (3.26),\ (3.30),\ \text{and } (3.36) \text{ hold}\}, \tag{3.35}$$

which incorporates zero initial conditions

$$x_{a_i}(0) = 0,\ \forall a_i \in A,\ h_p(0) = 0, \forall p \in P, \tag{3.36}$$

and the flow vectors $\mathbf{x}, \mathbf{h}, \mathbf{g}$ belong to some appropriate function spaces presumed in Section 3.3.1.

To facilitate the analysis of the necessary conditions for (3.34)-(3.35), it is helpful to restate the DVI as the following optimal control problem:

$$\min J = \sum_{\forall p \in P} \int_0^T \Psi_p(t, \mathbf{x}^*) h_p(t) dt, \tag{3.37}$$

subject to,

$$\frac{dx_{a_1}^p(t)}{dt} = h_p(t) - g_{a_1}^p(t)\ \ (\lambda_{a_1}^p)\ \forall p \in P, \tag{3.38}$$

$$\frac{dx_{a_i}^p(t)}{dt} = g_{a_{i-1}}^p(t) - g_{a_i}^p(t),\ \ (\lambda_{a_i}^p)$$
$$\forall p \in P,\ i \in [2,\ m(p)], \tag{3.39}$$

$$g_{a_1}^p\left(t + D_{a_1}\left(x_{a_1}(t)\right)\right)\left(1 + D'_{a_1}\left(x_{a_1}(t)\right)\dot{x}_{a_1}(t)\right) = h_p(t),\ \ (\gamma_{a_1}^p)\ \forall p \in P, \tag{3.40}$$

$$g_{a_i}^p\left(t + D_{a_i}\left(x_{a_i}(t)\right)\right)\left(1 + D'_{a_i}\left(x_{a_i}(t)\right)\dot{x}_{a_i}(t)\right) = g_{a_{i-1}}^p(t),\ \ (\gamma_{a_i}^p)$$
$$\forall p \in P,\ i \in [2, m(p)], \tag{3.41}$$

$$\frac{dE_w(t)}{dt} = \sum_{p \in P_w} h_p(t),\ \ (\mu_w)\ \forall w \in W, \tag{3.42}$$

$$-h_p(t) \leq 0,\ \ (\rho_{a_0}^p)\ \forall p \in P, \tag{3.43}$$

$$-g_{a_i}^p(t) \leq 0,\ \ (\rho_{a_i}^p)\ \forall p \in P, \tag{3.44}$$

$$-x_{a_i}^p(t) \leq 0,\ \ (\zeta_{a_i}^p)\ \forall p \in P, \tag{3.45}$$

$$E_w(T) = Q_w,\ \ (\phi_w)\ \forall w \in W, \tag{3.46}$$

$$x_{a_i}(t) \leq C_{a_i}(t),\ \ (\eta_{a_i})\ \forall a_i \in A, \tag{3.47}$$

$$x_{a_i}(0) = 0,\ \forall a_i \in A, E_w(0) = 0,\ \forall w \in W,\ h_p(0) = 0,\ \forall p \in P, \tag{3.48}$$

where Equations (3.42) and (3.46) define the flow conservation constraints, where $E_w(t)$ is an extended state (Friesz et al., 2001; Friesz, 2010), which are equivalent to (3.16).

(3.48) specifies the zero initial conditions, which mean that the network is empty at the beginning. (3.47) is the side constraint imposed, which is known as pure state variable inequality constraint in optimal control theory. The variables in brackets are Lagrange multipliers associated with the corresponding constraints.

The following proposition states the necessary condition for the DUE-SC with simultaneous departure-time-and-path-choice, which can be recognized as the type of equilibrium given by (3.18)-(3.19).

Proposition 3.3. Suppose an appropriate constraint qualification for the pure state constraint (3.47) is satisfied. The necessary condition for the DUE-SC with simultaneous departure-time-and-path-choice can be stated as follows:

$$h_p(t) \begin{cases} > 0 \Rightarrow \Psi_p(t, \mathbf{x}^*) + l_p(t) = \phi_w, \\ = 0 \Rightarrow \Psi_p(t, \mathbf{x}^*) + l_p(t) > \phi_w, \end{cases} \quad \forall p \in P_w,\ w \in W, \tag{3.49}$$

where ϕ_w is the travel cost of OD pair w under the DUE-SC condition, which is determined by the fixed total travel demand Q_w of the OD pair. $l_p(t) := \sum_{i=0}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \eta_{a_{i+1}}(u)\sigma_{a_{i+1}}^p du$, where η_{a_i} is the Lagrange multiplier associated with the side constraint imposed on link a_i which satisfies the following complementary slackness condition

$$\eta_{a_i}(t) \geq 0,\ x_{a_i}(t) - C_{a_i}(t) \leq 0,\ \eta_{a_i}(t)\left(x_{a_i}(t) - C_{a_i}(t)\right) = 0,\ \forall a_i \in A,$$

and t_i^p is the entry time to link a_{i+1} for vehicles traveling by path p that departed from the origin at time $t \triangleq t_0^p$. ∎

Proof of Proposition 3.3 Our intent is to motivate why the DUE condition (3.49) will be fulfilled when the postulated VI (3.32)-(3.33) or DVI (3.34)-(3.35) is solved. To this end it is enough to assume that a constraint qualification for the side constraint or the pure state variable inequality constraint (3.48) is satisfied and, therefore, the desired dual variables (or adjoint functions (equations)) are available. However, in optimal control theory, the proper foundation for knowing that the dual variables of a large system of multiple pure state constraints, like those dual variables needed to express the DUE condition (3.49), is a computational one. The appropriate dual variables and their trajectories are obtained when our computational approach based on the fixed point algorithm converges. In this proof, we derive the necessary condition for the DUE-SC by applying the Pontryagin Minimum Principle.

The Lagrangian for the optimal control problem (4.1)- (3.48) can be defined as:

$$
\begin{aligned}
Z^* &= \sum_{p\in P}\int_0^T \Big\{\Psi_p(t,\mathbf{x}^*)h_p(t) + \lambda^p_{a_1}(t)\left(h_p(t) - g^p_{a_1}(t) - \frac{dx^p_{a_1}(t)}{dt}\right) \\
&+ \sum_{i\in[2,m(p)]} \lambda^p_{a_i}(t)\left(g^p_{a_{i-1}}(t) - g^p_{a_i}(t) - \frac{dx^p_{a_i}(t)}{dt}\right) \\
&+ \gamma^p_{a_1}(t)\left(h_p(t) - g^p_{a_1}\left(t + D_{a_1}\left(x_{a_1}(t)\right)\right)\left(1 + D'_{a_1}\left(x_{a_1}(t)\right)\dot{x}_{a_1}(t)\right)\right) \\
&+ \sum_{i\in[2,m(p)]} \gamma^p_{a_i}(t)\left(g^p_{a_{i-1}}(t) - g^p_{a_i}\left(t + D_{a_i}\left(x_{a_i}(t)\right)\right)\left(1 + D'_{a_i}\left(x_{a_i}(t)\right)\dot{x}_{a_i}(t)\right)\right) \\
&+ \sum_{i\in[1,m(p)]} \eta_{a_i}(t)\left(x_{a_i}(t) - C_{a_i}(t)\right) - \rho^p_{a_0}(t)h_p(t) \\
&- \sum_{i\in[1,m(p)]} \rho^p_{a_i}(t)g^p_{a_i}(t) + \sum_{i\in[1,m(p)]} \zeta^p_{a_i}(t)x^p_{a_i}(t)\Big\}dt \\
&+ \sum_{w\in W}\int_0^T \mu_w(t)\left(\sum_{p\in P_w} h_p(t) - \frac{dE_w(t)}{dt}\right)dt \\
&+ \sum_{w\in W} \phi_w(T)\left(Q_w - E_w(T)\right), \qquad (3.50)
\end{aligned}
$$

where $\lambda^p_{a_i}(t)$ is the costate variable for the traffic dynamics of link a_i contributed by path p. $\mu_w(t)$, $\eta_{a_i}(t)$ and $\gamma^p_{a_i}(t)$ are the Lagrange multipliers associated with the cumulative inflow, the side constraint of link a_i, and the proper flow propagation constraint, respectively. $\phi_w(T)$ is the multiplier associated with the total throughput for each OD pair w. $\rho^p_{a_0}(t)$, $\rho^p_{a_i}(t)$, and $\zeta^p_{a_i}(t)$ are the Lagrange multipliers associated with the nonnegative flow constraints. To simplify the notation and proceed to the analysis, let us define

$$\tilde{g}^p_{a_i}(t) = g^p_{a_i}\left(t + D_{a_i}(x_{a_i}(t))\right),\ \forall p \in P,\ i \in [1, m(p)].$$

By a simple calculation in conjunction with the zero initial condition assumption, the Lagrangian can be specified as

$$
\begin{aligned}
Z^* &= \sum_{p\in P}\left(\sum_{i\in[1,m(p)]}\left(-\lambda^p_{a_i}(T)x^p_{a_i}(T)\right)\right) \\
&+ \sum_{w\in W}\left(-\mu_w(T)E_w(T)\right) + \sum_{w\in W}\phi_w(T)\left(Q_w - E_w(T)\right) \\
&+ \int_0^T\left(H(t) + \sum_{w\in W}\frac{d\mu_w(t)}{dt}E_w(t) + \sum_{p\in P}\left(\sum_{i\in[1,m(p)]}\frac{d\lambda^p_{a_i}(t)}{dt}x^p_{a_i}(t)\right)\right)dt, \qquad (3.51)
\end{aligned}
$$

where we define the Hamiltonian function for the optimal control problem as

$$\begin{aligned} H(t) &= \sum_{p\in P}\Big[\Psi_p(t,\mathbf{x}^*)h_p(t) + \lambda^p_{a_1}(t)\left(h_p(t) - g^p_{a_1}(t)\right) + \sum_{i\in[2,m(p)]}\lambda^p_{a_i}(t)\left(g^p_{a_{i-1}}(t) - g^p_{a_i}(t)\right) \\ &+ \gamma^p_{a_1}(t)\left(h_p(t) - \tilde{g}^p_{a_1}(t)\left(1 + D'_{a_1}\left(x_{a_1}(t)\right)\dot{x}_{a_1}(t)\right)\right) \\ &+ \sum_{i\in[2,m(p)]}\gamma^p_{a_i}(t)\left(g^p_{a_{i-1}}(t) - \tilde{g}^p_{a_i}(t)\left(1 + D'_{a_i}\left(x_{a_i}(t)\right)\dot{x}_{a_i}(t)\right)\right) \\ &+ \sum_{i\in[1,m(p)]}\eta_{a_i}(t)\left(x_{a_i}(t) - C_{a_i}(t)\right) - \rho^p_{a_0}(t)h_p(t) \\ &- \sum_{i\in[1,m(p)]}\rho^p_{a_i}(t)g^p_{a_i}(t) + \sum_{i\in[1,m(p)]}\zeta^p_{a_i}(t)x^p_{a_i}(t)\Big] \\ &+ \sum_{w\in W}\left(\mu_w(t)\left(\sum_{p\in P_w}h_p(t)\right)\right). \end{aligned} \tag{3.52}$$

The variation δZ^* of Z^* with respect to the control and state variables is given by

$$\begin{aligned} \delta Z^* &= \sum_{p\in P}\sum_{i\in[1,m(p)]}\left(-\lambda^p_{a_i}(T)\delta x^p_{a_i}(T)\right) + \sum_{w\in W}\left(-\mu_w(T)\delta E_w(T)\right) \\ &+ \sum_{w\in W}\phi_w(T)\left(-\delta E_w(T)\right) + \sum_{w\in W}\int_0^T\left(\frac{d\mu_w(t)}{dt}\delta E_w(t)\right)dt \\ &+ \sum_{p\in P}\int_0^T\left(\frac{\partial H(t)}{\partial h_p(t)}\delta h_p(t)\right)dt \\ &+ \sum_{p\in P}\sum_{i\in[1,m(p)]}\int_0^T\left(\frac{\partial H(t)}{\partial x^p_{a_i}(t)} + \frac{d\lambda^p_{a_i}(t)}{dt}\right)\delta x^p_{a_i}(t)dt \\ &+ \sum_{p\in P}\sum_{i\in[1,m(p)]}\int_0^T\left(\frac{\partial H(t)}{\partial g^p_{a_i}(t)}\delta g^p_{a_i}(t) + \frac{\partial H(t)}{\partial \tilde{g}^p_{a_i}(t)}\delta\tilde{g}^p_{a_i}(t)\right)dt. \end{aligned} \tag{3.53}$$

To simplify the notation and proceed to the analysis, let us further define $t^p_i = t^p_{i-1} + D_{a_i}\left(x^p_{a_i}(t^p_{i-1})\right)$, where t^p_{i-1} is the time of the flow $g^p_{a_{i-1}}$ entering link $a_i \in P$ and t^p_i is the time of the flow $g^p_{a_i}$ exiting the link. The partial derivatives of the Hamiltonian function are given as

$$\frac{\partial H(t)}{\partial h_p(t)} = \Psi_p(t,\mathbf{x}^*) + \lambda^p_{a_1}(t) + \mu_w(t) + \gamma^p_{a_1}(t) - \rho^p_{a_0}(t),\ \forall w\in W,\ p\in P_w, \tag{3.54}$$

$$\frac{\partial H(t)}{\partial g^p_{a_i}(t)} = -\lambda^p_{a_i}(t) + \lambda^p_{a_{i+1}}(t) - \rho^p_{a_i}(t) + \gamma^p_{a_{i+1}}(t),\ \forall p\in P, i\in[1,m(p)-1], \tag{3.55}$$

$$\frac{\partial H(t)}{\partial g^p_{a_{m(p)}}(t)} = -\lambda^p_{a_{m(p)}}(t),\ \forall p\in P, \tag{3.56}$$

$$\frac{\partial H(t)}{\partial \tilde{g}^p_{a_i}(t)} = -\gamma^p_{a_i}(t)\left(1 + D'_{a_i}(x_{a_i})\dot{x}_{a_i}\right),\ \forall p\in P,\ i\in[1,m(p)], \tag{3.57}$$

$$\frac{\partial H(t)}{\partial x^p_{a_i}(t)} = \zeta^p_{a_i}(t) + \eta_{a_i}(t)\sigma^p_{a_i},\ \forall p\in P,\ i\in[1,m(p)]. \tag{3.58}$$

In addition, a set of complementary slackness conditions for the nonnegative flow con-

straints and the side constraints can be obtained as:

$$h_p(t) \geq 0,\ \rho^p_{a_0}(t) \geq 0,\ \rho^p_{a_0}(t)h_p(t) = 0,\ \forall p \in P, \tag{3.59}$$

$$g^p_{a_i}(t) \geq 0,\ \rho^p_{a_i}(t) \geq 0,\ \rho^p_{a_i}(t)g^p_{a_i}(t) = 0,\ \forall p \in P,\ \forall i \in [1, m(p)], \tag{3.60}$$

$$x^p_{a_i}(t) \geq 0,\ \zeta^p_{a_i}(t) \geq 0,\ \zeta^p_{a_i}x^p_{a_i}(t) = 0,\ \forall p \in P,\ \forall i \in [1, m(p)], \tag{3.61}$$

$$\eta_{a_i}(t) \geq 0,\ x_{a_i}(t) - C_{a_i}(t) \leq 0,\ \eta_{a_i}(t)\left(x_{a_i}(t) - C_{a_i}(t)\right) = 0,\ \forall a_i \in A. \tag{3.62}$$

For an open path $p = \{a_1, a_2, \cdots, a_{i-1}, a_i, a_{i+1}, \cdots, a_{m(p)}\} \in P_w$, we have (Friesz et al., 2001)

$$\begin{aligned} g^p_{a_0}(t^p_0) &= h_p(t^p_0) > 0, \\ g^p_{a_i}(t^p_{i-1}) &> 0,\ \forall i \in [1, m(p)], \\ g^p_{a_i}(t^p_i) &> 0,\ \forall i \in [1, m(p)], \\ x^p_{a_i}(t) &> 0,\ \forall t \in [t^p_{i-1}, t^p_i],\ i \in [1, m(p)]. \end{aligned}$$

The Lagrange multipliers of the complementary slackness conditions (3.59)-(3.61) are relaxed to

$$\begin{aligned} \rho^p_{a_0}(t^p_0) &= 0, \\ \rho^p_{a_i}(t^p_{i-1}) &= \rho^p_{a_i}(t^p_i) = 0,\ \forall i \in [1, m(p)], \\ \zeta^p_{a_i}(t) &= 0,\ \forall t \in [t^p_{i-1}, t^p_i],\ i \in [1, m(p)]. \end{aligned}$$

With this in hand, we are ready to address the first order necessary condition for the optimal control problem. First, the following stationary conditions hold

$$\frac{\partial H}{\partial h_p}|_{t^p_0} = 0, \forall w \in W,\ p \in P_w, \tag{3.63}$$

$$\frac{\partial H}{\partial g^p_{a_i}}|_{t^p_i} + \left[\frac{\partial H}{\partial \tilde{g}^p_{a_i}} \frac{1}{1 + D'_{a_i}(x_{a_i})\dot{x}_{a_i}}\right]_{t^p_{i-1}} = 0, \forall p \in P,\ i \in [1, m(p)]. \tag{3.64}$$

The adjoint equations[2] are given by

$$\frac{d\lambda^p_{a_i}(t)}{dt} = -\frac{\partial H(t)}{\partial x^p_{a_i}(t)} = -\eta_{a_i}(t)\sigma^p_{a_i} - \zeta^p_{a_i}(t), \forall p \in P,\ i \in [1, m(p)], \tag{3.65}$$

$$\frac{d\mu_w(t)}{dt} = 0, \forall w \in W. \tag{3.66}$$

At terminal time T, the following boundary conditions hold:

$$\lambda^p_{a_i}(T) = 0,\ -\mu_w(T) - \phi_w(T) = 0,\ \forall w \in W,\ \forall p \in P_w,\ i \in [1, m(p)]. \tag{3.67}$$

[2] In Friesz et al. (2001); Friesz (2010), the adjoint equations are given as $\frac{d\lambda^p_{a_i}(t)}{dt} = -\frac{\partial H(t)}{\partial x^p_{a_i}(t)} + \frac{d}{dt}\frac{\partial H(t)}{\partial \dot{x}^p_{a_i}(t)}$. It is proven by Friesz (2010) that the term $\frac{d}{dt}\frac{\partial H(t)}{\partial \dot{x}^p_{a_i}(t)} = 0$.

Together with (3.66), the following equality is obtained:

$$\mu_w(t) = -\phi_w(T) = \mu_w, \forall w \in W. \tag{3.68}$$

Note that (3.63) is equal to

$$\Psi_p(t_0^p, \mathbf{x}^*) + \lambda_{a_1}^p(t_0^p) + \mu_w + \gamma_{a_1}^p(t_0^p) = 0, \forall w \in W,\ p \in P, \tag{3.69}$$

where we use the fact that $\mu_w(t) = \mu_w$, which is a constant and for an open path $\rho_{a_0}^p(t_0^p) = 0$. Now, by incorporating $\rho_{a_i}^p(t_i^p) = 0,\ \forall i \in [1, m(p)]$, the stationary condition (3.64) gives

$$\lambda_{a_{i+1}}^p(t_i^p) + \gamma_{a_{i+1}}^p(t_i^p) = \lambda_{a_i}^p(t_i^p) + \tilde{\gamma}_{a_i}^p(t_i^p), \forall p \in P, i \in [1, m(p)-1], \tag{3.70}$$

$$-\lambda_{a_{m(p)}}^p(t_{m(p)}^p) - \tilde{\gamma}_{a_{m(p)}}^p\left(t_{m(p)}^p\right) = 0,\ \forall p \in P, \tag{3.71}$$

where we adopt the following notation in line with Friesz et al. (2001); Friesz (2010):

$$\tilde{\gamma}_{a_i}^p(t_i^p) \doteq \gamma_{a_i}^p(t_{i-1}^p),\ \ \forall p \in P,\ \ i \in [1,\ m(p)]. \tag{3.72}$$

The adjoint equation (3.65) for an open path can then be given as

$$\frac{d\lambda_{a_i}^p(t)}{dt} = -\eta_{a_i}(t)\sigma_{a_i}^p,\ \ \forall p \in P,\ i \in [1, m(p)],\ t \in [t_{i-1}^p, t_i^p]. \tag{3.73}$$

Evaluating the adjoint equation (3.73) backward in time and space yields the following. To begin with, let us apply (3.73) to the last link of path p, i.e. link $a_{m(p)}^p$:

$$\lambda_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) - \lambda_{a_{m(p)}}^p\left(t_{m(p)}^p\right) = \int_{t_{m(p)-1}^p}^{t_{m(p)}^p} \eta_{a_{m(p)}}(t)\sigma_{a_{m(p)}}^p dt. \tag{3.74}$$

From (3.71) we have

$$\tilde{\gamma}_{a_{m(p)}}^p\left(t_{m(p)}^p\right) = \gamma_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) = -\lambda_{a_{m(p)}}^p(t_{m(p)}^p). \tag{3.75}$$

From (3.70), (3.73), and (3.75), the following relation is obtained

$$\begin{aligned} d_{m(p)-1} &= \lambda_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) + \gamma_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) = \lambda_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) - \lambda_{a_{m(p)}}^p\left(t_{m(p)}^p\right) \\ &= \int_{t_{m(p)-1}^p}^{t_{m(p)}^p} \eta_{a_{m(p)}}(t)\sigma_{a_{m(p)}}^p dt = \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-1}^p\right) + \tilde{\gamma}_{a_{m(p)-1}}^p\left(t_{m(p)-1}^p\right). \end{aligned} \tag{3.76}$$

Similarly,

$$\begin{aligned} d_{m(p)-2} &= \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-2}^p\right) + \gamma_{a_{m(p)-1}}^p\left(t_{m(p)-2}^p\right) \\ &= \lambda_{a_{m(p)-2}}^p\left(t_{m(p)-2}^p\right) + \tilde{\gamma}_{a_{m(p)-2}}^p\left(t_{m(p)-2}^p\right). \end{aligned} \tag{3.77}$$

From (3.77) we have

$$\begin{aligned} d_{m(p)-1} - d_{m(p)-2} &= \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-1}^p\right) - \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-2}^p\right) \\ &= -\int_{t_{m(p)-2}^p}^{t_{m(p)-1}^p} \eta_{a_{m(p)-1}}(t)\sigma_{a_{m(p)-1}}^p dt. \end{aligned} \tag{3.78}$$

By proceeding the induction, we have

$$d_i = d_{i+1} + \int_{t_i^p}^{t_{i+1}^p} \eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p dt, \ \forall\, i \in [1, m(p)-1]. \tag{3.79}$$

Finally, we have

$$d_1 = \lambda_{a_1}^p\left(t_1^p\right) + \gamma_{a_1}^p\left(t_0^p\right) = \sum_{i=1}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p dt. \tag{3.80}$$

Thus,

$$\begin{aligned} l_p(t_0^p) &= \lambda_{a_1}^p\left(t_0^p\right) - \lambda_{a_1}^p\left(t_1^p\right) + \lambda_{a_1}^p\left(t_1^p\right) + \gamma_{a_1}^p\left(t_0^p\right) = \int_{t_0^p}^{t_1^p} \eta_{a_1}(t)\sigma_{a_1}^p dt + d_1 \\ &= \sum_{i=0}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p dt. \end{aligned} \tag{3.81}$$

The dynamic user equilibrium condition with side constraints is as follows: for an open path $p \in P_w$, $h_p^*(t_0^p) > 0$,

$$\begin{aligned} &\Psi_p(t_0^p, \mathbf{x}^*) + l_p(t_0^p) = \phi_w(T), \\ &\text{and if} \\ &\Psi_p(t_0^p, \mathbf{x}^*) + l_p(t_0^p) > \phi_w(T),\ p \in P_w \Rightarrow h_p^*(t_0^p) = 0. \end{aligned} \tag{3.82}$$

As t_0^p is arbitrarily, without loss of generality, we represent (3.82) as

$$\begin{aligned} &\Psi_p(t, \mathbf{x}^*) + l_p(t) = \phi_w(T), \\ &\text{and if} \\ &\Psi_p(t, \mathbf{x}^*) + l_p(t) > \phi_w(T),\ p \in P_w \Rightarrow h_p^*(t) = 0, \end{aligned} \tag{3.83}$$

which is immediately recognized as the fundamental condition for the dynamic network user equilibrium described by (3.18)-(3.19). We also have the following complementary slackness conditions for the side constraints: (i) if $x_{a_i}(t) < C_{a_i}(t)$, the Lagrange multiplier $\eta_{a_i}(t) = 0$, (ii) if $x_{a_i}(t) = C_{a_i}(t)$, the Lagrange multiplier $\eta_{a_i}(t) \geq 0$. The value of $\phi_w(T)$ is determined by the total amount of traffic Q_w. As $\phi_w(T)$ is a constant, we drop the time script T to denote it as ϕ_w. $\Psi_p(t, \mathbf{x}^*)$ is interpreted as the effective path delay at time t on path p under travel condition $\mathbf{x}^*$. ■

Remark 3.1. To avoid the spillback effect of congestion, the side constraints $C_{a_i}(t),\ \forall a_i \in A$ are chosen as jam traffic volumes (i.e. storage capacities) of the links, e.g. the traffic volume K_a that corresponds to the jam density ρ_J in Figure 3.1. In this case, we have that $x_{a_i}(t) \leq K_{a_i},\ \forall a_i \in A$. Clearly, when $x_{a_i}(t) < K_{a_i},\ \forall a_i \in A,\ t \in [0, T]$, the problem

reduces to normal DUE traffic assignment. For the case when $x_{a_i}(t) = K_{a_i}$ for some links during several time intervals, we have the following interesting implication that

$$\dot{x}_{a_i} = \frac{d}{dt} x_{a_i} = \frac{d}{dt} K_{a_i} = 0.$$

Under this circumstance, the flow propagation becomes

$$g^p_{a_i}\left(t + D_{a_i}(x_{a_i}(t))\right) - g^p_{a_{i-1}}(t) = 0,$$

which implies that traffic behaves like an incompressible fluid when its density (or volume) approaches certain maximum, i.e. the flow is incompressible when traffic density (or volume) is equal to the jam density (or volume). This phenomenon is consistent with the findings in Ross (1988) where the LWR model was applied to describe the traffic dynamics. ∎

3.3.6 Interpretation of DUE with side constraints

For a path p with $h_p(t) > 0$, we have $\rho^p_{a_0}(t) = 0$. Furthermore, if there is no side constraint, i.e. $x_{a_i}(t) < C_{a_i}(t)$, $\forall a_i \in A$, $\forall t \in [0, T]$, we have $\eta_{a_i}(t) = 0$. The adjoint equation (3.73), implies that, $\frac{d\lambda^p_{a_i}(t)}{dt} = 0$, i.e. $\lambda^p_{a_i}(t)$ is a constant with respect to time. This, in conjunction with the boundary condition $\lambda^p_{a_i}(T) = 0$, implies that $\Psi_p(t, \mathbf{x}^*) = \phi_w(T)$, which is consistent with the normal DUE condition (Friesz et al., 2001).

We can also show that the DUE-SC can reduce to the SC-TAP. The system is operating in a steady state in the static case, i.e. $h_p(t) = h_p$, $\frac{dx_{a_i}(t)}{dt} = 0 \Rightarrow x_{a_i}(t) = x_{a_i}$, $\forall a_i \in A$, and all of the side constraints are constant. By defining the performance index (3.37) to be (3.1), the necessary condition for the DUE-SC reduce to the necessary condition for the SC-TAP. From Lemma 3.1, the SC-TAP can be written as

$$f_p \begin{cases} > 0 \Rightarrow c_p = \mu_w, \\ = 0 \Rightarrow c_p \geq \mu_w, \end{cases} \quad \forall p \in P_w,\ w \in W. \tag{3.84}$$

The necessary conditions for the DUE-SC and the SC-TAP are similar. The term

$$\int_{t^p_{i-1}}^{t^p_i} \eta_{a_i}(u)\sigma^p_{a_i} du,$$

can be regarded as an additional time penalty, or toll (access price) to be imposed on travelers during their presence on the controlled link a_i along their travel path p, which is similar to the interpretation of the term λ_{a_i} in the SC-TAP. Note that the additional time penalty λ_{a_i} in the static case is constant over time, whereas the additional time penalty in the DUE-SC is time-varying and a cumulative effect of the associated Lagrange multiplier over time.

3.4 Solution algorithm

3.4.1 The fixed point problem and its optimal control formulation

As Friesz et al. (2001) state the optimal control problem formulation cannot be used for the computation of the DUE/DUE-SC because its articulation presumes the knowledge of departure rates $h_p^*(t), \forall p \in P$, under the DUE/DUE-SC condition. The optimal control problem formulation is a mathematical convenience for analyzing the necessary conditions of the DUE/DUE-SC. There are several solution algorithms for the DUE without departure time choice, such as a quasi-VI based approach (Ban et al., 2008) or an equivalent gap function-based algorithm (Lu et al., 2009). However, only a few solution algorithms, such as those proposed by Huang and Lam (2002), Friesz and Mookherjee (2006) and Friesz et al. (2011), have been proposed for the DUE with simultaneous departure-time-and-path-choice. The DUE with simultaneous departure-time-and-path-choice has been shown to be equivalent to a fixed point problem. As explained in the previous section, the side constraints can be viewed as a restriction on the feasible region of the VI formulation of the DUE-SC, i.e. (3.32) and (3.33). Based on this VI formulation, the DUE-SC is equivalent to the following fixed point problem:

Lemma 3.3. (Huang and Lam, 2002) (Fixed point problem) When the conditions in Section 3.3.4, which guarantee the existence of the DUE-SC, hold, any solution of the fixed point problem

$$\mathbf{h} = \mathbf{P}_{\bar{\Lambda}} \left(\mathbf{h} - \alpha \Psi \left(\mathbf{h} \right) \right), \tag{3.85}$$

is also a solution of the DUE-SC, where $\mathbf{P}_{\bar{\Lambda}}(\mathbf{x}) = \text{argmin}_{\mathbf{z} \in \bar{\Lambda}} \|\mathbf{x} - \mathbf{z}\|$, i.e. the minimum norm projection onto $\bar{\Lambda}$ and $\alpha \in R_+^1$. ∎

In line with Friesz and Mookherjee (2006), the fixed point problem (3.85) requires that

$$\mathbf{h} = \arg\min_{\mathbf{v}} \left\{ \frac{1}{2} \left\| \mathbf{h} - \alpha \Psi \left(\mathbf{h} \right) - \mathbf{v} \right\|^2 : \mathbf{v} \in \bar{\Lambda} \right\}, \tag{3.86}$$

which is equivalent to seeking the solution of the following optimal control (infinite dimensional mathematical programming) problem

$$\min_{\mathbf{v} \in \bar{\Lambda}} J(\mathbf{v}) = \frac{1}{2} \int_0^T \left[\mathbf{h} - \alpha \Psi \left(\mathbf{h} \right) - \mathbf{v} \right]^T \left[\mathbf{h} - \alpha \Psi \left(\mathbf{h} \right) - \mathbf{v} \right] dt. \tag{3.87}$$

The advantage of the VI formulation of the DUE/DUE-SC is that it includes almost all DUE models regardless of the link traffic dynamics, flow propagation, and link travel time function employed. However, in this formulation, the effective path delay operators

$\Psi_p(t, \mathbf{h}),\ \forall p \in P$, generally cannot be expressed in a closed form whereas $\Psi_p(t, \mathbf{x}),\ \forall p \in P$ are well defined in a closed form by employing the link traffic dynamics (3.38)-(3.39), flow propagation constraints (3.40)-(3.41), and certain travel time function. By Lemma 3.3 and (3.87), the fixed point algorithm can be formulated and solved as an iterative optimal control problem as depicted in Figure 3.2. In the figure, the effective path delay operators $\Psi_p(t, \mathbf{x}),\ \forall p \in P$, in conjunction with (3.38)-(3.41) and (3.28) are applied to replace $\Psi(\mathbf{h})$ in (3.87), and to express the side constraints and the feasible region $\bar{\Lambda}$ explicitly rather than the abstract embedded formulation used in Section 3.3.4. In Figure 3.2, $\dot{\mathbf{x}} = \tilde{f}(\cdot)$ denotes the link traffic dynamics with the flow propagation constraints (3.38)-(3.41). We write $\dot{\mathbf{x}}$ explicitly because the link traffic volumes are the state variables of the optimal control problem. The flow propagation constraints (3.40)-(3.41) with travel time function defined by (3.28) are substituted into (3.38)-(3.39) to update the state variables. The equality constraint $M_1(\cdot)$ denotes the flow conservation equations (3.42) and (3.46) with zero initial condition (or (3.16)). The inequality constraint $M_2(\cdot)$ denotes the side constraints defined by (3.47). $\mathbf{v} \geq 0$ is a vector representation of (3.43). Detailed analysis of the convergence of such a fixed point algorithm is given by Friesz and Mookherjee (2006) for an abstract feasible region with certain properties. However, some conditions required to guarantee the convergence of the algorithm, such as the strong monotone of $F\left(\mathbf{x}^k, \mathbf{h}^k, \mathbf{v}\right)$, are unlikely to be verified for general traffic networks. Thus, the convergence of this fixed point algorithm is generally heuristic.

As Figure 3.2 indicates, the optimal control subproblem is what we need to solve. To the best of our knowledge, no effective algorithm has been reported to solve the optimal control subproblem with state dependent time lags and state constraints. However, several effective algorithms have been proposed to solve the optimal control problem with state and/or control constraints using nonlinear programming algorithms, such as sequential quadratic programming (SQP) (Buskens and Maurer, 2000). To solve this optimal control problem by nonlinear programming algorithms, it is necessary to approximate the functional differential equations (FDEs), which govern the traffic dynamics, by ordinary differential equations (ODEs). We use a typical method for handling transportation delays in control engineering to deal with these state dependent time lags. The idea is to use polynomial approximation of the time lags. A similar idea has been proposed by Astarita (1996) and Friesz and Mookherjee (2006). We use Padé approximation to approximate the state dependent time lags, which can be easily implemented using Matlab and Simulink.

After approximating the FDEs by ODEs, we denote the traffic dynamics as $\dot{\mathbf{x}}(t) =$

Algorithm Fixed point algorithm for the DUE-SC

1: Initialization. Identify an initial feasible solution $\mathbf{h}^1 \in \bar{\Lambda}$, $\alpha \in R_+$ and set $k = 1$

2: $k \leftarrow 2$

3: **while** $k \leq Iter_{max}$ and $\|\mathbf{h}^{k+1} - \mathbf{h}^k\| > \varepsilon$, **do**

4: Optimal control subproblem. Solve

$$\begin{aligned} \min_{\mathbf{v}} J^k(\mathbf{v}) &= \frac{1}{2}\int_0^T \left[\mathbf{h}^k - \alpha\Psi\left(\mathbf{x}(\mathbf{h}^k)\right) - \mathbf{v}\right]^T \left[\mathbf{h}^k - \alpha\Psi\left(\mathbf{x}(\mathbf{h}^k)\right) - \mathbf{v}\right] dt \\ &\triangleq \int_0^T F\left(\mathbf{x}^k, \mathbf{h}^k, \mathbf{v}\right) dt, \end{aligned}$$

subject to

$$\dot{\mathbf{x}} = \tilde{f}(\mathbf{x}, \mathbf{v}, t),\ M_1(\mathbf{v}, t) = 0,\ M_2(\mathbf{x}, t) \leq 0,\ -\mathbf{v} \leq 0, \forall t \in [0, T],\ \mathbf{x}(0) = 0$$

5: $k \leftarrow k + 1$, $\mathbf{h}^{k+1} = \mathbf{v}^*$

6: **end while**

7: **return** $\mathbf{h}^* \approx \mathbf{h}^{k+1}$

where $Iter_{max}$ is the maximum iteration number, and $\varepsilon \in R_+$ is the tolerance.

Figure 3.2: Fixed point algorithm for DUE-SC

$f_0(\mathbf{x}, \mathbf{v}, t)$. It is convenient for us to rewrite the optimal control subproblem as

$$\min_{\mathbf{v}} J^k(\mathbf{v}) = \int_0^T F\left(\mathbf{x}^k, \mathbf{h}^k, \mathbf{v}\right) dt, \tag{3.88}$$

subject to

$$\dot{\mathbf{x}}(t) = f_0(\mathbf{x}, \mathbf{v}, t),\ M_1(\mathbf{v}, t) = 0,\ M_2(\mathbf{x}, t) \leq 0,\ -\mathbf{v} \leq 0,\ \forall t \in [0, T], \mathbf{x}(0) = 0. \tag{3.89}$$

3.4.2 Solution algorithm for the DUE-SC

To apply an off-the-shelf nonlinear optimization algorithm to the optimal control problem (3.88)-(3.89), it is necessary to apply the time-discretization scheme to the problem. Here we present the recursive discretization approach based on Euler's method, which is commonly used in the literature (Buskens and Maurer, 2000). The planning time interval is divided into $N - 1$ segments uniformly, i.e. a fixed time step Δt is defined in the discretization as $\Delta t = \frac{T}{N-1}$, $t_l = (l-1)\Delta t$, $l = 1, 2, \cdots, N$. Applying Euler's method to the differential equation in (3.89) yields

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \Delta t \cdot f_0(\mathbf{x}_l, \mathbf{v}_l, t_l),\ \ l = 1, 2, \cdots, N-1. \tag{3.90}$$

Define the optimization variable $\mathbf{z} = [\mathbf{v}_1, \cdots, \mathbf{v}_N]^T$, and compute the state variables from (3.90) recursively as $\mathbf{x}_l = \mathbf{x}_l(\mathbf{z}, t) = \mathbf{x}_l(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{l-1}, t_{l-1})$, $l = 2, \cdots, N$, which are functions of the control variables with initial condition $\mathbf{x}_1 = \mathbf{x}(1) = 0$. The following NLP is defined

$$\min_{\mathbf{z}} y(\mathbf{z}) \triangleq \sum_{l=1}^{N-1} \Delta t \cdot F\left(\mathbf{x}^k, \mathbf{h}^k, \mathbf{v}_l\right), \tag{3.91}$$

subject to

$$\begin{aligned} &\mathbf{x}_l = \mathbf{x}_l(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{l-1}, t_{l-1}),\ l = 2, \cdots, N,\ \mathbf{x}_1 = 0, \\ &M_1(\mathbf{v}_l, t_l) = 0,\ M_2(\mathbf{x}_l, t_l) \leq 0,\ -\mathbf{v}_l \leq 0, l = 1, 2, \cdots, N. \end{aligned} \tag{3.92}$$

After the the optimal control subproblem is reformulated as an NLP problem, we can apply nonlinear optimization algorithms to solve it. The overall solution algorithm is summarized in Figure 3.3.

Step 1. Approximation of PDEs: Apply polynomial approximants (such as the Padé approximant) to approximate the state dependent time delays, which converts the original FDEs into ODEs.

Step 2. NLP formulation of the optimal control subproblem: Formulate a nonlinear programming problem of the optimal control problem (3.88)-(3.89) using the Euler's discretization approach.

Step 3. Initialization: Set initial feasible solutions of state and inflow $\mathbf{x}^1, \mathbf{h}^1 \in \bar{\Lambda}$, iteration counter $k = 1$.

Step 4. Solution of the optimal control subproblem: Apply any off-the-shelf nonlinear optimization algorithm to the equivalent NLP problem (e.g. SQP or GAUSS Pseudo Spectral Method).

Step 5. Check the convergence of the optimal control subproblem: If the preset tolerance ε_1 or the maximum iteration number $Iter_m$ of the NLP is achieved, declare the solution to be $\mathbf{h}^{k+1}$ and go to Step 6. Otherwise declare the solution to be the initial condition of the NLP and go to Step 4.

Step 6. Stopping test: If $\|\mathbf{h}^{k+1} - \mathbf{h}^k\| \leq \varepsilon$, where $\varepsilon \in R_+$ is a preset tolerance, stop and declare $\mathbf{h}^* \approx \mathbf{h}^{k+1}$. Otherwise, set $k = k + 1$ and go to Step 3.

Algorithm Solution algorithm for the DUE-SC

1: Approximation of PDEs

2: NLP formulation of the optimal control subproblem

3: Initialization: Set initial feasible solution $\mathbf{x}^1, \mathbf{h}^1 \in \bar{\Lambda}$, $\alpha \in R_+$ and $k = 1$

4: $k \leftarrow 2$

5: **while** $k \leq Iter_{max}$ and $\|\mathbf{h}^{k+1} - \mathbf{h}^k\| > \varepsilon$, **do**

6: Optimal control subproblem. Solve the following NLP:

6.1: Initialization of the NLP: Set $\mathbf{z}^{k,1} \in \bar{\Lambda}$, and $j = 1$

6.2: $j \leftarrow 2$

6.3: **while** $j \leq Iter_m$ and $\|\mathbf{z}^{k,j+1} - \mathbf{z}^{k,j}\| > \varepsilon_1$, **do**

$$\min_{\mathbf{z}} y(\mathbf{z}^{k,j}),$$

subject to

$$\mathbf{x}_l^{k,j} = \mathbf{x}_l^{k,j}\left(\mathbf{v}_1^{k,j}, \mathbf{v}_2^{k,j}, \cdots, \mathbf{v}_{l-1}^{k,j}, t_{l-1}\right), l = 2, \cdots, N,\ \mathbf{x}_1^{k,j} = 0,$$

$$M_1(\mathbf{v}_l^{k,j}, t_l) = 0,\ M_2(\mathbf{x}_l^{k,j}, t_l) \leq 0,\ -\mathbf{v}_l^{k,j} \leq 0,\ l = 1, 2, \cdots, N$$

6.4: $j \leftarrow j + 1$

6.5: **end while**

6.6: **declare** $\mathbf{z}^{k,*} \approx \mathbf{z}^{k,j+1}$

7: $k \leftarrow k + 1,\ \mathbf{h}^{k+1} \leftarrow \mathbf{z}^{k,*}$

8: **end while**

9: **return** $\mathbf{h}^* \approx \mathbf{h}^{k+1}$

where $Iter_m$ is the maximum iteration number for the NLP, and $\varepsilon_1 \in R_+$ is the tolerance for the NLP.

Figure 3.3: Solution algorithm for DUE-SC

3.5 Numerical examples

We have imposed upper bounds on path flows so as to prove the existence of equilibriums to DUE and the DUE-SC. In this section, the numerical examples are solved with the upper bounds on path flows set to $+\infty$.

3.5.1 A simple network case

Consider a network with a single OD pair connected by two parallel links as shown in Figure 3.4. The link delay functions are given by $D(x_1) = 0.27 + x_1/70$ unit-times and $D(x_2) = 0.3 + x_2/140$ unit-times, respectively. The overall travel demand is $J_{od} = 75$ units.

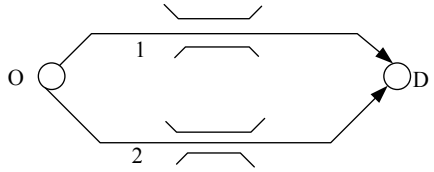


Figure 3.4: Network connected with parallel links

The planning time horizon is $T = 5$ unit-times. The desired arrival time is set as $t_{da} = 3$ unit-time. We consider the following symmetric early/late arrival penalty function:

$$\kappa[\chi] = \begin{cases} 0.1\left(t + D(x) - t_{de}\right)^2, & t < t_{de}, \\ 0, & t_{de} \le t \le t_{dl}, \\ 0.1\left(t + D(x) - t_{dl}\right)^2, & t > t_{dl}, \end{cases} \tag{3.93}$$

where $t_{de} = 2$ unit-time and $t_{dl} = 4$ unit-time. The effective delay is given by (3.93). Without loss of generality, we assume zero initial conditions. $\alpha = 0.01$ is chosen in the fixed-point algorithm as represented in Figure 3.2, and the fixed point stopping criteria is $\varepsilon = 0.0001$.

We first solve the DUE with simultaneous route/departure time choices of this network. Figure 3.5 shows that at the very beginning of the departure stage, the travel cost of link 1 is less than that of link 2 and not greater than the minimum travel cost. Thus, travelers all choose to use link 1. After some time, the costs of the two links tend to the same level, i.e., the DUE cost. Travelers then select their departure times to maintain the equilibrium state. Note that the inflows become zero whenever the travel times on the links are greater than the minimum travel time, which is consistent with the DUE condition. Figure 3.5b demonstrates the link traffic volumes of both links under the DUE condition.

Next, we restrict the traffic volume on link 1 to be less than or equals to 3 units during the whole planning horizon, i.e., $x_1 \le 3$. By applying the DUE-SC condition, we obtain the results that are depicted in Figure 3.6. When there is no traffic volume control, the traffic volume on link 1 exceeds the desired saturation, i.e., $x_1 = 3$, after one unit-time as demonstrated in Figure 3.5b. Figure 3.6b illustrates that the control parameter (or additional travel cost induced by the side constraint) for link 1 becomes positive at around one unit-time, which corresponds with the time that the side constraint is first violated. Note that the control parameter or additional travel cost imposed on link 1 is calculated following Equation (3.81). Figure 3.6a shows the travelers' responses to the additional travel cost imposed on link 1. The departure rate of link 1 suddenly decreases at around one unit-time (when the additional travel cost is first imposed) and then remains at the saturation level of around 9 units/time-steps (which is lower than the saturation level

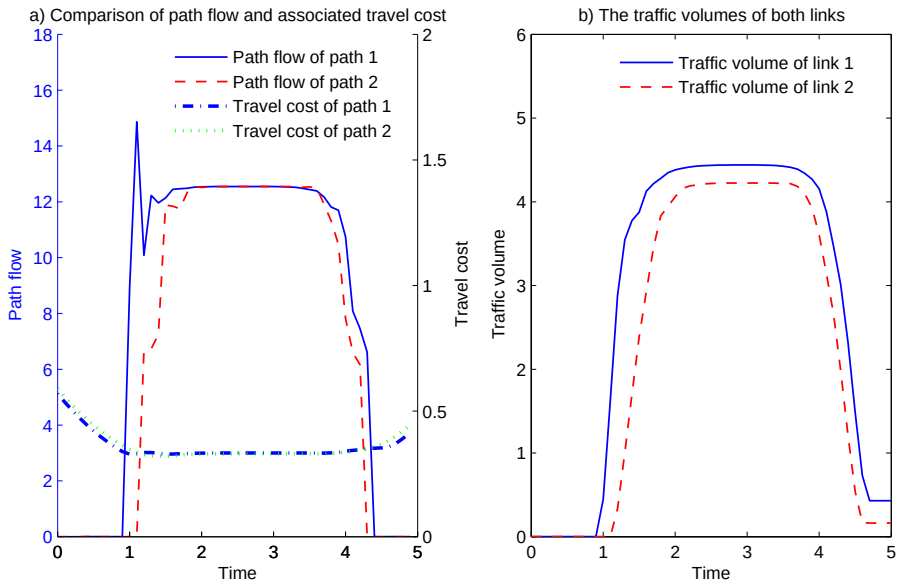


Figure 3.5: Inflow profiles, link traffic volumes and travel costs of both links under the DUE with symmetric penalty

of around 12 of the uncontrolled case as shown in Figure 3.5a). Figure 3.6a also shows an increase of the departure rate of link 2 as compared to the uncontrolled case (from the maximum flow rate of 12 to more than 14 units). This illustrates the diversion of traffic from link 1 to 2 to maintain the traffic volume of link 1 and the DUE-SC condition. We can also observe a slightly larger travel time window for those traveling by link 2 as compared to the uncontrolled case, which highlights the shift of travelers' departure times to maintain the DUE-SC condition. Figure 3.6b shows that the traffic volume of link 1 under DUE-SC condition satisfies the side constraint. The value of additional travel cost induced by the side constraint varies from 2.5 percent to 8.1 percent of the minimum effective travel time. Compared to the DUE case, the number of travelers using link 1 declines by about 23 percent during the planning horizon under the DUE-SC condition. The constant traffic volume control scheme in this example will yield a steady state to the departure rate of link 1 if the departure time window is long enough. As demonstrated in Figure 3.6a, the departure rate of link 1 gradually approaches a steady state controlled by the traffic volume control scheme. As we do not impose a side constraint on link 2, the additional travel cost of link 2 is zero throughout the planning horizon.

For a congested roadway, different traffic volume control schemes can be implemented to alleviate the congestion. For instance, the control may vary by time, or different restrictions on traffic volumes may be adopted during different peak hours. In the second test, we impose the traffic volume control function $x_1(t) \leq 3 \sin\left(\frac{(t+20)\pi}{100}\right)$ on link 1. This

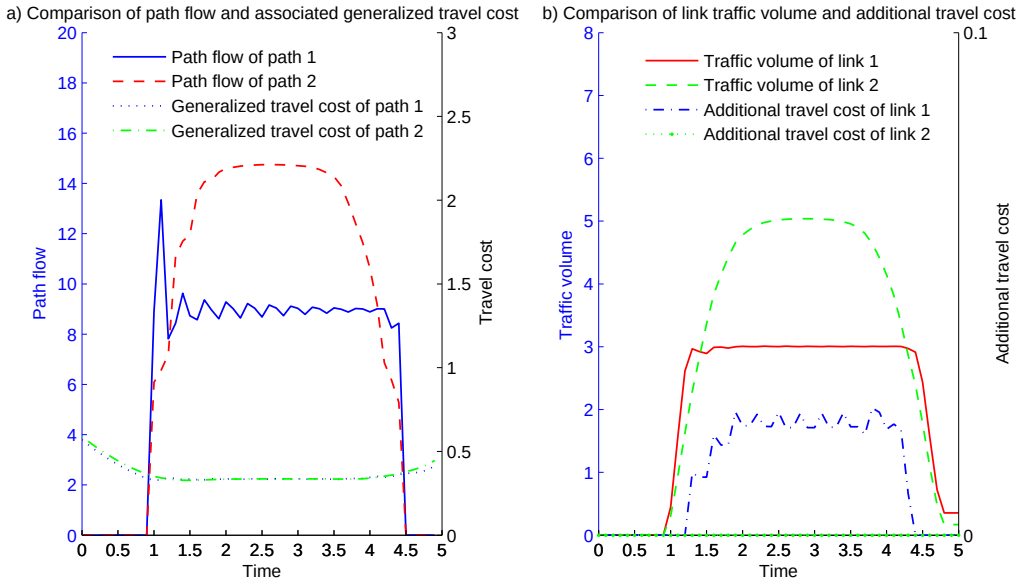


Figure 3.6: Inflow profiles, link traffic volumes and travel costs of both links under constant traffic volume control

constraint represents a kind of control that adjusts its amplitude responding to the peak hour traffic volume. In this example, the control has a small value at the beginning, and gradually increases as the time approaches the peak hour. After the peak period, the control then decreases. By applying the DUE-SC condition, we obtain the results depicted in Figure 3.7. Compared with the previous constant traffic volume control scheme, the travelers using link 1 are indirectly controlled to depart earlier to satisfy the new traffic volume control scheme. The inflow of link 2 does not change significantly. The inflow of the controlled link is shaped by the traffic volume control with some oscillations. This verifies that the control adjusts the departure rate of the controlled link to maintain its traffic volume and satisfy the time-varying traffic volume control scheme.

In the third test, a traffic volume control scheme with two step functions is tested. In our continuous time formulation, the traffic volume controls (or side constraints) are assumed to be continuously differentiable functions. This kind of discontinuous dynamic traffic volume control cannot be tested under the continuous time DUE-SC framework. However, due to the discretization scheme adopted in our solution algorithm, we can construct the following traffic volume control scheme, i.e.,

$$x_1(k) \leq 2, \forall k = 1, \cdots, 19, \text{ and } x_1(k) \leq 3, \forall k = 20, \cdots, 50. \tag{3.94}$$

The simulation results are plotted in Figure 3.8. By comparing Figure 3.8 and Figure 3.6, we can see that, some travelers who originally planned to use link 1 (see Figure 3.6) are

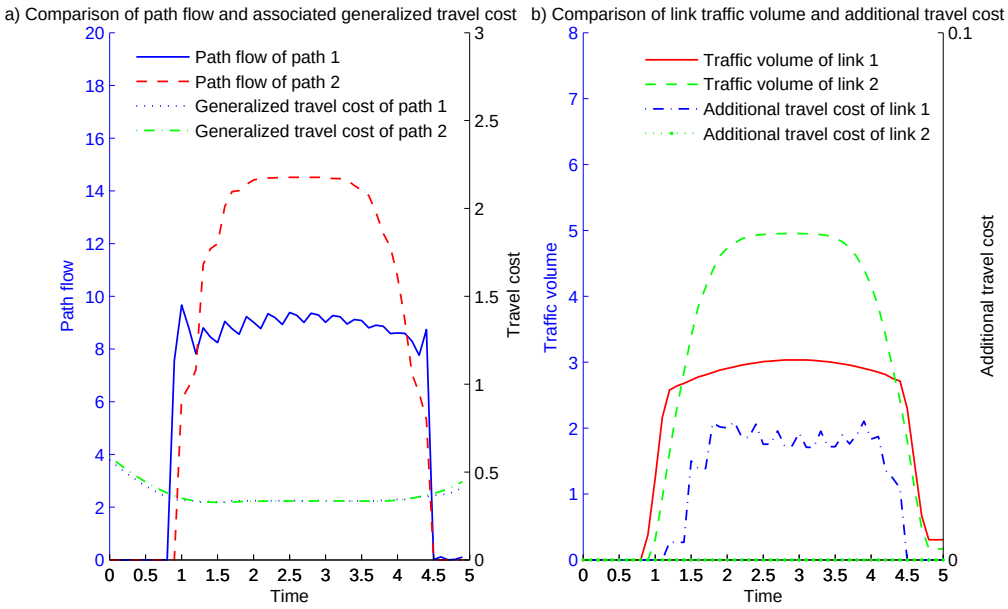


Figure 3.7: Inflow profiles, link traffic volumes and travel costs of both links under time-varying traffic volume control

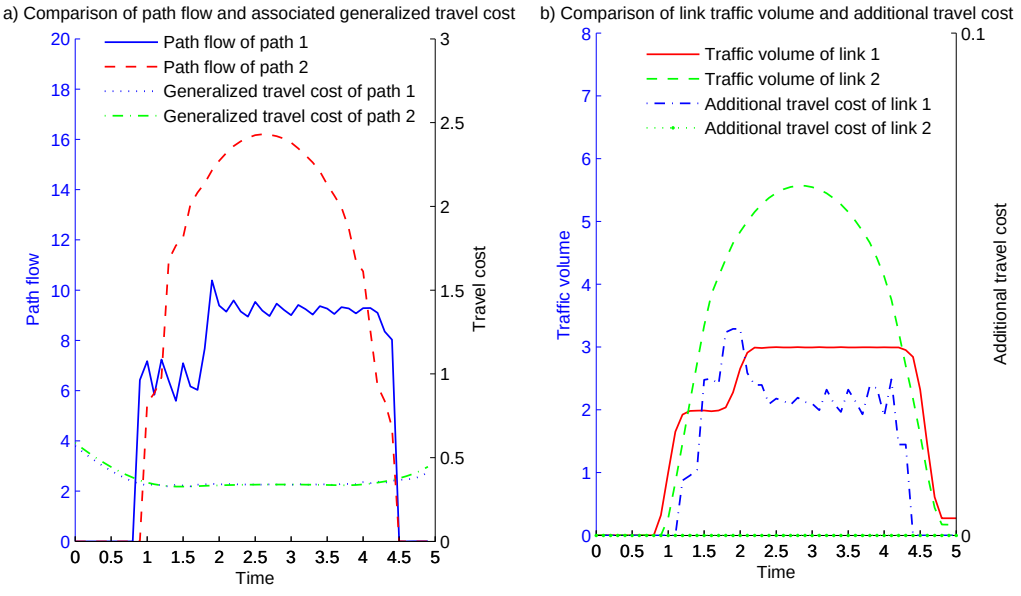


Figure 3.8: Inflow profiles, link traffic volumes and travel costs of both links under "step" traffic volume control

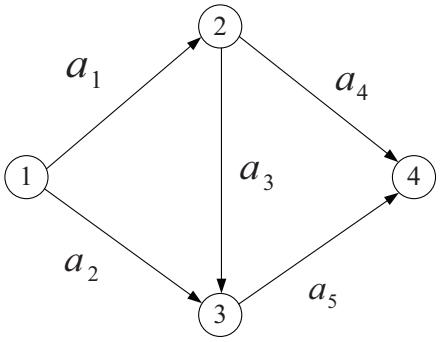


Figure 3.9: The Braess' network

now diverted to link 2. Some of the travelers on link 1 also depart earlier to avoid the additional travel cost. These effects arise because more stringent traffic volume control is imposed on link 1 during the first stage. The additional travel cost will be higher if the travelers persist in traveling during the same time period and on the same link. In the second stage, the traffic volume control forces the departure rate of link 1 to the steady state, which is determined by the side constraint. As demonstrated in Figure 3.8, the control forces the traffic volume on link 1 to meet the restriction in a smooth manner rather than a sudden switch, which is consistent with our assumption of continuity in the traffic volume control. The additional travel cost increases during the switch of these two traffic volume control strategies to depress the travel demand attracted by the relaxation of the side constraint. After the switch of the traffic volume control strategies the cost decreases, which in turn increases the travel demand on link 1. The control is then changed to adjust the departure rate of the controlled link so that the side constraint is satisfied.

3.5.2 The Braess' Network

Next, we will test the Braess' network depicted in Figure 3.9. The forward star array and link travel time functions of this network are summarized in Table 3.1. The planning horizon is $T = 6$ unit-times. The time incremental step is 0.1 unit-times. The desired arrival time is $t_{da} = 3$ unit-time. The overall travel demand is $J_{14} = 200$ units. The symmetric early/late arrival penalty function (3.93) is applied in this example with $t_{de} = 2$ unit-time and $t_{dl} = 5$ unit-time. We define the path set as $P_{14} \triangleq \{p_1,\ p_2,\ p_3\}$ with $p_1 \triangleq \{a_1,\ a_4\}$, $p_2 \triangleq \{a_1,\ a_3,\ a_5\}$ and $p_3 \triangleq \{a_2,\ a_5\}$. Without loss of generality, we assume zero initial conditions. We choose $\alpha = 0.01$ for the fixed point algorithm. The fixed point stopping criteria is $\varepsilon \in [0.00001, 0.0001]$.

First, the normal DUE with simultaneous route/departure time choices is solved. Figure 3.10 illustrates the path departure rates and the path travel costs under the DUE

Link name	From node	To node	Link delay function, $D_a(x_a(t))$
a_1	1	2	$0.5+\frac{x_{a_1}}{2000}$
a_2	1	3	$0.71+\frac{x_{a_2}}{2000}$
a_3	2	3	$0.18+\frac{x_{a_3}}{100}$
a_4	2	4	$0.71+\frac{x_{a_4}}{2000}$
a_5	3	4	$0.5+\frac{x_{a_5}}{2000}$

Table 3.1: Link configurations

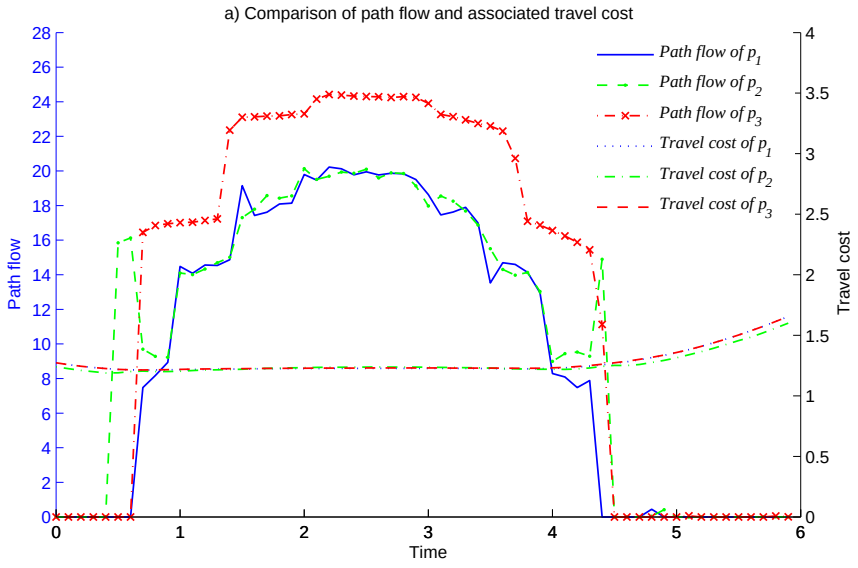


Figure 3.10: Path departure rates and travel costs of the Braess' network under the DUE condition

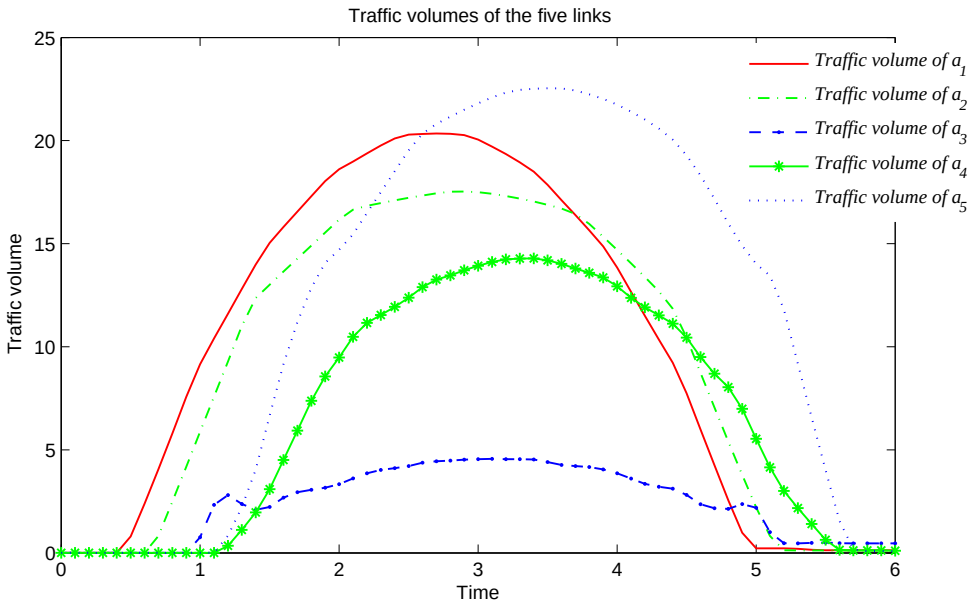


Figure 3.11: Traffic volumes of the five links under the DUE condition

condition for comparison. From Figure 3.10, at the very beginning none of the path is used because the travel costs are higher than the minimum travel cost. Path p_2 becomes active first due to its relatively lower free-flow travel cost. After a short time, the costs of the other two paths tend to the DUE cost. All three paths then become active, and these departure rates are adjusted to maintain the DUE condition. The path departure rates become zero whenever the travel costs of the paths are greater than the minimum travel cost, which is consistent with the DUE condition. Figure 3.11 demonstrates the link traffic volumes of the five links under the DUE condition.

Next, the traffic volume on link a_3 are controlled because it has a relatively smaller capacity. In this example, we impose the side constraint $x_{a_3} \leq 2$ on link a_3. By solving the DUE-SC, we obtain the results depicted in Figure 3.12. Compared with the previous DUE test, we note that the departure rate of the controlled path, i.e., path p_2, decreases throughout the departure time window. As demonstrated in Figure 3.12, this amount of traffic is mainly diverted to path p_1, whereas the departure rate of path p_3 does not change significantly. Travelers on all of these paths depart earlier to avoid the additional travel cost. Figure 3.13 illustrates that the additional travel cost for link a_3[3] becomes positive at around one unit-time, which corresponds to the time that the side constraint is first

[3]Note from (3.47) and the proof of Proposition 3.3 that the side constraints and the associated link additional travel costs (i.e., integrals of Lagrange multipliers associated with the side constraints over time) are defined for the links on the network, i.e. they are link based. They are then converted to path additional travel costs by Equation (3.81).

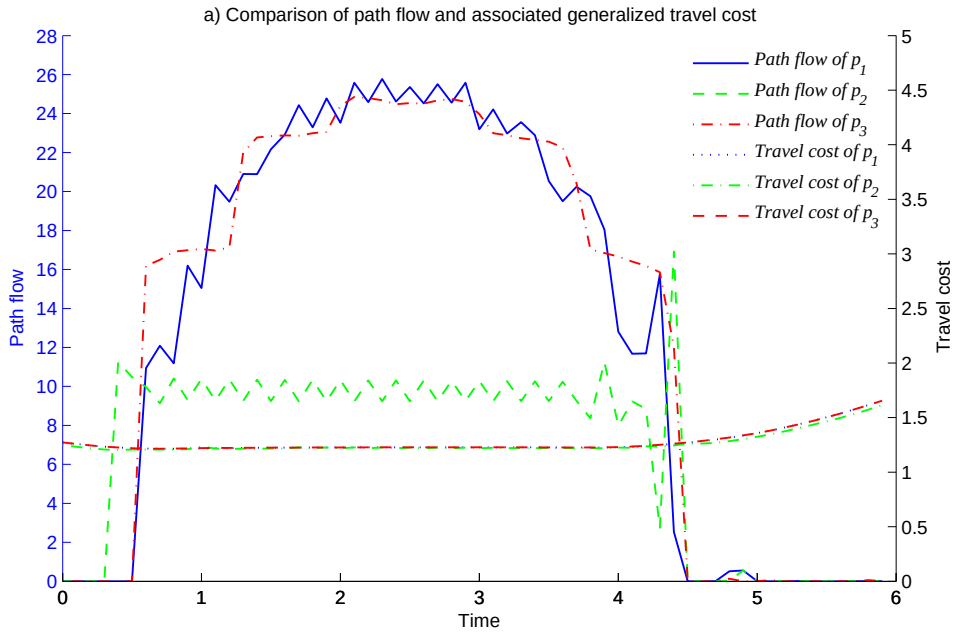


Figure 3.12: Path departure rates and travel costs of the Braess' network under constant traffic volume control

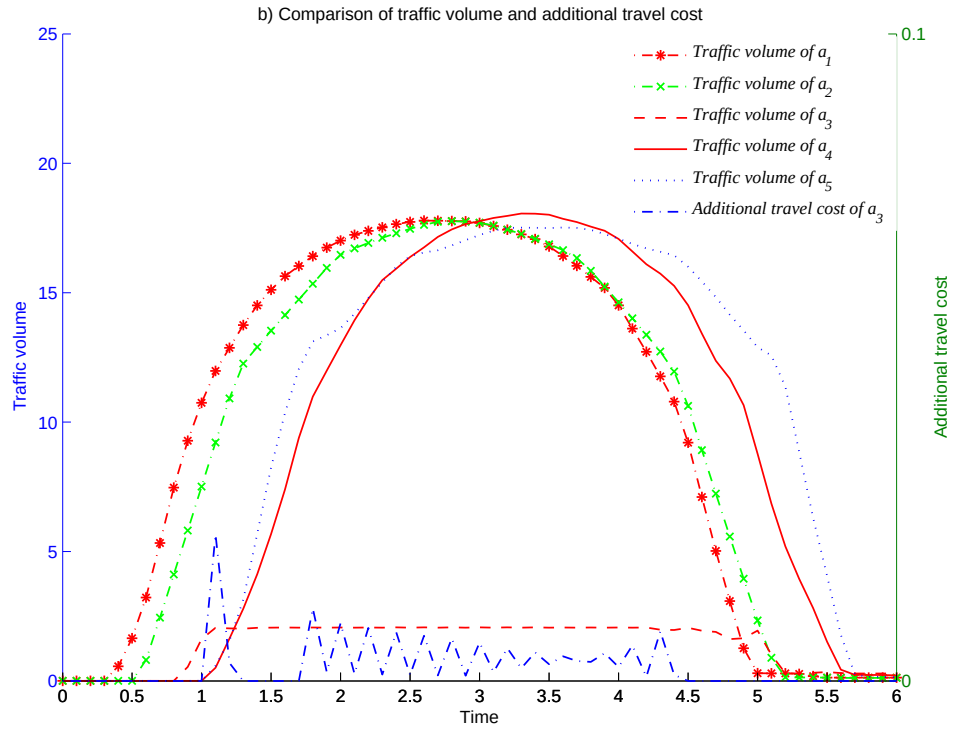


Figure 3.13: Traffic volumes of the five links and the additional travel cost under constant traffic volume control

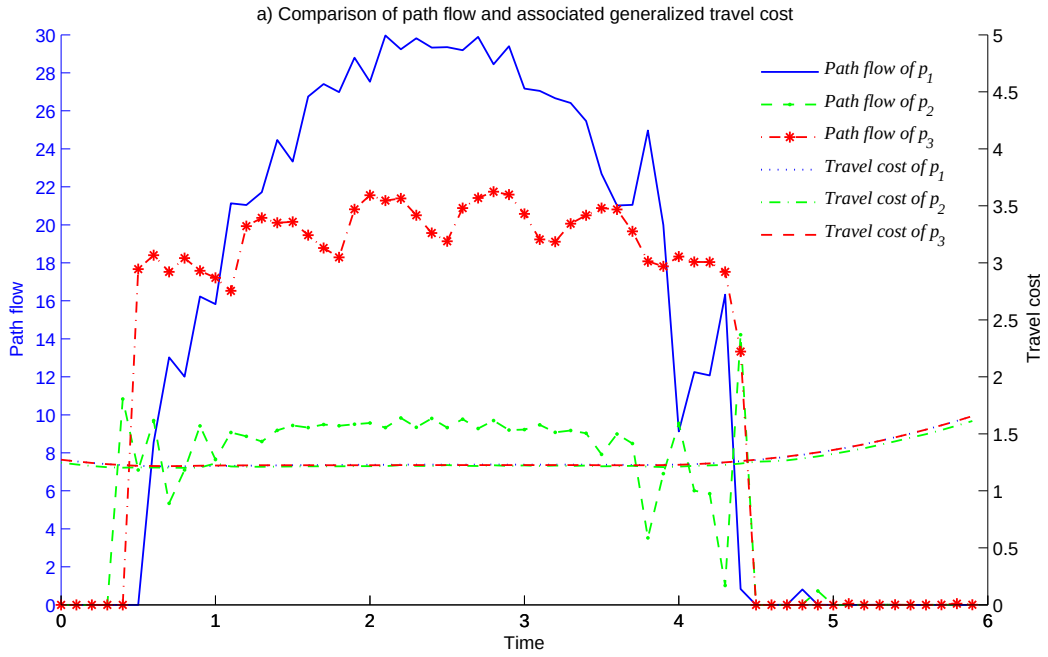


Figure 3.14: Path departure rates and travel costs of the Braess' network under time-varying traffic volume control

violated. The departure rate of path p_2 decreases sharply at the same time, compared with the DUE case in Figure 3.10. This verifies that the control (i.e., the additional travel cost) adjusts the departure rate to maintain the traffic volume of link a_3 to satisfy the side constraint. The amount of time-varying additional travel cost caused by the side constraint varies from 0.5 percent to 12.5 percent of the travel time of link a_3. The path flow of p_2 under the DUE-SC condition is about 64 percent of that under the DUE condition. As in the previous example, the departure rate of path p_2 is shaped by the the side constraint with some oscillations.

Finally, we test the DUE-SC for this network with time-varying side constraints on different links. To be more specific, we impose the following side constraints on links a_2 and a_3:

$$x_{a_2}(t) \leq 15 \sin\left(\frac{(t+20)\pi}{100}\right) \text{ units}, \; x_{a_3}(t) \leq 2 \sin\left(\frac{(t+20)\pi}{100}\right) \text{ units}. \tag{3.95}$$

Figure 3.14 depicts the path inflow rates of the three paths and the associated travel costs. Because we impose the side constraint on link a_2, paths p_1, and p_3 are no longer symmetric. Consequently, travelers on path p_3 depart earlier than in the case without side constraint to avoid the additional travel cost. Meanwhile, the amplitude of the path flow curve of p_3 decreases to satisfy the side constraint imposed on link a_2. The departure rate of path p_2 further deceases because the new side constraint imposed on link a_3 has relatively smaller values. Travelers switch to path p_1 to avoid the penalty caused by the

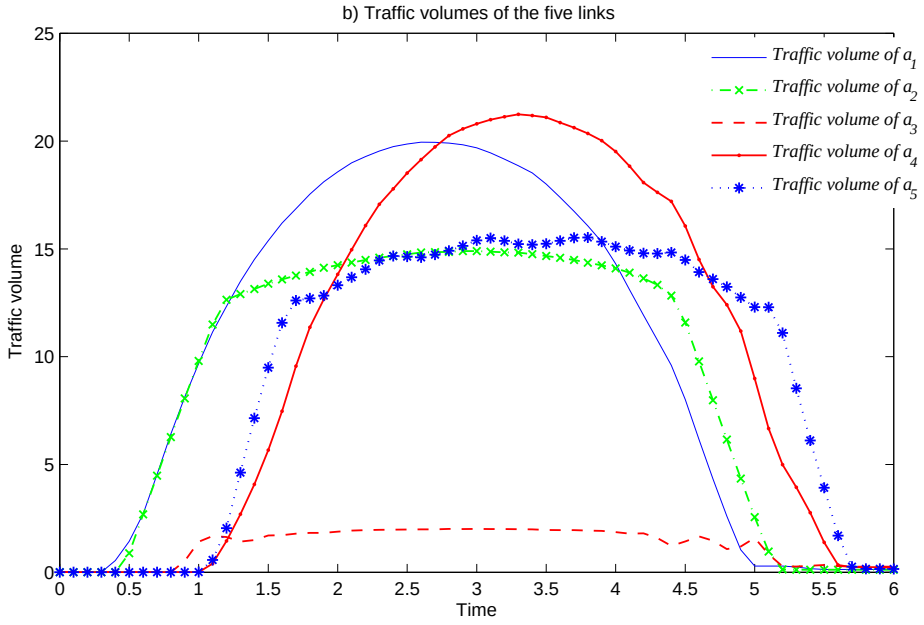


Figure 3.15: Traffic volumes of the five links under time-varying traffic volume control

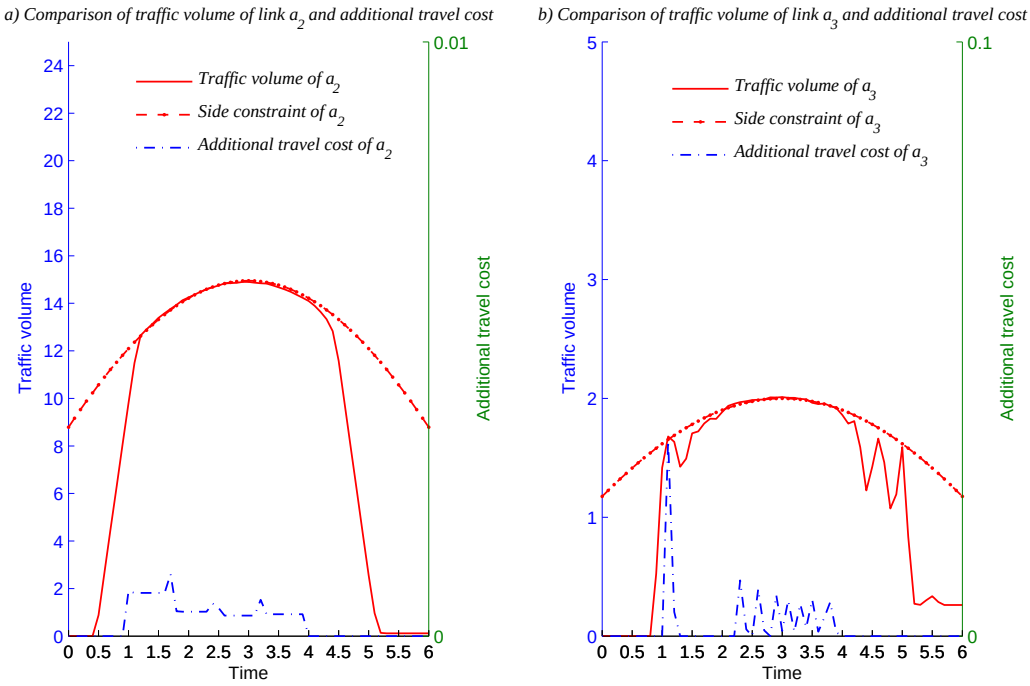


Figure 3.16: Link traffic volumes, time-varying side constraints and additional travel costs of the Braess' network under the DUE-SC condition

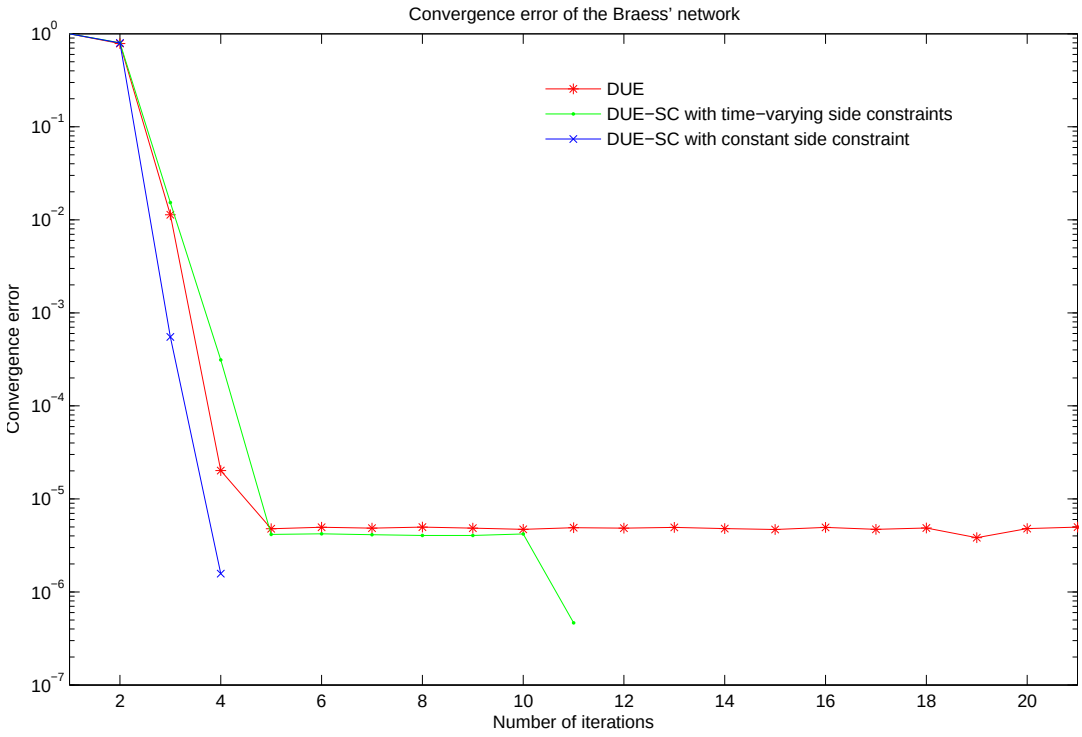


Figure 3.17: Change in the convergence error with iteration for the Braess' network

side constraints as illustrated in Figure 3.14. Figure 3.15 shows the traffic volumes of the five links. The comparison of link traffic volume, time-varying side constraint and the associated additional travel cost is depicted in Figure 3.16. The additional travel cost of link a_3 has larger value when x_{a_3} first reach the level specified by the side constraint. As a result, the traffic is diverted to other paths, which in turn decreases the traffic volume on link a_3. As indicated by Figure 3.16, the additional travel cost of link a_3 is generally larger than that of link a_2, perhaps because path p_3 has a larger free-flow time than path p_2. Travelers are more sensitive to the additional travel cost on that path. The path flows of paths p_2 and p_3 in this case are about 56 and 97 percent of those under the DUE condition, respectively. This verifies that the smaller the additional travel cost, the smaller the control effect it yielded.

The convergence errors for these three test with respect to iterations are depicted in Figure 3.17, where the error is defined as $e^{k+1} = \frac{\|\mathbf{h}^{k+1}-\mathbf{h}^k\|}{\|\mathbf{h}^k\|}$. As shown in the figure, the proposed numerical solution algorithm converges sharply for this example, and the proposed numerical solution method converges faster with the DUE-SC. This occurs because the DUE-SC has a relatively smaller search region (or feasible region) for the algorithm if it admits a solution.

One should note that the cumulative traffic volume on each link does not tend to zero as time tends to the end of the planning horizon. This phenomenon is also observed

by Friesz and Mookherjee (2006); Friesz et al. (2008). Numerical errors of the network loading may be one of the reasons. However, this phenomenon occurs because of the so-called double-counting-effect of the WLM Nie and Zhang (2002, 2005).

3.6 Conclusions

The traffic assignment with side constraint problem is extended to the dynamic case in this chapter to allow the study of a traffic volume control scheme. The side constraints are related to the desired temporal traffic volumes on certain links, which can be set according to the safety or environmental requirements. The dynamic user equilibrium problem with side constraints and simultaneous departure-time-and-path-choice is formulated as an infinite-dimensional variational inequality. We show the existence of equilibrium to the DUE-SC based on the VI formulation under certain assumptions. To analyze the necessary condition, we restate the problem as an equivalent optimal control problem. The optimality condition of the DUE-SC is obtained by applying the Pontryagin minimum principle to the optimal control problem. The equilibrium dynamic travel cost under the DUE-SC condition is shown to be the effective path delay function plus a term of additional travel cost induced by the side constraints. This additional travel cost term is governed by the accumulation of the Lagrange multipliers associated with the side constraints over time (unlike the static case). This additional travel cost term also represents the control parameter that allows the traffic volume control scheme to achieve the required link traffic volumes. If the side constraints are chosen as the link storage capacities, the additional cost can be viewed as the effect needed to prevent the network from spillback. In this sense, the chapter proposes a novel analytical approach to access the DTA considering the spillback effect while avoiding the drawbacks of physical queue models. Another meaningful implication of imposing the arc storage capacity constraint is that the flow is incompressible when the link traffic volume is equal to its storage capacity. This is consistent with the findings in Ross (1988) where the LWR model was applied to describe the traffic dynamics. The chapter has also highlighted the similarity between the additional delay terms from the static and dynamic cases.

We propose a solution algorithm for solving the DUE-SC by using the nonlinear programming approach with Euler's discretization scheme. Numerical examples are presented to illustrate the application of the theory. The numerical results confirm the solution algorithm's satisfactory performance against the small test networks. The results obtained in the constant traffic volume control test verify that given a long enough departure time

horizon, the inflow profile of the controlled link will converge to a steady state determined by the traffic volume control. This is consistent with the static case (Yang et al., 2004). Tests are also provided to verify that the proposed method is applicable to the time-varying traffic volume control scheme, which is one of the major contributions of this chapter and is the key difference between the DUE-SC and SC-TAP. The dynamic user equilibrium with strict capacity constraints, which has been intensively studied in the simulation-based DTA models, can also be included as a special case of the current DUE-SC framework by specifying the side constraints to be the link capacities.

As a Nash non-cooperative differential game, dynamic user equilibrium is used to represent the distribution of traffic that arises when travelers do not have knowledge about other travelers' strategies and compete with each other to minimize their own travel cost. Such distribution of traffic generally does not lead to the optimal usage of a traffic network. In some situations, such as traffic diversion under incidents (wherein queue control is always necessary to prevent the spillback effect), it would be more meaningful for the system manager to look for the best usage of the network under queue control. In the next chapter, we will address this issue and discuss some advanced issues can be achieved by the framework of dynamic traffic assignment with traffic volume control.

Appendix

Constraint qualifications for the side constraints

For a constrained optimal control problem, the constraints in the form of (3.47) are called pure state variable inequality constraints. Let us denote the side constraint imposed on link a_i as $\varpi_i(\mathbf{x}(t), t) \triangleq x_{a_i}(t) - C_{a_i}(t) \leq 0$. In this appendix, we denote these pure state variable inequality constraints in vector form as

$$\varpi(\mathbf{x}(t), t) \triangleq \mathbf{x}(t) - \mathbf{C}(t) \leq 0, \ \forall t \in [0, T], \tag{3.96}$$

where $\mathbf{x}(t) \in R^n$ is the vector of link traffic volumes at time t, which is defined in Section 3.3 and taken as the vector of state variables of the optimal control problem formulation of the DUE-SC. By assumption, $\varpi(\cdot)$ is a continuously differentiable multi-function.

To better understand the dynamic network loading procedure, let us define the link traffic dynamics, in line with Wie et al. (2002) and Friesz and Mookherjee (2006), as follows:

$$\frac{dx_{a_i}(t)}{dt} = u_{a_i}(t) - v_{a_i}(t), \tag{3.97}$$

where $u_{a_i}(t)$ represents the inflow rate of link a_i at time t, which is taken as the control variable, and $v_{a_i}(t)$ the outflow rate of this link at the same time. We denote the control vector as $\mathbf{u} = (u_{a_i} : \forall a_i \in A)$. The constraint $\varpi_i(\mathbf{x}(t), t) \leq 0$ is called a constraint of the r^{th} order if the r^{th} time derivative of $\varpi_i(\mathbf{x}(t), t)$ is the first time in which a term in control variable(s) u_{a_i} appears. We denote the r^{th} time derivative of $\varpi_i(\mathbf{x}(t), t)$ as $\varpi_i^r(\mathbf{x}(t), t)$. With respect to the i^{th} constraint $\varpi_i(\mathbf{x}(t), t) \leq 0$, an interval $[\pi_i^1, \pi_i^2] \subset [0, T]$ is called an interior or unconstrained interval if $\varpi_i(\mathbf{x}(t), t) < 0,\ \forall t \in [\pi_i^1, \pi_i^2]$. If the optimal trajectory "hits the boundary," i.e., $\varpi_i(\mathbf{x}(t), t) = 0$, for a particular i and an interval $[\varsigma_i^1, \varsigma_i^2] \subset [0, T]$, then $[\varsigma_i^1, \varsigma_i^2]$ is called a boundary or constrained interval. An instant ς_i^1 is called an entry time if there is an interior interval ending at time ς_i^1 and a boundary interval starting at time ς_i^1. Correspondingly, ς_i^2 is called an exit time if a boundary ends and an interior interval starts at time ς_i^2. If the trajectory touches the boundary at time ς_i, i.e., $\varpi_i(\mathbf{x}(t), \varsigma_i) = 0$ for a particular i and if the trajectory is in the interior just before and just after ς_i, then ς_i is called a contact time. These entry, exit, and contact times are called junction times.

It is easy for us to show that the pure state inequality constraints, which are the side constraints, are of the 1^{st} order. In fact, by evaluating the derivative of $\varpi_i(\mathbf{x}(t), t)$ with respect to time once, we obtain

$$\frac{d\varpi_i(\mathbf{x}(t), t)}{dt} = \frac{dx_{a_i}(t)}{dt} - \frac{dC_{a_i}(t)}{dt} = u_{a_i}(t) - v_{a_i}(t) - \frac{dC_{a_i}(t)}{dt}. \tag{3.98}$$

We thus observe that the first time derivative of $\varpi_i(\mathbf{x}(t), t)$ depends explicitly on the control $u_{a_i}(t)$. Therefore, the constraint is of the 1^{st} order. The same reasoning can be applied to other side constraints to conclude that all of the side constraints are of the 1^{st} order. In fact, the pure state constraints are of the 1^{st} order for many economic optimal control problems (Adida and Perakis, 2007; Hartl et al., 1995).

As for the pure state constraints (or the side constraints in the DUE-SC problem), the following full rank condition must be fulfilled on any boundary interval $[\varsigma_j^1, \varsigma_j^2]$:

$$\text{rank} \begin{bmatrix} \frac{\partial \varpi_1^1}{\partial \mathbf{u}} \\ \vdots \\ \frac{\partial \varpi_{\hat{b}}^1}{\partial \mathbf{u}} \end{bmatrix} = \hat{b}, \tag{3.99}$$

for $t \in [\varsigma_j^1, \varsigma_j^2]$, $\varpi_i(\mathbf{x}^*(t), t) = 0,\ i = 1, \cdots, \hat{b}$, $\varpi_i(\mathbf{x}^*(t), t) < 0,\ i = \hat{b}+1, \cdots, n$, where $\mathbf{x}^*(t)$ denotes the optimal state trajectory. It can be easily shown that the full rank condition

(3.99) holds for the DUE-SC problem. Let

$$\hat{\Theta} = \begin{bmatrix} \frac{\partial \varpi_1^1}{\partial \mathbf{u}} \\ \frac{\partial \varpi_2^1}{\partial \mathbf{u}} \\ \vdots \\ \frac{\partial \varpi_{\hat{b}}^1}{\partial \mathbf{u}} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \tag{3.100}$$

which clearly has rank $\hat{b}$, i.e., $\text{rank}(\hat{\Theta}) = \hat{b}$. We have thus shown that the constraint qualifications introduced by the side constraints are fulfilled by the proposed DUE-SC formulation.

As we have applied the direct adjoining approach[4], the costate trajectory (or adjoint function/adjoint equation) may have discontinuities for any time ς in a boundary interval and for any contact time ς. These discontinuities are governed by some jump conditions, which are determined by the pure state constraints (Hartl et al., 1995). However, as the pure state constraints are of the 1^{st} order, we have the following lemma.

Lemma A.1. (Hartl et al., 1995; Fattorini, 1999) Given that the pure state constraint ϖ is of the 1^{st} order, the adjoint function is continuous at junction time ς if the entry or exit is non-tangential, i.e., $\varpi^1\left(\mathbf{x}^*(\varsigma^-), \varsigma^-\right) > 0$ or $\varpi^1\left(\mathbf{x}^*(\varsigma^+), \varsigma^+\right) < 0$, where ς^+ and ς^- denote the left- and right- hand side limits, respectively.

The pure state constraints (or the proposed side constraints) satisfy the conditions in Lemma A.1. in the following sense. First, as proven previously, the side constraint ϖ_i is of the 1^{st} order. Let ς be a junction time for link a_i with the side constraint $\varpi_i\left(\mathbf{x}^*(t), t\right)$, then

$$x_{a_i}^*\left(\varsigma^-\right) < C_{a_i}(\varsigma^-),\ \varpi_i^1\left(\mathbf{x}^*(\varsigma^-), \varsigma^-\right) = \dot{x}_{a_i}^*\left(\varsigma^-\right) - \dot{C}_{a_i}(\varsigma^-). \tag{3.101}$$

Now we have the following two situations:

> Suppose that $\dot{C}_{a_i}(\varsigma^-) \leq 0$; to hit the constraint at time ς, i.e., $x_{a_i}^*(\varsigma) = C_{a_i}(\varsigma)$, $\dot{x}_{a_i}^*(\varsigma^-) > \dot{C}_{a_i}(\varsigma^-)$ must be satisfied, which implies that $\varpi_i^1\left(\mathbf{x}^*(\varsigma^-), \varsigma^-\right) > 0$;
>
> Suppose that $\dot{C}_{a_i}(\varsigma^-) > 0$; to render $x_{a_i}^*(\varsigma) = C_{a_i}(\varsigma)$, $\dot{x}_{a_i}^*(\varsigma^-) > \dot{C}_{a_i}(\varsigma^-)$ must be satisfied, which again implies that $\varpi_i^1\left(\mathbf{x}^*(\varsigma^-), \varsigma^-\right) > 0$;

[4]The approach derives its name from the fact that the constraints are directly adjoined to the Hamiltonian.

The same reasoning can be applied to other side constraints with boundary intervals. Therefore, the conditions in Lemma A.1. are satisfied by the proposed side constraints. The adjoint function, i.e., $\lambda^p_{a_i}$ in our DUE-SC case, is continuous at junction times.

Chapter 4

Dynamic marginal cost, access control, and pollution charge: a comparison of bottleneck and whole link models

In this chapter, we investigate theoretical constructions and properties of three interrelated travel demand management measures including marginal cost pricing, access control, and pollution charge under dynamic traffic assignment framework. For congested traffic networks modeled by the two vertical queue models, i.e. the whole link model (WLM) and the deterministic queuing model (DQM), on which flows are controlled, we derive dynamic marginal costs for paths and users' external costs for controlled links. As a strategy to implement the access control, the access pricing is formulated as a dynamic system optimal assignment with access (e.g. traffic volume, queue) control (DSO-AC) problem, wherein the access constraints represent the restrictions on the traffic volumes and/or environmental constraints (e.g. vehicle emission). For the WLM case, an optimal control problem formulation is adopted to investigate the dynamic traffic equilibrium. We derive and discuss the necessary condition for operating the transportation system with capacity/environmental constraints optimally, i.e. the total system travel cost is minimized. The Lagrange multipliers associated with the access constraints as derived from the optimality condition provide part of the users' external costs. For the DQM case, we utilize the formulation adopted by Yang and Meng (1998). The inflow to a bottleneck is saturated such that no queue would be formed. The access price is then given by the penalty associated with this constraint. Similar to the telecommunication network bandwidth allocation

scheme, the DSO-AC analysis reveals the variety of economic effect of a certain amount of road capacity with respect to its spatial and temporal allocation, e.g. decide which links can be used and how to use their available capacities as "holding" capacities for queues. The dynamic externalities of the two models are compared. It is found that different externality structures of the two models result in different tolling structures to achieve DSO. Based on this access pricing analysis and an "equivalent" environmental capacity that converts the environmental constraint into traffic volume restriction, we investigate the traffic induced air pollution pricing scheme. It is found that the traffic capacity based access price and traffic induced air pollution price would not become effective simultaneously for the DQM case. However, for the WLM case, we have a circumstance that both prices would be effective simultaneously. An algorithm based on Euler's discretization scheme and nonlinear programming is proposed to solve the WLM based optimal control problem. Numerical example is presented to illustrate the proposed method as a dynamic road pricing scheme.

4.1 Introduction

As explained in Chapter 1 that traffic congestion and environmental issues associated with vehicle use have been recognized as serious problems faced by modern cities for their negative effects on productivity, health and living conditions. Research has indicated that vehicles are responsible for at least 50% of the air pollution in urban areas (The Economist, 1996). Only about 20% of the town residents enjoy good enough air quality according to the estimation of the World Health Organization (WHO) in terms of the measured levels of emissions (Nagurney, 2000).

Nowadays, many governments would like to improve or maintain the air quality of the urban area by controlling the traffic volume in the network to achieve a more sustainable mobility, e.g. the temporary plate-number-based traffic rationing in Beijing and some long term implementations of road space rationing in Latin America, such as Mexico City, Santiago, Säo Paulo (Han et al., 2010). As a result, there has been a growing interest in the development of rigorous tools for both congestion and emission control management (see, e.g. Nagurney (2000) and the references therein). It has also been argued in the literature that pollution, especially the air pollution, caused by the traffic on the network should be considered as a kind of external cost which can be internalized by road pricing. Some economists and traffic engineers have advocated market-based approaches such as road pricing and congestion derivative (Lindsey, 2006; Friesz et al., 2008; Yao et al., 2010)

to reduce harmful traffic emissions. An extensive amount of research in road pricing for (deterministic/stochastic) static traffic network have concluded that road pricing is an efficient approach to internalize externalities such as congestion, air pollution, noise, and accidents (Sumalee and Xu, 2010; Yang et al., 2010). Side constrained traffic assignment is introduced to model the externality caused by link capacity constraint and the traffic induced pollution (Yang et al., 2010).

Although (static) road pricing has been extensively studied in literature, there are some drawbacks of the scheme, e.g. theoretically, the assumption of perfect information on the traffic network (both supply and demand sides) (Akamatsu, 2007; Tsekeris and Voß, 2009), and practically, despite advances in electronic tolling collection (ETC) technology, tolling is still costly (Tsekeris and Voß, 2009; Wang et al., 2010). The tradable network permit scheme, which is another traffic control scheme introduced by Akamatsu (2007) and Nagae and Sasaki (2009), uses an idea similar to road pricing. The scheme ensures the goal that there is no congestion in the network (which is the dynamic system optimum for deterministic queuing model) by imposing capacity (access) constraints on the bottlenecks. Mathematically, the combined problem can be formulated as a dynamic system optimal (DSO) assignment with access constraints on queues. The permit price to access a bottleneck in the network can be determined by solving the Lagrange multipliers associated with the bottleneck capacity constraints. This scheme can be interpreted as the system manager issuing a certain amount of access tickets for drivers to use certain bottlenecks. The drivers can then trade these tickets in an efficient and competitive market to access the competitive bottlenecks. Such kind of scheme is similar to the access pricing for competitive bottlenecks in the telecommunication and electricity networks, where networks operate in a competitive market for subscribers, and yet have a monopoly position for providing access to these subscribers (Armstrong et al., 1996). The access pricing is introduced to the nonlinear (static) road pricing by Wang et al. (2010) as a complementary scheme of the marginal cost (usage) pricing.

The DSO assignment analyzes the dynamic marginal cost, dynamic externality and the optimality condition under which the user surplus is maximized (Kuwahara, 2007; Chow, 2009a). The DSO assignment provides a bound on the best performance of a traffic network, which makes it as a benchmark for evaluating various transport policy measures, e.g. time-dependent pricing (Yang and Meng, 1998; Chow, 2009a,b), network access control (Smith and Ghali, 1990; Lovell and Daganzo, 2000; Shen and Zhang, 2009; Zhang and Shen, 2010), and road capacity allocation (Ghali and Smith, 1995). Therefore, in this chapter, we will formulate the underlying problems by DSO with access constraints to si-

multaneously consider dynamic congestion and environmental externalities. Two vertical (or point) queue models, i.e. the whole link model (WLM) and the deterministic queueing model (DQM) are applied as the network loading model to describe the traffic dynamics. We adopt the whole-link linear travel time model that guarantees the first-in-first-out (FIFO) discipline of dynamic flow. By using the WLM as network loading model, the DSO-AC is formulated as an optimal control problem for a class of dynamic systems with constraints on inputs and states. The Pontryagin minimum principle is then applied to derive the necessary condition of the DSO-AC. Two terms are presented as the dynamic external cost for usage pricing. One is the dynamic externality to a path, which is defined as the change in the value of the total system travel cost with respect to a perturbation in the path inflow profile (Kuwahara, 2007; Chow, 2009a). The other external cost is to be imposed on travelers during their presence on each link along each of their travel route. The Lagrange multipliers associated with the access constraints as derived from the optimality condition of the DSO-AC contribute to the access price, which can be interpreted as either the tolls or the permit prices imposed upon drivers for using the saturated links. With the DQM as network loading model, we utilize the formulation adopted by Yang and Meng (1998), wherein the inflow rates to bottlenecks are constrained to be less than or equal to the bottleneck capacities. The access price to a bottleneck is obtained by solving the Lagrange multiplier associated with the constraint. A comparison of the dynamic externality obtained from the WLM with that from the DQM will also be conducted.

The time-dependent concentrations of the traffic induced air pollution are estimated by the well known and widely applied traffic pollution model in environmental research and engineering—the operational street pollution model (OSPM) (Berkowicz, 1998; Vardoulakis et al., 2003, 2007). To analyze the effect of pollution control (or environmental constraint) and traffic induced air pollution pricing scheme under the umbrella of the access pricing analysis, we define an "equivalent" environmental traffic capacity that converts the environmental constraint on a link into traffic volume restriction on that link. Based on this effort, we investigate the traffic induced air pollution pricing scheme. To be more specific, we will address the following problems:

1. Which pricing scheme should be imposed on each link, the access control based or the environmental traffic capacity based pricing scheme?

2. Which is the dominant pricing scheme at a specific time instant?

3. How to determine the boundaries under which a traveler on a link should pay either an access toll or an extra pollution charge?

We will also include the dynamic pollution pricing as a special case of the access pricing.

Similar to the bandwidth allocation scheme developed in telecommunication network (see, e.g. Lazar and Semret (1999); Dramitinos et al. (2007)), the DSO with access constraints analysis also reveals the variety of economic effect of a certain amount of road capacity with respect to its spatial and temporal allocation. This provides a guideline to the roadway capacity allocation, e.g. decide which links can be used and how to use their available capacities (by imposing proper access constraints on them) as "holding" capacities for queues.

The remainder of this chapter is organized as follows. By using the WLM as network loading model, the DSO-AC is formulated and analyzed in Section 4.2. We study the DSO-AC for the case with DQM as network loading model in Section 4.3. We compare and discuss different results on the dynamic externality, dynamic marginal cost pricing, and access pricing in Section 4.4. The traffic induced pollution pricing scheme and its relationship between the access pricing scheme are discussed in Section 4.5. A solution algorithm for solving the DSO-AC problem is proposed in Section 4.6. Numerical examples are then provided to demonstrate the proposed method in Section 4.7. Section 4.8 concludes the chapter.

4.2 Problem formulation of the DSO-AC and its solution—the WLM case

As mentioned in the introduction and note that the road pricing is one of the means for capacity allocation (Johnston et al., 1995), we propose a dynamic equilibrium model for dynamic pricing in this chapter. The pricing scheme consists of a marginal cost (usage) pricing plus an access pricing which reveals the economic effect of a certain amount of road capacity with respect to its spatial and temporal allocation. In particular, the cost to access a bottleneck (or restricted link) in the network can be collected as a part of the toll (i.e. be implemented as a complementary scheme of the marginal cost (usage) pricing). Mathematically, the problem can be formulated as a dynamic system optimal (DSO) assignment with access constraints on link traffic volumes.

4.2.1 Problem formulation of the DSO-AC

The dynamic system optimal traffic assignment problem is initiated by Merchant and Nemhauser (1978a,b). The problem is reinvestigated in Chow (2009a) by applying the optimal control theory framework proposed by Friesz et al. (2001). In this chapter, we

consider a finite time planning horizon $T > 0$ and regard time $t \in [0, T]$ as a continuous variable. Let P denote the set of all paths in a network. An arbitrary path $p \in P$ of the network of interest is defined by a sequence of the links used by that path which is denoted by $p \doteq \{a_1, a_2, \cdots, a_{m(p)}\}$, where $m(p)$ is the number of links used by path p. To begin with, we list the problem formulation of dynamic system optimal assignment, which is formulated as the following optimal control problem. It seeks an optimal route inflow profile $h_p^*(t)$ to minimize the total system travel cost within the study period, T, given a fixed total amount of traffic Q_w to be served between each origin-destination (OD) pair w:

$$\min J = \sum_{\forall p \in P} \int_0^T \Psi_p(t, \mathbf{x}) h_p(t) dt, \tag{4.1}$$

subject to,

$$\begin{aligned}
\frac{dx_{a_1}^p(t)}{dt} &= h_p(t) - g_{a_1}^p(t),\ \forall p \in P,\ \forall t, && (4.2)\\
\frac{dx_{a_i}^p(t)}{dt} &= g_{a_{i-1}}^p(t) - g_{a_i}^p(t), && \\
&\quad \forall p \in P,\ i \in [2,\ m(p)],\ \forall t, && (4.3)\\
g_{a_1}^p\left(t + D_{a_1}\left(x_{a_1}(t)\right)\right)\left(1 + D'_{a_1}\left(x_{a_1}(t)\right)\dot{x}_{a_1}(t)\right) &= h_p(t),\ \forall p \in P,\ \forall t, && (4.4)\\
g_{a_i}^p\left(t + D_{a_i}\left(x_{a_i}(t)\right)\right)\left(1 + D'_{a_i}\left(x_{a_i}(t)\right)\dot{x}_{a_i}(t)\right) &= g_{a_{i-1}}^p(t), && \\
&\quad \forall p \in P,\ i \in [2, m(p)],\ \forall t, && (4.5)\\
\frac{dE_w(t)}{dt} &= \sum_{p \in P_w} h_p(t),\ \forall w \in W,\ \forall t, && (4.6)\\
E_w(T) &= Q_w,\ \forall w \in W, && (4.7)\\
h_p(t) &\geq 0,\ \forall p \in P, \forall t, && (4.8)\\
x_{a_i}(t) &\leq C_{a_i}(t),\ \ \forall a_i \in A,\ \forall t, && (4.9)\\
x_{a_i}(0) = 0,\ \forall a_i \in A, E_w(0) = 0,\ \forall w &\in W,\ h_p(0) = 0,\ \forall p \in P, && (4.10)
\end{aligned}$$

Equations (4.2)-(4.3) are the relevant link traffic dynamics, where $x_{a_i}^p(t)$ is the traffic volume on path p traversing link a_i at time t, $g_{a_i}^p(t)$ is the flow exiting link a_i and $g_{a_{i-1}}^p(t)$ is the flow entering link a_i of path $p \in P$ at time t. In addition, $g_{a_0}^p(t)$ is the flow exiting the origin of path p at time t which is referred to as the departure rate of path p at time t and is denoted by $h_p(t) = g_{a_0}^p(t)$. The total traffic volume $x_{a_i}(t)$ of link a_i at time t is defined by

$$x_{a_i}(t) = \sum_{p \in P} x_{a_i}^p(t) \delta_{a_i}^p,\ \forall a_i \in A, \tag{4.11}$$

and

$$\delta_{a_i}^p = \begin{cases} 1, \text{ if } a_i \in p, \\ 0, \text{ otherwise,} \end{cases}$$

is the Kronecker Delta function. Equations (4.4)-(4.5) define the proper flow propagation, where $G_{a_i}^p(t)$ denotes the cumulative link outflow of link a_i by path p up to time t, while $\tau_{a_i}^p(t)$ denotes the time of exit from link a_i for vehicles that enter path p at its origin at time t. The proper flow propagation conditions given by (4.4)-(4.5) are equivalent to Equations (62) and (63) of Friesz et al. (2001). However, as we can see in the proof of the DSO-AC, (4.4) and (4.5) are easier to handle in the analysis. Equations (4.6)-(4.7) define the flow conservation constraints, which are equivalent to:

$$\sum_{p \in P_w} \int_0^T h_p(t)dt = Q_w, \quad \forall w \in W, \tag{4.12}$$

where Q_w is the fixed total travel demand for OD pair $w \in W$, P_w denotes the set of paths connecting OD pair w. (4.8) defines the nonnegative constraints of path flows. Each link is characterized by a link travel time function, $D_{a_i}(x_{a_i}(t))$, which defines the link travel time as a function of the link traffic volume at the entry time to the link:

$$\begin{aligned} \tau_{a_1}^p(t) &= t + D_{a_1}(x_{a_1}(t)), \ \forall p \in P, \\ \tau_{a_i}^p(t) &= \tau_{a_{i-1}}^p(t) + D_{a_i}\left(x_{a_i}(\tau_{a_{i-1}}^p(t))\right), \ \forall p \in P, \ i \in [2, m(p)]. \end{aligned}$$

The following condition ensures the FIFO queue discipline for the DTA problems:

$$1 + D'_{a_i}(x_{a_i}(t))\dot{x}_{a_i}(t) \geq 0. \tag{4.13}$$

Under the FIFO queue discipline and given (4.8) as well as nonnegative initial conditions, the outflow $g_{a_i}^p(t)$ and traffic volume (or queue) $x_{a_i}^p(t)$ are nonnegative for all t. Without loss of generality, we assume zero initial conditions, that is $x_{a_i}(0) = 0$, $h_p(0) = 0$ and $E_w(0) = 0$. The dynamic link travel time model that we use is the whole-link linear travel time model proposed by Friesz et al. (1993). The model considers the link travel time to be a linear function of the traffic volume on the link. The exit time of a vehicle entering the link at time t can be calculated as:

$$\tau_a(t) = t + \psi_a + x_a(t)/R_a, \tag{4.14}$$

where ψ_a is a flow-invariant travel time (free-flow travel time) of link a and R_a is the link capacity. From (4.14), the definition of the whole link linear travel time model, the FIFO queue discipline, i.e., $\frac{d\tau_a(t)}{dt} > 0$, holds, and the strong FIFO condition also holds (Zhu and Marcotte, 2000).

The nested path delay operators proposed by Friesz et al. (1993) are defined as:

$$D_p(t,\mathbf{x}) := \sum_{i=1}^{m(p)} \delta_{a_i p}\Phi_{a_i}(t,\mathbf{x}), \forall p \in P, \tag{4.15}$$

where $\mathbf{x} = (x_{a_i} : \forall a_i \in A)$ and

$$\begin{aligned}
\Phi_{a_1}(t,\mathbf{x}) &= D_{a_1}\left(x_{a_1}(t)\right), \\
\Phi_{a_i}(t,\mathbf{x}) &= D_{a_i}\left(x_{a_i}(t+\Phi_{a_1}+\cdots+\Phi_{a_{i-1}})\right) = D_{a_i}\left(x_{ai}\left(t+\sum_{j=1}^{i-1}\Phi_{a_j}\right)\right), \\
&\quad \forall i \in [2, m(p)].
\end{aligned}$$

For the departure time choice, the schedule delay cost function (or early/late arrival penalty) $\kappa(\chi)$ is employed, whereby χ is defined as the difference between actual and preferred arrival time denoted by t^*: $\chi = t + D_p(t,\mathbf{x}) - t^*$. The path delays and schedule delay cost function are combined to obtain the effective path delay operators:

$$\Psi_p(t,\mathbf{x}) = D_p(t,\mathbf{x}) + \kappa(\chi), \quad \forall p \in P. \tag{4.16}$$

(4.9) defines the access constraints imposed on the links of the network. Let us define the following vector-valued function for the access constraints as $\mathbf{C} = (C_{a_i} : \forall a_i \in A)$, where $\mathbf{C} : [0,T] \to \Sigma \subset R_+^n$ is a continuously differentiable function, with Σ being a prescribed compact subset of R_+^n with known bounds. We rewrite the access constraints in a compact form as

$$\mathbf{x}(t) \le \mathbf{C}(t), \ \forall t \in [0,T]. \tag{4.17}$$

By construction, the set

$$\Sigma_x \triangleq \{\mathbf{x}(t) : 0 \le \mathbf{x}(t) \le \mathbf{C}(t)\},$$

is a compact subset of R^n.

4.2.2 Property of the DSO-AC

Proposition 4.1. The necessary condition for the DSO-AC can be stated as follows:

$$h_p(t) \begin{cases} > 0 \Rightarrow \Psi_p(t,\mathbf{x}^*) + \frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)} h_p(t) + l_p(t) = \phi_w, \\ = 0 \Rightarrow \Psi_p(t,\mathbf{x}^*) + \frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)} h_p(t) + l_p(t) > \phi_w, \end{cases} \quad \forall p \in P_w, \ w \in W, \tag{4.18}$$

where where $\mathbf{x}^*$ is an optimal state of the DSO-AC problem, ϕ_w is the travel cost of OD pair w under the DSO-AC condition, which is determined by the fixed total travel demand Q_w of the OD pair. The dynamic external cost consists of two parts, i.e. $\frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)} h_p(t)$,

and $l_p(t)$. The first part $\frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)} h_p(t)$ is the sensitivity value of the total system travel cost with respect to a perturbation in the path flow $h_p(t)$.

$$l_p(t) = \sum_{i=0}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \left(\varsigma_{a_{i+1}}^p(t) + \eta_{a_{i+1}}(t) \sigma_{a_{i+1}}^p \right) dt$$

is the other term of the dynamic external cost to be imposed on travelers during their presence on each link along their travel path. η_{a_i} is the Lagrange multiplier associated with the access constraint imposed on link a_i, $\int_{t_{i-1}^p}^{t_i^p} \eta_{a_i}(t) \sigma_{a_i}^p dt$ is the cost to access the restricted link (bottleneck) a_i. $\int_{t_{i-1}^p}^{t_i^p} \varsigma_{a_i}^p(t) dt$ represents part of the dynamic external cost to be imposed on travelers during their presence on link a_i along their travel path p, with $\varsigma_{a_i}^p(t) = h_p(t) \frac{\partial \Psi_p(t,\mathbf{x})}{\partial D_p(t,\mathbf{x})} \frac{\partial D_p(t,\mathbf{x})}{\partial x_{a_i}^p(t)}$. t_i^p is the entry time to link a_{i+1} for vehicles traveling by path p that departs from the origin at time $t \triangleq t_0^p$. ∎

Proof of Proposition 4.1. In this proof, we derive the necessary condition for the DSO-AC by applying the Pontryagin Minimum Principle. The Lagrangian for the optimal control problem (4.1)- (4.9) can be defined as:

$$\begin{aligned}
Z^* \;=\; & \sum_{p\in P} \int_0^T \Big\{ \Psi_p(t,\mathbf{x}) h_p(t) + \lambda_{a_1}^p(t) \left(h_p(t) - g_{a_1}^p(t) - \frac{dx_{a_1}^p(t)}{dt} \right) \\
& + \sum_{i\in[2,m(p)]} \lambda_{a_i}^p(t) \left(g_{a_{i-1}}^p(t) - g_{a_i}^p(t) - \frac{dx_{a_i}^p(t)}{dt} \right) \\
& + \gamma_{a_1}^p(t) \left(h_p(t) - g_{a_1}^p\left(t + D_{a_1}\left(x_{a_1}(t)\right)\right) \left(1 + D'_{a_1}\left(x_{a_1}(t)\right) \dot{x}_{a_1}(t)\right)\right) \\
& + \sum_{i\in[2,m(p)]} \gamma_{a_i}^p(t) \left(g_{a_{i-1}}^p(t) - g_{a_i}^p\left(t + D_{a_i}\left(x_{a_i}(t)\right)\right) \left(1 + D'_{a_i}\left(x_{a_i}(t)\right) \dot{x}_{a_i}(t)\right)\right) \\
& + \sum_{i\in[1,m(p)]} \eta_{a_i}(t) \left(x_{a_i}(t) - C_{a_i}(t) \right) - \rho_{a_0}^p(t) h_p(t) \Big\} dt \\
& + \sum_{w\in W} \int_0^T \mu_w(t) \left(\sum_{p\in P_w} h_p(t) - \frac{dE_w(t)}{dt} \right) dt \\
& + \sum_{w\in W} \phi_w(T) \left(Q_w - E_w(T) \right), \qquad (4.19)
\end{aligned}$$

where $\lambda_{a_i}^p(t)$ is the costate variable for the traffic dynamics of link a_i, i.e., (4.3). $\mu_w(t)$, $\rho_{a_0}^p(t)$, $\eta_{a_i}(t)$ and $\gamma_{a_i}^p(t)$ are the Lagrange multipliers associated with the cumulative inflow, the nonnegative flow constraint of path p, the access constraint of link a_i, and the proper flow propagation constraint, i.e., (4.5), respectively. $\phi_w(T)$ is the multiplier associated with the total throughput for each OD pair w. By a simple calculation in conjunction

with the zero initial condition assumption, the Lagrangian can be specified as

$$\begin{aligned} Z^* &= \sum_{p\in P}\left(\sum_{i\in[1,m(p)]}\left(-\lambda^p_{a_i}(T)x^p_{a_i}(T)\right)\right) \\ &+ \sum_{w\in W}\left(-\mu_w(T)E_w(T)\right)+\sum_{w\in W}\phi_w(T)\left(Q_w-E_w(T)\right) \\ &+ \int_0^T\left(H(t)+\sum_{w\in W}\frac{d\mu_w(t)}{dt}E_w(t)+\sum_{p\in P}\sum_{i\in[1,m(p)]}\frac{d\lambda^p_{a_i}(t)}{dt}x^p_{a_i}(t)\right)dt, \end{aligned} \quad (4.20)$$

where we define the Hamiltonian function for the optimal control problem as

$$\begin{aligned} H(t) &= \sum_{p\in P}\Big[\Psi_p(t,\mathbf{x})h_p(t)+\lambda^p_{a_1}(t)\left(h_p(t)-g^p_{a_1}(t)\right) \\ &+ \sum_{i\in[2,m(p)]}\lambda^p_{a_i}(t)\left(g^p_{a_{i-1}}(t)-g^p_{a_i}(t)\right) \\ &+ \gamma^p_{a_1}(t)\left(h_p(t)-g^p_{a_1}\left(t+D_{a_1}\left(x_{a_1}(t)\right)\right)\left(1+D'_{a_1}\left(x_{a_1}(t)\right)\dot{x}_{a_1}(t)\right)\right) \\ &+ \sum_{i\in[2,m(p)]}\gamma^p_{a_i}(t)\left(g^p_{a_{i-1}}(t)-g^p_{a_i}\left(t+D_{a_i}\left(x_{a_i}(t)\right)\right)\left(1+D'_{a_i}\left(x_{a_i}(t)\right)\dot{x}_{a_i}(t)\right)\right) \\ &+ \sum_{i\in[1,m(p)]}\eta_{a_i}(t)\left(x_{a_i}(t)-C_{a_i}(t)\right)-\rho^p_{a_0}(t)h_p(t)\Big] \\ &+ \sum_{w\in W}\left(\mu_w(t)\left(\sum_{p\in P_w}h_p(t)\right)\right). \end{aligned} \quad (4.21)$$

The variation δZ^* of Z^* with respect to the control and state variables is given by

$$\begin{aligned} \delta Z^* &= \sum_{p\in P}\sum_{i\in[1,m(p)]}\left(-\lambda^p_{a_i}(T)\delta x^p_{a_i}(T)\right)+\sum_{w\in W}\left(-\mu_w(T)\delta E_w(T)\right) \\ &+ \sum_{w\in W}\phi_w(T)\left(-\delta E_w(T)\right)+\sum_{p\in P}\int_0^T\left(\frac{\partial H(t)}{\partial h_p(t)}\delta h_p(t)\right)dt \\ &+ \sum_{p\in P}\sum_{i\in[1,m(p)]}\int_0^T\left(\frac{\partial H(t)}{\partial x^p_{a_i}(t)}+\frac{d\lambda^p_{a_i}(t)}{dt}\right)\delta x^p_{a_i}(t)dt \\ &+ \sum_{p\in P}\sum_{i\in[1,m(p)]}\int_0^T\left(\frac{\partial H(t)}{\partial g^p_{a_i}(t)}\delta g^p_{a_i}(t)+\frac{\partial H(t)}{\partial g^p_{a_i}\left(\tau^p_{a_i}(t)\right)}\delta g^p_{a_i}\left(\tau^p_{a_i}(t)\right)\right)dt \\ &+ \sum_{w\in W}\int_0^T\left(\frac{d\mu_w(t)}{dt}\delta E_w(t)\right)dt. \end{aligned} \quad (4.22)$$

To simplify the notation and proceed to the analysis, let us define

$$\tilde{g}^p_{a_i}(t)=g^p_{a_i}\left(\tau^p_{a_i}(t)\right)=g^p_{a_i}\left(t+D_{a_i}(x_{a_i}(t))\right),\ \forall p\in P,\ i\in[1,m(p)],$$

and $t^p_i = t^p_{i-1}+D_{a_i}\left(x^p_{a_i}(t^p_{i-1})\right)$, where t^p_{i-1} is the time of the flow $g^p_{a_{i-1}}$ entering link $a_i\in P$ and t^p_i is the time of the flow $g^p_{a_i}$ exiting the link. The partial derivatives of the

Hamiltonian function are given as

$$\begin{aligned}
\frac{\partial H(t)}{\partial h_p(t)} &= \Psi_p(t,\mathbf{x}) + \frac{\partial \Psi_p(t,\mathbf{x})}{\partial h_p(t)} h_p(t) + \lambda^p_{a_1}(t) + \mu_w(t) + \gamma^p_{a_1}(t) - \rho^p_{a_0}(t), \\
&\quad \forall w \in W,\ p \in P_w, && (4.23)\\
\frac{\partial H(t)}{\partial g^p_{a_i}(t)} &= -\lambda^p_{a_i}(t) + \lambda^p_{a_{i+1}}(t) + \gamma^p_{a_{i+1}}(t),\ \forall p \in P,\ i \in [1, m(p)-1], && (4.24)\\
\frac{\partial H(t)}{\partial g^p_{a_{m(p)}}(t)} &= -\lambda^p_{a_{m(p)}}(t),\ \forall p \in P, && (4.25)\\
\frac{\partial H(t)}{\partial \tilde{g}^p_{a_i}(t)} &= -\gamma^p_{a_i}(t)\left(1 + D'_{a_i}(x_{a_i})\dot{x}_{a_i}\right),\ \forall p \in P,\ i \in [1, m(p)], && (4.26)\\
\frac{\partial H(t)}{\partial x^p_{a_i}(t)} &= \eta_{a_i}(t)\sigma^p_{a_i} + \varsigma^p_{a_i}(t),,\ \forall p \in P,\ i \in [1, m(p)], && (4.27)
\end{aligned}$$

where

$$\varsigma^p_{a_i}(t) = h_p(t)\frac{\partial \Psi_p(t,\mathbf{x})}{\partial D_p(t,\mathbf{x})}\frac{\partial D_p(t,\mathbf{x})}{\partial x^p_{a_i}(t)}.$$

In addition, a set of complementary slackness conditions for the nonnegative flow constraints and the access constraints can be obtained as:

$$\begin{aligned}
h_p(t) &\geq 0,\ \rho^p_{a_0}(t) \geq 0,\ \rho^p_{a_0}(t)h_p(t) = 0,\ \forall p \in P, && (4.28)\\
\eta_{a_i}(t) &\geq 0,\ x_{a_i}(t) - c_{a_i}(t) \leq 0,\ \eta_{a_i}(t)\left(x_{a_i}(t) - c_{a_i}(t)\right) = 0, \forall a_i \in A. && (4.29)
\end{aligned}$$

For an open path p (Friesz et al., 2001), the Lagrange multiplier of the complementary slackness condition (4.28) is relaxed to $\rho^p_{a_0}(t^p_0) = 0$. At the optimality, the following stationary conditions hold

$$\begin{aligned}
\frac{\partial H}{\partial h_p}|_{t^p_0} &= 0, \forall w \in W,\ p \in P_w, && (4.30)\\
\frac{\partial H}{\partial g^p_{a_i}}|_{t^p_i} + \left[\frac{\partial H}{\partial \tilde{g}^p_{a_i}}\frac{1}{1 + D'_{a_i}(x_{a_i})\dot{x}_{a_i}}\right]_{t^p_{i-1}} &= 0, \forall p \in P,\ i \in [1, m(p)]. && (4.31)
\end{aligned}$$

The adjoint equations are given by

$$\begin{aligned}
\frac{d\lambda^p_{a_i}(t)}{dt} = -\frac{\partial H(t)}{\partial x^p_{a_i}(t)} &= -\eta_{a_i}(t)\sigma^p_{a_i}, \forall p \in P,\ i \in [1, m(p)], && (4.32)\\
\frac{d\mu_w(t)}{dt} &= 0, \forall w \in W. && (4.33)
\end{aligned}$$

At terminal time T, the following boundary conditions hold:

$$\lambda^p_{a_i}(T) = 0,\ -\mu_w(T) - \phi_w(T) = 0,\ \forall w \in W,\ \forall p \in P_w,\ i \in [1, m(p)]. \quad (4.34)$$

Together with (4.33), the following equality is obtained:

$$\mu_w(t) = -\phi_w(T) = \mu_w, \forall w \in W. \quad (4.35)$$

Note that (4.30) is equal to

$$\Psi_p(t_0^p, \mathbf{x}^*) + \lambda_{a_1}^p(t_0^p) + \mu_w + \gamma_{a_1}^p(t_0^p) = 0, \forall w \in W,\ p \in P, \tag{4.36}$$

where we use the fact that $\mu_w(t) = \mu_w$, which is a constant and for an open path $\rho_{a_0}^p(t_0^p) = 0$. Now, the stationary condition (4.31) gives

$$\begin{aligned} \lambda_{a_{i+1}}^p(t_i^p) + \gamma_{a_{i+1}}^p(t_i^p) &= \lambda_{a_i}^p(t_i^p) + \tilde{\gamma}_{a_i}^p(t_i^p), \\ &\quad \forall p \in P,\ i \in [1, m(p)-1], \end{aligned} \tag{4.37}$$

$$-\lambda_{a_{m(p)}}^p(t_{m(p)}^p) - \tilde{\gamma}_{a_{m(p)}}^p\left(t_{m(p)}^p\right) = 0,\ \forall p \in P, \tag{4.38}$$

where we adopt the following notation in line with Friesz et al. (2001):

$$\tilde{\gamma}_{a_i}^p(t_i^p) \doteq \gamma_{a_i}^p(t_{i-1}^p),\ \ \forall p \in P,\ \ i \in [1,\ m(p)]. \tag{4.39}$$

The adjoint equation (4.32) for an open path can then be given as

$$\frac{d\lambda_{a_i}^p(t)}{dt} = -\eta_{a_i}(t)\sigma_{a_i}^p - \varsigma_{a_i}^p(t),\ \ \forall p \in P,\ i \in [1, m(p)],\ t \in [t_{i-1}^p, t_i^p]. \tag{4.40}$$

Evaluating the adjoint equation (4.40) backward in time and space yields the following. To begin with, let us apply (4.40) to the last link of path p, i.e., link $a_{m(p)}^p$:

$$\lambda_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) - \lambda_{a_{m(p)}}^p\left(t_{m(p)}^p\right) = \int_{t_{m(p)-1}^p}^{t_{m(p)}^p} \left(\eta_{a_{m(p)}}(t)\sigma_{a_{m(p)}}^p + \varsigma_{a_{m(p)}}^p(t)\right) dt. \tag{4.41}$$

From (4.38) we have

$$\tilde{\gamma}_{a_{m(p)}}^p\left(t_{m(p)}^p\right) = \gamma_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) = -\lambda_{a_{m(p)}}^p(t_{m(p)}^p). \tag{4.42}$$

From (4.37), (4.40), and (4.42), the following relation is obtained

$$\begin{aligned} d_{m(p)-1} &= \lambda_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) + \gamma_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) = \lambda_{a_{m(p)}}^p\left(t_{m(p)-1}^p\right) - \lambda_{a_{m(p)}}^p\left(t_{m(p)}^p\right) \\ &= \int_{t_{m(p)-1}^p}^{t_{m(p)}^p} \left(\eta_{a_{m(p)}}(t)\sigma_{a_{m(p)}}^p + \varsigma_{a_{m(p)}}^p(t)\right) dt \\ &= \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-1}^p\right) + \tilde{\gamma}_{a_{m(p)-1}}^p\left(t_{m(p)-1}^p\right). \end{aligned} \tag{4.43}$$

Similarly,

$$\begin{aligned} d_{m(p)-2} &= \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-2}^p\right) + \gamma_{a_{m(p)-1}}^p\left(t_{m(p)-2}^p\right) \\ &= \lambda_{a_{m(p)-2}}^p\left(t_{m(p)-2}^p\right) + \tilde{\gamma}_{a_{m(p)-2}}^p\left(t_{m(p)-2}^p\right). \end{aligned} \tag{4.44}$$

From (4.44) we have

$$\begin{aligned} d_{m(p)-1} - d_{m(p)-2} &= \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-1}^p\right) - \lambda_{a_{m(p)-1}}^p\left(t_{m(p)-2}^p\right) \\ &= -\int_{t_{m(p)-2}^p}^{t_{m(p)-1}^p} \left(\eta_{a_{m(p)-1}}(t)\sigma_{a_{m(p)-1}}^p + \varsigma_{a_{m(p)-1}}^p(t)\right) dt. \end{aligned} \tag{4.45}$$

By proceeding the induction, we have

$$d_i = d_{i+1} + \int_{t_i^p}^{t_{i+1}^p} \left(\eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p + \varsigma_{a_{i+1}}^p(t)\right) dt, \ \forall\, i \in [1, m(p)-1]. \tag{4.46}$$

Finally, we have

$$d_1 = \lambda_{a_1}^p\left(t_1^p\right) + \gamma_{a_1}^p\left(t_0^p\right) = \sum_{i=1}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \left(\eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p + \varsigma_{a_{i+1}}^p(t)\right) dt. \tag{4.47}$$

Thus,

$$\begin{aligned} l_p(t_0^p) &= \lambda_{a_1}^p\left(t_0^p\right) - \lambda_{a_1}^p\left(t_1^p\right) + \lambda_{a_1}^p\left(t_1^p\right) + \gamma_{a_1}^p\left(t_0^p\right) = \int_{t_0^p}^{t_1^p} \left(\eta_{a_1}(t)\sigma_{a_1}^p + \varsigma_{a_1}^p(t)\right) dt + d_1 \\ &= \sum_{i=0}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \left(\eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p + \varsigma_{a_{i+1}}^p(t)\right) dt. \end{aligned} \tag{4.48}$$

The dynamic system optimal equilibrium condition with access constraints is as follows: for an open path $p \in P_w$, $h_p^*(t_0^p) > 0$,

$$\begin{aligned} &\Psi_p(t_0^p, \mathbf{x}^*) + \tfrac{\partial \Psi_p(t_0^p, \mathbf{x}^*)}{\partial h_p(t_0^p)} h_p(t_0^p) + l_p(t_0^p) = \phi_w(T), \quad \text{and if} \\ &\Psi_p(t_0^p, \mathbf{x}^*) + \tfrac{\partial \Psi_p(t_0^p, \mathbf{x}^*)}{\partial h_p(t_0^p)} h_p(t_0^p) + l_p(t_0^p) > \phi_w(T), p \in P_w \Rightarrow h_p^*(t_0^p) = 0. \end{aligned} \tag{4.49}$$

As t_0^p is arbitrarily and $\phi_w(T)$ is constant, without loss of generality, we write (4.49) as

$$\begin{aligned} &\Psi_p(t, \mathbf{x}^*) + \frac{\partial \Psi_p(t, \mathbf{x}^*)}{\partial h_p(t)} h_p(t) + l_p(t) = \phi_w, \quad \text{and if} \\ &\Psi_p(t, \mathbf{x}^*) + \frac{\partial \Psi_p(t, \mathbf{x}^*)}{\partial h_p(t)} h_p(t) + l_p(t) > \phi_w, \ p \in P_w \Rightarrow h_p^*(t) = 0, \end{aligned} \tag{4.50}$$

which is immediately recognized as the fundamental condition for the dynamic network system optimal equilibrium (Chow, 2009a). We also have the following complementary slackness conditions for the access constraints: (i) if $x_{a_i}(t) < C_{a_i}(t)$, the Lagrange multiplier $\eta_{a_i}(t) = 0$, (ii) if $x_{a_i}(t) = C_{a_i}(t)$, the Lagrange multiplier $\eta_{a_i}(t) \geq 0$. □

4.3 Problem formulation of the DSO-AC and its solution—the DQM case

To investigate the DSO-AC with the DQM as network loading model, we follow the problem formulation proposed by Akamatsu (2007) based on the following proposition (see e.g. Yang and Meng (1998); Akamatsu (2007); Kuwahara (2007); Chow (2009a,b); Varaiya (2008); Shen and Zhang (2009), and the references therein):

Proposition 4.2. At dynamic system optimum with departure time choice, there is no queue in the system. ∎

By this proposition, a possible way to achieve DSO for network modeled by the DQMs (or an alternative way to obtain optimal tolls from bottleneck models) is to constrain the inflow rates to the bottlenecks such that no queue will be formed. To be more specific, we restrict $x_a(t) = 0,\ \forall a \in A,\ \forall t$, which is equivalent to $h_a(t) \leq R_a, \forall a \in A,\ \forall t$. In Akamatsu (2007), the author showed that by designing a network permit system, wherein the permit prices are defined as the penalties associated with the constraints on the inflow rates, the equilibrium resource allocation is efficient (i.e. Pareto optimal) in the sense that the total transportation cost in a network is minimized. The author proved the proposed approach by formulating an equivalent constrained dynamic system optimal traffic assignment problem. He also showed that the feasibility of the constrained dynamic optimal assignment can be easily achieved by constructing a time-space extended network (Yang and Meng, 1998; Akamatsu, 2007). The DSO-AC with DQM as network loading model can be formulated as

$$\min J = \sum_{w \in W} \int_0^T q_w(t)\kappa_B(t)dt + \sum_{a \in A} \int_0^T \psi_a h_a(t)dt, \tag{4.51}$$

subject to

$$\frac{dE_w^B(t)}{dt} = q_w(t),\ q_w(t) \geq 0,\ \forall w \in W,\ \forall t, \tag{4.52}$$

$$E_w^B(T) = Q_w,\ \forall w \in W, \tag{4.53}$$

$$\sum_{k \in NI(i)} h_{k,i}\left(t - \psi_{k,i}\right) - \sum_{j \in NO(i)} h_{i,j}\left(t\right) = \sum_{w \in W} q_w(t)\delta_{i,w}, \tag{4.54}$$

$$0 \leq h_a(t) \leq R_a,\ \forall a \in A,\ \forall t. \tag{4.55}$$

The above DSO-AC with DQM as network loading model is a link-node formulation. The schedule cost function $\kappa_B(t)$ is defined similar to $\kappa\left(\chi\right)$ but is directly related to the arrival time t at destination. $q_w(t)$ is the arrival rate to the destination of OD pair w at time t. $h_a(t)$ is the inflow rate to link a. ψ_a is the free-flow travel time of link a. (4.52)-(4.53) are flow conservation for OD pair w. Equation (4.54) is introduced to ensure the flow conservation for a node [1], say node i, where $NI(i)$ is the set of upstream nodes

[1]Some of the authors, see e.g. Yang and Meng (1998); Nagurney et al. (2007), would ignore (4.54) in their formulations based on the argument that queues will be generally removed or replaced by the time-dependent tolls or network permits. As explained in Yang and Meng (1998), the reason is that in the case of constant exit capacities for all bottlenecks and (4.55), the system throughput is determined by the bottleneck capacities. If a queue was formed on a bottleneck, we could always replace (or remove) it by implementing a tolling or network permit scheme (Akamatsu, 2007) to affect commuters' departure times, thereby reducing the objective value at the system optimum. And hence, queuing is not socially-optimal. Under such circumstances and certain assumptions, the FIFO is satisfied (but may not hold in some special situations (Arnott et al., 1995)) and only flow conservation of OD pairs needs to be considered, i.e. (4.52)-(4.53).

of the links incident to node i and $NO(i)$ is the set of downstream nodes of the links incident from node i. Equation (4.54) is introduced also to ensure the FIFO principle (Akamatsu, 2007). $\delta_{i,w} = 1$ if node i is the destination of OD pair w, $\delta_{i,w} = -1$ if node i is the origin of OD pair w, and zero otherwise. (4.55) is the constraint proposed to restrict the inflow rate to access link a. The optimality condition for (4.51)-(4.55) has been studied by many authors, e.g. Yang and Meng (1998); Akamatsu (2007); Nagae and Sasaki (2009) in terms of mathematical (linear) programming, Nagurney et al. (2007) in terms of evolutionary variational inequality. Hence, we will not derive the optimality condition for this problem in this chapter. The optimal toll to access a bottleneck is obtained by solving the Lagrange multiplier associated with the access constraint imposed on the bottleneck (Yang and Meng, 1998). Rather than implementing the above penalty as toll, the authors identify it as the price for purchasing a permit to access the bottleneck (Akamatsu, 2007; Nagae and Sasaki, 2009), and as the access control (Shen and Zhang, 2009; Zhang and Shen, 2010). The equilibrium resource allocation is efficient (i.e. Pareto optimal) in the sense that the total transportation cost in a network is minimized.

4.4 DSO with access constraints, dynamic externality, and dynamic road pricing

In this section, we will further discuss the results obtained in the previous section. To be specific, we investigate the relations between the dynamic externality, road pricing, and the DSO-AC. We highlight the difference between the dynamic externality obtained from the whole link model and that obtained from the deterministic queuing model rather than their similarity as done in literature, e.g. (Chow, 2007a, 2009a) and the references therein. We also comment on the features of pricing schemes based on the dynamic externalities of these two models.

4.4.1 Consistency between the results on the sensitivity value of the total system travel cost

The first term of the dynamic external cost, i.e. $\Xi_p^1(t) = \frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)} h_p(t)$, is interpreted as the change in the value of the total system travel cost with respect to a perturbation in the path inflow profile h_p at time t. This term is obtained according to the definition of sensitivity in variational calculus (Friesz et al., 2007). This term has a different expression as defined in Chow (2009a), which is stated as $\Xi_p^2(t) = \int_0^T \left(\frac{\partial \Psi_p(s,\mathbf{x}^*)}{\partial u_t} \mid_s h_p(s) \right) ds$. The

term u_t is the perturbation in the inflow profile of path p, which is defined as

$$\frac{dh_p(s)}{du_t} = \begin{cases} 1, & \text{if } s \in [t, t+dt); \\ 0, & \text{otherwise.} \end{cases} \tag{4.56}$$

$\Xi_p^2(t)$ is interpreted as the change in the value of the total system travel cost with respect to a perturbation in the path inflow profile h_p during the time interval $[t, t+dt)$.

To show the consistency between $\Xi_p^1(t)$ and $\Xi_p^2(t)$, we expand $\Xi_p^2(t)$ as

$$\Xi_p^2(t) = \int_0^T \left(\frac{\partial \Psi_p(s, \mathbf{x}^*)}{\partial h_p(s)} \frac{\partial h_p(s)}{\partial u_t} \mid_s h_p(s) \right) ds. \tag{4.57}$$

According to (4.56), $\frac{dh_p(s)}{du_t} = 1$, only if $s = t$ and $dt \to 0$, the right hand side of (4.57) equals to $\frac{\partial \Psi_p(t, \mathbf{x}^*)}{\partial h_p(t)} h_p(t)$. The consistency between $\Xi_p^1(t)$ and $\Xi_p^2(t)$ is then shown. A method to evaluate this term in presence of nested delay operator is first proposed by Balijepalli and Watling (2005), and adopted by Chow (2009a).

4.4.2 Difference between the dynamic externalities of the whole link model and the deterministic queuing model

In this section, we will compare the dynamic externalities of the whole link model and the deterministic queuing model for one link only. The evaluation of the dynamic marginal cost for the deterministic queuing model is depicted in Figure 4.1. Zero free flow time, and constant service rate R_a (in congested condition) are presumed for the bottleneck. The bottleneck inflow rate at time t is denoted as $h_a(t)$. t_1 is the time when a queue is first formed and t_1^f corresponds to the time when this queue is dispersed. Figure 4.1 shows the arrival of an additional unit vehicle at time t_2, whose presence induces an additional delay to the overall queuing delay on this bottleneck. The time $t_1^{f_n}$ is the time when the queue is dispersed in presence of the inflow perturbation at time t_2. In Ghali and Smith (1995), this dynamic marginal delay (where the travel cost is the delay) is defined as the blue solid area in Figure 4.1, which is claimed (without a rigorously mathematical proof) to be equal to the green horizontal area in the same figure, i.e. $t_1^{f_n} - t_2$. The dynamic marginal cost is then obtained by $t_1^{f_n} - t_2$, i.e. the horizontal distance $(t_1^{f_n} - t_2) \times 1$, where 1 denotes the unit vehicle size. But $t_2^e - t_2$ is the link travel time encountered by individual vehicle. Thus the dynamic externality imposed on others is $t_1^{f_n} - t_2^e$

In Kuwahara (2007), based on the sensitivity analysis, he obtain the dynamic marginal cost for the bottleneck as:

$$f_a\left(\theta_a(t_2)\right) + \int_{t_2}^{t_1^{f_n}} \frac{df_a\left(\theta_a(s)\right)}{d\theta_a(s)} \frac{h_a(s)}{R_a} ds, \tag{4.58}$$

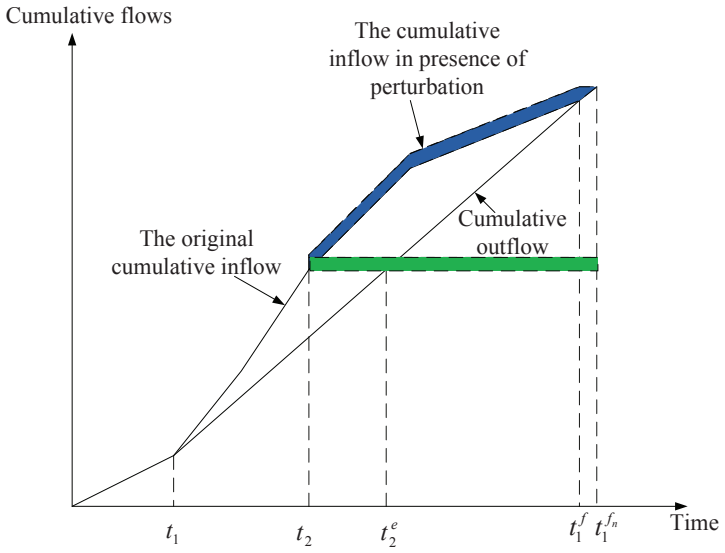


Figure 4.1: Dynamic marginal cost for the deterministic queuing model

where f_a is the travel cost function of the link, $\theta_a(t_2)$ is the queuing (travel) time of a user arriving at the bottleneck at time t_2, i.e. $\theta_a(t_2) = t_2^e - t_2$.

When there is no access constraint, the dynamic marginal cost of the whole link model for one link at time t_2 is given by

$$\Psi_a(t_2, x_a) + \frac{\partial \Psi_a(t_2, x_a)}{\partial h_a(t_2)} h_a(t_2) + \int_{t_2}^{\tau_a(t_2)} \frac{\partial \Psi_a(s, x_a)}{\partial \theta_a(s)} \frac{\partial \theta_a(s)}{\partial x_a(s)} h_a(s) ds, \tag{4.59}$$

where $\theta_a(t) = \tau_a(t) - t$ is the link travel time function. If we adopt linear travel time function, $\frac{\partial \theta_a(s)}{\partial x_a(s)} = 1/R_a$. (4.59) can be evaluated as

$$\Psi_a(t_2, x_a) + \frac{\partial \Psi_a(t_2, x_a)}{\partial h_a(t_2)} h_a(t_2) + \int_{t_2}^{\tau_a(t_2)} \frac{\partial \Psi_a(s, x_a)}{\partial \theta_a(s)} \frac{h_a(s)}{R_a} ds \tag{4.60}$$

To compare the two dynamic marginal cost, we define the following correspondence

$$\Psi_a(t_2, x_a) \triangleq f_a\left(\theta_a(t_2)\right), \tag{4.61}$$

$$\int_{t_2}^{t_1^{fn}} \frac{df_a\left(\theta_a(s)\right)}{d\theta_a(s)} \frac{h_a(s)}{R_a} ds \triangleq \int_{t_2}^{\tau_a(t_2)} \frac{\partial \Psi_a(s, x_a)}{\partial \theta_a(s)} \frac{h_a(s)}{R_a} ds. \tag{4.62}$$

The terms given by (4.61) are the link travel costs for the two models, which are physically equivalent. Comparing (4.58) and (4.60), we notice that one more term $\frac{\partial \Psi_a(t_2, x_a)}{\partial h_a(t_2)} h_a(t_2)$ is introduced in the whole link model. The other difference lies in the upper bounds of the integrals in (4.62). For the bottleneck model, the integral is up to the time the queue is dispersed, while for the whole link model, the integral is up to the time when the vehicle exists from the link. To see the first difference, we write the dynamics of the deterministic queuing model as

$$\dot{x}_a(t) = \begin{cases} h_a(t) - R_a, & \text{if } x_a(t) \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{4.63}$$

The dynamics of the whole link model is

$$\dot{x}_a(t) = h_a(t) - \frac{h_a\left(\varrho_a(t)\right)}{\dot{\tau}_a\left(\varrho_a(t)\right)}, \tag{4.64}$$

where $\varrho_a(t)$ is the inverse function of $\tau_a(t)$, which denotes the entry time to link a that leads to exit from the link at time t (Ban et al., 2008; Chow, 2009a). The dynamics of the deterministic queuing model (4.63) depends on the current system state $x_a(t)$ and the inflow rate $h_a(t)$ only, while the dynamics of the whole link model depends on the current inflow rate $h_a(t)$, the historical inflow rate $h_a\left(\varrho_a(t)\right)$ and historical system state which is included in $\dot{\tau}_a\left(\varrho_a(t)\right)$. That is to say the term $\frac{\partial \Psi_a(t,x_a)}{\partial h_a(t)} h_a(t)$ in the whole link model is introduced by the historical inflow (traced back to time $\varrho_a(t)$) that affects the current system state $x_a(t)$. To be more specific, we expand the term as

$$\frac{\partial \Psi_a(t,x_a)}{\partial h_a(t)} h_a(t) = \frac{\partial \Psi_a(t,x_a)}{\partial \theta_a(t)} \frac{\partial \theta_a(t)}{\partial x_a(t)} \frac{\partial x_a(t)}{\partial h_a(t)} h_a(t). \tag{4.65}$$

The term $\frac{\partial \Psi_a(t,x_a)}{\partial \theta_a(t)}$ depends on the detailed mapping of the cost function, which can be evaluated easily, and $\frac{\partial \theta_a(u)}{\partial x_a(u)} = 1/R_a$. The historical information involves the term $\frac{\partial x_a(t)}{\partial h_a(t)}$. To see this, we expand the system state as $x_a(t) = \int_{\varrho_a(t)}^{t} h_a(s)ds$. Thus,

$$\frac{\partial x_a(t)}{\partial h_a(t)} = \frac{\partial}{\partial h_a(t)} \int_{\varrho_a(t)}^{t} h_a(s)ds, \tag{4.66}$$

which depends on the historical inflow pattern during the time interval $[\varrho_a(t), t]$. A detailed analysis on this term is discussed by Chow (2007a, 2009a).

Remark 4.1. Chow (2007a) showed that the sensitivity of link travel time with respect to the perturbation in the inflow rate can be derived as

$$\begin{aligned} \frac{\partial \theta_a(t)}{\partial u_s} \mid_t &= \frac{\partial \theta_a(t)}{\partial x_a(t)} \frac{\partial x_a(t)}{\partial u_s} \mid_t \\ &= \frac{d\theta_a(t)}{dx_a(t)} \cdot \left(\int_{\varrho_a(t)}^{t} \frac{dh_a(v)}{du_s} dv + g_a(t) \frac{\partial \theta_a(t)}{\partial u_s} \mid_{\varrho_a(t)} \right). \end{aligned}$$

In fact, the term $g_a(t) \frac{\partial \theta_a(t)}{\partial u_s} \mid_{\varrho_a(t)}$ should be zero, otherwise this equation will be recursively traced back to origin (i.e. $t = 0$) in order to calculate the sensitivity value $\frac{\partial \theta_a(t)}{\partial u_s} \mid_t$. To show this, we refer to the proof of this statement in Chow (2007a):

$$\frac{\partial \theta_a(t)}{\partial u_s} \mid_t = \frac{\partial \theta_a(t)}{\partial x_a(t)} \cdot \left(\int_{\varrho_a(t)}^{t} \frac{dh_a(v)}{du_s} dv - h_a\left(\varrho_a(t)\right) \frac{\partial \varrho_a(t)}{\partial u_s} \right).$$

We can show that the term $h_a\left(\varrho_a(t)\right) \frac{\partial \varrho_a(t)}{\partial u_s} = 0$. To show this, we derive $h_a\left(\varrho_a(t)\right) \frac{d\varrho_a(t)}{dt} = g_a(t)$, i.e. $h_a\left(\varrho_a(t)\right) = g_a(t) \left(\frac{d\varrho_a(t)}{dt} \right)^{-1}$ from the definition of $\varrho_a(t)$. Since $\varrho_a(\cdot)$ is the inverse function of $\tau_a(\cdot)$, we have

$$\left(\frac{\partial \varrho_a(t)}{\partial t} \right)^{-1} = \frac{\partial \tau_a\left(\varrho_a(t)\right)}{\partial \varrho_a(t)}.$$

Thus,

$$h_a\left(\varrho_a(t)\right)\frac{\partial \varrho_a(t)}{\partial u_s} = g_a(t)\frac{\partial \tau_a\left(\varrho_a(t)\right)}{\partial \varrho_a(t)}\frac{\partial \varrho_a(t)}{\partial u_s}.$$

Since $\tau_a\left(\varrho_a(t)\right) = t$, the above equation is then $h_a\left(\varrho_a(t)\right)\frac{\partial \varrho_a(t)}{\partial u_s} = g_a(t)\frac{\partial t}{\partial u_s}$. We conclude $h_a\left(\varrho_a(t)\right)\frac{\partial \varrho_a(t)}{\partial u_s} = 0$ from $\frac{dt}{du_s} = 0$. Thus

$$\frac{\partial \theta_a(t)}{\partial u_s}\mid_t = \frac{d\theta_a(t)}{dx_a(t)} \cdot \left(\int_{\varrho_a(t)}^{t} \frac{dh_a(v)}{du_s} dv\right).$$

∎

Next, we will investigate the difference between the upper bounds of integrals in (4.62). Note from Figure 4.1 that the cumulative outflow curve will not change even if the inflow rate has been changed when the bottleneck is congested. This is due to the assumption on the constant service rate when the bottleneck is congested. The additional cost caused by the perturbation in the inflow actually lasts till the time the queue is dispersed as shown in Figure 4.1. However, this is not the case for the whole link model. For comparison, we depict the cumulative inflow and outflow curves for the whole link model, from the time when the perturbation occurred and onward, in Figure 4.2.

From the flow propagation equation of the whole link model, we have

$$g_a\left(\tau_a(t)\right) + \delta g_a\left(\tau_a(t)\right) = \frac{h_a(t) + \delta h_a(t)}{\dot{\tau}_a(t)}. \tag{4.67}$$

A change in the inflow rate at time t, i.e. $\delta h_a(t)$, results in a change in the outflow rate at time $\tau_a(t)$, i.e. $\delta g_a\left(\tau_a(t)\right)$. This implies that a perturbation in the cumulative inflow curve will cause a perturbation in the cumulative outflow curve automatically, see e.g. Figure 4.2. The amount of change in the cumulative outflow curve is given by (4.67). The effect of this perturbation is cumulated up to the time it exits from the link, i.e. $\tau_a(t)$. In this sense, we define it as the dynamic external cost imposed on travelers during their presence on the link.

4.4.3 Marginal cost pricing, access pricing, and the access constraint

Congestion pricing and network permits are recognized as qualitative and quantitative regulations for bottleneck traffic flow control, respectively. For dynamic congestion pricing, the optimal toll is the difference between the dynamic marginal cost imposed on everyone in the system and individual travel cost, i.e. $\frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)}h_p(t) + \sum_{i=0}^{m(p)-1}\int_{t_i^p}^{t_{i+1}^p} \varsigma_{a_{i+1}}^p(t)dt + \sum_{i=0}^{m(p)-1}\int_{t_i^p}^{t_{i+1}^p} \eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p dt$ for the whole link model. In the preceding analysis, we have defined the term

$$\frac{\partial \Psi_p(t,\mathbf{x}^*)}{\partial h_p(t)}h_p(t) + \sum_{i=0}^{m(p)-1}\int_{t_i^p}^{t_{i+1}^p} \varsigma_{a_{i+1}}^p(t)dt \tag{4.68}$$

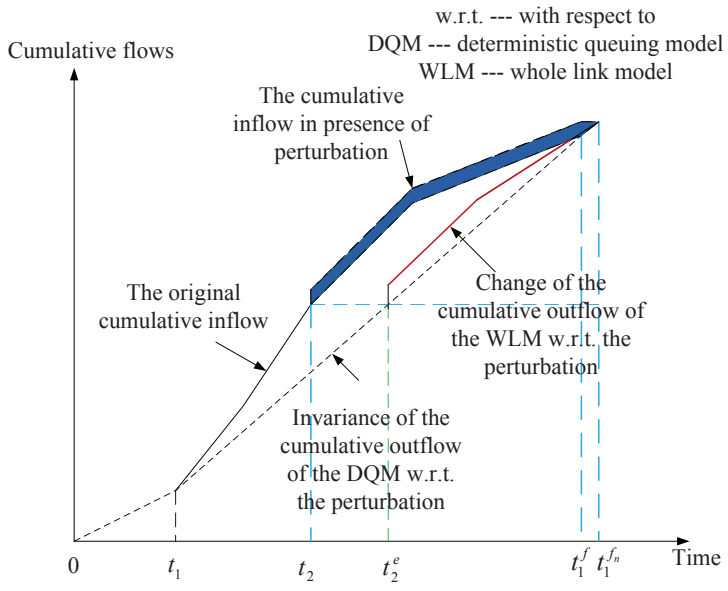


Figure 4.2: Comparison of the dynamic marginal costs for the two models

as the toll for dynamic marginal cost pricing, see e.g. (Chow, 2009a), and the term

$$\sum_{i=0}^{m(p)-1} \int_{t_i^p}^{t_{i+1}^p} \eta_{a_{i+1}}(t)\sigma_{a_{i+1}}^p dt \tag{4.69}$$

as the toll for access pricing.

4.4.3.1 Different toll structures for the whole link model and the deterministic queuing model

By the analysis in Section 4.4.2, the toll structures can be also different for these two models. For the whole link model, the toll for dynamic marginal cost pricing, i.e. (4.68), is nonnegative as long as the path inflow h_p is positive, while the toll for access pricing, i.e. (4.69), can be positive or zero depending on whether the access constraints are activated or not. However, this is not the case for the deterministic queuing model wherein both tolls for the marginal cost pricing and access pricing can be zeros. To see this, note from (4.58) that the toll for marginal cost pricing is zero if there is no queue on the bottleneck. The access price is zero if the access constraint is not activated.

Based on Proposition 4.2, we can apply dynamic marginal cost (usage) pricing only, or access pricing only, or both usage and access pricing to achieve a dynamic system optimal for networks modeled by the DQM. Different methods have been proposed to make a traffic network operating under its system optimum, e.g. dynamic marginal cost pricing (Arnott et al., 1993; Kuwahara, 2007; Chow, 2009a), tradable network permit (Akamatsu, 2007; Nagae and Sasaki, 2009), access control (Shen and Zhang, 2009), and congestion

derivative (Yao et al., 2010).

4.4.3.2 Some advanced issues

Theoretically, the optimal toll is to impose an dynamic toll that is equivalent to the queueing cost such that the queueing delay is completely eliminated. The temporal equilibrium can be maintained after the implementation of this toll, since users pay the same amount of additional cost as the queueing cost (Arnott et al., 1993; Kuwahara, 2007). However, the optimal toll is difficult to obtain in general. It is claimed that, the evaluation of (4.68) is difficult in both computational and analytical aspects (Balijepalli and Watling, 2005; Chow, 2009a). Similar viewpoint has been reported by Kuwahara (2007) for the evaluation of the second term of (4.58) for the deterministic queuing model. Closed-form solutions can be obtained for some special cases only, e.g. for signal bottleneck case (Arnott et al., 1993; Kuwahara, 2007; Yao et al., 2010). Note that when the WLM is applied as the network loading model, the toll for marginal cost pricing is path based, i.e. (4.68). Another difficulty in evaluating (4.68) is introduced by the flow propagation and the nested delay operator for dynamic traffic assignment. On the contrary, the toll for access pricing is obtained in a link based manner, e.g. (4.69). Compared with the marginal cost pricing, it is much easier to manipulate.

In static nonlinear pricing, the access pricing is found to be profitable to implement regardless of the severity of congestion. The profits from marginal cost (usage) only pricing and access-only pricing cannot be ranked in general. Access-only pricing may be more profitable if congestion is not too severe and demand is not too elastic. If access pricing is implemented, and congestion is moderately severe, it may be profitable to set the access fee high enough to exclude some travel demand in order to provide a better quality of service by reducing congestion (Wang et al., 2010).

As the toll for marginal cost pricing is positive whenever the inflow is positive, the dynamic pricing problem is more complex for the whole link model than the bottleneck model. For the WLM case, applying access pricing (4.69) only may not yield a DSO state. However, if the system planner specify a preferred (target DSO) state, i.e. $\mathbf{x}^*$, in advance, the problem reduces to a dynamic user equilibrium with side constraint (DUE-SC) problem proposed by Zhong et al. (2010). The tolls for access pricing to achieve this preferred (target DSO) state can be obtained by solving the DUE-SC problem. As the tolls for access pricing are link based, it is easy to manipulate. Since the optimal traffic state $\mathbf{x}^*$ is known in advance, for each departure time, the schedule delay cost can be calculated. For a given path and departure time, the entry time and exit time of a certain link can

also be obtained. The "tolls" obtained from solving the DUE-SC can be also traded in a competitive market similar to the tradable network permit case. This scheme maybe more efficient than on-line real-time pricing for achieving the preferred (target DSO) state $\mathbf{x}^*$.

4.5 Traffic induced air pollution pricing as a special case of the access pricing

As previously explained, to achieve a more sustainable mobility for urban areas, many governments would like to improve or maintain the air quality of the urban area by controlling the traffic volume in the network. Road pricing, as proposed to internalize various externalities caused by vehicle use in terms of tolling, should include the external cost caused by traffic induced air pollution as it has also been argued in the literature. An extensive amount of research on road pricing for (deterministic/stochastic) static traffic network have concluded that road pricing is an efficient approach to internalize externalities such as congestion, air pollution, noise, and accidents (Sumalee and Xu, 2010; Yang et al., 2010). In this section, based on the "equivalent" environmental traffic capacity concept, that converts the environmental constraints into restrictions on link traffic volumes, and a dynamic (time-dependent) traffic induced air pollution dispersion models, we will derive the dynamic pollution pricing as a special case of the access pricing. To be precise, we will address the following problems summarized at the end of Section 4.1 and repeated here for convenience:

> Which pricing scheme should be imposed on each link, the access control based or the environmental traffic capacity based pricing scheme?
>
> Which is the dominant pricing scheme at a specific time instant?
>
> How to determine the boundaries under which a traveler on a link should pay either an access toll or an extra pollution charge?

To begin with, let us review the two capacity concepts.

4.5.1 The traffic capacity and the environmental traffic capacity

The traffic capacity (HCM, 2000) as defined in transportation engineering means the capacity of a facility, which is the maximum hourly rate at which persons or vehicles can be reasonably expected to traverse a point or a uniform section of a lane or roadway during a given time period under the prevailing roadway, traffic, and control conditions. The

concept of environmental traffic capacity was first introduced by Buchanan (1963) wherein the capacity of a street (or an area) to accommodate vehicles under certain restrictions of environmental standards is defined as the environmental traffic capacity. Holdsworth and Singleton (1979, 1980) refined this definition as: "the maximum number of vehicles that may pass along a street in a certain period of time, under fixed physical conditions, without causing environmental detriment." Shiran (1997) further extended the above definition to an area wide environmental traffic capacity which can be stated as "the maximum amount of traffic activities that may occur in a given area during a certain period of time, under fixed physical conditions, without causing environmental detriment." Despite the difference between the microscopic (street level) and macroscopic (area wide level) environmental traffic capacity concepts, what they have in common is to control the traffic volume to restrict the traffic induced air pollution to an appropriate level defined by some physical and social economic conditions during certain time periods in some particular areas.

A review on various definitions of the two traffic capacity concepts is given by Li et al. (2009). In that paper, the authors try to find a connection between them and to utilize them to study the control of traffic induced air pollution. The authors combine the two concepts to study the relationship between transportation systems and the natural environment system. To control the traffic induced pollution, they first define the traffic-related environmental capacity (TREC) which is a limit to the traffic induced air pollution in a certain time under the physical and social economic conditions in a particular area. Based on the TREC, they calculate the maximum traffic volume can be accommodated by the network. The resulting maximum traffic volume is defined as the environmental traffic capacity (ETC). Rather than combining the two concepts, we try to sperate them for the dynamic road pricing purpose. We show in this chapter that the environmental traffic capacity restriction can be regarded as a special case of the access control. The traffic induced pollution pricing can be then included by the access pricing.

4.5.2 A brief review of the dynamic traffic pollution dispersion models

The study on the wind flow and air pollutant dispersion inside and over urban street canyons has attracted great concern during the past two decades mainly due to the increasing emission of air pollutants in urban areas and their adverse impacts on human health as mentioned in the introduction and Nagurney (2000). With the economic development, there are more and more high-rise buildings surrounding streets in urban areas, making the environment more and more unfavorable for traffic emissions to disperse. It is known that

both the flow and air pollutant dispersion inside street canyons are dominated by turbulent processes (Li, 2008). There are three main approaches to street-canyon-pollution research including field observations and measurements, and two kinds of numerical pollutant dispersion models, i.e. operational models, and computational fluid dynamics (CFD) models. Two reviews of these models are given by Vardoulakis et al. (2003, 2007). The dispersion of pollutants on street canyons depends on the following street canyon characteristics:

1. Street-canyon geometry;
2. Ambient wind flow;
3. Traffic-induced turbulence;
4. Pollutant transformation;
5. Population exposure.

To our purpose, we concentrate on the first three aspects in this brief review.

4.5.2.1 Street-canyon geometry

The wind flow pattern inside street canyons depends on their geometry, in particular the building-height-to-street-width (aspect) ratio (or AR) (h/b, where h is the building height and b is the street width). Based on the aspect ratio, a street canyon can be classified into three main categories: (i) regular street canyon with $0.7 < AR < 1.5$ (some authors proposed this to be $0.7 < AR < 2$), (ii) deep (or narrow) street canyon with $AR > 1.5$, (or $AR > 2$) and (iii) low-rise street canyon with $AR < 0.7$. The length L of the canyon is usually defined as the road distance between two major intersections. Similar to the AR, we can define street canyons into short $L/h \approx 3$; medium $L/h \approx 5$; and long canyons $L/h \approx 7$. Urban streets might be also classified in symmetric (or even) canyons, if the buildings flanking the street have approximately the same height, or asymmetric, if there are significant differences in building height (Vardoulakis et al., 2003).

4.5.2.2 Ambient wind flow

As indicated in Figure 4.3, the street canyon wind flow is determined by a wind vortex which is an interaction between the above roof-top wind conditions and the local wind flow within the cavity of the canyon. In the case of perpendicular flow, the up-wind side of the canyon is usually called leeward, and the downwind windward. Based on field measurements and mathematical modeling results, Oke (1988) identified three flow regimes

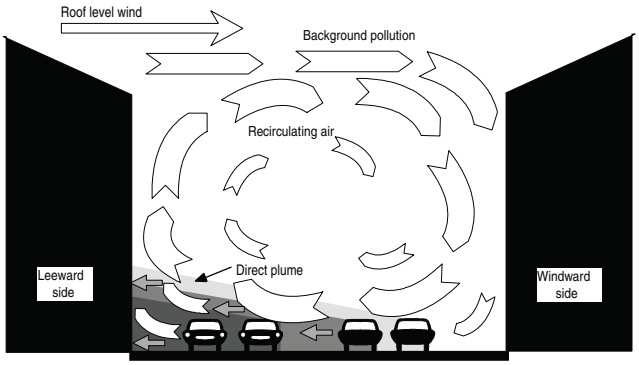


Figure 4.3: Schematic illustration of flow and dispersion conditions in street canyons (Berkowicz et al., 2008)

for wind direction perpendicular to the street axis, see e.g. Figure 4.4. For wide canyons ($AR < 0.3$), the flow fields associated with the buildings do not interact, which results in the isolated roughness flow (IRF) regime. As buildings become more closely spaced ($0.3 < AR < 0.7$), the disturbed wind flow behind the upwind building is disturbed by the recirculation created in front of the windward building. This is the wake interference flow (WIF) regime. By further reducing the spare space between buildings ($AR > 0.7$), the bulk of the synoptic flow skims over the canyon which results in the skimming flow (SF) regime. In this case a stable recirculation is developed inside the canyon and the ambient flow is decoupled from the street flow. Under this circumstance, the vehicular pollutants at the street level could not be easily ventilated and would accumulate, resulting in high pollutant concentration and poor air quality. The strength of the wind-induced recirculations inside street canyons mainly depends on the strength of roof-level wind and is also influenced mildly by the mechanical turbulence induced by moving vehicles.

The strength of the wind vortices inside the canyon mainly depends on wind speed at roof-top level. The local wind flow is also affected by the mechanical turbulence induced by moving vehicles and the roughness elements within the street, e.g. trees, kiosks, balconies, slanted building roofs, etc. (Vardoulakis et al., 2007). In relatively deep canyons $AR \approx 1.3$, the main wind vortex is usually displaced towards the upper part of the cavity. As the aspect ratio increases $AR \approx 2$, a weak counter-rotating secondary vortex maybe observed at street level. For even higher aspect ratios $AR \approx 3$, a third weak vortex might be also formed.

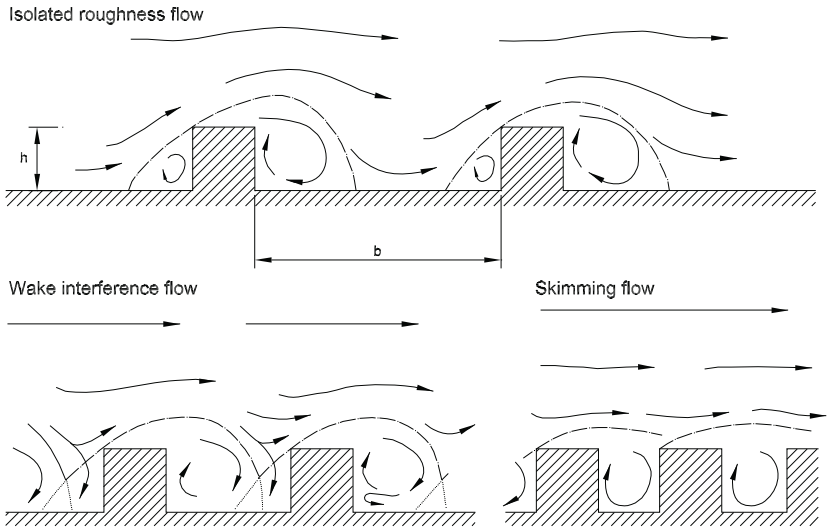


Figure 4.4: Three flow regimes associated with different building-height-to-street-width ratios h/b (Oke, 1988)

4.5.2.3 Numerical traffic-induced pollution dispersion models–CFD based models

As mentioned, there are mainly two categories of numerical traffic dispersion models, i.e. the CFD based models and operational models. In the current section, we give a brief review on the CFD based models. A CFD based traffic-induced pollution dispersion model consists of two functional blocks: a numerical wind field model, and a numerical pollutant transport model.

Regarding to the wind field model, Reynolds-averaged Navier-Stokes (RANS) models, e.g. the $k-\epsilon$ turbulence closure schemes, are extensively employed to simulate turbulence in urban street canyons. However, these models have some inherent limitations, e.g. they cannot handle the unsteadiness and intermittency of the street-canyon flow and air pollutant dispersion, as the nature of RANS models is a steady-state methodology (Li, 2008). Therefore, RANS models cannot accurately predict the transient wind field and hence are unable to precisely model the turbulent pollutant transport processes. The empirical evidence reveals that there are complicated processes beyond the reach of RANS models. To this end, large-eddy simulation (LES) has been recently proposed and become a popular approach to investigate the turbulence inside street canyons hand in hand with advances in computer science and technology. Compared with the RANS modes, the major advantages of LES are its capabilities of handling the unsteadiness and intermittency of the flow as well as providing detailed information on the turbulence structure. The accuracies of the LES technique are, however, greatly affected by the negligence of transport processes

in the spanwise direction. Moreover, calculations of the LES were solely numerical without support from empirical validation.

CFD based dispersion models mainly employ two approaches: Lagrangian approach and Eulerian approach. The main difference of these two approaches is the choice of coordinate (Li, 2008). The concentration of a pollutant is described for a particular fluid element as it travels with the flow in the Lagrangian approach. In this approach, the coordinates are dependent variables, and the fluid element is identified by its position in the field relative to that at some arbitrary time. However, in the Eulerian approach, the concentration of a pollutant is described at a given position and time, with the coordinates fixed in space and time. Therefore, in the Eulerian approach, the coordinate position and time are independent variables. Conventionally, the Lagrangian approach employs particle tracking methods to simulate the dispersion characteristics of pollutants in a variety of fluid flow fields. Such kind of models are further refereed as the Lagrangian particle dispersion (LPD) models which are also known as random walk models, random flight models, or Lagrangian Monte Carlo models. In the LPD approach, the turbulent transport is modeled by tracing the trajectories of a large number of particles as they are transporting with the air flow (or driven by the wind field), which is generated in-prior by a numerical wind field model. The release of particles may be either sequential (as a plume) or simultaneous (as a puff). Concentration fields are determined from the spatial distribution of particles. Note that the LPD models are useful when the time dependent wind field data can be obtained.

Compared with the Lagrangian approach, the Eulerian models are more appropriate for describing long-range transport with chemical reactions and transformations. To simulate the pollutant dispersion these models solves an advection-diffusion equation of conserved scalars (e.g. mean concentration or mass fraction) for a set of receptors in 2D or 3D computational domains. Eulerian models can also address the production and loss terms, which may include exchanges with the surrounding grid elements, emissions, chemical transformations, and dry and wet deposition (Li, 2008). Despite the expensively computational effort, the CFD based models contain numerous empirical parameters that may lead to uncertainties in the modeling accuracy. The comparisons did not always indicate that results from CFD are more accurate than those from the simpler operational models, e.g. the Operational Street Pollution Model (OSPM).

4.5.2.4 Numerical traffic-induced pollution dispersion models–operational models

The operational models include box models, Gaussian plume models, operational street pollution model, and Lagrangian puff models, etc. Operational models usually need some empirical or semi-empirical parameters from observation and make several crude simplifications. Despite their inaccuracy, they are very useful for environmental monitoring and air quality assessment, where a quick and reasonable estimation of pollutant concentration is required. Reviews on the performances of the operational models are given by Vardoulakis et al. (2003, 2007). Among the various operational models, the OSPM is one of the most popular traffic-induced pollution dispersion models. An huge amount of research works has been dedicated to calibrate and validate this model. It is claimed that when the aspect ratio $AR < 3$, there is no direct evidence that the CFD based models performs better than the OSPM (Li, 2008)[2]. To this end, in this chapter we apply the OSPM to model the concentration of traffic induced pollution.

The OSPM is a semi empirical dispersion model which combines Gaussian plume theory with empirical box model techniques to calculate concentration of gas pollutants in a street canyon say link a. The model assumes three different contributions to the concentration of pollutants on link i at time t: the direct impact of pollutants from the source to the receptor $\mathcal{C}_{i,d}(t)$ at time t, the recirculation component $\mathcal{C}_{i,r}(t)$ due to the flow of pollutants around the horizontal vortex generated within the recirculation zone of the canyon, and the urban background contribution $\mathcal{C}_{i,b}(t)$. The overall concentration of pollutants on link i at time t is given by

$$\mathcal{C}_i(t) = \mathcal{C}_{i,d}(t) + \mathcal{C}_{i,r}(t) + \mathcal{C}_{i,b}(t). \tag{4.70}$$

The emission density for a line source[3] is

$$d\mathcal{Q}_i(t) = \frac{\mathcal{Q}_i(t)}{b_i} dy,$$

where the emission field is treated as a number of infinitesimal line sources aligned perpendicular to the wind direction at the street level and with thickness dy, b_i is the width of

[2]Some authors also claim that the OSPM only performs well when $AR < 2$ or even when $AR \approx 1$. However, the performance of such kind of model depends on the calibration (or modification of the model to adapt the study area) and the climate (mainly the wind speed and direction) of the study area. For example the wind speed in Hong Kong is faster than those of some inland cities in mainland China or some Mediterranean dense cities. The aspect ratios for the well performance of the OSPM would be quite different.

[3]A line source is a source of roadway air pollution that emanates from a linear (one-dimensional) geometry, e.g. a link.

the canyon, and $\mathcal{Q}_i(t)$ is the emissions from vehicles on link i at time t. The contribution to the concentration at a point located at a distance y from the line source is given by,

$$d\mathcal{C}_{i,d}(t) = \sqrt{\frac{2}{\pi}} \frac{d\mathcal{Q}_i(t)}{r_i \sigma_{z,i}(y)}, \tag{4.71}$$

where r_i is the street-level wind speed, $\sigma_{z,i}(y)$ is the vertical dispersion parameter at a downwind distance y. To evaluate the concentration of the pollutants induced by the traffic, (4.71) is integrated along the wind path at the street level. The integration path depends on wind direction, extension of the recirculation zone and the street length. Conventionally, under some assumptions, the integration can be approximated by a Gaussian plume model

$$\mathcal{C}_{i,d}(t) = \sqrt{\frac{2}{\pi}} \frac{\mathcal{Q}_i(t)}{b_i \sigma_{w,i}} \varpi_{i,1} \triangleq \mathcal{Q}_i(t) \tilde{\varpi}_{i,1}, \tag{4.72}$$

where $\sigma_{w,i}$ is the standard deviation of the vortex plume velocity (also refereed to as the vertical velocity fluctuation due to mechanical turbulence generated by wind and vehicle traffic in the street) which is thus defined

$$\sigma_{w,i} = \sqrt{(\alpha r_i)^2 + \sigma_{wo,i}^2}, \tag{4.73}$$

where α is a proportionality constant (given the empirically value as 0.1), and $\sigma_{wo,i}$ is the traffic-induced turbulence defined as

$$\sigma_{wo,i} = \vartheta \left(\frac{N_i \bar{v}_i S^2}{b_i} \right)^{1/2},$$

where ϑ is an aerodynamic drag coefficient (given empirically the value of 0.3), N_i the number of vehicles using the street per time unit, $\bar{v}_i$ the average vehicle speed, S^2 the road surface occupied by a single vehicle. $\varpi_{i,1}$ is a parameter to be calibrated, which depends on the (average) height of the roadside buildings, the street-level wind speed, the (average) height of vehicle exhaust, the standard deviation of velocity at vortex roof level, the correction factor for low wind, the angle between street and wind, etc. The reader can refer to Berkowicz (1998) for the details on how to calibrate this parameter.

The contribution from the recirculation is calculated using a simple box model, which assumes that the pollutants are well mixed inside the box:

$$\mathcal{C}_{i,r}(t) = \frac{\mathcal{Q}_i(t)}{b_i} \varpi_{i,2} \triangleq \mathcal{Q}_i(t) \tilde{\varpi}_{i,2}, \tag{4.74}$$

where $\varpi_{i,2}$ is another parameter to be calibrated, which depends on the geometry of the canyon, the ventilation velocity of the canyon, the roof-level wind speed, the extension of the recirculation zone, the relation between street and roof-level winds, etc. Detailed description on these parameters can be found in Berkowicz (1998, 2000a,b); Vardoulakis et al. (2002, 2003).

Remark 4.2. Strictly speaking (Vardoulakis et al., 2002), concentrations should be calculated as the sum of the direct and recirculation contributions on the leeward side of the street, while on the windward side only the direct contribution of emissions generated outside the recirculation zone need to be taken into account. If the recirculation zone extends throughout the whole canyon, then the windward concentrations are calculated from only the recirculation component. For near parallel flow, emissions from outside the recirculation zone may contribute to the leeward concentrations. When the wind speed is near zero or parallel to the street axis, the concentrations on both sides of the canyon become equal. In all cases, the background contribution should be added to obtain the final result. ∎

Remark 4.3. The OSPM is designed to produce time series of pollutant concentrations within street canyons, which requires calibration of several model parameters, an amount of input information and computational resources. On the other hand, they are based on a number of empirical assumptions and parameters that might not be applicable to all urban environments, e.g. very deep street canyons with high wind speed. In the introduction of OSPM, we implicitly assume the model parameters are constant, i.e. the time-average values. Nevertheless, these model parameters can be also calibrated as time-dependent values, e.g. with respect to different time-of-day. ∎

4.5.3 Environmental traffic capacity constraint and pollution charge

Since we have applied the whole link model and the deterministic queueing model, which do not distinguish the leeward side and the windward side, we just calculate the concentration of the pollutants by (4.70). Moreover, since we consider the traffic-induced air pollution only, we assume in this chapter that the restriction on the maximal concentration ($\mathcal{C}_i^e$) of pollutants that can be afforded by link i is separated from the background contribution, i.e. $\mathcal{C}_{i,d}(t)+\mathcal{C}_{i,r}(t) \leq \mathcal{C}_i^e$. Given the number of vehicles $V_i(t)$ on link i at time t, the length of the link l_i, and the weighted mean of vehicular emission factors $\bar{E}$, the rate of release of emissions from vehicles on link i at time t can be obtained by

$$\mathcal{Q}_i(t) \approx \frac{\bar{v}_i(t)}{l_i} V_i(t) \bar{E}.$$

The constraint $\mathcal{C}_{i,d}(t)+\mathcal{C}_{i,r}(t) \leq \mathcal{C}_i^e$ implies

$$V_i(t) \leq \frac{\mathcal{C}_i^e}{\bar{E}\left(\tilde{\varpi}_{i,1}+\tilde{\varpi}_{i,2}\right)} \frac{l_i}{\bar{v}_i(t)}. \tag{4.75}$$

Definition 4.1. The environmental traffic capacity of link i, i.e. κ_i^e, is defined as

$$\kappa_i^e = \frac{\mathcal{C}_i^e}{\bar{E}\left(\tilde{\varpi}_{i,1}+\tilde{\varpi}_{i,2}\right)}.$$

4.5.3.1 The whole link model case

For the whole link model, the number of vehicles $V_i(t)$ on link i at time t is the link traffic volume at time t, i.e. $x_i(t)$. (4.75) is equivalent to

$$x_i(t) \leq \kappa_i^e \frac{l_i}{\bar{v}_i(t)}.$$

By approximating $\frac{l_i}{\bar{v}_i(t)}$ by the link travel time $\theta_i(t)$, the above equation is then

$$x_i(t) \leq \kappa_i^e \theta_i(t).$$

As we have adopted the linear travel time function $\theta_i(t) = \psi_i + x_i(t)/R_i$, the above equation can be further represented as

$$x_i(t) \leq \frac{\kappa_i^e \psi_i}{\left(1 - \frac{\kappa_i^e}{R_i}\right)}. \tag{4.76}$$

By the access control, we have

$$x_i(t) \leq C_i(t). \tag{4.77}$$

Now we are ready to state the following proposition:

Proposition 4.3. If $\kappa_i^e \geq R_i$, no pollution charge should be implemented. Otherwise, we have the following scenarios:

If

$$\frac{\kappa_i^e \psi_i}{1 - \frac{\kappa_i^e}{R_i}} > C_i(t), \tag{4.78}$$

the constraint for the access control, i.e. (4.77), will be violated before the violation of (4.76), which implies that the Lagrange multiplier associated with the environmental constraint (4.76) is zero. If (4.78) hold $\forall t \in [0, T]$, then we do not need to charge pollution price.

If

$$\frac{\kappa_i^e \psi_i}{1 - \frac{\kappa_i^e}{R_i}} \leq C_i(t), \tag{4.79}$$

the constraint (4.76) for the pollution charge will be violated before the violation of access constraint (4.77), which implies that the Lagrange multiplier associated with the access control constraint is zero. If (4.79) hold $\forall t \in [0, T]$, then no access price will be charged.

In any case, the Lagrange multipliers associated with (4.76) and (4.77) will not be positive simultaneously for any time instance $t \in [0, T]$. ∎

Proof of Proposition 4.3. The proof of this proposition is straightforward. If $\kappa_i^e \geq R_i$, we have $1 - \frac{\kappa_i^e}{R_i} \leq 0$. No violation of (4.76) will be observed, which implies that the pollution charge is not required. The other two scenarios can be proven by comparing the amplitudes of the two constraints, i.e. the right hand sides of (4.76) and (4.77). □

Remark 4.4. Although the Lagrange multipliers associated with (4.76) and (4.77) cannot be positive simultaneously, the integration of either of them over a time interval can. Therefore, the access price and pollution price, defined by

$$\int_t^{\tau_i(t)} (\eta_i^{ac}(s) + \eta_i^{ap}(s))\, ds,$$

where $\eta_i^{ac}(s)$ denotes the Lagrange multiplier associated with the access control constraint (4.77) and $\eta_i^{ap}(s)$ the Lagrange multiplier associated with the pollution constraint (4.76), can exist simultaneously. ∎

4.5.3.2 The deterministic queueing model case

In the formulation for access control for the network with deterministic queueing model, we restrict the inflow to the bottleneck to be less or equal to its capacity, i.e.

$$h_i(t) \leq R_i. \tag{4.80}$$

We restrict the concentration of traffic induced pollutants on link i to be less than or equal to the environmental constraint, i.e. (4.75). As we impose (4.80), no queue will be formed on the link. The traffic volume on the link can be evaluated as $V_i(t) = h_i(t)\psi_i$. (4.75) can be then represented as

$$h_i(t) \leq \kappa_i^e. \tag{4.81}$$

Similar to the whole link model case, we have the following proposition:

Proposition 4.4. For the deterministic queueing model case, we have the following two scenarios:

> If $\kappa_i^e \geq R_i$, the constraint for the access control, i.e. (4.80), will be violated before the violation of (4.81), which implies that the Lagrange multiplier associated with the environmental constraint (4.81) is zero. In this case, we do not need to implement pollution charge.

If $\kappa_i^e < R_i$, the constraint (4.81) for the pollution pricing will be violated before the violation of (4.80), which implies that the Lagrange multiplier associated with the access control constraint is zero. In this case, no access price should be charged.

In any case, the pollution charge and the access charge will not be activated simultaneously.

∎

For both the WLM and DQM cases, if the link environmental capacity κ_i^e is greater than or equal to the corresponding traffic capacity R_i, then no pollution pricing should be implemented. The boundary of the two pricing schemes is given by $\min(R_i, \kappa_i^e)$ for the DQM case. For the WLM case, another boundary is defined by $\min\left\{\frac{\kappa_i^e \psi_i}{1-\frac{\kappa_i^e}{R_i}},\ C_i(t)\right\}$, that determines which Lagrange multiplier is the positive. From (4.76) and (4.81), we can conclude that the pollution based pricing is, in fact, a special case of the access control.

4.6 Solution algorithm for the DSO-AC with the WLM as network loading model

As explained in Section 4.3, when the DQM is applied as the network loading model, the DSO-AC has been studied by several authors. Several solution algorithms are available to access the problem, e.g. linear programming, evolutionary variational inequality, etc. Therefore, in this section, we study the solution algorithm for the DSO-AC with the WLM as network loading model case only.

4.6.1 Reformulation of the optimal control problem and functional approximation

We represent the optimal control formulation of the DSO-AC in Section 4.2 in a compact form as depicted in Figure 4.5. In the figure, $\Psi(t, \mathbf{x}(\mathbf{h})) = (\Psi_p(t, \mathbf{x}) : \forall p \in P)$ is the effective network delay operator, which is a column vector. In Figure 4.5, $\dot{\mathbf{x}} = \tilde{f}(\cdot)$ denotes the link traffic dynamics with the flow propagation constraints (4.2)-(4.5). We write $\dot{\mathbf{x}}$ explicitly because the link traffic volumes are the state variables of the optimal control problem. The flow propagation constraints (4.4)-(4.5) with travel time function defined by (4.14) are substituted into (4.2)-(4.3) to update the state variables. The equality constraint $M_1(\cdot)$ denotes the flow conservation equations (4.6)-(4.7) or (4.12). The inequality constraint $M_2(\cdot)$ denotes the access constraints defined by (4.9). $\mathbf{v} \geq 0$ is a vector representation of (4.8).

The optimal control problem for the DSO-AC

Solve

$$\min_{\mathbf{h}} J = \int_0^T \Psi\left(t, \mathbf{x}(\mathbf{h})\right)^T \mathbf{h}(t) dt \triangleq \int_0^T F\left(\mathbf{x}, \mathbf{h}\right) dt,$$

subject to

$$\dot{\mathbf{x}} = \tilde{f}(\mathbf{x}, \mathbf{h}, t),\ M_1(\mathbf{h}, t) = 0,\ M_2(\mathbf{x}, t) \leq 0,\ -\mathbf{h} \leq 0, \forall t \in [0, T],\ \mathbf{x}(0) = 0$$

Figure 4.5: optimal control formulation of the DSO-AC in a compact form

To obtain the necessary condition of the DSO-AC, we need to solve the optimal control problem. To the best of our knowledge, no effective algorithm has been reported to solve the optimal control problem with state dependent time lags and state constraints. However, several effective algorithms have been proposed to solve the optimal control problem with state and/or control constraints using nonlinear programming algorithms, such as sequential quadratic programming (SQP) (Buskens and Maurer, 2000; Betts, 2010; Subchan and Żbikowski, 2009). To solve this optimal control problem by nonlinear programming algorithms, it is necessary to approximate the functional differential equations (FDEs), which govern the traffic dynamics, by ordinary differential equations (ODEs). We use a typical method for handling transportation delays in control engineering to deal with these state dependent time lags. The idea is to use polynomial approximation of the time lags. A similar idea has been proposed by Astarita (1996) and Friesz and Mookherjee (2006). We use Padé approximation to approximate the state dependent time lags, which can be easily implemented using Matlab and Simulink.

After approximating the FDEs by ODEs, we denote the traffic dynamics as $\dot{\mathbf{x}}(t) = f_0(\mathbf{x}, \mathbf{v}, t)$. It is convenient for us to rewrite the optimal control problem as

$$\min_{\mathbf{h}} J = \int_0^T F\left(\mathbf{x}, \mathbf{h}\right) dt, \tag{4.82}$$

subject to

$$\dot{\mathbf{x}}(t) = f_0(\mathbf{x}, \mathbf{h}, t),\ M_1(\mathbf{h}, t) = 0,\ M_2(\mathbf{x}, t) \leq 0,\ -\mathbf{h} \leq 0,\ \forall t \in [0, T], \mathbf{x}(0) = 0. \tag{4.83}$$

4.6.2 Solution algorithm for the DSO-AC

To apply an off-the-shelf nonlinear optimization algorithm to the optimal control problem (4.82)-(4.83), it is necessary to apply the time-discretization scheme to the problem. Here we present the recursive discretization approach based on Euler's method, which is commonly used in the literature (Buskens and Maurer, 2000; Betts, 2010; Subchan and

Żbikowski, 2009). The planning time interval is divided into $N-1$ segments uniformly, i.e., a fixed time step Δt is defined in the discretization as $\Delta t = \frac{T}{N-1}$, $t_l = (l-1)\Delta t$, $l = 1, 2, \cdots, N$. Applying Euler's method to the differential equation in (4.83) yields

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \Delta t \cdot f_0(\mathbf{x}_l, \mathbf{h}_l, t_l), \quad l = 1, 2, \cdots, N-1. \tag{4.84}$$

Define the optimization variable $\mathbf{z} = [\mathbf{h}_1, \ \cdots, \ \mathbf{h}_N]^T$, and compute the state variables from (4.84) recursively as $\mathbf{x}_l = \mathbf{x}_l(\mathbf{z}, t) = \mathbf{x}_l\left(\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{l-1}, t_{l-1}\right), \ l = 2, \cdots, N$, which are functions of the control variables with initial condition $\mathbf{x}_1 = \mathbf{x}(1) = 0$. The following NLP is defined

$$\min_{\mathbf{z}} y(\mathbf{z}) \triangleq \sum_{l=1}^{N-1} \Delta t \cdot F\left(\mathbf{x}, \mathbf{h}_l\right), \tag{4.85}$$

subject to

$$\begin{aligned} &\mathbf{x}_l = \mathbf{x}_l\left(\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{l-1}, t_{l-1}\right), \ l = 2, \cdots, N, \ \mathbf{x}_1 = 0, \\ &M_1(\mathbf{h}_l, t_l) = 0, \ M_2(\mathbf{x}_l, t_l) \leq 0, \ -\mathbf{h}_l \leq 0, l = 1, 2, \cdots, N. \end{aligned} \tag{4.86}$$

After the the optimal control problem is reformulated as an NLP problem, we can apply nonlinear optimization algorithms to solve it (Buskens and Maurer, 2000; Betts, 2010; Subchan and Żbikowski, 2009).

4.7 Numerical example

Consider a network with a single OD pair connected by two parallel links as shown in Figure 4.6. The link delay functions are given by $D(x_1) = 0.28 + x_1/70$ unit-times and $D(x_2) = 0.3 + x_2/140$ unit-times, respectively. The overall travel demand is $J_{od} = 100$ units. The planning horizon is 6 unit-times. The desired arrival time is set as $t_{da} = 3$ unit-time. We consider the following early/late arrival penalty function:

$$\kappa[\chi] = \begin{cases} 0.1\left(t + D(x) - t_{de}\right)^2, & t < t_{de}, \\ 0, & t_{de} \leq t \leq t_{dl}, \\ 0.1\left(t + D(x) - t_{dl}\right)^2, & t > t_{dl}, \end{cases} \tag{4.87}$$

where $t_{de} = 2$ unit-time and $t_{dl} = 5$ unit-time. The effective delay is given by (4.16).

We first solve the DSO for this network. Figure 4.7(a) depicts the path (link) inflow against the corresponding travel cost over time. By internalizing the dynamic externality to the dynamic marginal cost, the DSO can be represented as a generalized DUE condition (4.18) in which, for each origin-destination pair, all travelers have the same dynamic

marginal cost for all routes and departure time windows. The figure shows that the dynamic marginal costs are the same for the two paths during their departure time windows, the DSO is well achieved in this example. Travelers select their departure times to maintain the equilibrium state. The inflows are zero whenever the dynamic marginal costs on the links are greater than the equilibrium cost. Figure 4.7(b) depicts the link traffic volumes of both links against the additional travel costs (introduced by the access constraints) under the DSO condition. As we do not impose access constraint on the links in the case, the additional travel costs are zero as shown in Figure 4.7(b). Figure 4.8 depicts the dynamic externalities for the two paths, i.e. $\frac{\partial \Psi_a(t,x_a)}{\partial h_a(t)} h_a(t) + \int_t^{\tau_a(t)} \frac{\partial \Psi_a(s,x_a)}{\partial \theta_a(s)} \frac{\partial \theta_a(s)}{\partial x_a(s)} h_a(s) ds$. The dynamic externality of link 1 is generally larger than that of link 2. This is because the capacity of link 2 is much larger than link 1. The dynamic externality of link 1 is about 13 % of its travel cost, while it is about 10 % for link 2.

Next, we impose the following time-varying access constraints on the two links, i.e.

$$x_1(t) \leq 3.5 \sin\left(\frac{(t+15)\pi}{100}\right) \text{ units}, \; x_2(t) \leq 4.5 \sin\left(\frac{(t+15)\pi}{100}\right) \text{ units}.$$

These constraints represent a kind of access control that adjusts its amplitude responding to the peak hour traffic volume. In this example, the control has a relative small value at the beginning, and gradually increases as the time approaches the peak hour. After the peak period, the control then decreases.

Meanwhile, we assume that the environmental capacities of the two links are

$$\kappa_1^e = 10, \text{ and } \kappa_2^e = 20,$$

respectively. By (4.76), we obtain the correspondingly equivalent restrictions on link traffic volumes as

$$x_1(t) \leq 3.15 \text{ units}, \text{ and } x_2(t) \leq 7 \text{ units}.$$

Figure 4.9(a) shows the travelers' responses to the additional travel costs imposed on the two links. Compared with the DSO case, the departure rates of the two links decrease. The departure time windows become wider for both links under the DSO-AC condition as compared to the uncontrolled case. These illustrate the shift of travelers' departure times to fulfill the access constraints and to maintain the new equilibrium condition under the access control.

Figure 4.9(b) shows that the traffic volumes on both links under DSO-AC condition satisfies the constraints. As explained in Section 4.2, the additional travel cost imposed on link 1 is calculated following the first part of Equation (4.48), to be more specific,

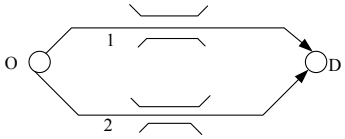


Figure 4.6: Network connected with parallel links

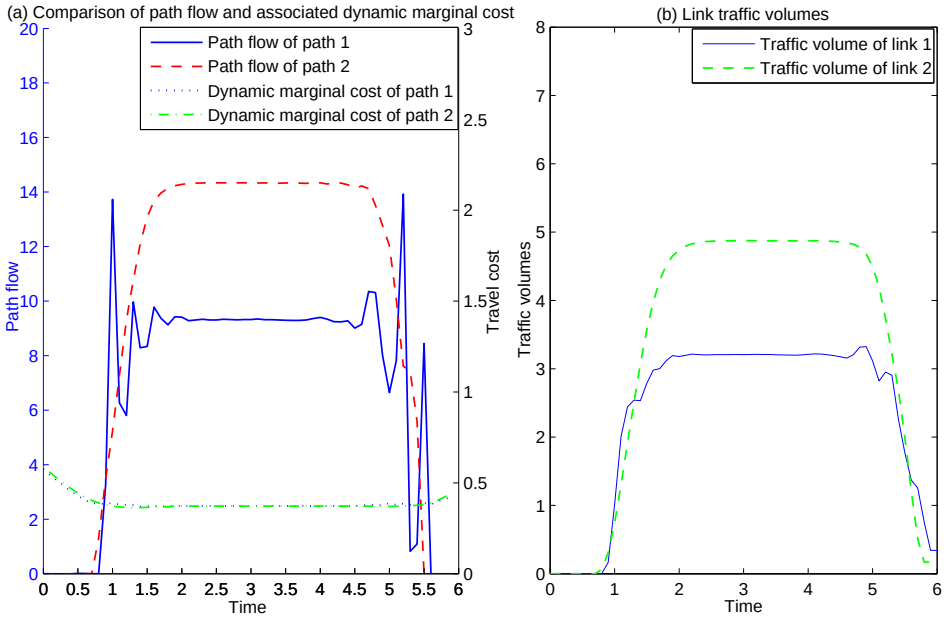


Figure 4.7: Inflow profiles, link traffic volumes and dynamic marginal costs of both links under the DSO condition

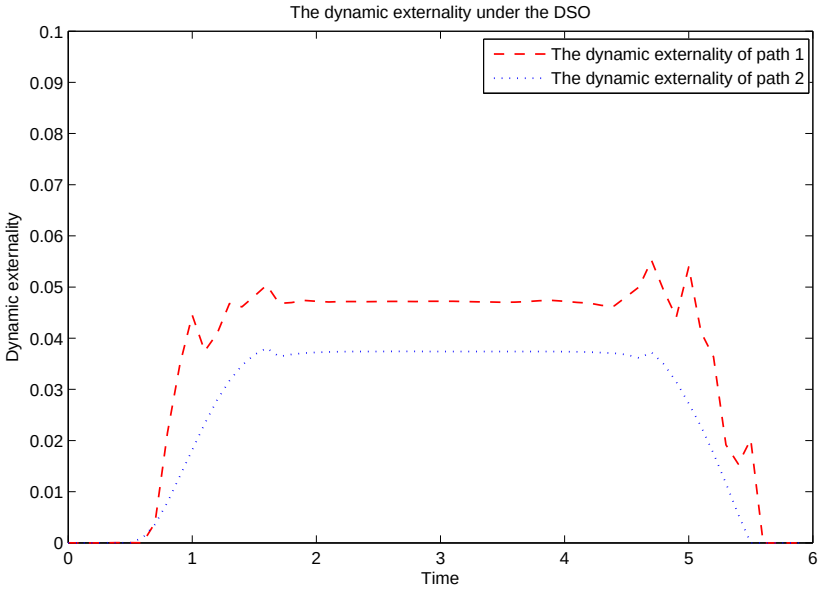


Figure 4.8: The externalities of the two links

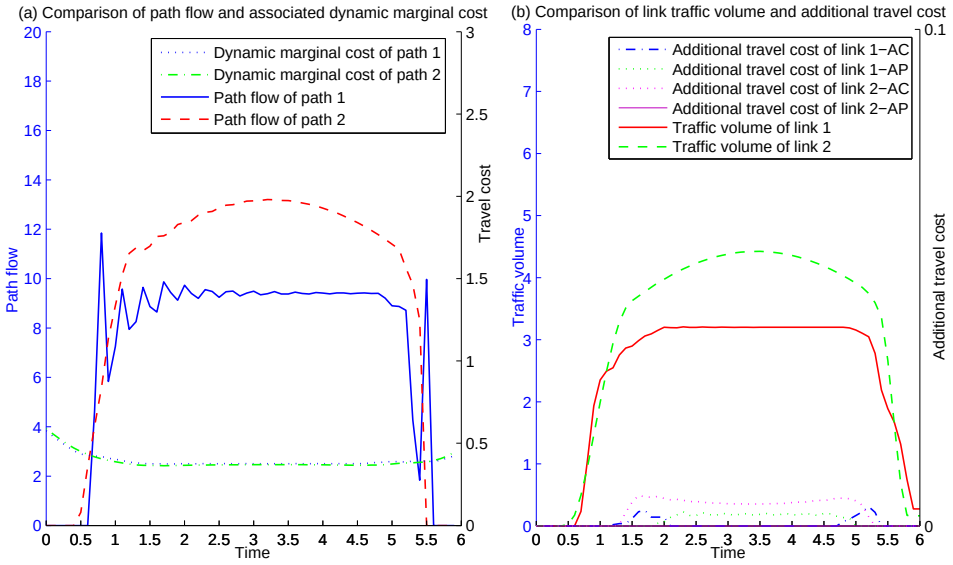


Figure 4.9: Inflow profiles, link traffic volumes, dynamic marginal costs, and additional costs of both links under the DSO-AC condition

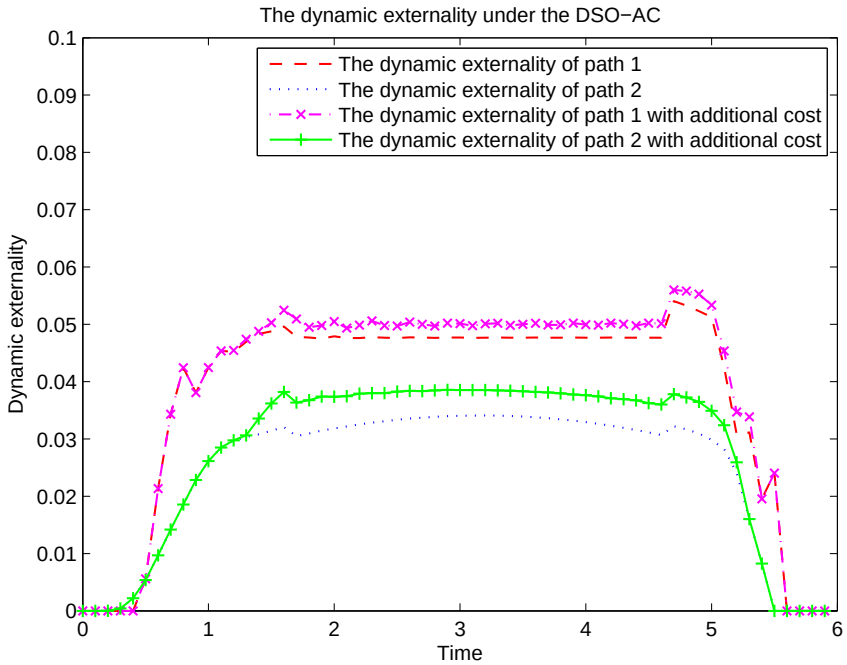


Figure 4.10: The externalities of the two links

the integral $\int_t^{\tau_a(t)} \eta_a(s)ds$ in this example. As the environmental capacity constraint is greater than the access control restriction for link 2, $\forall t \in [0, T]$, no pollution pricing should be implemented on this link, which is reflected by Figure 4.9(b)[4] wherein the additional travel cost (integral of the corresponding Lagrange multiplier) caused by the environmental capacity constraint is zero throughout the entire planing horizon. As the access control is activated, an amount of access price is charged to maintain the traffic volume on the link. For link 1, the access pricing is first activated because the access control constraint is smaller than the environmental capacity constraint at the initial stage. In this stage, no pollution pricing is implemented. After a time, the access control constraint becomes larger than the environmental capacity constraint. The pollution pricing is then activated. During the transition period, both access pricing and pollution pricing are activated as depicted in Figure 4.9(b). To be more specific, we assume the first transition time to be T_s in this example. For a time instant t_1 satisfying $t_1 < T_s < \tau_1(t_1)$, the overall access price is given by

$$ACP(t_1) = \int_{t_1}^{\tau_1(t_1)} \left(\eta_1^{ac}(s) + \eta_1^{ap}(s)\right) ds = \int_{t_1}^{T_s} \eta_1^{ac}(s)ds + \int_{T_s}^{\tau_1(t_1)} \eta_1^{ap}(s)ds.$$

The first part is contributed by the Lagrange multiplier associated with access control constraint while the second part is contributed by the Lagrange multiplier associated with environmental capacity constraint. Similar reasoning can be applied to the second transition period. The overall access pricing structure of this link is (by their time sequence):

1. access pricing only,
2. mixture of access pricing and pollution pricing;
3. pollution pricing only;
4. mixture of pollution pricing and access pricing;
5. access pricing only.

The dynamic externalities of the two links under DSO-AC condition are depicted in Figure 4.10. The values of additional travel costs induced by the access constraints vary from 5 % to 19.3 % of the dynamic externalities of the links.

4.8 Conclusions

We study the dynamic marginal cost pricing and the access pricing in this chapter. We derive dynamic system marginal costs for paths and dynamic access costs for controlled links

[4] In the legend of this figure, "AC" means the access control while "AP" means the air-pollution control.

for the two point queue models, i.e. the DQM and WLM. The problem is formulated as a dynamic system optimal traffic assignment problem with access constraints, wherein the access constraints represent the restrictions on the traffic volumes and/or environmental constraints (e.g. vehicle emission). We discuss the necessary condition for operating the transportation system with capacity/environmental constraints optimally. The Lagrange multipliers associated with the access constraints as derived from the optimality conditions provide the tolls for access pricing. We compare the difference between the dynamic externalities obtained from the whole link model and the deterministic queuing model. This difference results in different toll structures for these two models to achieve DSO. To be more specific, the queue can be constrained to zero in the DQM such that no dynamic externality will be introduced for marginal cost (usage) pricing. In this case, only access pricing exists, wherein the tolls are obtained from the access constraints. For the WLM, the link traffic volume is not zero as there is a positive inflow rate to the link. In this sense, we cannot constrain the link traffic volume (as queue in the bottleneck model) to zero to avoid the dynamic marginal cost (usage) pricing and to obtain the access price only. In other words, when we apply the WLM, the pricing scheme consists of both marginal cost (usage) pricing and access pricing (if the access constraint is activated) so as to achieve DSO. Compared with the dynamic marginal cost pricing, the access pricing is found to be much easier to manipulate.

By defining an environmental traffic capacity that converts the environmental constraint into traffic volume restriction, we show that the traffic induced air pollution pricing can be included as a special case of the access pricing. It is found that the traffic capacity based access price and traffic induced air pollution price would not become effective simultaneously for the DQM case. The boundary to determine the dominant price is defined by the minimum of the two capacities. However, for the WLM case, we have a circumstance that both prices would be effective simultaneously.

By far, we have considered the traffic volume (queue) control for traffic networks under two route choice behavior assumptions, i.e. the DUE and the DSO. The queue control is important especially under over-saturated conditions. Meanwhile, there are some important issues we have not yet considered. For instance, during an incident with lane blockage, congestion forms when the time-varying travel demand exceeds the reduced roadway capacity. Meanwhile, the growing incident induced lane changes and queue spillbacks significant interrupt the traffic flows among the adjacent lanes and exacerbate the incident induced congestion. Usually, traffic information is provided to drivers by ATMIS through various information media. The information is broadcasted in order to support

traveler's decision making which in turn influences their travel choices and consequently reduce the (total/individual) travel time and improve efficiency of the traffic network. However, divers behave in a very different way in face of the information provided to them due to their different degrees of risk aversion and perception errors on travel times, which in turn affects their routing decisions and the travel demand. Because of these unexpected events, traffic flows and travel times on the roadways are uncertain. Stochastic models are urgently needed to predict the traffic flows and travels time under these uncertain demand and supply conditions. In the forthcoming two chapters, we will develop a macroscopic stochastic dynamic traffic flow model to describe traffic flows on networks influenced by both demand and supply uncertainties.

Chapter 5

Stochastic cell transmission model (SCTM): a stochastic dynamic traffic model for freeway corridor traffic state surveillance

The chapter proposes a first-order macroscopic stochastic dynamic traffic model, namely the Stochastic Cell Transmission Model (SCTM), to model traffic flow density on freeway segments with stochastic demand and supply. The SCTM consists of five modes of traffic states on the freeway segment. Each mode is formulated as a discrete time bilinear stochastic system. A set of probabilistic conditions is proposed to characterize the probability of occurrence of each mode. The overall effect of the five modes is estimated by the joint traffic density which is derived from the theory of finite mixture distribution. The proposed model possesses the Markovian property which allows a practical implementation. The SCTM captures not only the mean and standard deviation (SD) of density of the traffic flow, but also the propagation of SD over time and space. The SCTM is tested with a hypothetical highway corridor simulation and an empirical study. The simulation results are compared against the means and SDs of traffic densities obtained from the Monte Carlo Simulation (MCS) of the modified cell transmission model (MCTM). An approximately two-miles freeway segment of Interstate 210 West (I-210W) in Los Ageles, Southern California, is chosen for the empirical study. Traffic data is obtained from the Performance Measurement System (PeMS). The stochastic parameters of the SCTM are calibrated against the flow-density empirical data of I-210W. Both the SCTM and the MCS of the MCTM are tested. A discussion of the computational efficiency and the ac-

curacy issues of the two methods is provided based on the empirical results. Both the numerical simulation results and the empirical results confirm that the SCTM is capable of accurately estimating the means and SDs of the highway densities as compared to the MCS.

5.1 Introduction and motivation

Dynamic traffic flow models are one of the key components of dynamic traffic assignment (DTA) as well as real-time traffic control and management. To model the complex freeway traffic, many efforts have been made to establish and validate both microscopic (*e.g.* car-following) and macroscopic (*e.g.* hydro-dynamics based) models. However, many of these models are too computationally demanding for online estimation of traffic states for a large-scale road network. A comparative study of the four macroscopic link models that are widely used in DTA is given by Nie and Zhang (2005). It is found in the paper that these macroscopic link models would produce the same traffic assignment result unless there is a shockwave.

Among the macroscopic traffic flow models, Lighthill-Whitham-Richards (LWR) model would be the most popular and most-cited one. In terms of fluid dynamics, the traffic dynamics of a freeway segment modeled by the LWR model is governed by the following two equations.

$$\begin{aligned} \frac{\partial \rho(x,t)}{\partial t} + \frac{\partial f(x,t)}{\partial x} &= \nu_+(x,t) - \nu_-(x,t), \\ f(x,t) &= F\left(\rho(x,t)\right), \end{aligned} \tag{5.1}$$

where x, t represents position (measured in the direction of traffic flow) and time, respectively. $\rho(x,t)$ and $f(x,t)$ denote the traffic density and the traffic flow (as functions of location x and time t), respectively. $\nu_\pm(x,t)$ are the source terms which may be due to ramp flows with the plus sign denotes on-ramps and the minus sign denotes off-ramps (Schönhof and Helbing, 2007). The first equation of (5.1) is the principle of conservation of vehicles, which is followed from fluid mechanics. The second equation of (5.1) is a flow-density relationship which is also known as the "fundamental diagram". There are several ways to introduce stochastic elements to the LWR modeling framework, e.g.

1. stochastic initial and boundary conditions,
2. stochastic source terms, and
3. stochastic speed-density relationship or fundamental diagram.

Some dynamic traffic flow models (*e.g.* the Cell Transmission Model (CTM) proposed by Daganzo (1994, 1995a), the Modified Cell Transmission Model (MCTM) proposed by Muñoz et al. (2003), and the Enhanced Lagged CTM proposed by Szeto (2008)), which discretize the LWR model (or its simplified version) in both time and space, were shown to be computationally efficient and easy to analyze yet capture many important traffic phenomena, such as queue build-up and dissipation, backward propagation of congestion waves, *etc.* In general, the original LWR model and other first-order macroscopic traffic flow models derived from it, *e.g.* CTM, have a common assumption of a steady-state speed-density relationship which does not allow fluctuations around the equilibrium (nominal) fundamental flow-density diagram, or these models adopt a number of deterministic parameters (*e.g.* free-flow speed, jam-density, capacity, *etc.*).

However, research and empirical studies on the fundamental flow-density diagram have revealed that the fundamental flow-density diagram admits large variations (see Figure 5.1) due to the congestion, driving behavior, *etc.* (*e.g.* Kim and Zhang (2008); Wang et al. (2009); Li et al. (2009) and the references therein). Microscopic and macroscopic modeling approaches have been proposed to model and interpret this phenomenon. In the microscopic approach, this phenomenon has been interpreted as the effects of anticipation, strong correlations in the vehicle motion on different lanes, delay in the driving adaptation or safe time-gap variations (*e.g.* Ngoduy (2009) and the references therein). In the macroscopic approach, the phenomenon has been modeled as a diffusion coefficient to reproduce significant elements of the synchronized traffic flow, the interactions between several vehicle classes (*e.g.* trucks and cars), randomness in driving behavior, and adverse weather conditions, etc., (Chen et al., 2001; Ngoduy, 2009). As mentioned by Geistefeldt and Brilon (2009), the stochastic features of freeway capacity can be revealed by analyzing the transition of traffic flow from free flow to congested conditions, which is referred to as a traffic breakdown. A traffic breakdown indicates that the traffic demand has exceeded the capacity, the variability of these breakdown volumes indicates the randomness of freeway capacity. Therefore, "stochastic" traffic flow models, *e.g.* Boel and Mihaylova (2006); Kim and Zhang (2008) and Li et al. (2009), were developed to capture random traffic states of freeways. This can be considered as attempts to introduce the stochasticity into the speed-density relationship (or fundamental diagram) which corresponds to the third component of the stochastic LWR modeling discussed earlier.

In fact, the other kind of uncertainty is travel demand variability, which is always regarded as recurrent uncertainty or disturbance in traffic flow dynamics. Corresponding to the LWR model, the demand uncertainty would represent the stochastic boundary

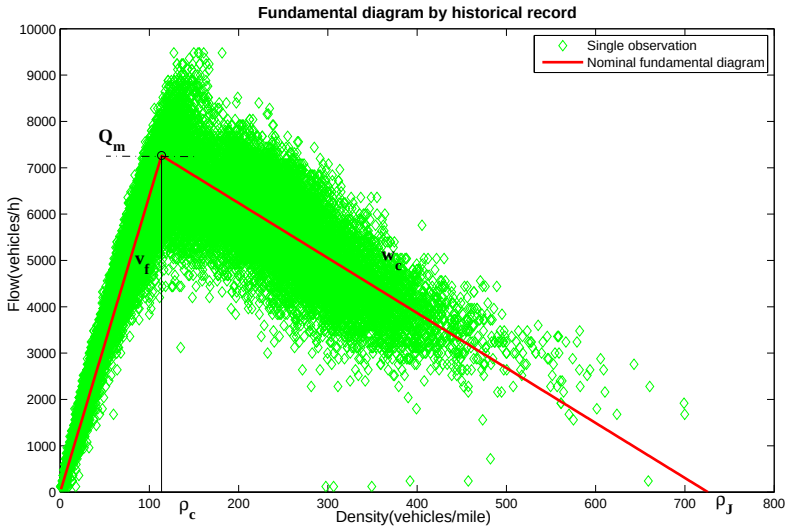


Figure 5.1: A fundamental flow-density diagram of traffic flow

conditions and source terms (this corresponds to the first and second components of the stochastic LWR modeling discussed earlier). By allowing the stochastic demand input, the dynamic link model can better capture the possible future uncertainty of the travel demand which can enhance the application of the model with short and medium term operation/planning. Therefore, this chapter aims to extend the CTM to estimate the stochastic freeway traffic states under stochastic fundamental flow-density diagrams as well as the stochastic travel demand, wherein the deterministic LWR and its extensions fail.

The CTM proposed by Daganzo (1994, 1995a) defines piecewise affine sending and receiving functions of traffic flow to describe interactions between adjacent freeway cells as well as shockwaves. For the purposes of developing surveillance, assignment and control strategies on freeways, it is also important to explicitly model the randomness of the traffic state evolution (*e.g.* Peeta and Zhou (2006), and Friesz et al. (2008) and the references therein). This randomness can be reflected in the model via some stochastic process of the parameters governing the sending and receiving functions (Boel and Mihaylova, 2006). To extend the CTM to deal with the stochastic elements, the simplest approach is to apply the Monte Carlo Simulation method to the CTM. Based on the switching-mode model (SMM) (*i.e.* a simplified version of MCTM distinguishing between the free-flow mode and the congestion mode) and a sequential Monte Carlo algorithm (*i.e.* the so-called mixture Kalman filtering), Sun et al. (2003) proposed a freeway traffic estimator. As a drawback of the Monte Carlo Simulation (MCS), the model may suffer from high computational cost.

A stochastic compositional model for freeway traffic flows was proposed by Boel and Mihaylova (2006). This model extends the CTM by defining sending and receiving func-

tions as random variables, and specifying the dynamics of the average speed in each cell. In this model, the traffic states are divided into two extreme cases: very light traffic conditions and extremely congested conditions where the sending functions are assumed to follow Binomial and Gaussian distributions respectively. However, the intermediate cases between very light traffic and very dense traffic are not well defined. Based on the above dynamic traffic model, a particle filtering (PF) framework was proposed to estimate both traffic density and speed (Mihaylova et al., 2007). The implemented PF performs well with a small number of particles (which can be regarded as samples in the MCS) in the case of the light traffic condition. However, obtaining a good estimation in the case of the dense traffic condition can be computational expensive. It is also difficult to characterize in general the PF accuracy and complexity because they highly depend on the road structure and the traffic conditions (Mihaylova et al., 2007).

In this chapter, we develop a stochastic cell transmission model (SCTM) to describe the macroscopic dynamics of the traffic flow under demand and supply uncertainties. The SCTM extends the CTM by defining the parameters governing the sending and receiving functions explicitly as random variables. The stochascities of the sending and receiving functions are governed by the random parameters of the piecewise flow-density diagram, *i.e.* free-flow speed, jam-density, and backward wave speed. In addition, the SCTM also allows the inflow demand to be stochastic. The stochastic elements in our framework are described by some wide sense stationary, second-order processes consisting of uncorrelated random vectors with known mean and variance. These elements can vary with time depending on the availability of on-line measurements and the locations of the cells.

The proposed model avoids the non-linearities in the original CTM by using the SMM[1] with five possible traffic modes (or states) previously proposed by Muñoz et al. (2003), and Sun et al. (2003). The SMM is a simplified version of the MCTM and will be described in detail later. Each of the traffic modes (or states) of the SCTM is then redefined as a discrete time stochastic bilinear system (*e.g.* Mohler (1973) and Tuan (1985)). Since the SCTM operates under a stochastic environment, all five modes are possible at each time step. This will cause a problem of the curses of dimensionality, *i.e.* the dimension of the problem increases exponentially with respect to time, if we track all the modes at each time step. To this end, a set of probabilistic conditions is defined for approximating the joint traffic density following the theory of finite mixture distribution to avoid the curses

[1] A simpler version of the SMM was first proposed by Zhang et al. (1996) from a traffic control context. In that paper, the traffic flow were modeled by different modes without specifying the various types of waves systematically.

of dimensionality.

Freeway traffic data is often available in the form of occupancy and volume measurements collected from loop detectors embedded in the pavement. In conjunction with effective vehicle length data, these measurements can be converted into macroscopic quantities such as traffic density and speed. Loop detector data sets are often incomplete, or contain bad samples. However, for the purpose of dynamic traffic assignment (DTA) and ramp metering control strategies, such as ALINEA (*e.g.* Gomes and Horowitz (2006)), accurate traffic OD information and density information are required in order to effectively direct traffic and regulate on-ramp inflows to the freeway. It is thus essential to reconstruct the missing traffic measurement data. The SCTM provides us a tool to reconstruct the traffic data which is adaptive to changing stochastic external conditions (supply and demand uncertainties) such as: weather and lighting conditions, percentage of trucks, variable speed limits applied, and variation of travel demand, *etc.* Numerical and empirical tests involve comparing the means and standard deviations (SDs) of the dynamic traffic densities as approximated by the SCTM and the Monte Carlo simulation with the density-based equivalent of CTM. In addition, a numerical test is also conducted to illustrate the feature of the proposed model in capturing the propagation of the uncertainty through space and time. All the tests give satisfactory results which prove that the SCTM is computationally efficient and is suitable for real-time traffic monitoring and control applications.

The outline of the chapter is as follows: Section 5.2 gives a brief review on the MCTM and the SMM. The SCTM is formulated in Section 5.3. Numerical tests of the SCTM are conducted in Section 5.4. The empirical study is provided in Section 5.5. Lastly, conclusions are presented in Section 5.6.

5.2 The MCTM and the SMM

The modified cell transmission model (MCTM) was developed by Muñoz et al. (2003). This model uses cell densities instead of cell occupancies which permits the CTM to adopt non-uniform cell lengths and leads to greater flexibility in partitioning freeways. In the MCTM, the density of cell i evolves according to the conservation of vehicles:

$$\rho_i(k+1) = \rho_i(k) + \frac{T_s}{l_i}\left(q_{i,in}(k) - q_{i,out}(k)\right), \tag{5.2}$$

where $\rho_i(k)$ is the vehicle density of cell i at time index k, $q_{i,in}(k)$ and $q_{i,out}(k)$ are the total flows (in vehicles per unit time) entering and leaving cell i during the time interval $[kT_s,\ (k+1)T_s)$ respectively, T_s is the sampling duration, and l_i is the length of cell i. The model parameters, including the free-flow speed v_f, the backward congestion wave

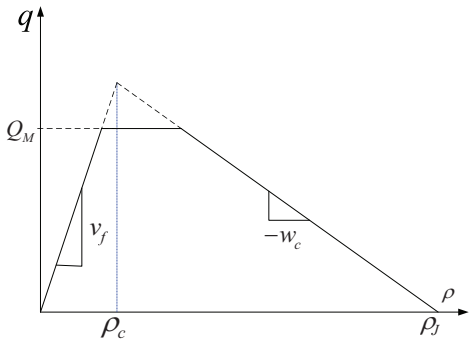


Figure 5.2: A trapezoidal fundamental diagram for the modified cell transmission model

speed w_c, the maximum allowable flow Q_M, the jam density ρ_J, and the critical density ρ_c, are depicted in the trapezoidal fundamental diagram of Figure 5.2. These parameters can vary from cell to cell over time.

Following Daganzo (1994, 1995a), $q_i(k)$ is determined by taking the minimum of two quantities:

$$q_{i,in}(k) = \min(S_{i-1}(k), R_i(k)), \tag{5.3}$$

where

$$S_{i-1}(k) = \min\left(v_{f,i-1}(k)\rho_{i-1}(k), Q_{M,i-1}(k)\right),$$

is the maximum flow supplied by cell $i-1$ under the free-flow condition, over the interval $[k,\ k{+}1)$, and

$$R_i(k) = \min\left(Q_{M,i}(k), w_{c,i}(k)(\rho_{J,i}(k) - \rho_i(k))\right),$$

is the maximum flow received by cell i under the congested condition over the same time interval. (5.2) and (5.3) are the density-based equivalents of those described in Daganzo (1994).

Although the MCTM is much simpler than many other higher order hydrodynamics-based partial differential models, the nonlinear nature of the flow-density relationship due to (5.3) still makes the MCTM difficult to be analyzed and used as a basis for the design of traffic controllers (Muñoz et al., 2003; Gomes et al., 2008). To avoid the nonlinearity, the switching mode model (SMM) was proposed by Muñoz et al. (2003). The SMM is a hybrid system (or switched linear system) that switches among different sets of linear difference equations (representing different traffic states of the freeway), depending on the mainline boundary data and the congestion status of the cells in a freeway segment. The SMM formulation avoids the nonlinearity of the CTM at the cost of using the same triangular

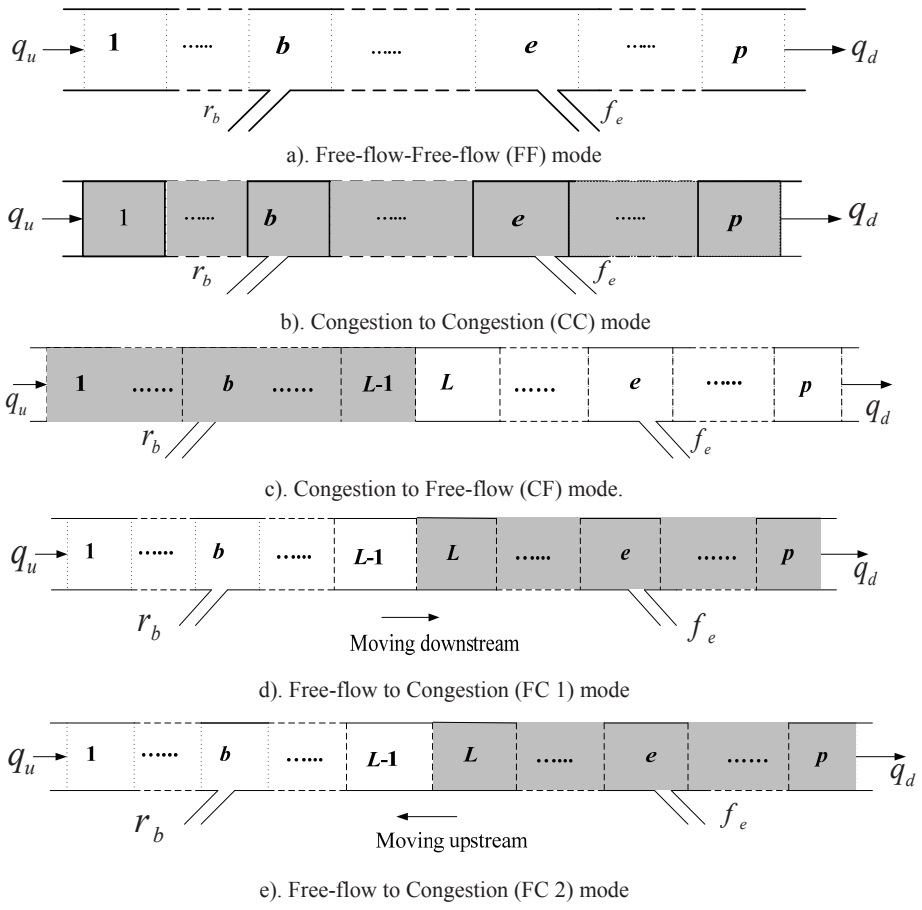


Figure 5.3: Five traffic operational modes for a freeway segment with p cells

flow-density relationship for all the cells along the whole freeway segment, and introducing the switching condition based on the following atmost-one-wavefront assumption:

Assumption 5.1. (Muñoz et al., 2003) There is at most one wavefront in the freeway segment.

Based on the above assumption, five modes are defined in the state space representation (see Figure 5.3):

1. "Free flow-Free flow (FF)" (Figure 5.3a),
2. "Congestion - Congestion (CC)" (Figure 5.3b),
3. "Congestion - Free flow (CF)" (Figure 5.3c),
4. "Free flow-Congestion 1 (FC1)" (Figure 5.3d),
5. "Free flow - Congestion 2 (FC2)" (Figure 5.3e).

A wavefront is assumed to be located at the boundary between the two cells at time k. Among these five modes, the FF and CC modes are steady-state modes while the others are transient modes. The two modes of "Free flow - Congestion" are determined by the relative magnitudes of the supplied flow of the last uncongested cell upstream of the wave-front and the receiving flow of the first congested cell downstream of the wave-front. If the former is smaller, the SMM is in the FC1 mode, otherwise it is in the FC2 mode. In the SMM, the mode of the model is determined following a set of traffic density based switching criteria in which only one mode is activated at each time step.

5.3 The stochastic cell transmission model

5.3.1 The overall framework of the SCTM

As previously described, in Muñoz et al. (2003); Sun et al. (2003), the MCTM has been piecewise-linearized to obtain the SMM with five operational modes for a freeway segment based on Assumption 5.1. From the traffic control context, the linear structure of the SMM lends the advantage of simplifying control analysis, control design, and data-estimation design methods. From the traffic flow simulation context, Assumption 5.1 simplifies the traffic state of the freeway segment which increases the computational efficiency. We follow the concept of operational modes used in the SMM. However, due to the stochastic supply and demand, the wavefront is uncertain, which implies that within one subsystem all the five modes are possible (hence five probabilistic events) but with different probabilities

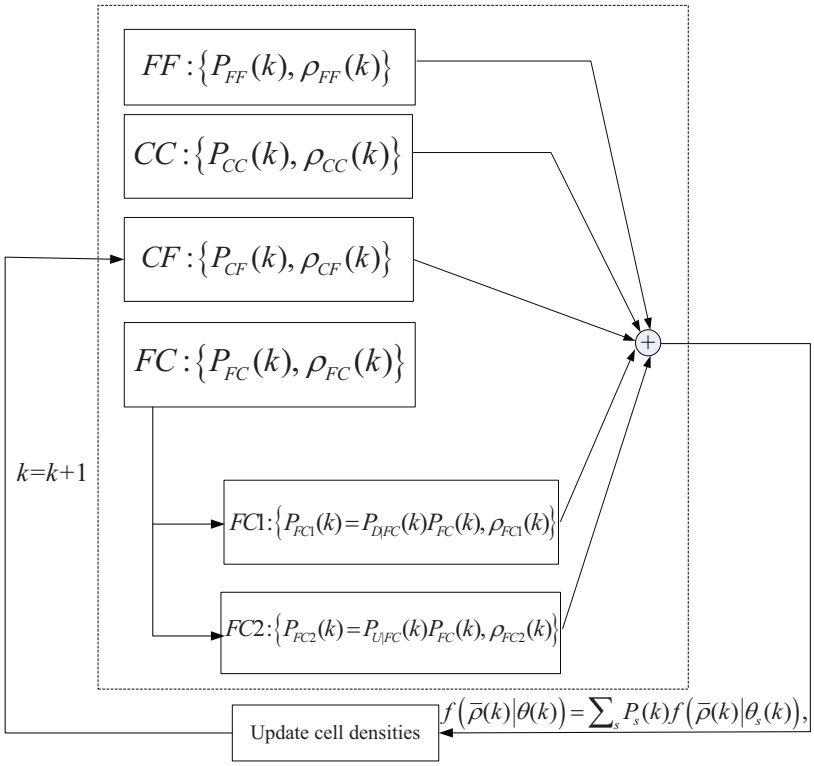


Figure 5.4: The overall framework of the stochastic cell transmission model

of occurrence. We denote these probabilities as: $P_{FF}(k), P_{CC}(k), P_{CF}(k), P_{FC1}(k)$, and $P_{FC2}(k)$, where $P_s(k)$ is the probability of mode $s \in \{FF,\ CC,\ CF,\ FC1,\ FC2\}$ to occur at time index k. To this end, we update the dynamics of the SCTM as depicted in Figure 5.4, where the overall effect of the five modes is defined as the joint (or "actual") traffic density. The probabilities of occurrence in conjunction with the density vectors of the five modes are used to define the probabilistic density function (PDF) of the joint traffic density vector $\bar{\rho}(k)$, which is approximated by a finite mixture approximation of the probabilistic density functions of the five modes. Its mean $E(\bar{\rho}(k)|\theta(k))$ and covariance matrix $Var(\bar{\rho}(k)|\theta(k))$ can be obtained by the theory of finite mixture distribution which will be explained in Section 5.3.3, where $\theta(k) = \{\theta_s(k)\}$, $\theta_s(k) = (\rho_s(k),\ P_s(k))$, and $\rho_s(k)$ denotes the vector of cell densities of mode s.

To sum up, the SCTM accepts the random inflows (uncertain demand) as well as random parameters of the fundamental flow-density diagram (uncertain supply functions) with known means and variances of the freeway segment as exogenous inputs, and then calculates the means and variances of the joint traffic densities, outflow of the freeway segment, and probabilities of its operational modes. With respect to the above framework, five key issues are needed to be addressed. The fist issue is to define the demand and supply uncertainties. The second issue is to define the probabilities of occurrence of the

five operational modes and to approximate the PDF of the joint traffic density vector by a finite mixture distribution of the PDFs of the five operational modes. The third issue is to model the dynamics of the five operational modes. The means and auto-correlations of the dynamics of the five operational modes are needed to be evaluated. Finally, we need to define the implementation of the SCTM for traffic state estimation.

5.3.2 Formulation of demand and supply uncertainties

From the analysis in the previous subsection, to evaluate the stochastic traffic dynamics, we need to define the probabilities and the traffic flow dynamics for the five modes. To begin with, let's first specify the demand and supply uncertainties considered in this chapter. Consider a freeway segment consisting of p cells and one on-ramp and one off-ramp[2] as depicted in Figure 5.3. We denote the traffic state at time index $k \geq 0$ as the traffic density $\rho(k) = (\rho_1(k), \ldots, \rho_p(k))^T$, and $u(k) = (q_u(k),\ r_b(k),\ f_e(k),\ q_d(k))^T$ is comprised of system inflow and outflow at time index k. $r_i(k)$ and $f_j(k)$ are the measured on-ramp and off-ramp flows entering cell i and leaving cell j at time index k respectively. $q_u(k)$ and $q_d(k)$ are respectively the (measured) upstream and downstream boundary flows at time index k, and $\rho_u(k)$, and $\rho_d(k)$ are densities defined correspondingly. To save notations, the notations for the five parameters in Figure 5.2 are also used for representing the corresponding five vectors in the SCTM when there is no ambiguity. To be more specific, $v_f = (v_{f,1}, \ldots, v_{f,p})^T$ is the vector of free-flow speeds, $\rho_c = (\rho_{c,1}, \ldots, \rho_{c,p})^T$ is the vector of critical densities, $w_c = (w_{c,1}, \ldots, w_{c,p})^T$ is the vector of backward congestion wave speeds, $\rho_J = (\rho_{J,1}, \ldots, \rho_{J,p})^T$ is the vector of jam densities, and $Q_M = (Q_{M,1}, \ldots, Q_{M,p})^T$ is the vector of maximum flow rates. According to the triangular fundamental diagram of a given cell, only three among the five system parameters are independent variables. We denote an independent set of the system parameters in a compact form as $\Gamma = (v_f,\ w_c,\ Q_M)^T$. In the real world, the system parameter vector Γ admits uncertainties. We assume that the system parameter vector is perturbed by certain noise sequence as follows:

$$\Gamma(k) = \Gamma_0 + \xi^{\Gamma}(k), \tag{5.4}$$

where $\Gamma(k)$ is the system parameter vector for time k, Γ_0 is the nominal value of the system parameters, and $\xi^{\Gamma}(k)$ is the noise vector for system parameters at time index k. Note

[2]In this chapter, we do not consider the dynamics of the on-/off- ramps, i.e. we do not consider the merge and diverge operations. The on-/off- ramp flows considered here are the measured on-/off- ramp flows.

that $\{\xi^{\Gamma}(k)\}_{k\in N}$ is a second-order wide-sense stationary (WSS) process[3] to be specified later. Also, we assume the travel demand to be a random vector in the form

$$u_d(k) = u_0(k) + \xi_u(k), \tag{5.5}$$

where $u_d(k) = (q_u(k), r_b(k))^T$, $u_0(k)$ is the nominal calibrated travel demand vector for time index k, and $\xi_u(k)$ is the demand noise at time index k. $\{\xi_u(k)\}_{k\in N}$ is a second-order WSS process to be specified later. Without loss of generality, all the noise sequences and initial conditions are assumed to follow Gaussian (white-noise) processes.

For the demand side, we assume the noise sequence $\{\xi_u(k)\}_{k\in N}$ in the control input to be a zero-mean Gaussian random process:

$$E(\xi_u(k)) = 0, E\left(\xi_u(k)\xi_u^T(t)\right) = \begin{cases} \Omega_u, \text{ if } k = t; \\ 0, \text{ otherwise,} \end{cases} \tag{5.6}$$

where k and t are time indices. Similarly, for the supply side we assume that the noise $\xi^{\Gamma}(k)$ and the initial state $\rho(0)$ of the system satisfy the following conditions:

1. The noise $\xi^{\Gamma}(k)$ can be described by a zero-mean Gaussian random process. For any $k \geq 0$ and $t \geq 0$, the following equations are satisfied:

$$E(\xi^{\Gamma}(k)) = 0, E\left(\xi^{\Gamma}(k)\left(\xi^{\Gamma}(t)\right)^T\right) = \begin{cases} \Omega_{\Gamma}, \text{ if } k = t; \\ 0, \text{ otherwise.} \end{cases} \tag{5.7}$$

   We also assume that, the components of the vector $\xi^{\Gamma}(k)$ are mutually independent for any $k \geq 0$, or the matrix Ω_{Γ} is a diagonal semi-positive definite matrix.

2. The components of the initial traffic density vector $\rho(0)$ are mutually independent and normally distributed.

3. $\rho(0)$ and $\xi^{\Gamma}(k)$ are uncorrelated to each other for any $k \in N$.

Remark 5.1. As mentioned before, only three among the five system parameters are independent. For illustration purposes, consider cell i with a triangular flow-density relationship and let $(v_{f,i}, w_{c,i}, Q_{M,i})$ be the independent set of the parameters for cell i. $\rho_{c,i}$ and $\rho_{J,i}$ can then be determined by $v_{f,i}$, $w_{c,i}$, and $Q_{M,i}$. Let $x_i = (v_{f,i},\ w_{c,i},\ Q_{M,i})^T$, then $\rho_{c,i} = g(x_i) = \frac{Q_{M,i}}{v_{f,i}}$. Applying Taylor expansion to $g(x_i)$ at x_0 yields

$$\rho_{c,i} = g(x_i) = g(x_0) + (x_i - x_0)^T \nabla g(x_0) + \frac{1}{2}(x_i - x_0)^T H(x_0)(x_i - x_0) + \cdots,$$

[3] A random process $x(k)$ is said to be wide-sense stationary (WSS) if $E(x(k)) = c$ and $E\left(x(l)x^T(k)\right) = \Omega_x(k-l) = \Omega_x(\tau)$, where c is a constant vector and $\Omega_x(\cdot)$ is the correlation matrix of the process, and $\tau = k - l$ is the time lag.

where $\nabla g(x_0)$ is the gradient of g at x_0 and $H(x_0)$ is the corresponding Hessian matrix. Take

$$g(x_i) \approx g(x_0) + (x_i - x_0)^T \nabla g(x_0) + \frac{1}{2}(x_i - x_0)^T H(x_0)(x_i - x_0). \tag{5.8}$$

Since x_i is a vector with its components mutually independent and its mean and variance are given in the above assumptions, we can approximate the mean of $\rho_{c,i}$ by taking expectation on both sides of (5.8), and the variance can be obtained respectively. Notice that if the first-order approximation is used in (5.8), $\rho_{c,i}$ is normally distributed if we assume x_i is governed by a normal distribution. Similarly, we can approximate the mean and variance for $\rho_{J,i} = \frac{Q_{M,i}}{w_{c,i}} + \frac{Q_{M,i}}{v_{f,i}}$. ∎

In what follows, we denote the vector of system parameters as

$$\Phi(k) = col\left(v_f(k), w_c(k), Q_M(k), \rho_c(k), \rho_J(k)\right).$$

5.3.3 Dynamic process of the SCTM and probabilistic conditions

As mentioned, we need to specify the probabilities of the five modes at each time step to evaluate the stochastic traffic flow. The probabilities of the FF, and CC modes to occur can be determined as follows: FF mode:

$$P_{FF}(k) \triangleq \Pr\left(\rho_u(k) < \rho_{c,1}(k) \bigcap \rho_d(k) < \rho_{c,p}(k)\right), \text{and} \tag{5.9}$$

CC mode:

$$P_{CC}(k) \triangleq \Pr\left(\rho_u(k) \geq \rho_{c,1}(k) \bigcap \rho_d(k) \geq \rho_{c,p}(k)\right). \tag{5.10}$$

CF mode: As mentioned before, the shockwave exists only in the three transient modes: CF, FC1, and FC2. Due to the stochastic environment, the location of wavefront is uncertain. Thus we define the following probability to capture the probability of the CF mode to occur with the wavefront located at a specific location:

$$P_{CF,L}(k) \triangleq \Pr\left(\begin{array}{c} (\rho_u(k) \geq \rho_{c,1}(k)) \bigcap \left(\bigcap_{i=2}^{L-1} \bar{\rho}_i(k) \geq \rho_{c,i}(k)\right) \\ \bigcap \left(\bigcap_{j=L}^{p-1} \bar{\rho}_j(k) < \rho_{c,j}(k)\right) \bigcap (\rho_d(k) < \rho_{c,p}(k)) \end{array}\right), \tag{5.11}$$

where $P_{CF,L}(k)$ denotes the probability of the CF mode occurring at time step k with the wavefront located at the boundary between cells $L-1$ and L. To this end, we have the probability of the CF mode to occur for the whole freeway segment:

$$P_{CF}(k) = \sum_{L=2}^{p} P_{CF,L}(k). \tag{5.12}$$

Let $P_{FC}(k)$ be the probability of the FC mode occurring at time k. Then,

$$P_{FC}(k) \triangleq 1 - (P_{FF}(k) + P_{CC}(k) + P_{CF}(k)). \tag{5.13}$$

Assume the probability that the wavefront is located at the boundary between cells $L-1$ and L, and is moving downstream (event D) as

$$P_{D|FC,L}(k) \triangleq \Pr\left(\begin{array}{c}(\rho_u(k) < \rho_{c,1}(k)) \bigcap \left(\bigcap_{i=2}^{L-1} \bar{\rho}_i(k) < \rho_{c,i}(k)\right) \\ \bigcap \left(\bigcap_{j=L}^{p-1} \bar{\rho}_j(k) \geq \rho_{c,j}(k)\right) \bigcap (\rho_d(k) \geq \rho_{c,p}(k)) \\ \bigcap (v_{f,L-1}(k)\bar{\rho}_{L-1}(k) \leq w_L(k)(\rho_{J,L}(k) - \bar{\rho}_L(k)))\end{array}\right). \tag{5.14}$$

The notation $FC1$ is now restricted as the whole freeway segment is in the FC mode and the wavefront is moving downstream, and $FC2$ is similarly defined. In this sense, we define

$$P_{FC1}(k) \triangleq P_{D|FC}(k) = \sum_{L=2}^{p} P_{D|FC,L}(k), \text{and} \tag{5.15}$$

$$P_{FC2}(k) \triangleq P_{FC}(k) - P_{FC,1}(k). \tag{5.16}$$

With the above definitions of probabilities of occurrence of the five operational modes, we need to address the problem that how to estimate (or approximate) the overall effect of the five possible operational modes (or the joint traffic density) given their PDFs. In this chapter, we provide a finite mixture distribution approach to solve the above question, i.e. the overall effect of the five possible operational modes is estimated (or approximated) by a finite sum of known PDFs. The probability density function (PDF) of the joint traffic density $f(\bar{\rho}(k)|\theta(k))$ can be approximated by the following finite mixture distribution (Frühwirth-Schnatter, 2006):

$$f(\bar{\rho}(k)|\theta(k)) = \sum_s P_s(k) f(\bar{\rho}(k)|\theta_s(k)), \tag{5.17}$$

where f is the PDF of the joint traffic density $\bar{\rho}(k)$, the parameter set is defined as $\sum_s P_s(k) = 1$, $\{\theta(k)\} = \{\theta_s(k)\}$, $\theta_s(k) = (P_s(k),\ \rho_s(k))$, $\rho_s(k)$ denotes the vector of cell densities of mode s at time k, and $P_s(k)$ is defined by (5.9)-(5.16).

Under the mixture model (5.17), the expectation $E(\bar{\rho}(k)|\theta(k))$ is obviously given by

$$E(\bar{\rho}(k)|\theta(k)) = \sum_s P_s(k) E(\bar{\rho}(k)|\theta_s(k)) = \sum_s P_s(k) E(\rho_s(k)). \tag{5.18}$$

Let $\mu_s(k) = E(\rho_s(k))$ and $\mu(k) = E(\bar{\rho}(k)|\theta(k))$. Then we have $\mu(k) = \sum_s P_s(k)\mu_s(k)$. To evaluate the covariance matrix $Var(\bar{\rho}(k)|\theta(k))$, we define the covariance matrix of $\rho_s(k)$ as

$$\psi_s(k) = E\left((\rho_s(k) - \mu_s(k))(\rho_s(k) - \mu_s(k))^T\right).$$

Then the covariance matrix $Var\left(\bar{\rho}(k)|\theta(k)\right)$ can be evaluated as:

$$Var\left(\bar{\rho}(k)\,|\theta(k)\,\right) = \sum_{s} P_s(k)\left(\psi_s(k) + \mu_s(k)\mu_s^T(k)\right) - \mu(k)\mu^T(k). \tag{5.19}$$

If the mean and covariance matrix are defined, the "joint traffic density" is well defined for a second-order random process. As the probabilities of the five modes are already defined, we need to obtain the mean of the cell traffic density vector $\mu_s(k)$ and the covariance matrix $\psi_s(k)$ for each mode s at each time step k.

To end this section, we would like to give some comments on the wavefront concept, at-most one wavefront assumption, and their relations with the SCTM. To utilize the SMM, we need to ensure that the traffic dynamics of a freeway segment can be accurately described by the five modes. However, Assumption 5.1 cannot be fulfilled for general freeway segments except some special cases. We emphasize here that:

First, the current framework does not rely on Assumption 5.1. This assumption is only used to define the operational modes for the SCTM. We will provide a simple solution to the case that there are several uncertain wavefronts on a freeway corridor in Chapter 6.

Second, the uncertain wavefront concept is converted into several probabilistic operational modes in the current framework. The uncertain wavefront is described by the probabilities of occurrence of these operational modes.

5.3.4 The SCTM as a class of stochastic bilinear system

To allow the analysis to be more systematic and compact, we define the density update equations of each mode in the form of a dynamic system. Due to the multiplicative effect of the system parameters in the SCTM, such as v_f, w_c, and ρ_J, our system is no longer a linear system. Nevertheless, the SCTM can be reformulated as a class of discrete time stochastic bilinear system in the form of (5.20) below. However, instead of specifying the system parameter vector $\Phi(k)$ of the freeway segment as internal dynamics (or system matrices) as what has been done in Muñoz et al. (2003), we take the system parameter vector as an exogenous input to the system together with the inflow vector $u(k)$.

$$\rho(k+1) = \left(A_0 + \sum_{i=1}^{p} A_i\omega_i(k)\right)\rho(k) + \left(B_0 + \sum_{i=1}^{p} B_i\omega_i(k)\right)\lambda(k) + Bu(k), \tag{5.20}$$

where $B,\ A_i, B_i, i = 0, 1, \cdots, p$ are constant matrices to be defined later, $\omega_i(k),\ \forall k \in N$ are p second-order processes consisting of mutually uncorrelated real-valued random variables. For fixed $k \in N$, the random variables $\omega_1(k), \ldots, \omega_p(k)$ are not necessarily independent.

The sequence of random vectors $\lambda(k)$, $\forall k \in N$ in (5.20) is viewed as a disturbance signal. The disturbance of the system equations in (5.20) consists of two parts, $B_0\lambda(k)$ and $\sum_{i=1}^{p} B_i\omega_i(k)\lambda(k)$. We call the first term the drift component and the second the diffusion component of the disturbance. The presence of both types of multiplicative disturbances in (5.20) (*i.e.,* the drift and the diffusion terms) is an essential feature of our SCTM. As to be shown later, it allows for parameter excitations in both the state and the disturbance input matrices. The actual formulation of (5.20) under each mode of the SCTM is presented next.

In the FF mode, we set $\omega_i(k)$ to be the free flow speed $v_{f,i}(k)$ in (5.20), and the state equation can be defined as:

$$\rho(k+1) = \left(I + \sum_{i=1}^{p} A_i v_{f,i}(k)\right)\rho(k) + Bu(k), \tag{5.21}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 & \cdots & 0 \\ \frac{T_s}{l_2} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_{p\times p}, \; A_i = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_i} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_{i+1}} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times p}, \forall i \in \mathbb{Z} \cap [2, p-1],$$

$$A_p = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_p} \end{bmatrix}_{p\times p}, \; B = \begin{bmatrix} \frac{T_s}{l_1} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{T_s}{l_b} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\frac{T_s}{l_{e+1}} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{p\times 4}.$$

Equation (6.1) is a special case of (5.20) with B_i be a null matrix and $\lambda(k)$ be a null vector. Notice that in (6.1), the free-flow speed $v_{f,i}(k)$ is no longer the internal dynamics but the exogenous noise sequence. Similarly, we can define the other four modes as follows.

In the CC mode, we define $\omega_i(k) = w_{c,i}(k)$ and $\lambda(k) = (\rho_{J,1}(k), \ldots, \rho_{J,p}(k))^T$. The state equation is then

$$\rho(k+1) = \left(I + \sum_{i=1}^{p} A_i w_{c,i}(k)\right)\rho(k) + \sum_{i=1}^{p} B_i w_{c,i}(k)\lambda(k) + Bu(k), \tag{5.22}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{p\times p}, \; A_i = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{T_s}{l_{i-1}} & \cdots & 0 \\ 0 & \cdots & -\frac{T_s}{l_i} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{p\times p} \quad \forall i \in \mathbb{Z} \cap [2, p],$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{T_s}{l_{b-1}} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\frac{T_s}{l_e} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\frac{T_s}{l_p} \end{bmatrix}_{p\times 4}, B_i = -A_i, \forall i \in \mathbb{Z} \cap [1, p].$$

In the definition of the three transient modes, we are concerned with the case that the wavefront is located at the boundary between cells $L-1$ and L at time k. In the CF mode, we can define $\omega_i(k) = w_{c,i}(k),\ \forall i \in \mathbb{Z} \cap [1, L-1]$, $\omega_j(k) = v_{f,j}(k), \forall j \in \mathbb{Z} \cap [L, p]$ and the vector $\lambda(k) = (\rho_{J,1}(k), \cdots, \rho_{J,L-1}(k), Q_{M,L}(k), 0, \cdots, 0)^T$. The state equation is then

$$\rho(k+1) = \left(I + \sum_{i=1}^{p} A_i \omega_i(k)\right) \rho(k) + \left(B_0 + \sum_{i=1}^{p} B_i \omega_i(k)\right) \lambda(k) + Bu(k), \qquad (5.23)$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{p\times p}, \; A_i = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{T_s}{l_{i-1}} & \cdots & 0 \\ 0 & \cdots & -\frac{T_s}{l_i} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{p\times p},$$

$$\forall i \in \mathbb{Z} \cap [2, L-1],$$

$$A_j = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_j} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_{j+1}} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{p\times p}, \ \forall j \in \mathbb{Z} \cap [L, p-1],$$

$$A_p = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_p} \end{bmatrix}_{p\times p}, B_0 = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_{L-1}} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_L} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{p\times p},$$

$$B_i = \begin{cases} -A_i, & \forall i \in \mathbb{Z} \cap [1, L-1] \\ 0, & \forall j \in \mathbb{Z} \cap [L, p] \end{cases}, \quad B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{T_s}{l_{b-1}} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\frac{T_s}{l_{e+1}} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{p\times 4}.$$

In the FC1 mode, we define $\omega_i(k) = v_{f,i}(k),\ \forall i \in \mathbb{Z} \cap [1, L-1]$, $\omega_j(k) = w_{c,j+1}(k)$, $\forall j \in \mathbb{Z} \cap [L, p-1]$ and $\lambda(k) = (\rho_{J,L+1}(k), \cdots, \rho_{J,p})^T$. The state equation is then

$$\rho(k+1) = \left(I + \sum_{i=1}^{p-1} A_i \omega_i(k)\right) \rho(k) + \sum_{j=L+1}^{p} B_j w_{c,j}(k) \lambda(k) + Bu(k), \tag{5.24}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 & \cdots & 0 \\ \frac{T_s}{l_2} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_{p\times p} \quad A_i = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_i} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_{i+1}} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times p},$$

$$\forall i \in \mathbb{Z} \cap [2, L-1],$$

$$A_j = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{T_s}{l_{j-1}} & \cdots & 0 \\ 0 & \cdots & -\frac{T_s}{l_j} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times p}, \ \forall j \in \mathbb{Z} \cap [L, p-1],$$

$$B_{L+1} = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ -\frac{T_s}{l_L} & \cdots & 0 & \cdots & 0 \\ \frac{T_s}{l_{L+1}} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{p\times(p-L)}, B_j = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_{j-1}} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_j} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times(p-L)},$$

$$\forall j \in \mathbb{Z} \cap [L+2, p], B = \begin{bmatrix} \frac{T_s}{l_1} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{T_s}{l_b} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\frac{T_s}{l_e} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\frac{T_s}{l_p} \end{bmatrix}_{p\times 4}.$$

In the FC2 mode, we define $\omega_i(k) = v_{f,i}(k),\ \forall i \in \mathbb{Z} \cap [1, L-2]$, $\omega_j(k) = w_{c,j+1}(k)$, $\forall j \in \mathbb{Z} \cap [L-1, p-1]$ and $\lambda(k) = (\rho_{J,L}(k), \cdots, \rho_{J,p}(k))^T$. The state equation is

$$\rho(k+1) = \left(I + \sum_{i=1}^{p-1} A_i \omega_i(k)\right)\rho(k) + \sum_{j=L}^{p} B_j w_{c,j}(k)\lambda(k) + Bu(k), \tag{5.25}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 & \cdots & 0 \\ \frac{T_s}{l_2} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_{p\times p}, \ A_i = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_i} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_{i+1}} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times p},$$

$$\forall i \in \mathbb{Z} \cap [2, L-2],$$

$$A_j = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{T_s}{l_{j-1}} & \cdots & 0 \\ 0 & \cdots & -\frac{T_s}{l_j} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times p}, \ \forall j \in [L-1, p-1],$$

$$B_L = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ -\frac{T_s}{l_{L-1}} & \cdots & 0 & \cdots & 0 \\ \frac{T_s}{l_L} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times(p-L+1)}, B_j = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{T_s}{l_{j-1}} & \cdots & 0 \\ 0 & \cdots & \frac{T_s}{l_j} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}_{p\times(p-L+1)},$$

$$\forall j \in \mathbb{Z} \cap [L+1, p], \ B = \begin{bmatrix} \frac{T_s}{l_1} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{T_s}{l_b} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\frac{T_s}{l_e} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\frac{T_s}{l_p} \end{bmatrix}_{p\times 4}.$$

Notice that L is an arbitrary integer in $[2, p]$. The case that the wavefront is located at other places can be similarly defined by changing the value of L. Thus far, we have represented all the five modes as discrete time bilinear stochastic systems. Since these systems are influenced by second-order random processes, we need to find the means and variance matrices to characterize the traffic density vectors. Each of these state equation systems is associated with each mode of the SCTM as shown in Figure 5.4. In order to obtain an analytical approximation of the mean and variance of the mixture distribution of the traffic density at each time step, it is necessary to investigate the statistical properties of the cell density under each mode which will be discussed next.

5.3.5 Mean and auto-correlation of stochastic traffic densities

The dynamics of $\rho(k)$ can be represented by a discrete time bilinear stochastic system of the form (5.20), which can be further simplified into the following Markovian representation

(Tuan, 1985):

$$\rho(k+1) = (A_0 + W(k))\,\rho(k) + D(k)\lambda(k) + Bu(k), \tag{5.26}$$

where the following notations are adopted:

$$W(k) = \sum_{i=1}^{p} A_i\omega_i(k),\ D(k) = B_0 + \sum_{i=1}^{p} B_i\omega_i(k),$$

where p represents the number of cells within one SCTM subsystem. Equation (5.26) exhibits the Markovian property. Thus we can represent the state vector as:

$$\rho(k) = \Phi_w(k,0)\rho\,(0) + \sum_{t=0}^{k-1} \Phi_w(k,t+1)\,(Bu(t) + D(t)\lambda(t))\,, \tag{5.27}$$

for every $k \geq 1$, with $\Phi_w(k,k) = I$ and $\Phi_w(\tau,t) = [A_0 + W(\tau-1)]\ldots[A_0 + W(t)]$ for $\tau > t$. Consider the state sequence generated by (5.26), and define the mean and the auto-correlation matrix for each $k \geq 0$:

$$\varphi(k) = E\{\rho(k)\},\quad \Omega(k) = E\{\rho(k)\rho^T(k)\}.$$

The existence of $\varphi(k)$ and $\Omega(k)$ for each $k \geq 0$ can then be guaranteed by the independence and second-order assumptions. Therefore, by Equation (5.27) we obtain the mean as:

$$\begin{aligned}\varphi(k) &= E\left(\Phi_w(k,0)\right)\varphi\,(0)\\ &+ \sum_{t=0}^{k-1} E\left(\Phi_w(k,t+1)\right)\left(BE\left(u(t)\right) + E\left(D(t)\right)E\left(\lambda(t)\right)\right), \forall k \geq 1,\end{aligned} \tag{5.28}$$

where $E\left(\Phi_w(k,0)\right) = [A_0 + \sum_{i=1}^{p} E\left(\omega_i(0)\right)A_i]\ldots[A_0 + \sum_{i=1}^{p} E\left(\omega_i(k-1)\right)A_i]$, and the term $E\left(\Phi_w(k,t+1)\right)$ can be similarly defined. Regarding the mixed terms involving both disturbances and states, for each $k \geq 0$, by using the independent assumptions, we have:

$$\begin{aligned}G_1(k) &= E\left([A_0 + W(k)]\varphi(k)u^T(k)B^T\right)\\ &= \left[A_0\varphi(k)\left(E\left(u(k)\right)\right)^T + \sum_{i=1}^{p} A_i\varphi(k)\left(E\left(\omega_i(k)u(k)\right)\right)^T\right]B^T,\end{aligned} \tag{5.29}$$

$$\begin{aligned}G_2(k) &= E\left(D(k)\lambda(k)u^T(k)B^T\right)\\ &= \left[B_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)B_i\right]\left(E\left(\lambda(k)\right)\right)\left(E\left(u(k)\right)\right)^T B^T,\end{aligned} \tag{5.30}$$

$$\begin{aligned}G_3(k) &= E\left([A_0 + W(k)]\,\varphi(k)\lambda^T(k)D^T(k)\right)\\ &= \left[A_0\varphi(k) + \sum_{i=1}^{p} E\left(\omega_i(k)\right)A_i\varphi(k)\right]\left(E\left(\lambda(k)\right)\right)^T\cdot\\ &\quad \left[B_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)B_i\right]^T + \sum_{i=1}^{p}\gamma_i A_i\varphi(k)\left(E\left(\lambda(k)\right)\right)^T B_i^T,\end{aligned} \tag{5.31}$$

where $\gamma_i = E\left(\omega_i(k)\omega_i(k)\right) - \left(E\left(\omega_i(k)\right)\right)^2$. Furthermore, let $G(k) = G_1(k) + G_2(k) + G_3(k)$, and

$$\begin{aligned} V(k) &= G(k) + G^T(k) + B\left(E\left(u(k)u^T(k)\right)\right)B^T \\ &+ \left[B_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)B_i\right]\left(E\left(\lambda(k)\lambda^T(k)\right)\right)\left[B_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)B_i\right]^T \\ &+ \sum_{i=1}^{p} \gamma_i B_i\left(E\left(\lambda(k)\lambda^T(k)\right)\right)B_i^T. \end{aligned}$$

It can be verified by the independent argument that

$$\begin{aligned} \Omega(k+1) &= E\left(\left[A_0 + W(k)\right]\rho(k)\rho^T(k)\left[A_0 + W(k)\right]^T\right) + V(k), \\ &= \left[A_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)A_i\right]\Omega(k)\left[A_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)A_i\right]^T \\ &+ \sum_{i=1}^{p} \gamma_i A_i \Omega(k) A_i^T + V(k),\ \forall k \geq 0, \end{aligned} \tag{5.32}$$

If we define $F_0 = [A_0 + \sum_{i=1}^{p} E\left(\omega_i(k)\right)A_i]$, and $F_i = \sqrt{\gamma_i}A_i$, then (5.32) is equivalent to

$$\Omega(k+1) = \sum_{i=0}^{p} F_i \Omega(k) F_i^T + V(k),\ \forall k \geq 0. \tag{5.33}$$

The solution of (5.33) can be obtained by induction as:

$$\Omega(k) = L^k[\Omega(0)] + \sum_{t=0}^{k-1} L^{k-t-1}[V(t)],\ \forall k \geq 1, \tag{5.34}$$

where the operator $L[.]$ is defined as

$$L[X] = \sum_{i=0}^{p} F_i X F_i^T, \text{ and } L^t[X] = \sum_{i_t=0}^{p} \cdots \sum_{i_1=0}^{p} F_{i_t} \ldots F_{i_1} X F_{i_t}^T \ldots F_{i_1}^T. \tag{5.35}$$

To illustrate the mean and variance update of traffic density by the SCTM, we provide a small analytical numerical example in the Appendix of the chapter.

5.3.6 The one wavefront assumption, an interconnected SCTM approach to model a freeway corridor, and its implementation

The definitions of probabilities and the corresponding state space equations of the five modes developed in Section 5.3.3 and 5.3.4 are somehow complicated. Nevertheless, even if we use only the five modes, it is hard for us to cover the whole probabilistic space by defining the probabilities of different modes in terms of equations (5.9)-(5.12). To render the overall probability equal to 1, we define the probability of FC mode by (5.13). This definition may over-estimate the probability of the FC mode to occur, which in turn

renders the traffic state estimation inaccurate for the transient modes. Also, as pointed out in Section 5.3.3, the uncertain multiple wavefronts case is not perfectly addressed due to the complexity of the transient modes and the corresponding definitions of probabilities of occurrence. Remind that the uncertain wavefront concept has been converted into probabilistic operational modes of the freeway. To handle the above problems, we need to ensure that the traffic dynamics of a freeway segment can be accurately described by the five modes and the probabilities of occurrence of the five modes are properly defined. A simple solution is to divide a freeway corridor into several short segments wherein each of these segments is modeled by one SCTM subsystem consisting of two cells. The traffic state of each subsystem then can be covered by the five modes (i.e., the at-most one wavefront assumption is satisfied under the deterministic environment). The freeway corridor is then modeled by cascading these SCTM subsystems. We will discuss this in detail in the next chapter. In the empirical study of this chapter, we will compare the performances of the "simplified" SCTM proposed in this chapter and the interconnected SCTM subsystems approach to be introduced in the next chapter.

Figure 5.5 depicts a flow chart for implementation of the SCTM for freeway traffic state estimation. As mentioned previously in the section, we first divide a freeway corridor into several segments with each of the segments modeled by one SCTM subsystem with appropriate cells[4]. Then a calibration of the model is conducted, which gives the statistics of the boundary variables[5] with respect to time. After initializing the SCTM, we are ready to run the simulation by specifying the statistics of boundary variables as inputs to the SCTM.

5.4 Numerical example

To demonstrate the proposed method, we conduct the following numerical example. Consider a freeway segment consisting of four cells with neither on- nor off-ramp as depicted in Figure 5.6. We assume that the first three cells of this freeway segment are of 4 lanes and the last cell consists of only 3 lanes. The cell length is set to be 100 meters, and the time interval is T=5 seconds.

It is assumed that the nominal flow-density relationships of all the four cells are characterized by triangular fundamental diagrams. The nominal fundamental diagrams of the first three cells and the last cell are shown in Figures 5.7(a) and 5.7(b), respective-

[4]Please refer to Figure 6.3(b) of Chapter 6.

[5]Please refer to Section 5.3.2 of this chapter and Section 6.3.1 of Chapter 6 for the definition of boundary variables.

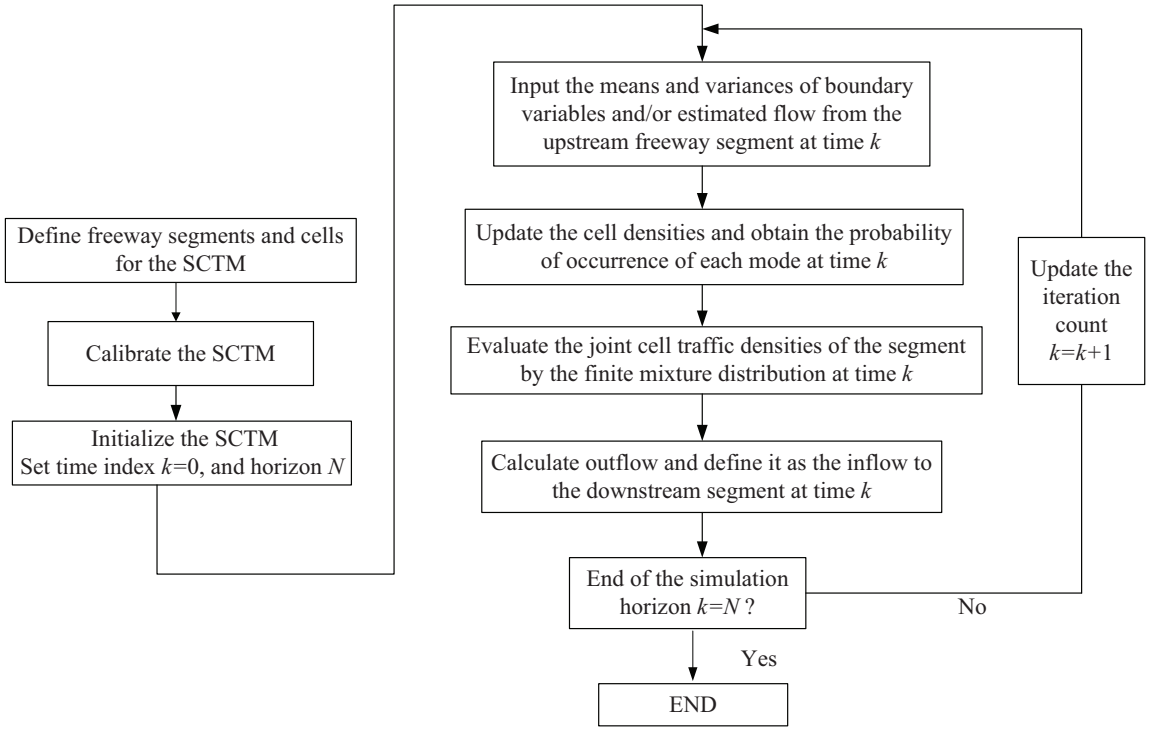


Figure 5.5: The flow chart for implementation of the SCTM for a freeway segment

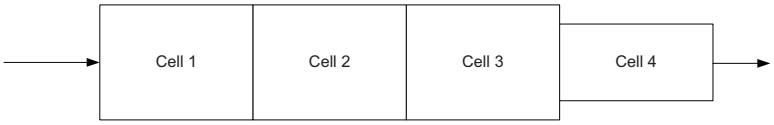


Figure 5.6: Freeway segment consisting of 4 cells

ly. To illustrate the properties of the SCTM, the noise vector of the supply parameters $\{\xi^P(k),\ k \in N\}$ is as follows:

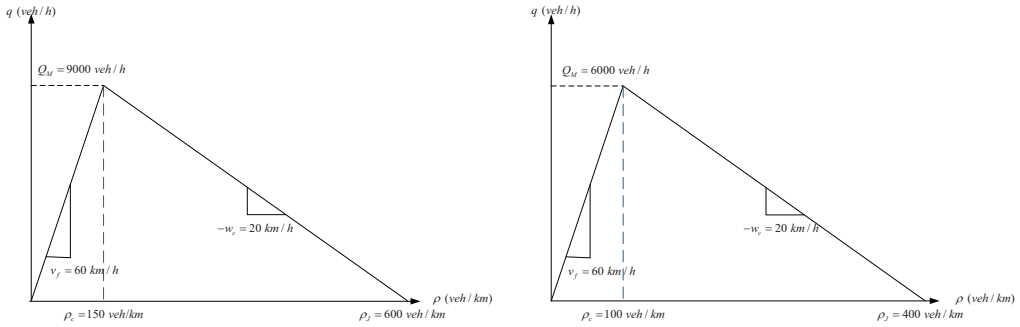
$$\xi^P(k) = \begin{bmatrix} v_f(k) \\ w_c(k) \\ \rho_J(k) \end{bmatrix} = \begin{bmatrix} v_f \\ w_c \\ \rho_J \end{bmatrix} + \begin{bmatrix} \xi_1(k) \\ \xi_2(k) \\ \xi_3(k) \end{bmatrix}, \tag{5.36}$$

where $\{\xi_i(k)\}, i = 1, 2, 3$ are mutually independent Normal distributed random sequences, and $v_f(k)$, $w_c(k)$ and $\rho_J(k)$ are respectively the vectors of the free-flow speed, the backward wave speed, and the jam density of all cells at time index k. The standard deviation of the uncertainties are assumed to be 10 percent of their nominal values. The sequences of $Q_M(k)$ and $\rho_c(k)$ can then be obtained by using the method given in Remark 5.1.

To test the build-up of congestion, we consider the following deterministic inflow profile:

$$q_u(k) = \begin{cases} 3000 \text{ vph}, & k \leq 50 \text{ time increment}; \\ 8000 \text{ vph}, & k \geq 50 \text{ time increment}. \end{cases} \tag{5.37}$$

This inflow profile yields two steady states: the first one is a free-flow state with ρ=50



a) Fundamental diagram of the first three cells

b) Fundamental diagram of the fourth cell

Figure 5.7: The nominal fundamental diagrams

veh/km for all four cells while the second one is a congestion state with ρ=300 *veh/km* for the first three cells and ρ=100 *veh/km* for the last cell.

By applying the SCTM to this example, the means and SDs of the traffic densities on all four cells over time can be obtained as shown in Figure 5.8. This figure plots the mean of the traffic density and the values of mean density plus and minus the SD. From the result, the traffic states during the low demand period (*i.e.*, time interval $[0, 50\,T]$) in the first three cells have relatively low variability. However, during the same time interval, there is a high variability of the traffic density in cell 4. In fact, the SD in this time interval increases as the flow moves to downstream cells. The gradual increase in SD and the high SD in cell 4 during this early period is due to the accumulation effect of the supply uncertainty. The downstream cells will therefore experience a higher level of uncertainties compared to the upstream cells. For the period with a higher probability of the occurrence of the congestion mode (*i.e.* $k > 50$), the variabilities of the traffic states, in the contrary, do not seem to increase as the flow moves downstream. On the other hand, in this case we can observe the propagation of the uncertainty in the reverse direction of the traffic flow. Under the congested condition, the bottleneck cell (cell 4) has a high probability to be congested due to the significant undersupply condition. Thus, the stochastic element of the backward wave becomes influential in which the uncertainty also propagates backward with the end of the queue (or backward wave).

For comparison, the Monte Carlo Simulation (MCS) with 5000 trials is also applied to the modified cell transmission model. The results in terms of the means and SDs of traffic densities are shown in Figure 5.9. By comparing Figures 5.8 and 5.9, we observe that the two methods provide similar results in terms of the mean traffic densities. However, the SDs of the densities in cells 1-3 for $k > 50$ as computed by the SCTM are significantly

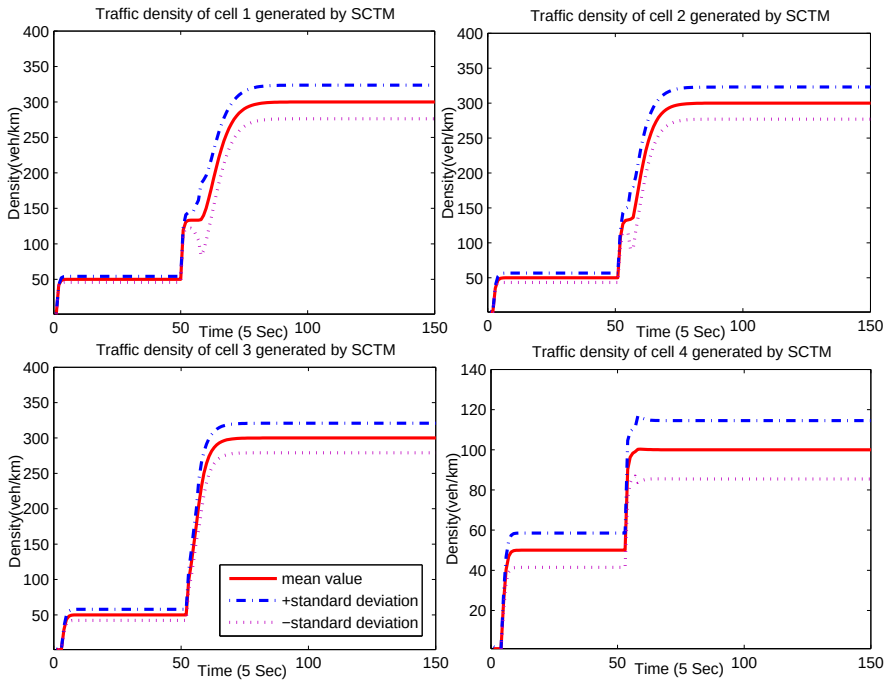


Figure 5.8: Traffic density generated by the SCTM for the test case

lower than those calculated from the MCS. The SDs from the MCS have smooth trends and transition particularly between the two steady states of the traffic condition (at k = 50). On the contrary, the SDs from the SCTM particularly in the last cell increases suddenly at the transition state.

Nevertheless, the trends of the SDs and means of the traffic densities in both cases are similar. It is not clear why the SD from the SCTM is lower than that from the MCS. However, it is noteworthy that the MCS is subject to the sampling error, which normally over-approximates the variance (or similarly SD). Various techniques for variance-reduction sampling have thus been proposed in the literature. On the other hand, the SCTM does not face this random sampling error. For the computational time, the SCTM only requires around 1 percent of the time taken by the MCS. In addition, the computer memory used by the SCTM is significantly less than that of the MCS.

The second test is set up to illustrate the propagation of SD over time and space. In this test, the same freeway corridor as shown in Figure 5.6 is adopted but we assume that only the first cell admits uncertainties (and other three downstream cells have no supply uncertainties). The same inflow demand pattern as in (5.37) is adopted. The results are shown in Figure 5.10. Despite the deterministic inflow pattern and supply characteristics of the last three cells, we can observe some uncertainties in the traffic densities in these three cells during the free-flow and transition states. When the traffic is in the free-flow steady state (*i.e.*, flow moving downstream), the SD of traffic density in the first cell propagates

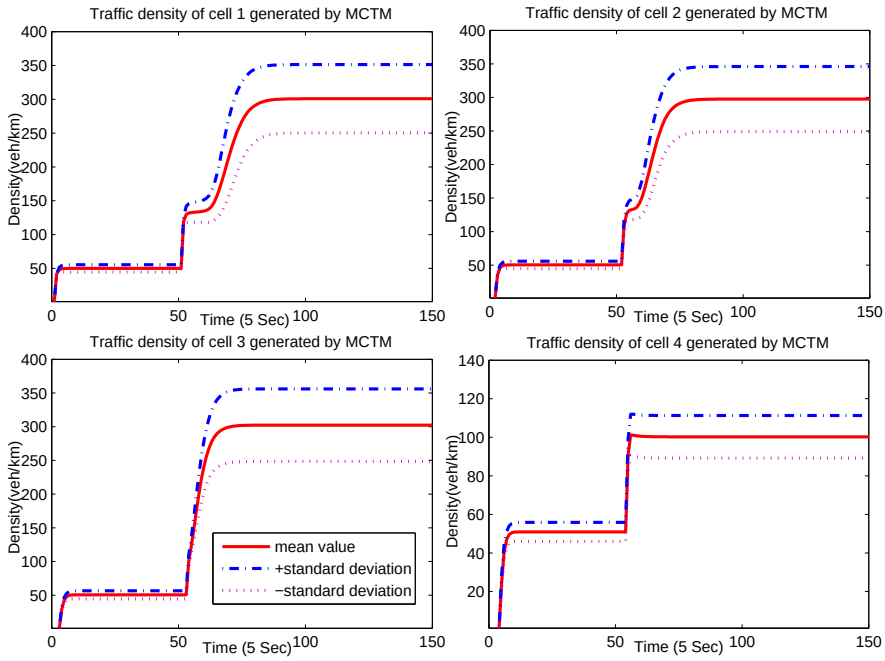


Figure 5.9: Traffic density generated by the Monte Carlo Simulation of MCTM

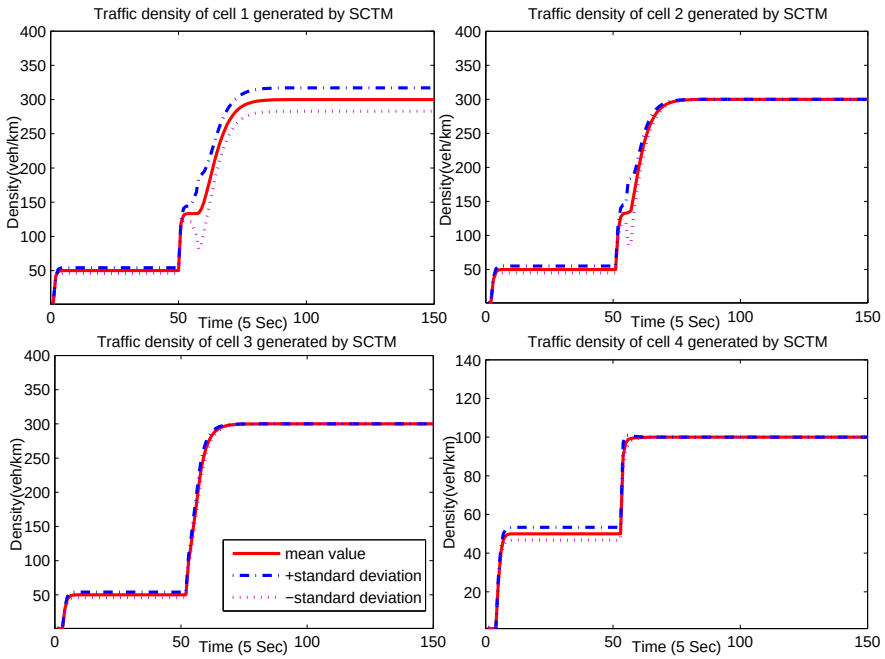


Figure 5.10: Propagation of SD of the traffic density

downstream to the following three cells. The traffic densities during the transition period between the free-flow and congested states also have certain levels of uncertainties (SD). This is due to the influence of the supply variability of the first cell which determines the time period that the downstream cells will become congested. However, once the freeway enters the state with a high probability of having congestion, the queue from the last cell builds up, spills backward and causes the last three cells to be in the definite congested state. The SDs of the last three cells during the congestion state are zero since there is no supply uncertainty and the cells are fully occupied by the vehicles.

5.5 An empirical study

In this section, we will validate the SCTM by two scenarios with empirical traffic data. To compare their performance, two SCTM approaches, i.e. the SCTM based on the one wavefront assumption proposed in this chapter (which will be denoted as the "simplified" SCTM in this test) and the interconnected SCTM subsystems approach proposed in Chapter 6, are utilized in this empirical study.

The first scenario is to test the proposed model against the supply uncertainty, i.e. only uncertain supply functions are considered. In this case, the demand pattern is chosen from a particular day. The utilized traffic flow data of 24 hours were collected on April 22, 2008 from the Performance measurement system (PeMS)[6]. In this study, we will compare the performance of the SCTM against those obtained from the MCTM and MCS of the MCTM to validate the proposed model.

The second scenario is to validate the SCTM against both demand and supply uncertainties. In this case, the demand pattern is obtained from a statistical analysis of the historical data. Traffic flow data of 7 hours (4:00 am-11:00 am) collected on Tuesday, Wednesday and Thursday of April 2008 and April 2009 from the PeMS is utilized in this test.

5.5.1 Test site description and model parameters calibration

The region of interest is a section of Interstate 210 West, approximately two miles in length, as shown in Figure 5.11. This section, located in Los Angeles, stretches from S

[6]The Freeway Performance Measurement System (PeMS: https://pems.eecs.berkeley.edu/) is conducted by the Department of Electrical Engineering and Computer Sciences at the University of California, at Berkeley, with the cooperation of the California Department of Transportation, California Partners for Advanced Transit and Highways, and Berkeley Transportation Systems.

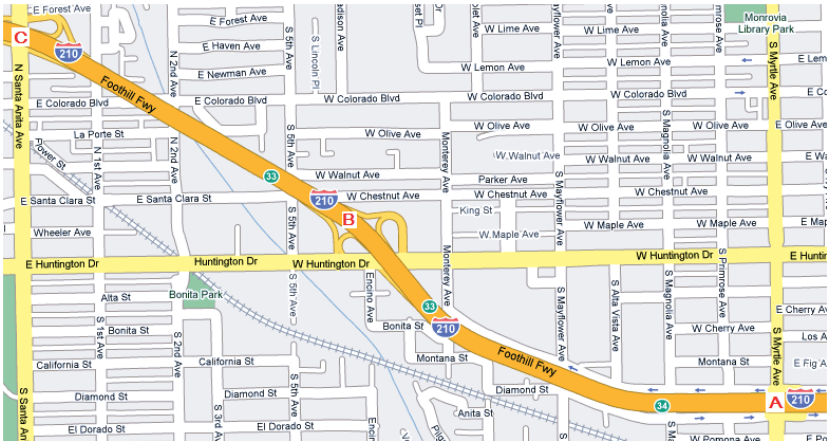


Figure 5.11: Map of the test site (Source: Google map)

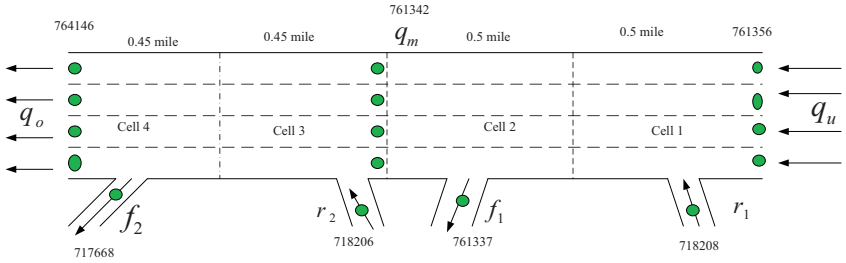


Figure 5.12: A section of I210-W divided into 4 cells and its detector configuration

Myrtle Ave (A) through W. Huntington Dr(B) to N Santa Anita Ave(C), and contains 2 on-ramps and 2 off-ramps. The section is instrumented with single-loop inductance detectors, which are embedded in the pavement along the mainline, HOV lane, on-ramps, and off-ramps. Typically, on I-210, mainline loop detectors are situated slightly upstream of on-ramp merge points. This segment of freeway is chosen here for the following reasons:

1. The high level of recurrent congestion within the section can be observed in the early morning period (6 am-10 am).

2. The segment possesses necessary infrastructure and traffic detectors embedded in the on-ramps and mainline lanes for data collection.

Figure 5.12 depicts the test section partitioned into four cells with lengths range from 0.45 to 0.5 miles. The green points along the freeway segment denote where and how many loop detectors are installed. Each loop detector group is assigned a signature of six digital numbers. q_u denotes the inflow profile of the freeway segment while q_o is the outflow profile. r_1 and r_2 denote the two on-ramps while f_1 and f_2 denote the two off-ramps. q_m denotes the flow detected by the detector installed on the boundary between cells 2 and 3.

	$\hat{v}_f$	$\sigma_{\hat{v}_f}$	$\hat{w}_c$	$\sigma_{\hat{w}_c}$	$\hat{\rho}_c$	$\hat{\rho}_J$	$\sigma_{\hat{\rho}_J}$	$\hat{Q}_m$	$\sigma_{\hat{Q}_m}$
Cell 1	64.2	5.5	15.3	4.4	132.4	686.2	172.8	8500	725.6
Cell 2	63.8	5.7	19.3	5.4	133.3	573.0	133.5	8500	764.1
Cell 3	63.2	5.5	16.6	4.9	134.6	645.4	164.0	8500	746.2
Cell 4	63.2	4.9	16.2	4.9	126.5	619.6	162.1	8000	753.4

Table 5.1: Calibration results of the four cells against the traffic flow data collected on April 22, 2008

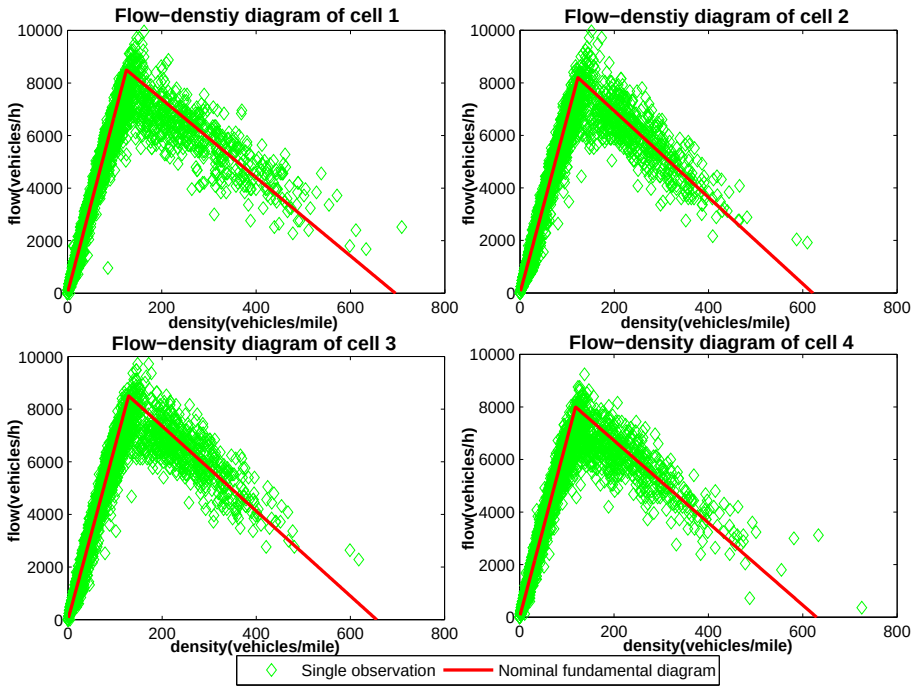


Figure 5.13: The fundamental diagrams of the four cells calibrated from the traffic flow data collected on April 22, 2008

Each loop detector gives volume (veh/time-step) and occupancy measurements every 30 seconds. Densities could then be computed for each lane using the occupancy divided by the g-factor, where the g-factor is the effective vehicle length, in miles, for the detector. A necessary condition for the numerical stability of CTM is that vehicles traveling at the maximum speed may not cross multiple cells in one time step, that is, $v_{f,i} \cdot T_s \leq l_i$. This in conjunction with the aforementioned cell lengths prohibits a simulation time step as large as 30 seconds. Thus a $zero^{th}$-order interpolation is applied to the PeMS data to yield data with $T_s = 5$ sec in order to make $v_{f,i} \cdot T_s \leq l_i$ holds for almost all the time.[7] As it was mentioned in Muñoz et al. (2003), one difficulty in selecting a test section is that it is rare for all the loop detectors in a section to be functioning properly at the same time. In the cases where detectors were not functional, the data was corrected using information from neighboring sensors or data from similar days. The interpolated, filtered, and corrected data sets were used as simulation inputs. As shown in Figure 5.13, by assuming that all the parameters must satisfy the triangular fundamental relationship and using the least square method, calibration was conducted for the four cells. Calibration results are listed in Table 5.1. The notations with hats denote the mean values of the parameters. As for example, $\hat{v}_f$ denotes the mean values of free flow speeds for the four cells. The notations σ with the mean notations as subscripts denote the standard deviations of the corresponding parameters. Compared with previous studies, such as Muñoz et al. (2003), we can verify that our calibration results are reasonable.

5.5.2 Test results against the supply uncertainty

In this subsection, three models, namely the MCTM, the Monte Carlo Simulation of MCTM and the SCTM, are used to simulate the traffic flow pattern for the calibrated section, between 4 am-12 am, during some of which the morning rush-hour congestion normally occurs. This time interval is chosen also for the reason that all the five modes would be active during this time interval. First, the MCTM is applied to the test site with the PeMS data and the calibration results. The measured and simulated mainline densities are depicted in Figure 5.14. As it was a normal day with a good calibration, the MCTM gives a quite satisfactory result. However, as we can see from the figure, good results are obtained for cells 1 and 3 only under the free flow condition. The congestion

[7] In fact, the FIFO condition, *i.e.*, $v_f \cdot T_s \leq l_i$, proposed in the CTM to ensure numerical stability can not be always satisfied in our formulation, since the free-flow speed v_f can be anything along its distribution. The concept we used here is the probabilistic FIFO which can be roughly defined as $Pr(v_{f,i} \cdot T_s \leq l_i) \geq \chi$, where χ is a positive real number which satisfies $1 - \epsilon < \chi \leq 1$ for a small real number $\epsilon > 0$.

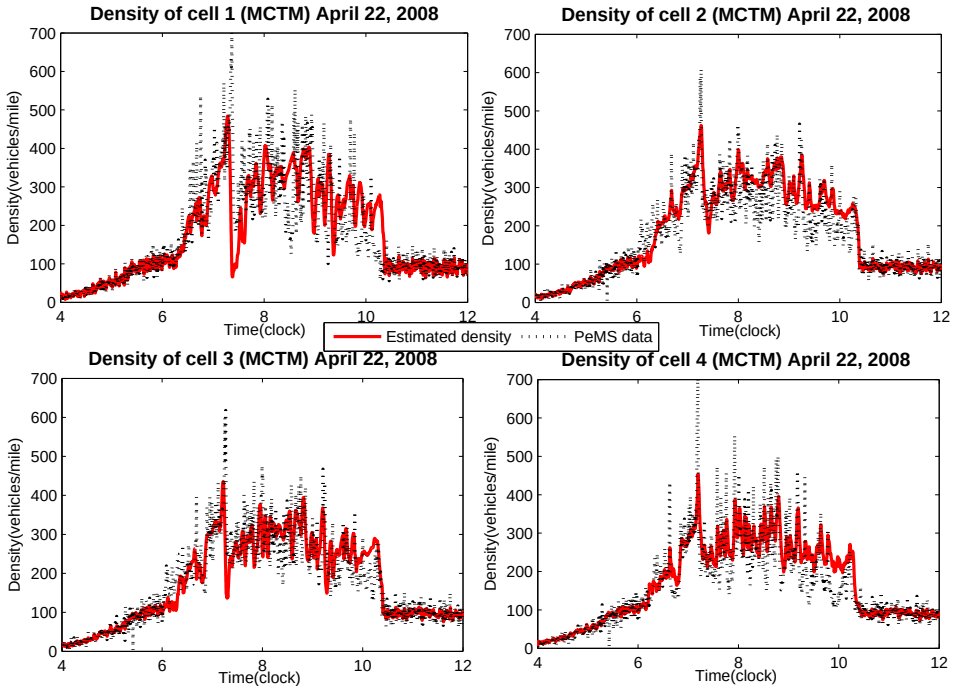


Figure 5.14: Measured densities and the MCTM's estimated densities for a segment of I-210W on April 22, 2008

states for these two cells are not well estimated by the MCTM. This may be due to the fact that congestion state introduces more supply uncertainties to our fundamental diagram than the free flow state, as demonstrated in Figure 5.13. To verify this, the Monte Carlo Simulation (MCS) of MCTM is conducted. By assuming the uncertainties obey normal distribution with means and standard deviations given in Table 5.1, the MCS of MCTM is conducted with 500 samples. The mean values of the simulated traffic densities are plotted against the measured traffic densities in Figure 5.15. As expected, some improvement is achieved by the MCS of MCTM, but not very significant. Figure 5.16[8] depicts the mean values of the simulated densities and its 68 percent confidence interval, *i.e.* $[\bar{\rho}-\sigma_{\bar{\rho}}, \bar{\rho}+\sigma_{\bar{\rho}}]$, generated by the MCS of MCTM against the measured traffic densities. Almost all the measured traffic densities including the sharp impulse points fall in this interval. We can conclude from Figure 5.16 that the MCS of MCTM with 500 samples over-estimates the means and variances of the traffic densities (the 68 percent confidence interval covers almost all the data), which is consistent with our previous simulation results. Regardless of the accuracy, this MCS of MCTM is already computational and memory demanding.

Next, the SCTM is applied to this test case. We first apply equations (5.9)-(5.16) to the definitions of probabilities of the five modes. Simulation results are depicted in Figures

[8]In the figures involving SDs, we plot the PEMS raw data every 15 minutes to reduce the resolution to make the figure clearer and more readable, while the simulated results are plotted every 5 minutes.

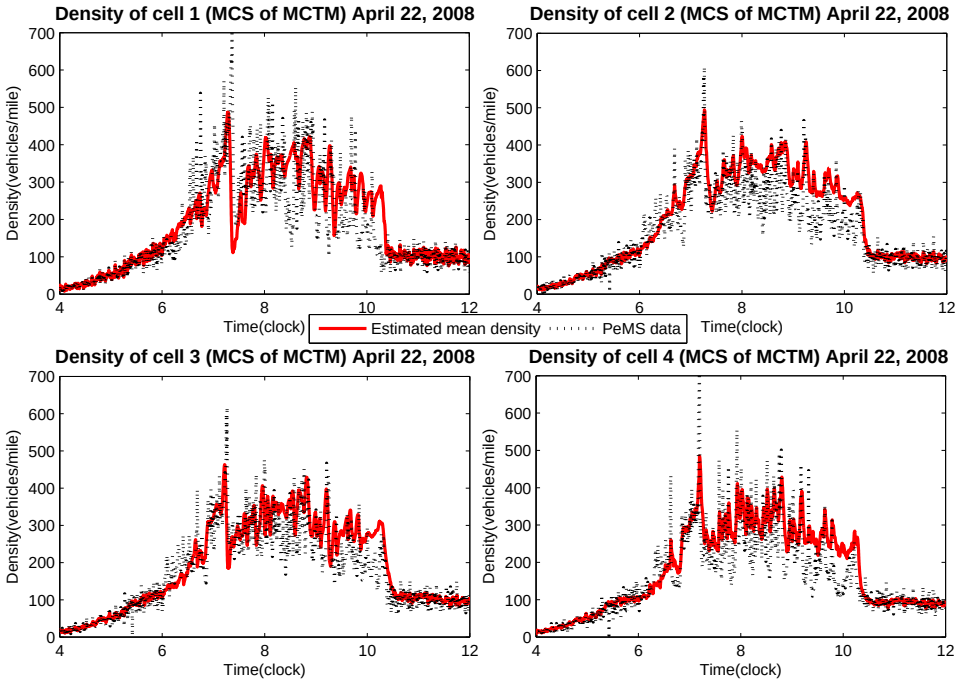


Figure 5.15: Measured densities, simulated mean densities obtained by the MCS of MCTM for a segment of I-210W on April 22, 2008

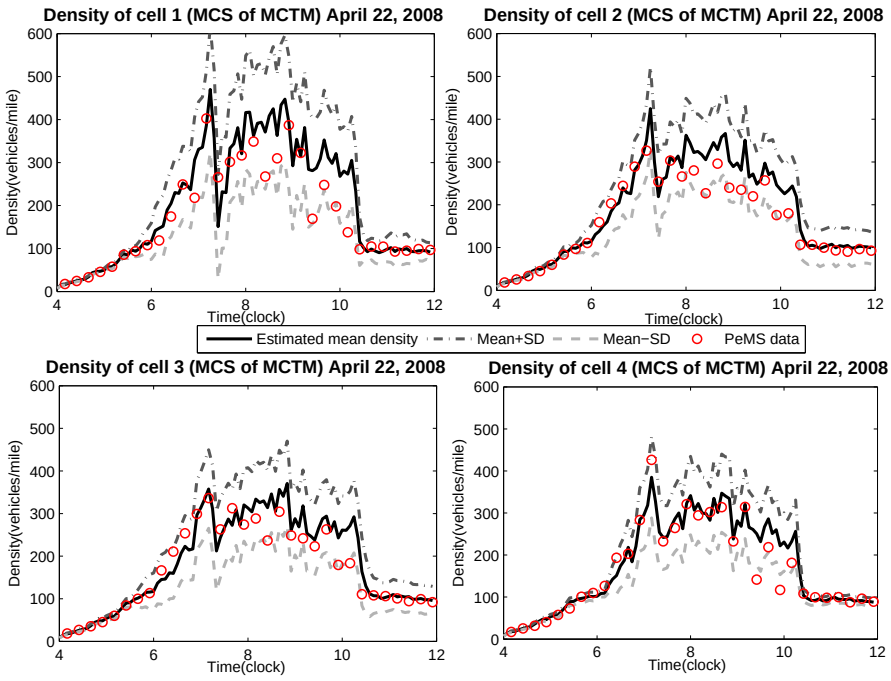


Figure 5.16: Measured densities, simulated mean densities, and the 68 percent confidence interval obtained by the MCS of MCTM for a segment of I-210W on April 22, 2008

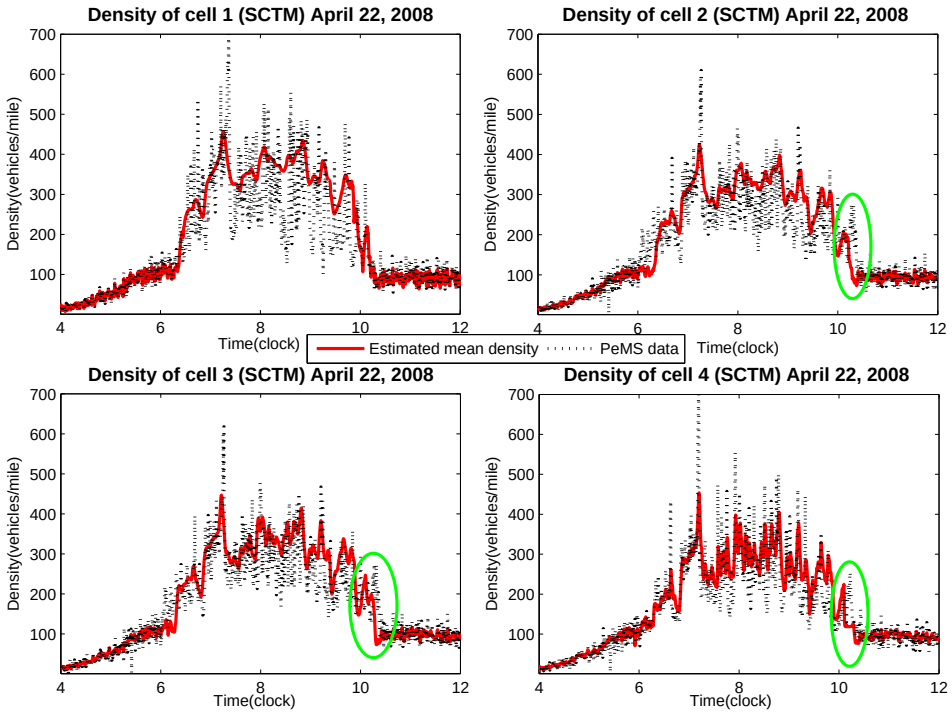


Figure 5.17: Measured densities and the SCTM's estimated mean densities for a segment of I-210W on April 22, 2008

5.17-5.19. By comparing Figure 5.17 with Figure 5.15, the SCTM produces more accurate estimated mean values than the MCS of MCTM. Figure 5.18 depicts the mean values of the simulated densities and its 68 percent confidence interval generated by the SCTM against the measured traffic densities. One can conclude from this figure that the SCTM produces more reasonable variances when compared with the MCS of MCTM. However, as the freeway segment has two pairs of on-/off- ramps, the five modes would not be enough to capture the traffic dynamics. Note that the error of the mean density is propagating upstream (or backward) in congested state. This is because the queue spills backward in congested state, which is consistent with the numerical simulation. Figure 5.19 shows the probability distributions of the five modes over time. As pointed out in Section 5.3.6, since we assign the probability of the FC mode to be the difference between 1 and the sum of the probabilities of the FF, CC, and CF modes, we may suffer from the over-estimation of the probability of the FC mode. A direct consequence of this over-estimation is that the estimated traffic densities will approach to the FF state rapidly, as illustrated in Figure 5.17.

To solve the above problem, we apply the interconnected SCTM approach proposed in Section 5.3.6. We further divide the segment into two interconnected subsystems with each subsystem having two cells. The results are shown in Figures 5.20-5.22. The mean values

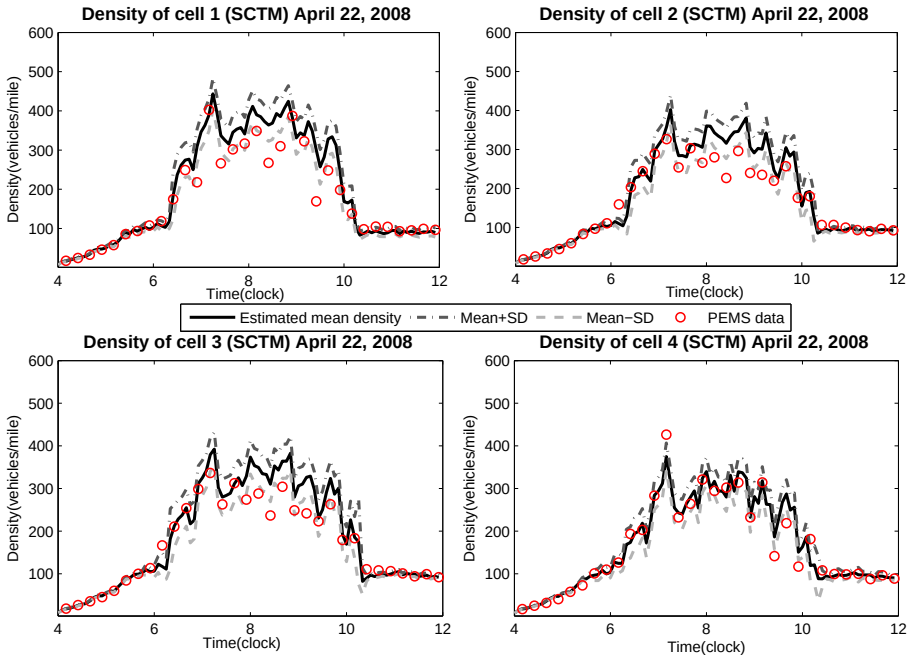


Figure 5.18: Measured densities, the SCTM's estimated mean densities and the 68 percent confidence interval for a segment of I-210W on April 22, 2008

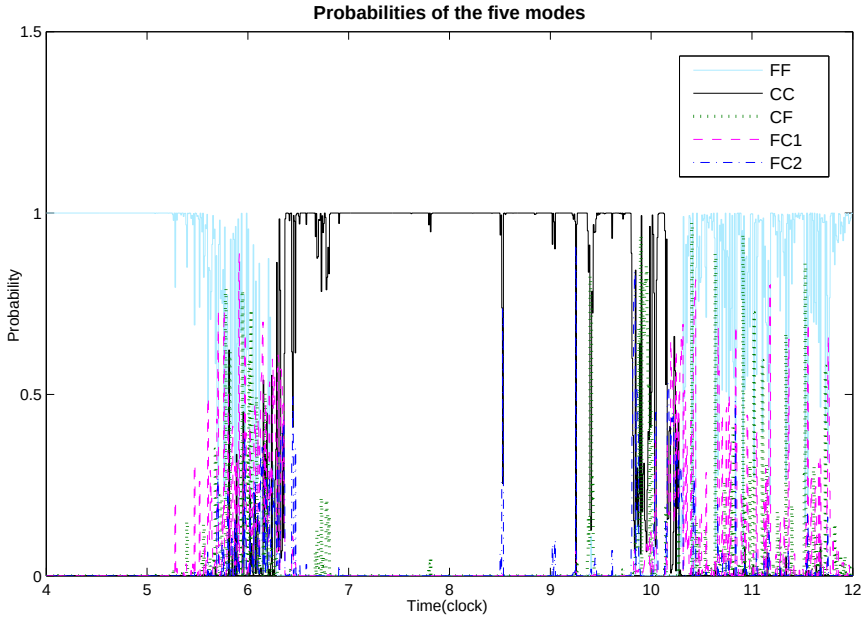


Figure 5.19: Probability distributions of different modes in the "simplified" SCTM approach over time

of the simulated traffic densities generated by the SCTM are plotted against the measured traffic densities in Figure 5.20. The figure demonstrates that the SCTM outperforms the other three techniques, *i.e.* the MCTM, the MCS of MCTM, and the previous "simplified" SCTM, in this test example. The mean values generated by the SCTM follow the measured data closely but in a smoother way, especially in the morning peak. Figure 5.21 depicts the mean values of the simulated densities and its 68 percent confidence interval generated by the SCTM against the measured traffic densities. About 60 percent of the measured traffic density data falls in the interval excluding almost all the sharp impulse points. All these sharp impulse points are taken as noise in the PeMS 30-sec data. To counteract the noise, a 1^{st}-order Butterworth low-pass filter was applied to the data using a zero-phase forward-and-reverse filtering technique, see Muñoz et al. (2003). From this example, the SCTM is found to be adaptive to the noise. The probability distributions for all five modes over time are depicted in Figure 5.22. At the beginning, *i.e.*, from 4:00 am to 5:30 am, the FF mode dominates the stochastic traffic states. After the traffic densities increase to the critical densities, the transient modes become active. The CC mode dominates the states after the transient modes. Due to the fast varying measured traffic data, all the three transient modes are active, without one dominating the simulation. From 10:30 am onward, the measured traffic densities are sliding near to the critical densities. The FF mode and its transient modes become active again. The better performance of this approach is also due to the fact that in the two-cell subsystem approach, the definitions of the probabilities of the five modes cover the whole probability space, while the "simplified" four-cell version only assigns the probability of the FC mode to be the difference between 1 and the sum of the probabilities of the other modes.

5.5.2.1 Reproducing missing data

It is assumed that the upstream and downstream mainline data (q_u, q_d), as well as the ramp flow data, are known, whereas the middle density, ρ_m, is considered to be "missing", which must be estimated. The purpose of this test is to determine whether the models can accurately reproduce ρ_m. By applying the PeMS data to the SCTM, the following simulation result is obtained. The flow data, q_m, which is assumed to be missing is reproduced by the SCTM and plotted against the measured data in Figure 5.23. From the results, the missing flow and density data is reproduced in a satisfactory manner.

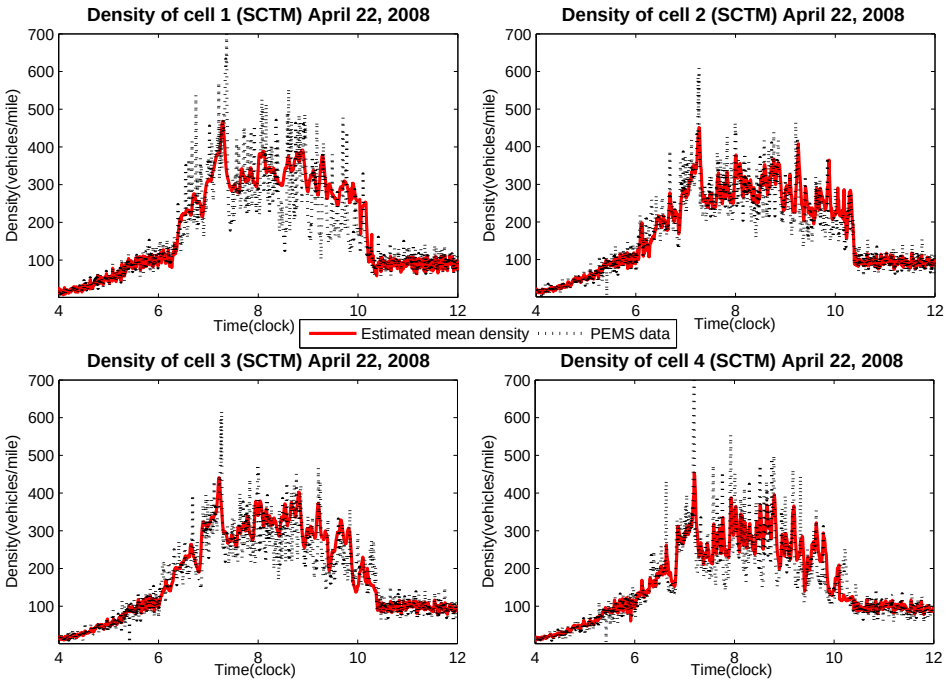


Figure 5.20: Measured densities and estimated mean densities by the interconnected SCTM for a segment of I-210W on April 22, 2008

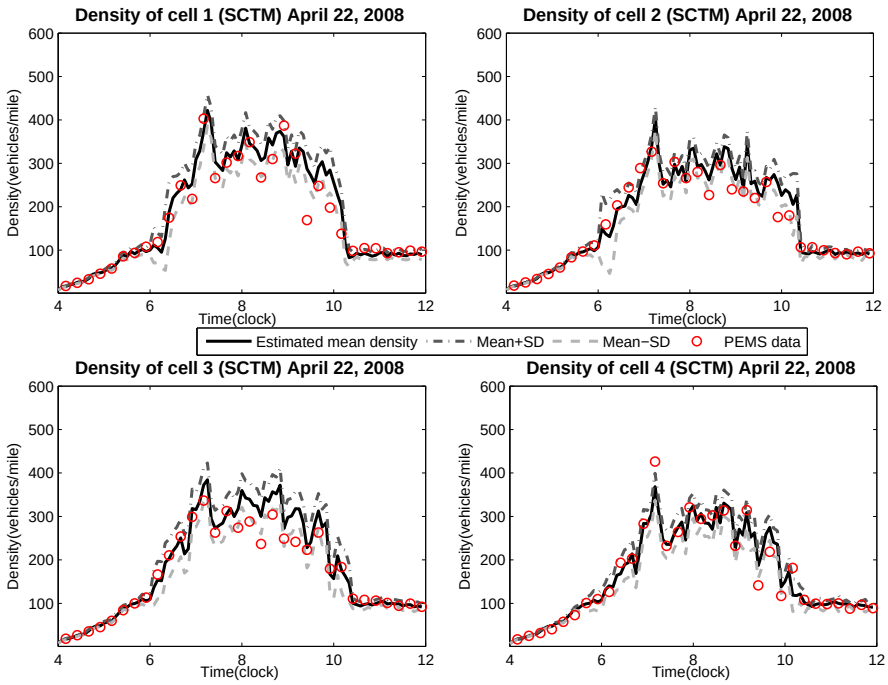


Figure 5.21: Measured densities, estimated mean densities and the 68 percent confidence interval by the interconnected SCTM for a segment of I-210W on April 22, 2008

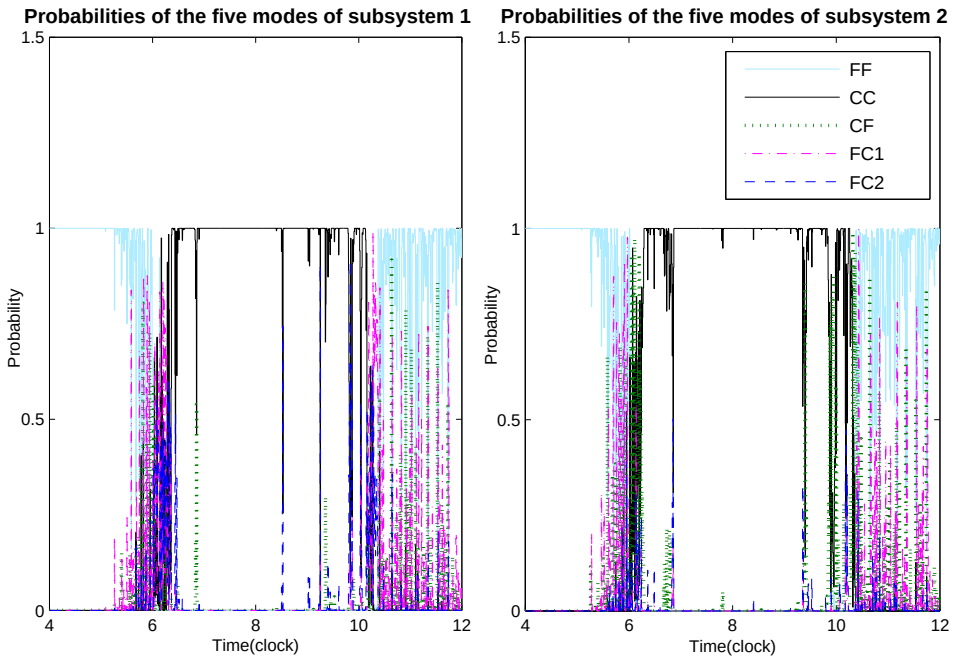


Figure 5.22: Probability distributions of different modes in the interconnected SCTM approach over time

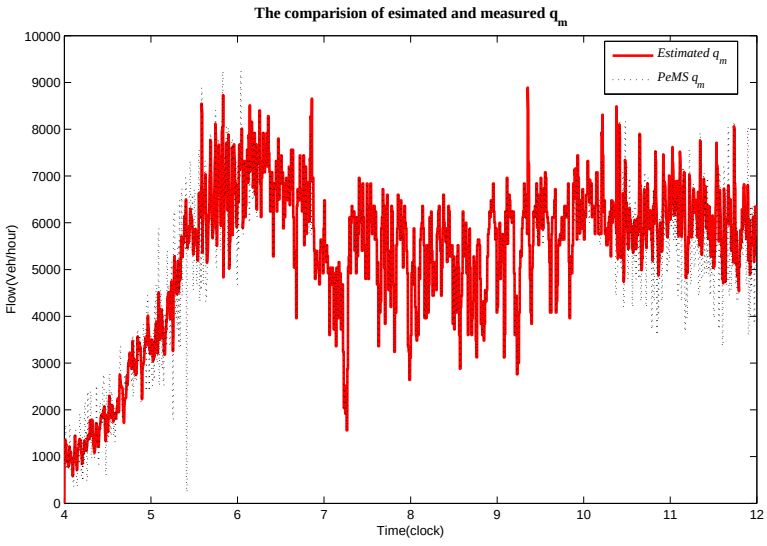


Figure 5.23: The measured "missing" flow q_m and its estimated value

	$\hat{v}_f$	$\sigma_{\hat{v}_f}$	$\hat{w}_c$	$\sigma_{\hat{w}_c}$	$\hat{\rho}_c$	$\sigma_{\hat{\rho}_c}$	$\hat{\rho}_J$	$\sigma_{\hat{\rho}_J}$	$\hat{Q}_m$	$\sigma_{\hat{Q}_m}$
Cell 1	63.6	8.95	25.72	9.12	149.36	21	518.68	149.83	9500	1336.5
Cell 2	62.46	6.69	21.69	7.16	137.68	14.75	534.22	146.80	8600	921.46
Cell 3	62.46	6.69	23.25	7.87	140.88	15.09	519.30	144.67	8800	942.89
Cell 4	63.29	7.01	20.30	6.31	127.97	14.17	526.96	137.20	8100	897.06

Table 5.2: Calibration results of the four cells against the the historical data over the selected days

	Cell 1	Cell 2	Cell 3	Cell 4	Average
MAPE	7.47 %	6.12 %	7.77 %	10.2 %	7.89 %

Table 5.3: The mean absolute percent errors of the four cells

5.5.3 Test results against both demand and supply uncertainties

This test aims to validate the SCTM against both demand and supply uncertainties. We use the traffic flow data of 7 hours (4:00 am-11:00 am) collected on Tuesday, Wednesday and Thursday of April 2008 and April 2009 from the PeMS in this test. The calibration of the stochastic triangular fundamental diagram is conducted for the four cells by using the historical data over the selected days. The results are shown in Table 5.2 and Figure 5.24. As illustrated in Table 5.2 and Figure 5.24, the supply functions admit significant uncertainties. The calibrated variances of the supply functions are lager than those shown in Table 5.1 in the previous test. Statistical analysis on the collected traffic data is also conducted for the demand side. The observed raw data of the inflow to the upstream of the segment, its mean and standard deviation with respect to time are depicted in Figure 5.25. From this figure, we can observe that the inflow profile admits significant uncertainty. In a word, both the demand[9] and supply sides admit significant uncertainties.

We input the calibrated means and variances to simulate the SCTM. The estimated traffic densities are depicted against the the historical data over the selected days in Figures 5.26-5.27. The corresponding mean absolute percent error (MAPE) between estimated mean traffic densities and the observed mean traffic densities of the four cells are reported in Table 5.3. It can be seen from Figure 5.26 that the SCTM produces an accurate estimation of mean traffic densities. The corresponding average MAPE of the four cells is about 7.9% as indicated in Table 5.3. We can observe from Figure 5.27 that the estimated variances are smaller than the observed ones especially when it approaches to the tail

[9] As we utilize the measured traffic flow data, the demand here is the statistics of historical detected inflow(s) to the segment.

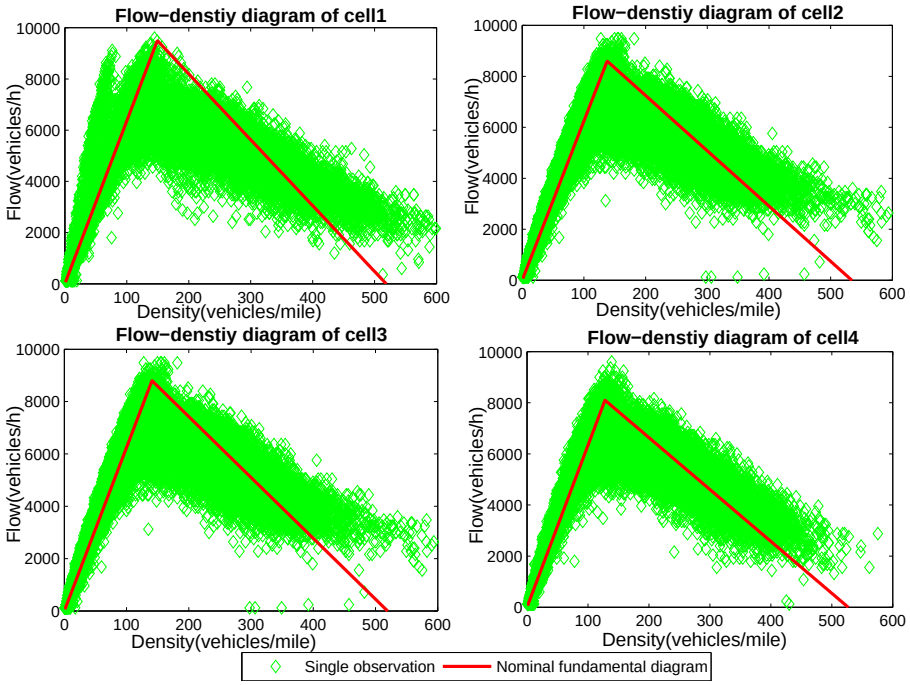


Figure 5.24: The fundamental diagrams of the four cells calibrated from the historical data over the selected days

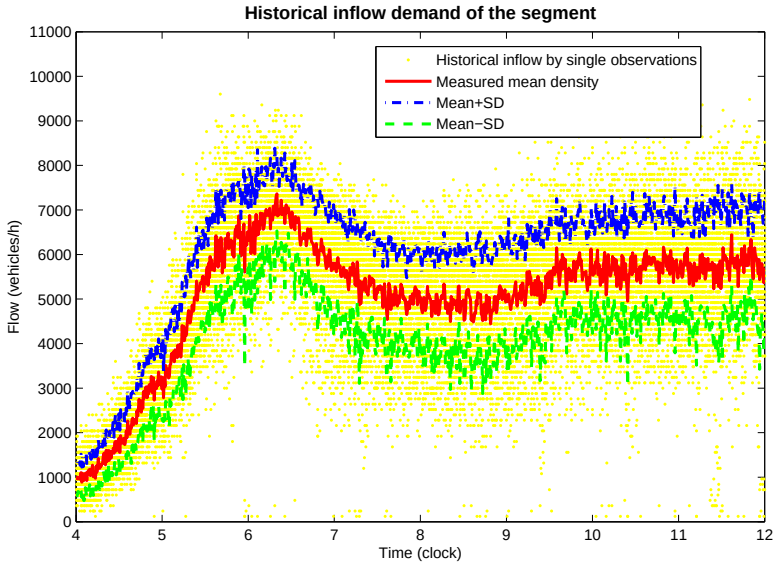


Figure 5.25: A demonstration of the demand uncertainty

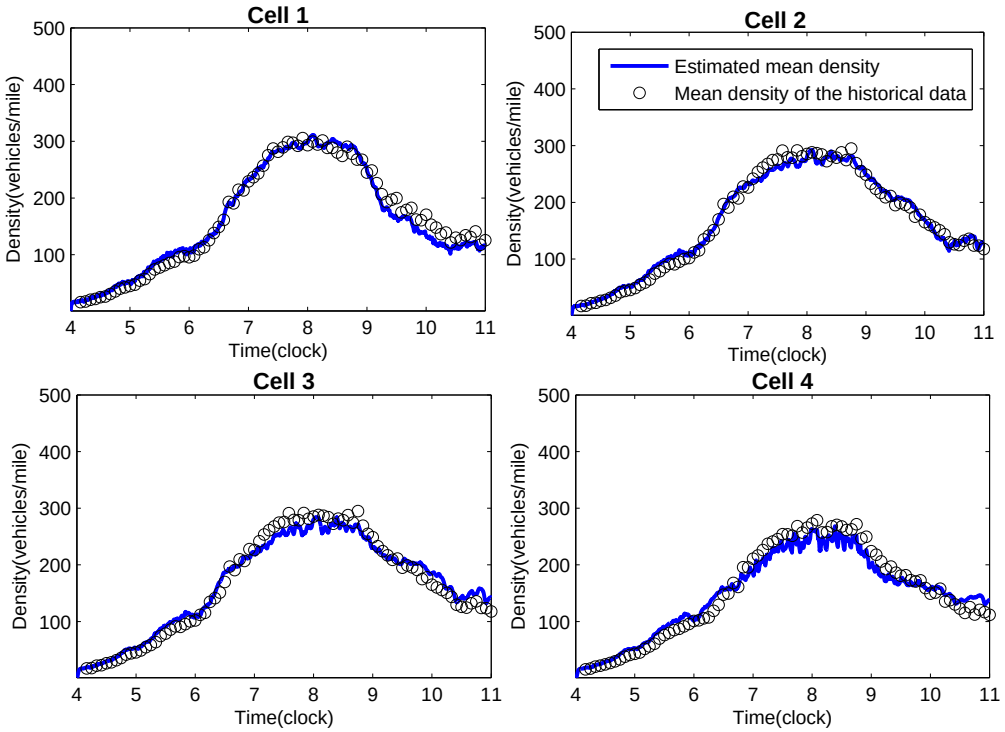


Figure 5.26: The estimated mean densities against the historical mean densities

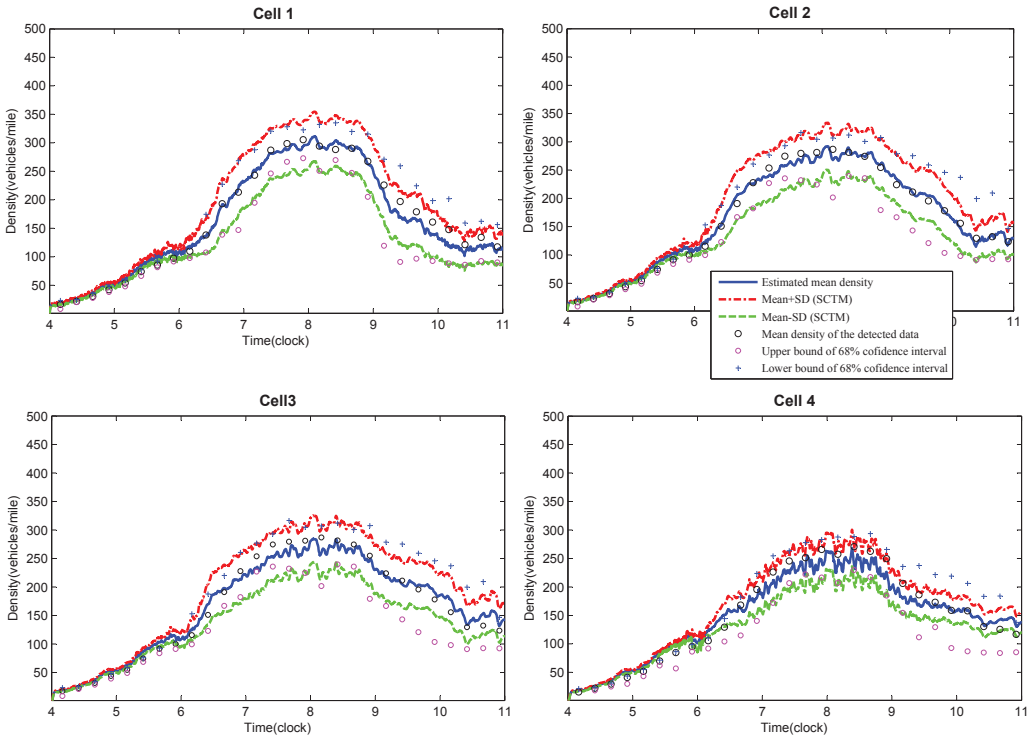


Figure 5.27: The estimated mean densities and the 68 percent confidence against the historical data over the selected days

end of the simulation horizon. Besides the error introduced by the LWR model (and its discretized version—the CTM) to approximate the traffic dynamics, this under-estimation of variance may be due to the following two reasons:

First, as illustrated in the numerical example, the SCTM itself under-estimates the variance of traffic density. This may be caused by the finite mixture (Gaussian sum) approach we utilized to approximate any possible random distribution of the traffic density.

The second reason would be the noises and errors introduced by the data detection and conversion of PeMS. The overall average error is reported to be about 16% (Chen, 2003; Chen et al., 2003).

5.6 Conclusion

In this chapter, a stochastic cell transmission model (SCTM) is proposed for simulating the traffic density of a freeway section under stochastic demand and supply. The uncertainty terms are assumed to be wide sense stationary, second-order processes consisting of uncorrelated random vectors with known means and variances. The stochasticities of the sending and receiving functions in the SCTM are governed by the random parameters of the fundamental flow-density diagrams, including the capacities, backward wave speeds, and the free-flow speeds. The model also permits random demand inflow patterns. The switching mode model of the CTM is adopted to avoid the nonlinearity of the original CTM caused by the "min" operator. The SCTM is formulated as a class of discrete time stochastic bilinear systems. A set of probabilistic switching conditions between different traffic modes for the SCTM is introduced. The chapter then provides analytical approximations of the means and SDs of the traffic densities. Numerical examples and an empirical study are carried out to illustrate the advantages of the SCTM over the Monte Carlo Simulation approach in terms of computation time and memory requirement. The illustration of the propagation of uncertainties of traffic states over time and space is also provided. However, there are some discrepancies between the SDs of traffic densities from the SCTM and Monte-Carlo simulation which may be due to the sampling error of the Monte-Carlo simulation. Empirical results from I-210W freeway case in Southern California conclude that the MCS of MCTM overestimates the SDs of the traffic densities while the SCTM underestimates the SDs a bit. The empirical study confirms that the SCTM performs well for all traffic conditions ranging from light to very dense traffic conditions. This is an advantage of the proposed model over the previous

proposed macroscopic stochastic dynamic traffic models, (*e.g.* Boel and Mihaylova (2006); Kim and Zhang (2008)). The empirical study also reveals that the SCTM outperforms the MCTM, and the MCS of MCTM. The SCTM proposed in this chapter can only apply to the freeway segment case. To capture the uncertain traffic flows on a general traffic network, we have to extend the model to consider the traffic merge and diverge as well as the interrupted infrastructures. In the next chapter, we will address these problems by connecting certain simplified SCTM subsystems.

Appendix: a small analytical numerical example

In this Appendix, we give a small analytical numerical example to illustrate the implementation of the SCTM. We consider a freeway segment consisting of two cells without on- nor off-ramp. The simulation time step is 5 seconds. The two cells are 100 meters long. We assume that the two cells admit the same nominal fundamental diagram as depicted in Figure 5.7(b). The SDs of the parameters are assumed to be 10 percent of their nominal values. The SCTM has an constant inflow rate of 5000 veh/h. Then the system matrices for the FF mode, i.e. Equation (6.1), are given as

$$A_1 = \begin{bmatrix} -0.0139 & 0 \\ 0.0139 & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & 0 \\ 0 & -0.0139 \end{bmatrix}, \; B = \begin{bmatrix} 0.0139 & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.38}$$

The system matrices for the CC mode, i.e. Equation (5.22), are obtained as

$$\begin{aligned} A_1 &= \begin{bmatrix} -0.0139 & 0 \\ 0 & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & 0.0139 \\ 0 & -0.0139 \end{bmatrix}, \\ B_1 &= -A_1, \; B_2 = -A_2, \; B = \begin{bmatrix} 0 & 0 \\ 0 & -0.0139 \end{bmatrix}. \end{aligned} \tag{5.39}$$

The system matrices for the CF mode, i.e. Equation (5.23), are obtained as

$$\begin{aligned} A_1 &= \begin{bmatrix} -0.0139 & 0 \\ 0 & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & 0 \\ 0 & -0.0139 \end{bmatrix}, \\ B_0 &= \begin{bmatrix} 0 & -0.0139 \\ 0 & 0.0139 \end{bmatrix}, \; B_1 = \begin{bmatrix} 0.0139 & 0 \\ 0 & 0 \end{bmatrix}, \; B_2 = 0. \end{aligned} \tag{5.40}$$

The system matrices for the FC1 mode, i.e. Equation (5.24), are:

$$A_1 = \begin{bmatrix} -0.0139 & 0 \\ 0.0139 & 0 \end{bmatrix}, \; B = \begin{bmatrix} 0.0139 & 0 \\ 0 & -0.0139 \end{bmatrix}. \tag{5.41}$$

The system matrices for the FC2 mode, i.e. Equation (5.25), are:

$$A_2 = \begin{bmatrix} 0 & 0.0139 \\ 0 & -0.0139 \end{bmatrix}, \; B_2 = \begin{bmatrix} 0 & -0.0139 \\ 0 & 0.0139 \end{bmatrix}, \; B = \begin{bmatrix} 0.0139 & 0 \\ 0 & -0.0139 \end{bmatrix}. \quad (5.42)$$

Other matrices which are not specified here are null matrices. To illustrate the evolution, the mean traffic densities of the two cells at time step $k = 30$ equal to 83.2870 and 83.0789 veh/km are obtained from the SCTM. The corresponding SDs are 10.3376 and 15.2070 veh/km for the two cells, respectively. The auto-correlation matrix is $Q(30) = \begin{pmatrix} 7043.6 & 6876.4 \\ 6876.4 & 7133.4 \end{pmatrix}$. Then we can obtain the probabilities of occurrence of the five modes as $P_{FF}(30) = 0.7225$, $P_{CC}(30) = 0.0217$, $P_{CF}(30) = 0.1011$, $P_{FC1}(30) = 0.1278$, and $P_{FC2}(30) = 0.0270$ according to the method proposed in Section 5.3.3. The means and auto-correlation matrices of traffic densities of the five modes at $k = 31$ can be calculated according to the dynamics proposed in Section 5.3.4 and the corresponding evaluations developed in Section 5.3.5. For example, the mean and auto-correlation matrix of the FF mode are thus calculated:

$$\begin{aligned} \rho_{FF}(31) &= (I + 60A_1 + 60A_2) \begin{pmatrix} 83.2870 \\ 83.0789 \end{pmatrix} + B \begin{pmatrix} 5000 \\ 0 \end{pmatrix} = \begin{pmatrix} 83.3256 \\ 83.2523 \end{pmatrix}, \\ Q_{FF}(31) &= (I + 60A_1 + 60A_2)\, Q(30)\, (I + 60A_1 + 60A_2)^T \\ &+ 36A_1Q(30)A_1^T + 36A_2Q(30)A_2^T \\ &+ G_1(30) + G_1^T(30) + B \begin{pmatrix} 5000 \\ 0 \end{pmatrix} \begin{pmatrix} 5000 \\ 0 \end{pmatrix}^T B^T = \begin{pmatrix} 6995.0 & 6901.8 \\ 6901.8 & 7098.1 \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} G_1(30) &= \left(I \times \begin{pmatrix} 83.2870 \\ 83.0789 \end{pmatrix} [5000\ 0] + (A_1 + A_2) \times \begin{pmatrix} 83.2870 \\ 83.0789 \end{pmatrix} [300000\ 0] \right) B^T \\ &= \begin{pmatrix} 964 & 0 \\ 5781.4 & 0 \end{pmatrix}. \end{aligned}$$

The means and auto-correlation matrices of other modes can be similarly obtained, we list them here as

$$\begin{aligned} \rho_{CC} &= \begin{pmatrix} 83.2292 \\ 87.7792 \end{pmatrix}, \; Q_{CC} = \begin{pmatrix} 7388 & 7127.4 \\ 7127.4 & 8028.2 \end{pmatrix}, \; \rho_{CF} = \begin{pmatrix} 87.9295 \\ 97.1798 \end{pmatrix}, \\ Q_{CF} &= \begin{pmatrix} 8059 & 8470.3 \\ 8470.3 & 9569.3 \end{pmatrix}, \; \rho_{FC1} = \begin{pmatrix} 83.3256 \\ 69.1514 \end{pmatrix}, \; Q_{FC1} = \begin{pmatrix} 6995 & 5720.8 \\ 5720.8 & 5064.6 \end{pmatrix}, \\ \rho_{FC2} &= \begin{pmatrix} 64.6978 \\ 87.7792 \end{pmatrix}, \; Q_{FC2} = \begin{pmatrix} 4489 & 5492.1 \\ 5492.1 & 8028.2 \end{pmatrix}. \end{aligned}$$

By substituting these means and auto-correlation matrices into Equations (5.18)-(5.19), we obtain the means of the joint traffic densities for $k = 31$ as

$$\begin{aligned} \mu(31) &= \sum_s P_s(30)\rho_s(31) = \begin{pmatrix} 83.2870 \\ 83.0789 \end{pmatrix}, \\ diag\left\{Var\left(\bar{\rho}(31) \mid \theta(31)\right)\right\} &= diag\left\{\sum_s P_s(30)Q_s(31) - \mu(31)\mu^T(31)\right\} \\ &= \begin{pmatrix} 106.865 & 0 \\ 0 & 231.2515 \end{pmatrix}. \end{aligned}$$

The corresponding SDs are 10.3376 and 15.2070 veh/km.

Chapter 6

Stochastic cell transmission model for traffic networks with demand and supply uncertainties

This chapter extends the stochastic cell transmission model (SCTM) to simulate traffic flows on networks with stochastic demand and supply. The SCTM divides a roadway segment into cells and accepts the means and variances of the stochastic travel demand and supply functions as exogenous inputs, and produces the corresponding cell traffic densities over time. This chapter defines the rules of flow propagation for freeway corridors; traffic merges/diverges; and signalized junctions. In the numerical studies, we simulate the network SCTM with a hypothetical network. We apply the SCTM to estimate the queues and delays at signalized intersections. Compared with some well-known delay and queue estimation formulas, e.g., Webster, Beckmann, McNeil, and Akcelik, the results show good consistency between the SCTM and these formulas. In addition, the SCTM describes the temporal behavior of the queue and delay distributions at signalized junctions with stochastic supply functions and (non-stationary) arrivals.

6.1 Introduction

Due to stochastic queues at signalized junctions (or interrupted facilities) and traffic states of uninterrupted facilities, predicting traffic congestion for an urban traffic network is a difficult and complex task. To capture the randomness in both demand and supply sides of the traffic network, we extend the SMM to consider stochastic parameters of the fundamental flow-density diagram as well as the stochastic travel demand. The proposed

model is entitled "the stochastic cell transmission model (SCTM)". In the SCTM, the supply uncertainties are governed by the random parameters of the triangular fundamental flow-density diagram, *e.g.,* free-flow speed, jam-density, and backward wave speed, etc. The stochastic demand is also modeled as stochastic exogenous input to the SCTM. The empirical studies in Chapter 5 and Sumalee et al. (2010b) validate the performance of the SCTM in estimating and predicting the stochastic traffic densities and dynamic travel time distribution against empirical freeway traffic data. However, the SCTM proposed in the previous chapter can only represent a single highway corridor without any interrupted facility.

Estimation and/or prediction of traffic densities on freeways and arterials are critical to traffic control and management. Most of the existing traffic control strategies, e.g. the generic advanced motorway optimal control (AMOC) tool (Kotsialos et al., 2002; Kotsialos and Papageorgiou, 2004), assumes a perfect model calibration of the traffic flow model applied. However, inaccurate calibration may occur if the underlying spatial-temporal traffic flow phenomena are not appropriately considered. In particular, the area around the capacity flow of a fundamental diagram is properly visible in real data only at active bottleneck locations (Carlson et al., 2010). These in turn introduce supply uncertainty to the traffic flow model. Perfect information of the future disturbances, i.e. travel demand, turning ratios, and exit-rate profiles of the network, are also required in the AMOC. However, due to demand and supply uncertainties, these approaches become moderate (suboptimal) when directly applied to the freeway traffic network. The efficiency of the optimal control based strategies deteriorates with increasing disturbance prediction and modeling errors. For instance, in view of the uncertain roadway capacity, any optimal control based ramp metering strategy attempting to achieve a pre-specified capacity flow value, will either lead to overload and congestion or to under utilization of the infrastructure.

Perfect model calibration and disturbance prediction are also required in the signal control strategies, e.g. traffic-responsive urban control (TUC) (Kosmatopoulos et al., 2006a). Moreover, such kind of signal control strategies requires the queue lengths (or traffic densities) of the signalized arterials to be accurately estimated or modeled. The estimation and prediction of queue overflows and delays at signalized junctions also has long been recognized as an important issue to signal performance measures and optimization in transportation engineering and operations research. Nevertheless, a general theory for such queue and delay estimation/prediction problem is still insufficient (Viti and Van Zuylen, 2009, 2010).

For better implementation of traffic control strategies, a method is needed to capture the traffic dynamics on the freeways and arterials under demand and supply uncertainties, and to measure the queues at signalized junctions by assuming any temporal distribution of the arrivals. The objective of this chapter is to extend the SCTM to network case for the purposes of traffic state surveillance, prediction and control. Specifically, this chapter proposes a SCTM framework to simulate the stochastic dynamics of a traffic network with mixed freeways and urban arterials. Detailed objectives are:

1. To model the stochastic traffic dynamics of freeways;
2. To model the on-/off- ramps of freeways or traffic merge/diverge operations;
3. To model the signal effects and the turning movements of urban arterials.

Parallel to Daganzo (1995a) and to meet our purpose, we extend the SCTM for networks by assuming that a temporal OD table (with mean and variance for each element) is given and the statistics of the temporal turning ratios are known for every junction.

After the theoretical development, we conduct simulations to demonstrate the application of the model. The first simulation illustrates the feasibility to apply the model as a stochastic dynamic traffic network model. We demonstrate the stochastic flow propagation along the links by calculating the means and variances of traffic densities. As we have converted the possible wave-fronts to the probabilities of occurrence of different operational modes of the SCTM, the possible wave-fronts are captured by tracing these probabilities. In the second simulation, we focus on estimating the stochastic delays and queues at signalized junctions. By applying the proposed signalized SCTM and the stochastic dynamic travel time estimation method proposed by Sumalee et al. (2010b) to a signalized junction, we obtain the time average delays with respect to different levels of saturation. The results are then compared with the delay estimations obtained from the traditional methods, e.g., Webster's, McNeill's and Akcelik's formulas. A comparison of the queue length is also carried out.

The rest of the chapter is organized as follows: The next section introduces a general structure of a traffic network and basic functional blocks which are required for representing the network. The dynamics of a basic SCTM subsystem is presented in Section 6.2 for completeness. The third and fourth sections illustrate the modifications of the SCTM to represent the four basic functional blocks of a traffic network. Numerical simulations are then carried out in Section 6.5 to illustrate the application of the model. The final section concludes the chapter.

6.2 Descriptions of a traffic network and the basic SCTM

A traffic network is usually composed of freeways and urban arterials. Freeways interact with urban arterials through on-ramps and/or off-ramps. On-ramp flows from urban arterials to freeways are usually controlled by the on-ramp metering, while off-ramp flows from freeways to urban arterials are usually controlled by the signals installed on arterials. We identify four basic functional blocks of a traffic network, i.e. (i) a freeway corridor, (ii) a traffic diverge (off-ramp), (iii) a traffic merge (on-ramp), and (iv) a signalized junction. Blocks (i)-(iii) can be further recognized as uninterrupted facilities and block (iv) be considered as interrupted facility. In what follows, we will model these basic functional blocks by using the basic SCTM subsystem. A SCTM representation of a typical traffic network can then be defined by connecting these basic functional blocks.

As mentioned in Chapter 5, to utilize the SMM, we need to ensure that the traffic dynamics of a freeway segment can be accurately described by the five modes (i.e. the at-most one wavefront assumption is satisfied under the deterministic environment) and the probabilities of occurrence of the five modes are properly defined. However, Assumption 5.1 cannot be fulfilled for general freeway segments except some special cases. A simple solution is to divide a freeway corridor into several short segments wherein each of these segments is modeled by one SCTM subsystem consisting of two cells. In this chapter, we will build up an SCTM representation of a traffic network by connecting these basic SCTM subsystems. Therefore, it is necessary for us to refine the dynamics of a basic SCTM subsystem.

Figure 6.1 shows the application of the basic SCTM to represent a freeway segment without on-/off- ramps, which is divided into two cells. This basic SCTM consisting of two cells is defined as a basic subsystem of a traffic network as depicted in Figure 6.2. The basic SCTM accepts the random inflows (uncertain demand) as well as random parameters of the fundamental flow-density diagram (uncertain supply functions) with known means and variances of the freeway segment as exogenous inputs, and then calculates the means and variances of the traffic densities, outflow of the freeway segment, and probabilities of its operational modes. We specify the dynamics of each mode of the SCTM by the bilinear system formulation of (5.20) as:

In the FF mode, we set $\omega_i(k)$ to be the free flow speed $v_{f,i}(k)$ in (5.20), and the state equation can be represented as:

$$\rho(k+1) = \left(I + \sum_{i=1}^{2} A_i v_{f,i}(k) \right) \rho(k) + Bu(k), \tag{6.1}$$

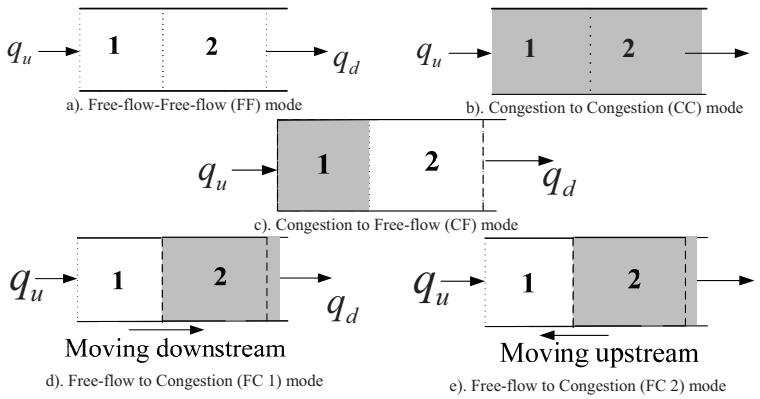


Figure 6.1: Five traffic operational modes for a freeway segment with 2 cells

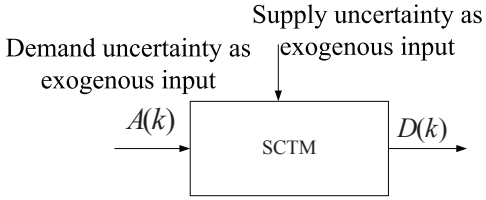


Figure 6.2: A block diagram of the basic SCTM subsystem

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 \\ \frac{T_s}{l_2} & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{T_s}{l_2} \end{bmatrix}, \; B = \begin{bmatrix} \frac{T_s}{l_1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Equation (6.1) is a special case of (5.20) with $B_i,\ i = 1, 2$ be null matrices and $\lambda(k)$ be a null vector.

In the CC mode, we define $\omega_i(k) = w_{c,i}(k)$ and the vector $\lambda(k) = (\rho_{J,1}(k),\ \rho_{J,2}(k))^T$. The state equation is then

$$\rho(k+1) = \left(I + \sum_{i=1}^{2} A_i w_{c,i}(k) \right) \rho(k) + \sum_{i=1}^{2} B_i w_{c,i}(k) \lambda(k) + Bu(k), \tag{6.2}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 \\ 0 & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & \frac{T_s}{l_1} \\ 0 & -\frac{T_s}{l_2} \end{bmatrix}, \; B = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{T_s}{l_2} \end{bmatrix}, \; B_i = -A_i, i = 1, 2.$$

In the CF mode, we can define $\omega_1(k) = w_{c,1}(k)$, $\omega_2(k) = v_{f,2}(k)$, and the vector $\lambda(k) = (\rho_{J,1}(k), Q_M(k))^T$. The state equation is then

$$\rho(k+1) = \left(I + \sum_{i=1}^{2} A_i \omega_i(k) \right) \rho(k) + \left(B_0 + \sum_{i=1}^{2} B_i \omega_i(k) \right) \lambda(k) + Bu(k), \tag{6.3}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 \\ 0 & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{T_s}{l_2} \end{bmatrix}, \; B_0 = \begin{bmatrix} 0 & -\frac{T_s}{l_1} \\ 0 & \frac{T_s}{l_2} \end{bmatrix}, \; B_1 = -A_1, \; B_2 = 0, \; B = 0.$$

In the FC1 mode, we define $\omega_1(k) = v_{f,1}(k)$, $\omega_2(k) = 0$, and $\lambda(k)$ as a null vector. The state equation is then

$$\rho(k+1) = (I + A_1\omega_1(k))\,\rho(k) + Bu(k), \tag{6.4}$$

where

$$A_1 = \begin{bmatrix} -\frac{T_s}{l_1} & 0 \\ \frac{T_s}{l_2} & 0 \end{bmatrix},\ B = \begin{bmatrix} \frac{T_s}{l_1} & 0 \\ 0 & -\frac{T_s}{l_2} \end{bmatrix}.$$

In the FC2 mode, we define $\omega_1(k) = 0,\ \omega_2(k) = w_{c,2}(k)$, and $\lambda(k) = (0, \rho_{J,2}(k))^T$. The state equation is

$$\rho(k+1) = (I + A_2\omega_2(k))\,\rho(k) + B_2\omega_2(k)\lambda(k) + Bu(k), \tag{6.5}$$

where

$$A_1 = 0,\ A_2 = \begin{bmatrix} 0 & \frac{T_s}{l_1} \\ 0 & -\frac{T_s}{l_2} \end{bmatrix},\ B_1 = 0,\ B_2 = \begin{bmatrix} 0 & -\frac{T_s}{l_1} \\ 0 & \frac{T_s}{l_2} \end{bmatrix}, B = \begin{bmatrix} \frac{T_s}{l_1} & 0 \\ 0 & -\frac{T_s}{l_2} \end{bmatrix}.$$

The corresponding probabilities of occurrence of the five operational modes can be defined as:

FF mode: $P_{FF}(k) \triangleq \Pr\left(\rho_u(k-1) < \rho_{c,1}(k-1) \cap \rho_d(k-1) < \rho_{c,2}(k-1)\right)$,

CC mode: $P_{CC}(k) \triangleq \Pr\left(\rho_u(k-1) \geq \rho_{c,1}(k-1) \cap \rho_d(k-1) \geq \rho_{c,2}(k-1)\right)$,

CF mode: $P_{CF}(k) \triangleq \Pr\left(\rho_u(k-1) \geq \rho_{c,1}(k-1) \cap \rho_d(k-1) < \rho_{c,2}(k-1)\right)$, and

FC mode: $P_{FC}(k) \triangleq 1 - (P_{FF}(k) + P_{CC}(k) + P_{CF}(k))$,

with the wave-front, which is located at the boundary between cells 1 and 2, moving downstream (event D) as

$$P_{D|FC}(k) \triangleq \Pr\left(v_{f,1}(k-1)\bar{\rho}_1(k-1) \leq w_2(k-1)(\rho_{J,2}(k-1) - \bar{\rho}_2(k-1))\right),$$

and the wave-front moving upstream (event U) as $P_{U|FC}(k) = 1 - P_{D|FC}(k)$. Then the probabilities of the FC1 and FC2 to occur at time step k are: FC1 mode: $P_{FC1}(k) \triangleq P_{D|FC}(k)P_{FC}(k)$, and FC2 mode: $P_{FC2}(k) \triangleq P_{U|FC}(k)P_{FC}(k)$, where $\rho_{c,i}$ is the critical density, w_i the backward congestion wave speed, $\rho_{J,i}$ the jam density of cell i, respectively. $\bar{\rho}_i$ is the joint density of cell i, which is defined as a finite mixture distribution of the five modes. The joint traffic density vector, its mean and covariance matrix can be defined and evaluated according to (5.17)-(5.19).

6.3 Uninterrupted facilities

In the CTM, flow propagation is defined by solving the flow cross two adjacent cells, i.e. the minimum of the sending and receiving function of the two cells. The basic elements of the CTM, i.e. the cells and wave-fronts, are now the cells, basic SCTM subsystems and the probabilities of operational modes in the SCTM framework. It is validated by the empirical study of Chapter 5 that a simple way obtain an accurate traffic state estimation for a long freeway corridor with possible multiple wavefronts by the SCTM framework is to divide the corridor into several short segments with each segment modeled by one SCTM subsystem. The long freeway corridor under consideration is then represented by connecting these basic SCTM subsystems as demonstrated in Figure 6.3. In this section, we will discuss this interconnected basic SCTM subsystems approach in detail and extend it to model the uninterrupted facilities including the freeway corridor, traffic merge and diverge.

6.3.1 Freeway corridor

As previously explained, we model a long freeway corridor by cascading several SCTM subsystems as depicted in Figure 6.3(b). Each SCTM subsystem admits several exogenous inputs as shown in Figure 6.2 and Figure 6.3(a). To furnish the modeling, we need to define a number of boundary variables to simulate the stretch SCTM system for a freeway corridor:

1. flow at the stretch origin $q_{u,1}$,
2. flow at the stretch destination $q_{d,y}$,
3. measured on-ramp flows $r_{on,j}$ (if any), and measured off-ramp flows $r_{off,j}$ (if any),
4. the uncertain supply functions of each cells.

Similar idea has been adopted in Wang et al. (2007). Inside the stretch SCTM system, each of these subsystems accepts the outflow from the upstream segment as inflow.

To capture the stochastic flow propagation, we need to define the flow propagation law between two neighboring SCTM subsystems besides the above flow propagation law within one SCTM subsystem. Note that the concept of wavefront in the original CTM framework is converted into five operational modes (or five probabilistic events) and their probabilities of occurrence in the SCTM framework. The calculation of flow between two neighboring SCTM subsystems is similar to the calculation of flow between two adjacent

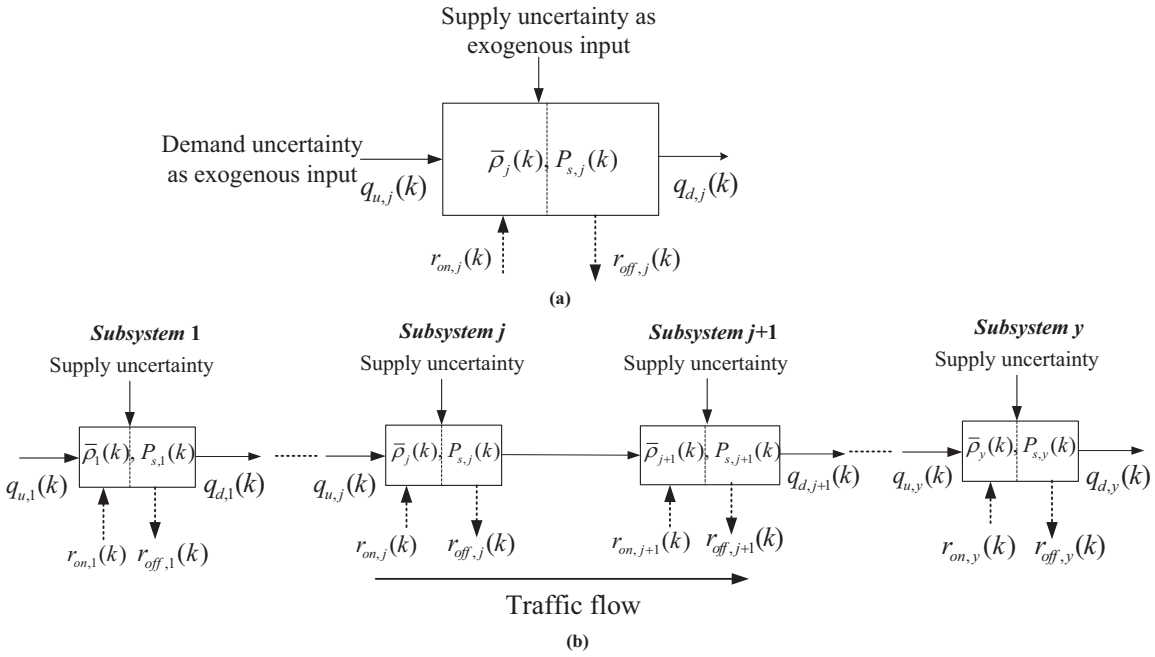


Figure 6.3: An interconnected SCTM approach to model a freeway corridor: (a) a short segment as one SCTM subsystem, segment variables, and segment inputs; (b) a freeway corridor as interconnected SCTM subsystems.

cells without wavefront that can be represented by a finite mixture distribution similar to (5.17). As depicted in Figure 6.4, let subsystems $j-1$ and j are two neighboring SCTM subsystems with two adjacent cells $i-1$ and i. Let $S_{j-1}(k)$ denote the sending function of subsystem $j-1$ (which is one of the outputs of the SCTM subsystem). Then

$$S_{j-1}(k) = mix\left(v_{f,i-1}(k)\bar{\rho}_{i-1}(k), Q_{i-1}(k)\right), \tag{6.6}$$

where mix denotes the finite mixture distribution defined similar to (5.17). The finite mixture distribution definition of the sending function (6.6) means that: if the last cell of subsystem $j-1$, is free flowing at time k, the amount of traffic to be sent out is $v_{f,i-1}(k)\bar{\rho}_{i-1}(k)$, if the last cell of subsystem $j-1$ is congested, the amount to be sent out is $Q_{i-1}(k)$. The probabilities for these two events are $P_1^S(k) = \left(P_{FF,j-1}(k) + P_{CF,j-1}(k)\right)$, $P_2^S(k) = \left(P_{FC,j-1}(k) + P_{CC,j-1}(k)\right)$, respectively. The mean and variance of (6.6) can be evaluated by (5.18)-(5.19). To determine the flow received by the downstream SCTM subsystem, we compare this flow profile with the receiving flow of the downstream SCTM subsystem and define the following four events:

1. The first cell of the downstream SCTM subsystem is free-flow (FF_i) and the sending function $S_{j-1}(k)$ is less than its capacity. In this case, $S_{j-1}(k)$ will be loaded onto the first cell of subsystem j. The corresponding probability is defined as: $P_1(k) =$

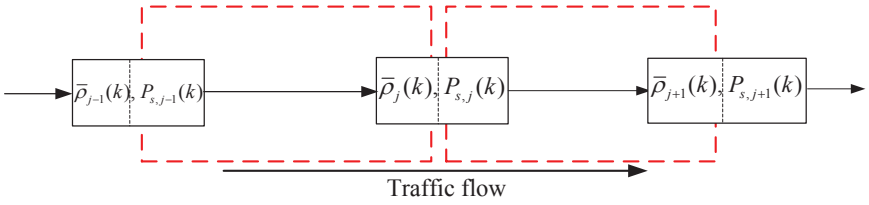


Figure 6.4: The interconnected SCTM subsystems approach as paired up two neighboring cells

$\Pr\left(FF_i(k) \cap (S_{j-1}(k) < Q_i(k))\right)$, where $Q_i(k)$ is the capacity of the first cell of the downstream SCTM subsystem;

2. The first cell of the downstream SCTM subsystem is free-flow (FF_i) and the sending function $S_{j-1}(k)$ is not less than its capacity. In this case, an amount of vehicles equals to $Q_i(k)$ will be loaded. The probability for this event is defined as: $P_2(k) = \Pr\left(FF_i(k) \cap (S_{j-1}(k) \geq Q_i(k))\right)$;

3. The first cell of the downstream SCTM subsystem is congested (CC_i) and the sending function $S_{j-1}(k)$ is less than its available space $w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right)$. Then, $S_{j-1}(k)$ will be loaded onto the first cell of subsystem j. The probability is $P_3(k) = \Pr\left(CC_i(k) \cap (S_{j-1}(k) < w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right))\right)$, where $w_i, \rho_{J,i}, \bar{\rho}_i$ are the backward wave speed, the jam density, and the density of the first cell of subsystem j, respectively;

4. The first cell of the downstream SCTM subsystem is congested (CC_i) and the sending function $S_{j-1}(k)$ is not less than its available space. In this case, an amount of vehicles which equals to the available space of the first cell of subsystem j will be loaded, The probability for this event is thus defined: $P_4(k) = \Pr\left(CC_i(k) \cap (S_{j-1}(k) \geq w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right))\right)$.

According to the definitions of probabilities of occurrence of the five modes, $FF_i(k)$ and $CC_i(k)$ are determined by the the traffic condition of the subsystem j at time k-1. To simplify the calculation, we assume $FF_i(k)$ and $CC_i(k)$ are independent of the events $(S_{j-1}(k) < Q_i(k))$, $(S_{j-1}(k) < w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right))$, $(S_{j-1}(k) \geq Q_i(k))$, and $\left(S_{j-1}(k) \geq w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right)\right)$. Then the probabilities can be calculated as:

$$
\begin{aligned}
P_1(k) &= \left(P_{FF,j}(k) + P_{FC,j}(k)\right) \Pr\left(S_{j-1}(k) < Q_i(k)\right), \\
P_2(k) &= \left(P_{FF,j}(k) + P_{FC,j}(k)\right) \Pr\left(S_{j-1}(k) \geq Q_i(k)\right), \\
P_3(k) &= \left(P_{CC,j}(k) + P_{CF,j}(k)\right) \Pr\left(S_{j-1}(k) < w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right)\right), \\
P_4(k) &= \left(P_{CC,j}(k) + P_{CF,j}(k)\right) \Pr\left(S_{j-1}(k) \geq w_i(k)\left(\rho_{J,i}(k) - \bar{\rho}_i(k)\right)\right),
\end{aligned}
$$

with $\sum_y P_y(k) = 1$. We thus define the PDF for the traffic flow received by subsystem j, $R_j^a(k)$ as a finite mixture of the four probabilistic events:

$$g_R\left(R_j^a(k)|\lambda(k)\right) = \sum_y P_y(k) g_R\left(R_j^a(k)|\lambda_y(k)\right), \tag{6.7}$$

where $\lambda(k) = \{\lambda_y(k)\}$, $\lambda_y(k) = \left(P_y(k),\ R_y(k)\right)$. The set λ contains the four events defined previously, with P_y, $R_y(k)$, $y = 1, 2, 3, 4$, the probabilities and receiving flows of the four events. The mean and variance of the joint receiving flow (6.7) can be evaluated according to (5.18) and (5.19), respectively.

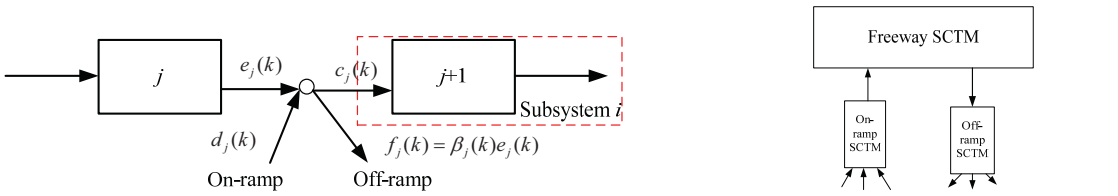
The interconnected SCTM approach calculates the flow propagation by pairing up two neighboring cells, which can be viewed as an extension of the approach used in the CTM. Consider the example depicted in Figure 6.4. First, two cells are chosen to form a basic SCTM subsystem. By the basic SCTM subsystem, random traffic state (including the traffic density and the possible wavefront in terms of probabilities of occurrence of operational modes) of the segment is calculated. Then, the last cell of the upstream subsystem and the first cell of the downstream subsystem is paired up to calculate the flow across these two subsystems.

6.3.2 On-/off- ramps, traffic merge and diverge

Assume that the freeway segment with on-/off- ramps can be represented by a link-node formulation as depicted in Figure 6.5(a)[1]. Denote the on-ramp as on_j, and the off-ramp as off_j, where the subscript j represents the node of the freeway segment (or SCTM subsystem) to which the ramps are connected. We consider the ramps as SCTM subsystems as depicted in Figure 6.5(b). The on-/off- ramp flows depend on the states of the freeway and the ramps.

To calculate the stochastic on-ramp flow, we first explain the CTM representation of the ramp under the deterministic case. The actual on-ramp flow $r_j(k)$ is determined by the state of the freeway segment to which the on-ramp merges (Kurzhanskiy, 2007;

[1] This kind of link-node CTM representation of freeway segment, introduced by Kurzhanskiy (2007) and Muralidharan and Horowitz (2009), calculates the traffic flows in traffic networks (by flow conservation of a node) in a simpler manner. The simulation software Aurora, a simulation tool developed by the Tools for Operations Planning (TOPL) in University of California Berkeley, is based on this CTM implementation. A critical overview of macroscopic node models is provided by Tampère et al. (2011), wherein some shortcomings of state-of-the-art node models are summarized. It would be interesting for us to further extend our model following the generic class of first order macroscopic node models proposed by Tampère et al. (2011).



(a) A link-node model of a freeway segment

(b) On-/off- ramps as SCTM subsystems

Figure 6.5: A link-node model of a freeway segment

Muralidharan and Horowitz, 2009), i.e.,

$$r_j(k) = \frac{\min\left(c_j(k), w_{j+1}(k)\left(\rho_{J,j+1} - \rho_{j+1}(k)\right)\right)}{c_j(k)} d_j(k). \tag{6.8}$$

To calculate $r_j(k)$, we first have to obtain the total demand for cell j+1, $c_j(k)$. Assume the off-ramp demand $f_j(k)$ is a fraction of the flow on the freeway segment, the travel demand $c_j(k)$ of node j at time k is given as (see Figure 6.5(a)):

$$c_j(k) = e_j(k) + d_j(k) - f_j(k) = \left(1 - \beta_j(k)\right) e_j(k) + d_j(k), \tag{6.9}$$

where $\beta_j(k)$ denotes the turning fraction, $e_j(k)$ is the outflow of cell j. For the stochastic case, we assume that the nominal function of $\beta_j(k)$ is known and perturbed by a Gaussian noise process with zero mean and known variance. $\beta_j(k)e_j(k)$ is taken as the sending function to the off-ramp (off-ramp demand). Whether this amount of flow can be received by the off-ramp depends on the traffic condition of the off-ramp. This "actual" off-ramp flow can be defined as joint off-ramp flow similar to (6.7). To this end, we define the four events following the flow propagation logic defined in Section 6.3.1:

1. The first cell of the off-ramp is in free-flow condition (F_j) and the off-ramp demand is less than its capacity. In this case, the off-ramp demand $f_j(k)$ will be loaded onto the off-ramp. The corresponding probability is defined as:

$$P_1^{f_j}(k) = \Pr\left(F_j(k) \cap \left(f_j(k) < Q_1^{f_j}(k)\right)\right);$$

2. The first cell of the off-ramp is in free-flow condition (F_j) and the off-ramp demand is greater than or equal to than the capacity of the off-ramp. In this case, an amount of vehicles equals to the off-ramp capacity will be loaded onto the off-ramp. The probability for this event is defined as:

$$P_2^{f_j}(k) = \Pr\left(F_j(k) \cap \left(f_j(k) \geq Q_1^{f_j}(k)\right)\right);$$

3. The first cell of the off-ramp is in congested condition (C_j) and the off-ramp demand is less than its available space. In this case, the off-ramp demand $f_j(k)$ will be loaded onto the off-ramp. This probability is:

$$P_3^{f_j}(k) = \Pr\left(C_j(k) \cap \left(f_j(k) < w_1^{f_j}(k)\left(\rho_{J,1}^{f_j}(k) - \rho_1^{f_j}(k)\right)\right)\right);$$

4. The first cell of the off-ramp is in congested condition (C_j) and the off-ramp demand is greater than or equal to the available space of the off-ramp. In this case, an amount of vehicles which equals to the available space of the off-ramp $w_1^{f_j}(k)\left(\rho_{J,1}^{f_j}(k) - \rho_1^{f_j}(k)\right)$ will be loaded onto the off-ramp, The probability for this event is defined as:

$$P_4^{f_j}(k) = \Pr\left(C_j(k) \cap \left(f_j(k) \geq w_1^{f_j}(k)\left(\rho_{J,1}^{f_j}(k) - \rho_1^{f_j}(k)\right)\right)\right).$$

By applying the independent argument similar to that in Section 6.3.1, the probabilities of these events can be defined as:

$$\begin{aligned}
P_1^{f_j}(k) &= \left(P_{FF}^{f_j}(k) + P_{FC}^{f_j}(k)\right)\Pr\left(f_j(k) < Q_1^{f_j}(k)\right), \\
P_2^{f_j}(k) &= \left(P_{FF}^{f_j}(k) + P_{FC}^{f_j}(k)\right)\Pr\left(f_j(k) \geq Q_1^{f_j}(k)\right), \\
P_3^{f_j}(k) &= \left(P_{CC}^{f_j}(k) + P_{CF}^{f_j}(k)\right)\Pr\left(f_j(k) < w_1^{f_j}(k)\left(\rho_{J,1}^{f_j}(k) - \rho_1^{f_j}(k)\right)\right), \\
P_4^{f_j}(k) &= \left(P_{CC}^{f_j}(k) + P_{CF}^{f_j}(k)\right)\Pr\left(f_j(k) \geq w_1^{f_j}(k)\left(\rho_{J,1}^{f_j}(k) - \rho_1^{f_j}(k)\right)\right),
\end{aligned}$$

with $\sum_y P_y^{f_j}(k) = 1$. We define the PDF for the joint off-ramp flow ("actual" stochastic off-ramp flow) $f_j^a(k)$ as a finite mixture of the four probabilistic events:

$$g_f\left(f_j^a(k)|\phi(k)\right) = \sum_y P_y^{f_j}(k) g_f\left(f_j^a(k)|\phi_y(k)\right), \tag{6.10}$$

where $\phi(k) = \{\phi_y(k)\}$, $\phi_y(k) = \left(P_y^{f_j}(k),\ f_y^{f_j}(k)\right)$.

Next we will address the on-ramp case (or merge operation). This on-ramp flow pattern depends on the freeway traffic state to which the on-ramp belongs. From the link-node formulation, we define the total travel demand to cell $j+1$, i.e. $c_j(k)$, as the sending function of cell j. Then the "actual" on-ramp flow can be determined by comparing this flow pattern with the receiving flow function of cell $j+1$, following the rules defined in Section 6.3.1. Similar to the off-ramp case, we define the PDF of the joint on-ramp flow (the "actual" stochastic on-ramp flow) as:

$$g_r\left(r_j(k)|\varphi(k)\right) = \sum_x P_x^{r_j}(k) g_r\left(r_j(k)|\varphi_x(k)\right), \tag{6.11}$$

where $\varphi(k) = \{\varphi_x(k)\}$, $\varphi_x(k) = \left(P_x^{r_j}(k),\ d_x^{r_j}(k)\right)$, with φ contains four events, which can be defined in line with those in Section 6.3.1 by replacing $S_{j-1}(k)$ and the receiving functions of cell i with $c_j(k)$ and the receiving functions of cell $j+1$, respectively. We omit the details here for brevity. The on-ramp flows of the first and third events are the same since the total travel demand to cell $j+1$, i.e. $c_j(k)$ can be received by the cell implies the freeway can accommodate the on-ramp demand $d_j(k)$. Since $c_j(k)$ cannot be received by cell $j+1$ for the second and the fourth events, the available space of cell $j+1$ is assigned to the on-ramp flow according to its proportion to the total demand $c_j(k)$. To be more specific, the on-ramp flows for the four events are:

$$d_1^{r_j}(k) = d_j(k),\ d_2^{r_j}(k) = \frac{d_j(k)}{c_j(k)} Q_{j+1}(k),\ d_3^{r_j}(k) = d_j(k),\ \text{and}$$
$$d_4^{r_j}(k) = \frac{d_j(k)}{c_j(k)} w_{j+1}(k)\left(\rho_{J,j+1}(k) - \rho_{j+1}(k)\right),$$

where $Q_{j+1}(k)$, $w_{j+1}(k)$, and $\rho_{J,j+1}(k)$ are capacity, backward wave speed and jam density of cell $j+1$, respectively. The probabilities for these four events can be defined in line with those in Section 6.3.1. The mean and variance of $d_2^{r_j}(k)$, $d_4^{r_j}(k)$ can be approximated by Taylor series given the means and variances of its definitional variables are known as explained in Chapter 5. The mean and variance of the joint on-ramp flow (6.11) can be calculated by utilizing (5.18)-(5.19) with the total travel demand $c_j(k)$ for cell $j+1$ given by (6.9). Daganzo (1995a) pointed out that a merge can be in one of the following three possible causality regimes:

1. Forward: if the flow on both approaches is dictated by conditions upstream (i.e. waves move forward);

2. Backward: if the flow on both approaches is dictated by conditions downstream (i.e. waves move backward);

3. Mixed: if the flow is dictated by conditions upstream for one approach and downstream for the other.

The first two cases are common and well addressed by the SCTM. Case 3 is not common, and arises when one approach has higher priority than the other approach under congested condition. This case can be handled by the current SCTM framework by assigning priority ratios to the receiving function of the downstream cell as done by Daganzo (1995a). We will also investigate this case in the numerical simulation.

The "actual" inflow to the on-ramp is saturated by the on-ramp SCTM subsystem and can be similarly evaluated as the actual on-ramp flow, i.e. (6.11). The flows which

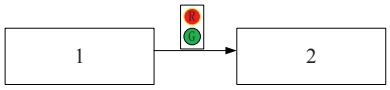


Figure 6.6: The model of a signalized cell

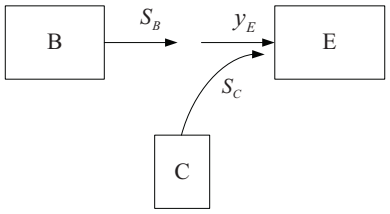


Figure 6.7: A signalized merge

divert to the arterials through the off-ramp depend on the states of arterials and can be similarly calculated using (6.10). Since the basic SCTM accepts the mean and variance of the inflow as inputs, we need to obtain them in advance to simulate the basic SCTM. If detected flow data is available, it can also be used as input to the basic SCTM.

6.4 Model of signalized junctions

In this section, we will model the the signalized intersections by the interconnected SCTM subsystem approach. We start with the SCTM model of a signalized cell depicted in Figure 6.6. When the signal is green, the sending function of cell 1 is given by

$$S_1(k) = mix\left(v_{f,1}(k)\rho_1(k), Q_1(k)\right),$$

where mix denotes the finite mixture distribution defined similar to (5.17) (see Section 6.3.1). Whether this amount of flow can be received by cell 2 can be determined by (6.7). When the signal is red, cell 1 should send out nothing. The two cells update their traffic states independently. As pointed out by Lo (1999b) and Lo et al. (2001), the congested situation in Hong Kong renders virtually all turnings to have protected, rather than permitted, signals, see e.g. Figure 6.7. Due to the signalization, the flows from cell C and cell B do not flow into cell E at the same time. This simplification yields:

1. either cell C or cell B is flowing;

2. neither cell C nor cell B is flowing.

A common signalized junction in Hong Kong is depicted in Figure 6.8. To represent the signalized junction by the link-node model, we assume that the junction area is covered by a node. Figure 6.9 depicts a link-node representation of the junction in green and

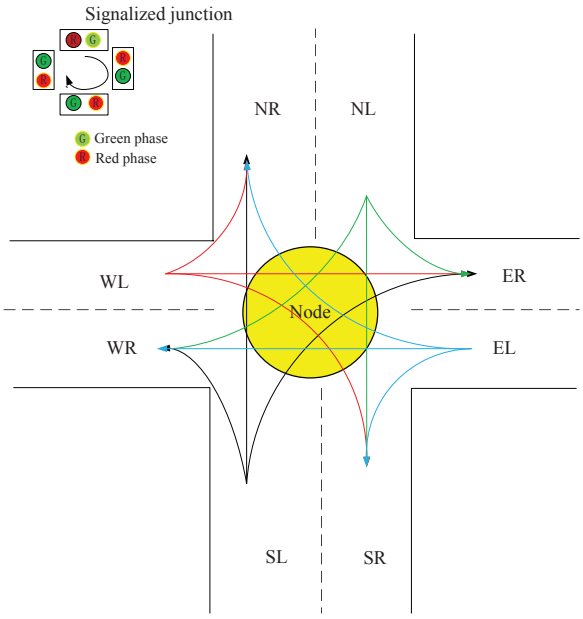


Figure 6.8: A typical signalized junction

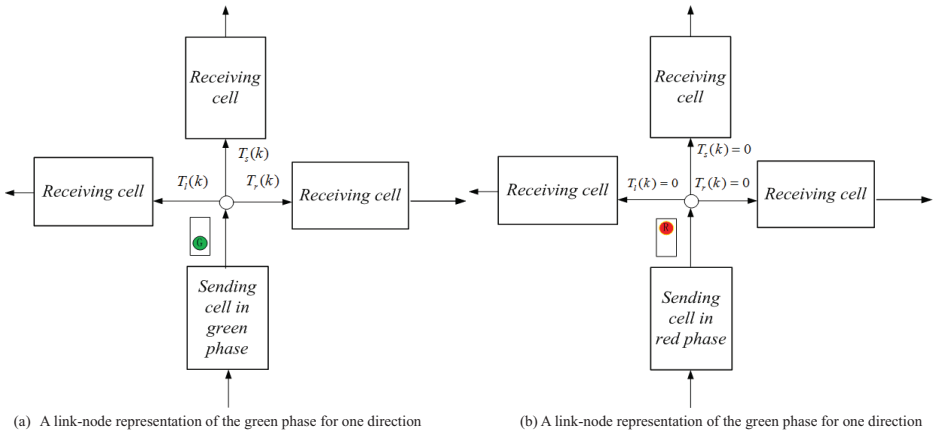


Figure 6.9: A link-node representation of the junction in green and red phases for one direction

red phases for one direction. When the signal is red for this direction, no flow should be sent out from the sending cell. In this case, the sending cell has no connection with the downstream receiving cells. When the signal is green, the sending cell is connected to a diverging node. The diverge logic is applied to this case, i.e. the sending cell sends out the flows, whether these flows can be received by the downstream cells depends on the traffic conditions of the receiving cells. By the above observation, we model the turning movements by "virtual" merge and diverge operations (or on-/off- ramps). By the word of "virtual", we mean that the length of the ramp is zero. To simplify and proceed to the analysis, we make the following assumptions:

Assumption 6.1. The queue will build up at the cell sending out the turning vehicles if the receiving cell does not have available space, i.e. no vehicles will queue up at the node (or the "virtual" ramps).

Assumption 6.2. Assume that the nominal value functions of left-/right- turning ratios, denoted as $\gamma_L(k)$, $\gamma_R(k)$, are known, and the noise terms of $\gamma_L(k), \gamma_R(k)$ are known Gaussian white noises. The nominal value functions $\gamma_L(k), \gamma_R(k)$ will be 0 if the signal is in the red phase.

Assumption 6.3. Assume that the signal phases are known and there is no amber time between red and green phases.

With these assumptions, we are ready to specify the dynamics of this signalized junction. Firstly, the signalized junction is divided into four SCTM subsystems as depicted in Figure 6.10. Due to the protected signal assumption and Assumption 6.1, the four SCTM subsystems do not overlap each other. Each subsystem has three phases corresponding to the signal phases as demonstrated in Figure 6.11(a). When subsystem 1 is in green phase, it can be modeled as a normal SCTM subsystem with two off-ramps which correspond to the two turning movements. This phase is named as PH1. When subsystem 2 is in green phase, the right turning movement of subsystem 2 can be modeled as an on-ramp of subsystem 1. Since subsystem 1 is in red phase under this condition, there is no straight movement for subsystem 1, i.e. there is no connection between the two cells of subsystem 1. Similar analysis can be applied when subsystem 3 is in green phase. Since in the SCTM, we do not distinguish the detailed locations of the ramps inside one cell, these two cases are grouped into one phase, i.e. PH2. When subsystem 4 is in green phase, neither turning movement nor straight movement is valid for subsystem 1. The two cells of this subsystem update their states independently, which is denoted as PH3. Notice that PH3 is in fact a special case of PH2 when the on-ramp flow equals to 0, these two phases are further grouped into one.

Detailed interconnection between two SCTM subsystems is depicted in Figure 6.11(b). The upstream cell of subsystem 1 sends out the vehicles that want to turn right via a "virtual" off-ramp (or diverge), while the downstream cell of subsystem 2 receives the flow via a "virtual" on-ramp (or merge). Dynamics of PH1 can be described by the basic SCTM subsystem in which a virtual off-ramp flow is defined as the summation of the two

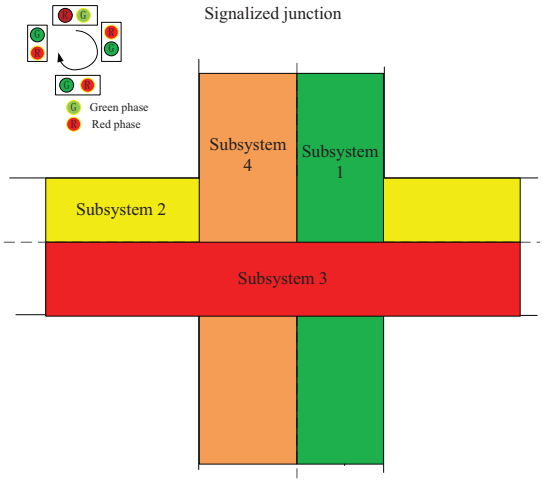
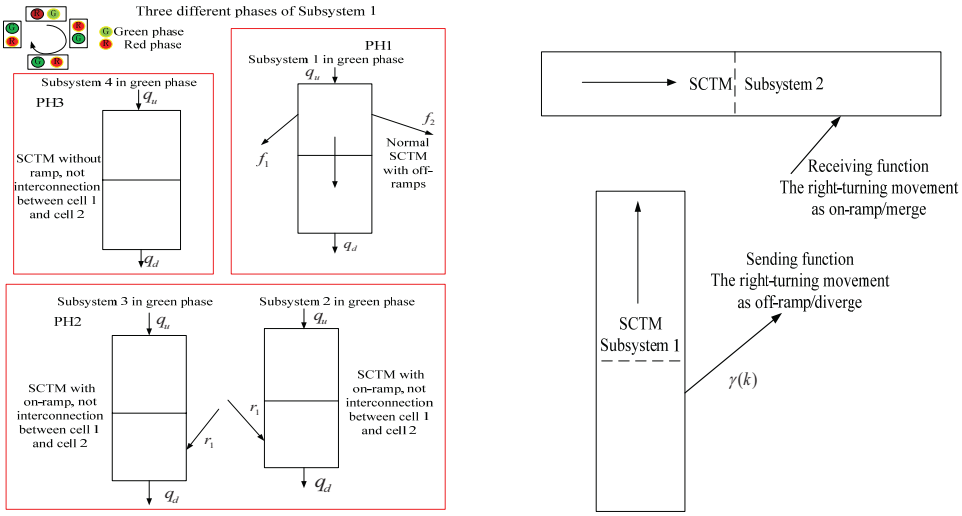


Figure 6.10: A typical signalized junction represented by four SCTM subsystems



(a) Three different phases of Subsystem 1

(b) Two signalized movements

Figure 6.11: Three phases of Subsystem 1 and two signalized movements

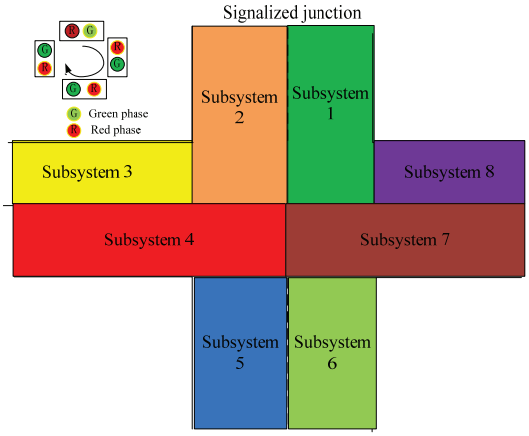


Figure 6.12: The SCTM subsystems representation under permitted signal

off-ramp flows. The dynamic equations for PH1 of subsystem 1 are:FF mode:

$$\begin{aligned}\rho_1(k+1) &= \rho_1(k)+\frac{T}{l_1}\left(q_u(k)-v_{f,1}(k)\,\rho_1(k)\right),\\ \rho_2(k+1) &= \rho_2(k)+\frac{T}{l_2}\left(v_{f,1}(k)\,\rho_1(k)-\left(f_1(k)+f_2(k)\right)-v_{f,2}(k)\,\rho_2(k)\right);\end{aligned}$$

CC mode:

$$\begin{aligned}\rho_1(k+1) &= \rho_1(k)+\frac{T}{l_1}\left(w_{c,1}(k)\left(\rho_{J,1}(k)-\rho_1(k)\right)-\left(f_1(k)+f_2(k)\right)\right.\\ &\quad \left.-w_{c,2}(k)\left(\rho_{J,2}(k)-\rho_2(k)\right)\right),\\ \rho_2(k+1) &= \rho_2(k)+\frac{T}{l_2}\left(w_{c,2}(k)\left(\rho_{J,2}(k)-\rho_2(k)\right)-q_d(k)\right);\end{aligned}$$

CF mode:

$$\begin{aligned}\rho_1(k+1) &= \rho_1(k)+\frac{T}{l_1}\left(w_{c,1}(k)\left(\rho_{J,1}(k)-\rho_1(k)\right)-Q_2(k)\right),\\ \rho_2(k+1) &= \rho_2(k)+\frac{T}{l_2}\left(Q_2(k)-\left(f_1(k)+f_2(k)\right)-v_{f,2}(k)\,\rho_2(k)\right);\end{aligned}$$

FC1 mode:

$$\begin{aligned}\rho_1(k+1) &= \rho_1(k)+\frac{T}{l_1}\left(q_u(k)-v_{f,1}(k)\,\rho_1(k)\right)\\ \rho_2(k+1) &= \rho_2(k)+\frac{T}{l_2}\left(v_{f,1}(k)\,\rho_1(k)-\left(f_1(k)+f_2(k)\right)-q_d(k)\right);\end{aligned}$$

FC2 mode:

$$\begin{aligned}\rho_1(k+1) &= \rho_1(k)+\frac{T}{l_1}\left(q_u(k)-w_{c,2}(k)\left(\rho_{J,2}(k)-\rho_2(k)\right)-\left(f_1(k)+f_2(k)\right)\right),\\ \rho_2(k+1) &= \rho_2(k)+\frac{T}{l_2}\left(w_{c,2}(k)\left(\rho_{J,2}(k)-\rho_2(k)\right)-q_d(k)\right);\end{aligned}$$

where the triangular fundamental diagram is presumed, $f_1(k)$ and $f_2(k)$ are the off-ramp flows of the segment, respectively. One can easily convert the above equations into bilinear system representation in the form of (5.20).

Note that there is no interconnection between the two cells in PH2, the probability for each mode should be revised accordingly. To be more specific, the definitions of the steady states are retained, whereas the transient modes need to be revised. The case that the upstream cell is in free-flow condition and the downstream cell is in congested condition is taken as FC mode since there is no connection between the two cells, which implies no wave-front would exist at the boundary of the two cells in this phase. The same logic is applicable to CF mode. The state space equations for PH2 are:

FF mode:

$$\begin{pmatrix}\rho_1(k+1)\\ \rho_2(k+1)\end{pmatrix}=\begin{pmatrix}\rho_1(k)\\ \rho_2(k)\end{pmatrix}+\begin{pmatrix}\frac{T}{l_1}q_u(k)\\ \frac{T}{l_2}\left[r_1(k)-v_{f,2}(k)\,\rho_2(k)\right]\end{pmatrix};$$

CC mode:

$$\begin{pmatrix} \rho_1(k+1) \\ \rho_2(k+1) \end{pmatrix} = \begin{pmatrix} \rho_1(k) \\ \rho_2(k) \end{pmatrix} + \begin{pmatrix} \frac{T}{l_1}\left[w_{c,1}(k)\left(\rho_{J,1}(k) - \rho_1(k)\right)\right] \\ \frac{T}{l_2}\left[r_1(k) - q_d(k)\right] \end{pmatrix};$$

CF mode:

$$\begin{pmatrix} \rho_1(k+1) \\ \rho_2(k+1) \end{pmatrix} = \begin{pmatrix} \rho_1(k) \\ \rho_2(k) \end{pmatrix} + \begin{pmatrix} \frac{T}{l_1}\left[w_{c,1}(k)\left(\rho_{J,1}(k) - \rho_1(k)\right)\right] \\ \frac{T}{l_2}\left[r_1(k) - v_{f,2}(k)\,\rho_2(k)\right] \end{pmatrix};$$

FC mode:

$$\begin{pmatrix} \rho_1(k+1) \\ \rho_2(k+1) \end{pmatrix} = \begin{pmatrix} \rho_1(k) \\ \rho_2(k) \end{pmatrix} + \begin{pmatrix} \frac{T}{l_1} q_u(k) \\ \frac{T}{l_2}\left[r_1(k) - q_d(k)\right] \end{pmatrix};$$

where $r_1(k)$ is the on-ramp flow of the segment. The dynamics of PH3 is a special case of PH2 by specifying $r_1(k) = 0$, we omit it here for brevity.

Figure 6.10 is not the unique representation of the signalized junction depicted in Figure 6.8. Generally speaking, a signalized junction can be represented by different means of connecting the SCTM subsystems. In the case that the protected signal assumption does not hold, we need more SCTM subsystems to represent the signalized junction depicted in Figure 6.8 based on the link-node formulation. A feasible representation is illustrated in Figure 6.12. Analysis of the dynamics and phases of the SCTM subsystems can be conducted similar to those of protected signal case.

6.5 Numerical examples

6.5.1 An application of the network SCTM as a stochastic dynamic traffic network model

A simple hypothetical network with 8 links, one OD pair, 6 paths is depicted in Figure 6.13. All links are one mile long and admit the same nominal fundamental flow-density diagram as depicted in Figure 6.14. There is a signalized junction at the center of the network. The turning movements and the corresponding turning ratios are shown in the figure. The turning ratios are assumed to be constant. We choose 5 sec as the simulation time increment and divide each link into two cells such that each link can be modeled by one basic SCTM subsystem. We apply priority turning ratios for link M if it is operating under the critical condition, i.e. the traffic density of the link is close to its critical density. Under this situation, 1/3 and 2/3 of the available space of the first cell of link M will be assigned to the flows from links Y and T, respectively. The means of the parameter

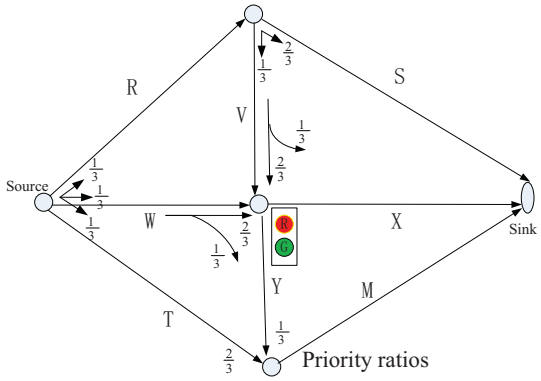


Figure 6.13: Specification of the test network

uncertainties of the fundamental flow-density diagrams are assumed to be 0, and the SDs of the parameter uncertainties are assumed to be 5% of their nominal values. The signal has a cycle time of 2 min, which is equally assigned to the green and the red phases. Zero initial condition is presumed. There is one inflow profile $q_u(k)$ enters the network from the source. The nominal function of the inflow is given by

$$q_u(k) = \begin{cases} 6000\ veh/h, & 0 \leq t \leq 20\,\text{min}; \\ 21000\ veh/h, & 20 < t \leq 40\,\text{min}; \\ 0 \qquad veh/h, & 40 < t \leq 60\,\text{min}\,. \end{cases}$$

We assume that the SD of the inflow profile is 5% of its nominal value. The inflow is chosen such that all the states ranging from free-flow to congested can be activated. The network operates under the free-flow condition in the first stage of the inflow, i.e. from 0 to 20 min. Congestion onset for some links will be observed when the inflow switches to the second stage. Then the traffic will tend to a steady-state of congestion for some links. Some time after the inflow switches to 0, congestion dissolve will be observed for the congested links. Finally, all the traffic will be clear from the network. By applying the proposed SCTM to the network with uncertainties, we obtain the following results as depicted in Figures 6.15 to 6.22.

The traffic densities of links R and S are shown in Figure 6.15. Since links R and S do not involve the signal (strictly, not affected by the signalization), the obtained results are similar to a freeway corridor test. At the first stage of the inflow profile, i.e. from 0 to 20 min, link R is operating under the free-flow condition as depicted in Figures 6.15 and 6.16 as the inflow rate to link R has a mean value of 2000 *veh/h,* which is far from the saturated flow of the link. The variance of the traffic density in this stage is small. We are quite certain about the estimation when the variance of the inflow and the traffic density is low (the supply uncertainty under free-flow condition is small). The inflow to

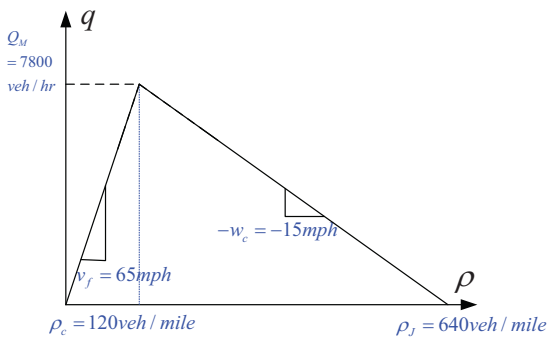


Figure 6.14: Nominal fundamental diagram

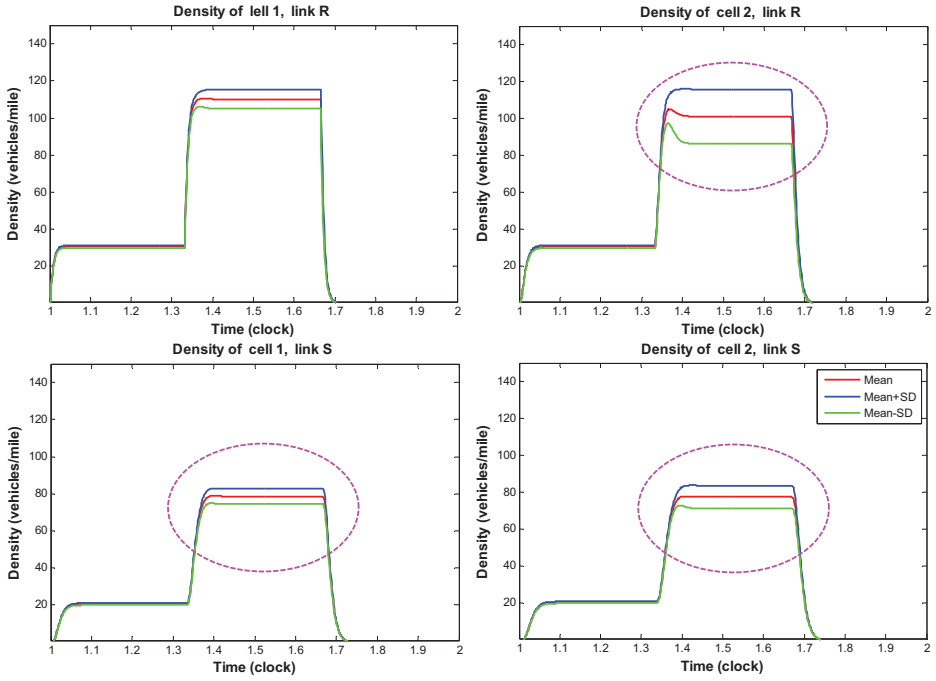


Figure 6.15: Traffic densities of links R and S, and the 68% confidence interval

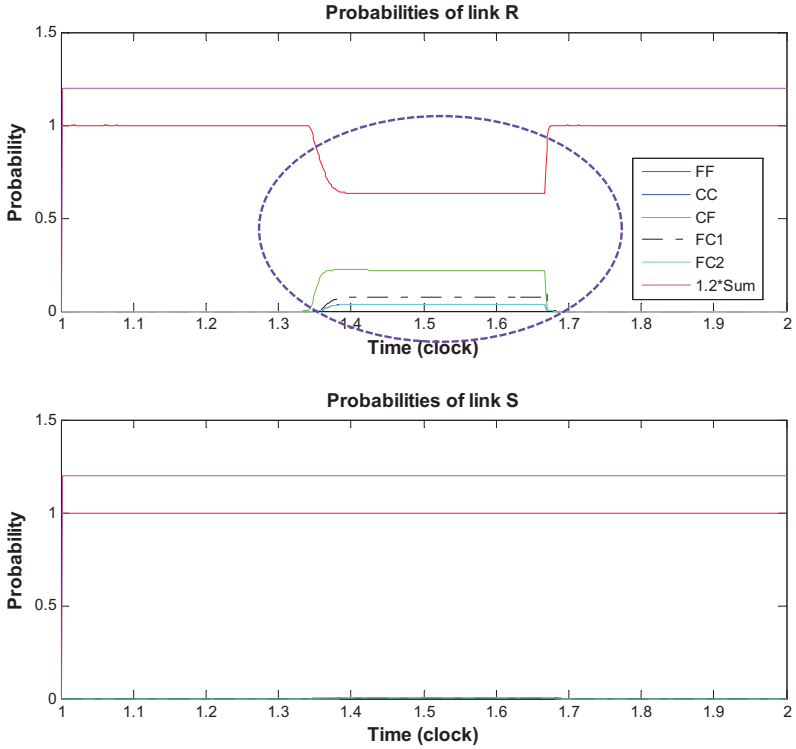


Figure 6.16: Probabilities of occurrence of the five modes of links R and S

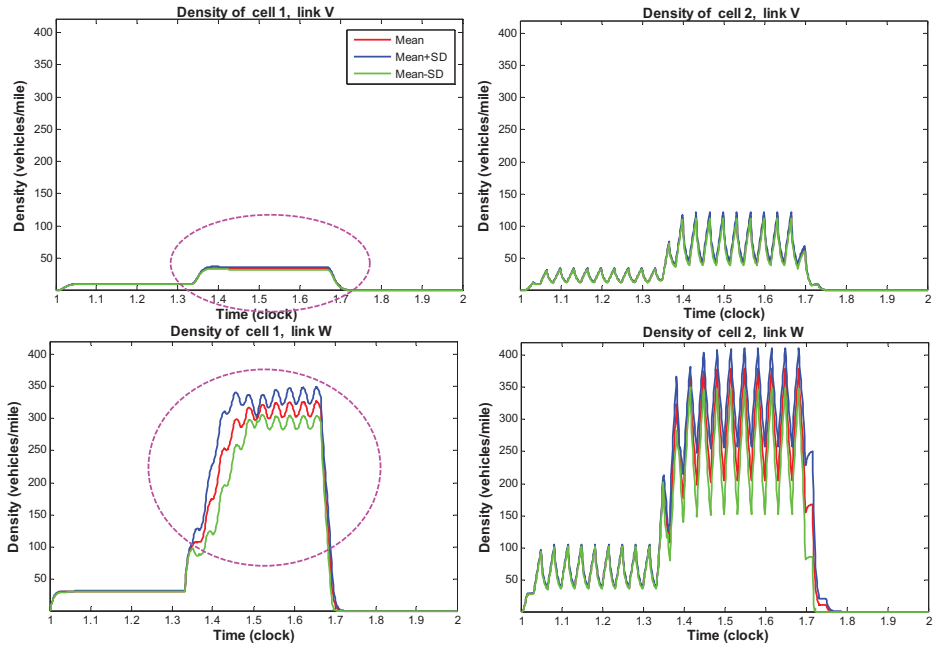


Figure 6.17: Traffic densities of links V and W, and the 68% confidence interval

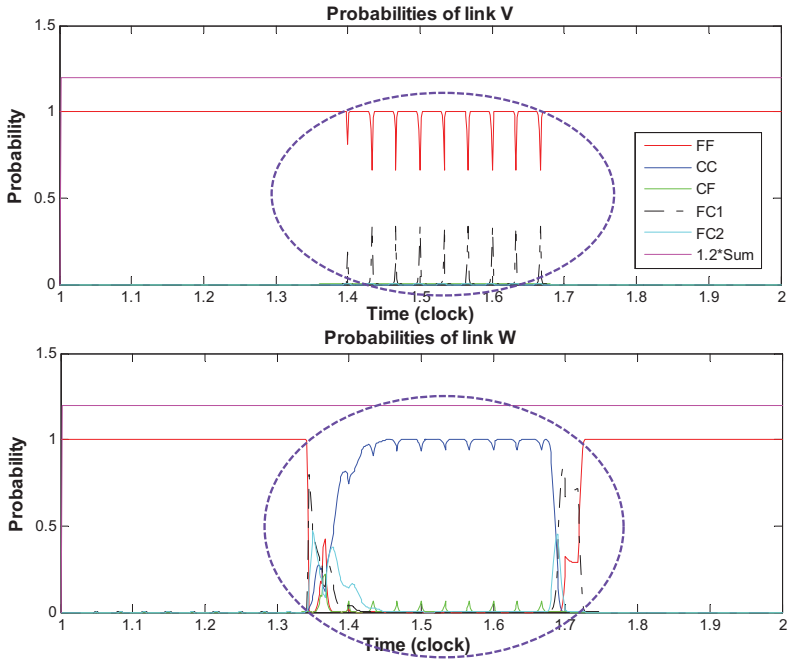


Figure 6.18: Probabilities of occurrence of the five modes of links V and W

link R at the second stage of the network inflow has a mean value of 7100 veh/hr, which is close to its link capacity. Therefore at this stage, link R is operating close to the critical condition, which is verified by the part circled in Figure 6.16. Compared with the first stage, the variance increases because the supply uncertainties under the critical condition and congested condition are much larger than that under free-flow condition, and the variance of the inflow also increases in this stage. Since there is no congestion spillback, the variance is propagating downstream as shown by Figure 6.15. The gradual increase in SD and the high SD in cell 2 of link R during this period is due to the accumulation effect of the uncertainties. The downstream cells will therefore experience a higher level of uncertainties compared to the upstream cells. Note that the variance of link S circled is much smaller than that of link R. This is because: First, there is a turning movement, link S receives only a part of the outflow from link R, therefore the variance of the inflow to link S reduces. Second, at the second stage, link R is operating close to the critical condition while link S is almost operating under free-flow condition. As we know, traffic states under critical and/or congested condition are much more uncertain than those under free-flow condition. By definition, the summation of the probabilities of occurrence of the five modes is equal to 1. In the figures involving the probabilities, we plot 1.2 times of the summation to distinguish the summation from the probability of an individual mode that equals to 1.

Next, we will go through the results for the upstream links of the signal, i.e. links W

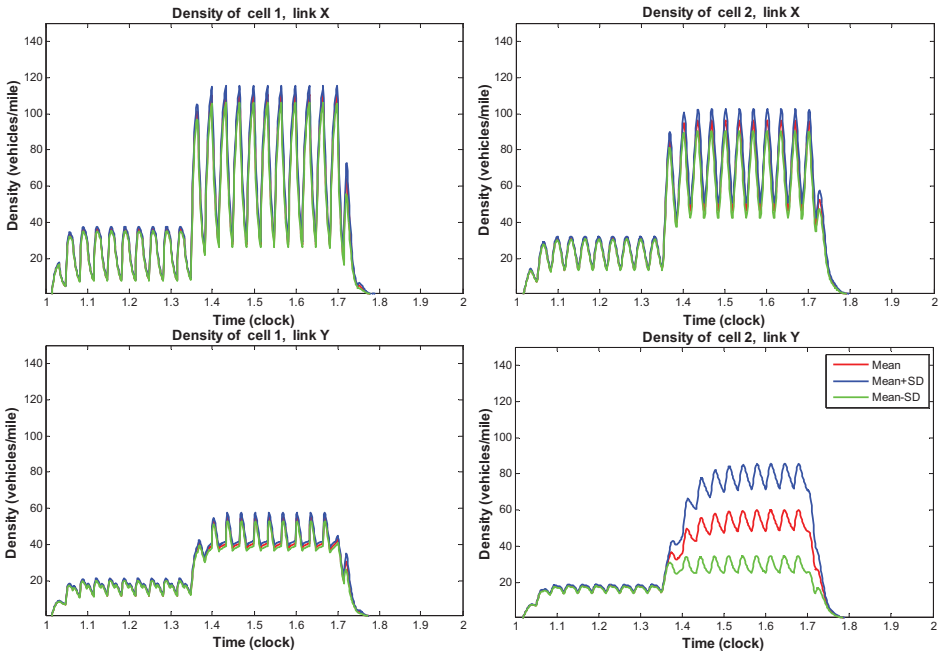


Figure 6.19: Traffic densities of links X and Y, and the 68% confidence interval

and V. Due to the light traffic condition, link V does not have congestion. The first cell of the link will not be affected by the signal as demonstrated in Figure 6.17. On the other hand, due to the heavy traffic condition and the signal, congestion is first formed on the second cell of link W at the second stage and then spills back to its first cell as illustrated in Figure 6.17. Different from the variance propagation of link R, the variability of the traffic states does not seem to increase as the flow moves downstream. On the other hand, we can observe the propagation of the uncertainty in the reverse direction of the traffic flow (or propagates backward). The cell densities of link W are shaped by the signal cycle. The corresponding probability diagram is shown in Figure 6.18.

The results for the downstream links of the signal are depicted in Figures 6.19-6.20. Figure 6.19 illustrates that these two links are operating under free-flow condition most of the time. The variance of cell 2 of link Y at the second stage is larger as shown in Figure 6.19. Besides the variance propagation, it is due to also the fact that the priority turning ratios are activated because the saturated traffic condition on link M. The variance is largely affected by the traffic condition (to be more specific, the variance of uncertain available space) of link M, which causes large variance to the traffic density of the second cell of link Y. Figure 6.20 depicts the corresponding probabilities of the five operational modes.

Finally, we will go through the results for links T and M. Since at the first stage the links are in free-flow conditions, we concentrate our analysis on the second stage of the

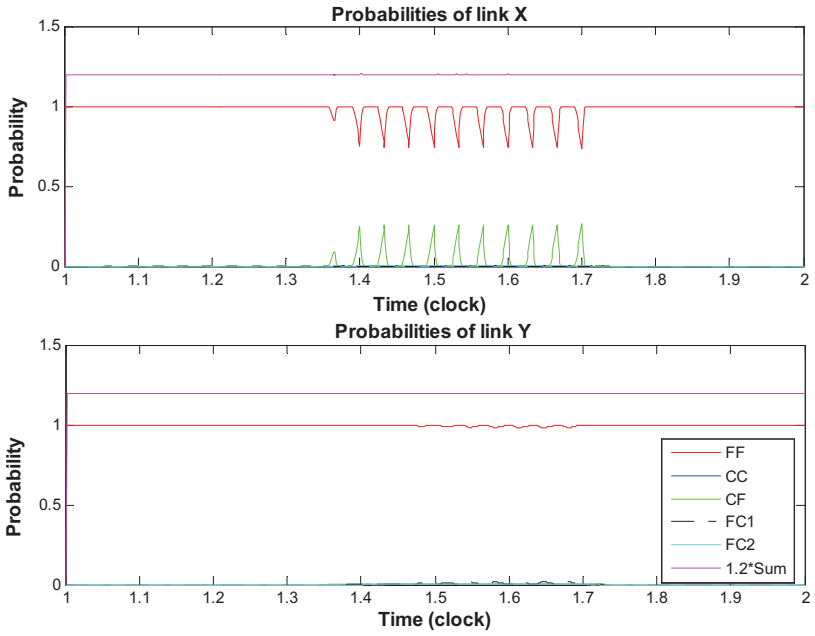


Figure 6.20: Probabilities of occurrence of the five modes of links X and Y

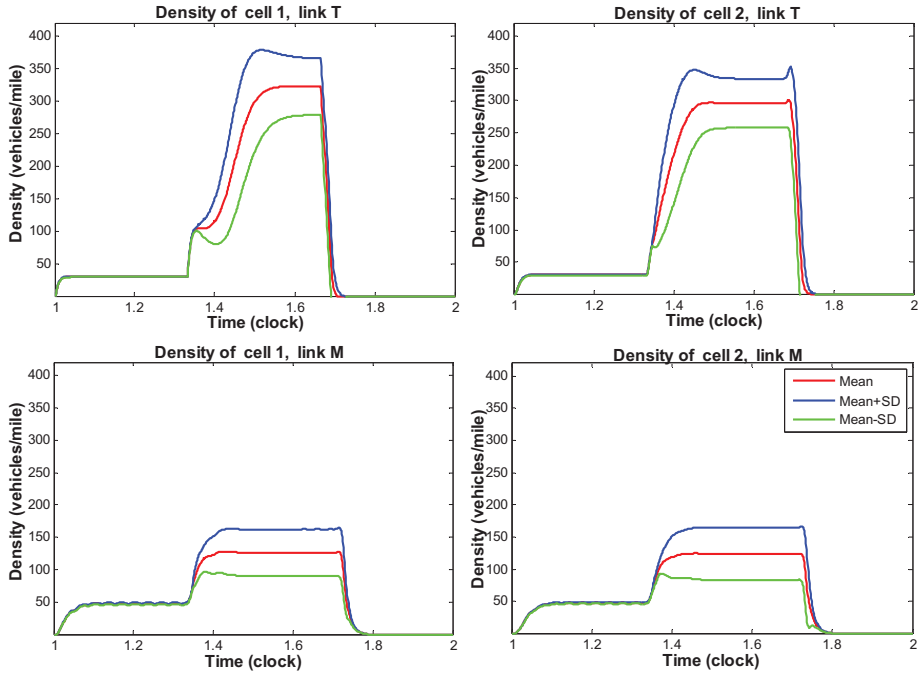


Figure 6.21: Traffic densities of links T and M, and the 68% confidence interval

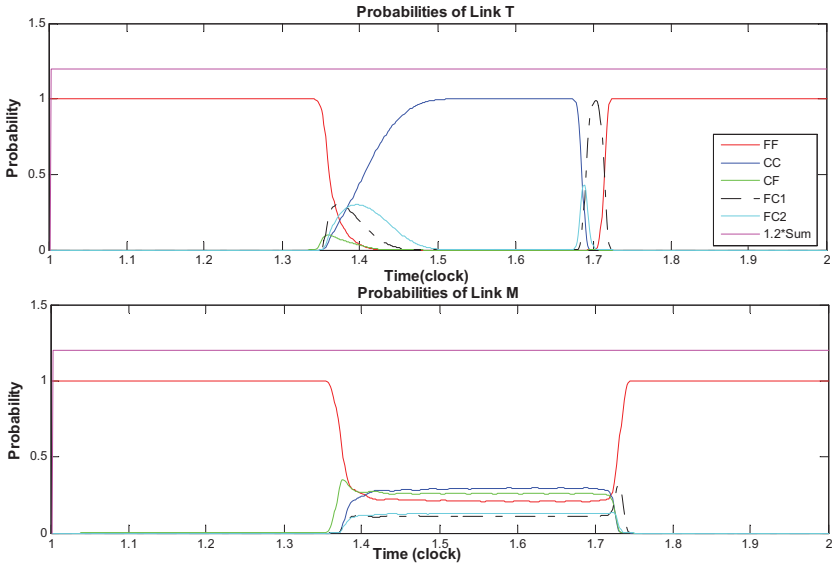


Figure 6.22: Probabilities of occurrence of the five modes of links T and M

inflow. Unlike links R and S, these two links involve heavy traffic conditions during this time period as depicted in Figure 6.22. However, as link M is connected to the sink, no congestion will be formed on this link. The variance is propagating downstream as demonstrated in Figure 6.21. The critical traffic condition on link M activates the priority turning ratios, which in conjunction with the heavy traffic condition on link T creates congestion on link T. The congestion then spillback to the first cell of link T. Figure 6.21 shows both the mean and variance of link T at the second stage is propagating backward.

6.5.2 Queues and delays at a signalized junction

6.5.2.1 Background

In this section, we will compare the delays and queues obtained by the SCTM with those obtained by the following formulas. Webster (1958) proposed the following approximate expression for delay estimation at a signalized junction:

$$E(W) = \frac{\tau_c \left(1 - \frac{\tau_g}{\tau_c}\right)^2}{2\left(1 - \frac{\tau_g}{\tau_c} x\right)} + \frac{x^2}{2q_a(1-x)} - 0.65\left(\frac{\tau_c}{q_a^2}\right)^{\frac{1}{3}} x^{2+5\frac{\tau_g}{\tau_c}}, \tag{6.12}$$

where τ_g and τ_r are the effective green time and red time, respectively. τ_c is the cycle time. q_a is the average arrival rate (veh/s). $x = \frac{q_a}{Q_c}$ is the degree of saturation. Q_c is average signal capacity which is defined by $Q_c = Q\frac{\tau_g}{\tau_c}$, where Q is average service rate (or saturation flow rate (veh/s). The first term represents the analytical expression of the uniform delay, while the second is a characterization of the random delay, which is derived by assuming Poisson arrivals and deterministic service rate. The last term is introduced

to reduce the discrepancy with results observed from simulations and is assumed to be about 10% of the sum of the first two terms (Viti and Van Zuylen, 2010).

McNeill (1968) proposed a formula for the expected signal delay for a general arrival process, under the assumption of constant departure time. The average delay $E(W)$ is calculated as:

$$E\left(W\right)=\frac{\tau_r}{2\tau_c\left(1-\frac{q_a}{Q}\right)}\left(\tau_r+\frac{2}{q_a}E\left(Q_0\right)+\frac{1}{Q}\left(1+\frac{I}{1-\frac{q_a}{Q}}\right)\right), \tag{6.13}$$

where $E(Q_0)$ is the expected overflow queue from previous cycles, $I=\frac{\sigma(q_a)}{q_a}\tau_c$ indicates the index of dispersion for the arrivals. For a binomial arrival process, $I=1-\frac{q_a}{Q}$. Regarding the computation of delays at signalized junctions, Akcelik proposed another formula:

$$E\left(W\right)=\begin{cases}\frac{\tau_c\left(1-\frac{\tau_g}{\tau_c}\right)^2}{2\left(1-\frac{q_a}{Q}\right)}+\frac{E(Q_0)}{Q_c}, & x<1;\\ \frac{\tau_c-\tau_g}{2}+\frac{E(Q_0)}{Q_c}, & x\geq 1;\end{cases} \tag{6.14}$$

To calculate the expected value of the overflow queue, Miller's (1968) formula represents one of the most popular expressions:

$$E\left(Q_0\right)=\frac{\exp\left(-1.33\sqrt{Q\,\tau_g\,\frac{1-x}{x}}\right)}{2\left(1-x\right)} \tag{6.15}$$

Akcelik (1980) further simplified this expression with the following expression:

$$E\left(Q_0\right)=\frac{1.5\left(x-x_0\right)}{1-x}, \tag{6.16}$$

where $x_0=0.67+\frac{Q\tau_g}{600}$ represents the value above which overflow queues become non-negligible. This expression is valid for $x_0<x<1$, while $E[Q_0]=0$ for $x\leq x_0$. Both Miller's and Akcelik's formulas assume the expectation value for the overflow to be in equilibrium state. However, the equilibrium state may take a long time to be reached when the demand is close to the signal capacity. This time can be so long that the arrival distribution is unlikely to remain stationary. For this reason, static models are often not consistent with real observations and their application is restricted to planning and design of uncongested traffic systems. Akcelik (1980) formulated an expression that is widely utilized for the estimation of expectation value of the temporal overflow queue. He evaluated the temporal queue evolution by using the coordinate transformation technique:

$$E\left(Q_0\right)=\begin{cases}\frac{Q_cT}{4}\left(x-1+\sqrt{\left(x-1\right)^2+\frac{12(x-x_0)}{Q_cT}}\right), & x>x_0;\\ 0, & x\leq x_0;\end{cases} \tag{6.17}$$

This model gives time dependency to the expectation of the overflow queue in a fixed time period T in both under-saturated and over-saturated conditions.

Figure 6.23: Delay and queue evaluation scenario

6.5.2.2 Test scenario

To evaluate the consistency of delay and queue estimations from the various models presented in this chapter, we carry out the delay and queue evaluations in this section for a roadway segment consisting of one lane as depicted in Figure 6.23. There is a fixed-time traffic signal with a 120 seconds cycle length and a 30 seconds effective green interval on the segment. The roadway segment admits a nominal fundamental diagram as illustrated in Figure 6.24. We assume that the SDs of the parameters are 10% of their nominal values. For each model, delay and queue evaluations are carried out for different degrees of saturation ratios x varying from 0.2 to 1.5. This allows evaluations to be carried out for a range of traffic conditions extending from highly under-saturated to highly over-saturated conditions. For each scenario, vehicle arrivals are further assumed to follow a random process with a constant average arrival rate. As the SCTM provides the time-dependent queues and delays, to compare with the results obtained by the above formulas, we will evaluate the expected queue lengths and delays for 15 minutes time interval for all degree of saturation ratios x considered, which has been frequently adopted by the time-dependent models (Viti and Van Zuylen, 2009, 2010).

As we have applied the CTM based model, the condition $v_f T_s \leq l_i$ is required to ensure the issue of numerical stability. This condition cannot always be satisfied in our formulation, since the free-flow speed v_f can be anything along its distribution. The probabilistic version of the above condition is roughly defined as $\Pr(v_f T_s \leq l_i) \geq \chi$, where χ is a positive real number which satisfies $1-\epsilon < \chi < 1$ for a small real number $\epsilon > 0$. For physical consideration, the simulation time increment T_s and the cell length cannot be too small. For example, if the average length of vehicles is assumed to be 8.33 meters long, then the cell length has to be longer than 8.33 meters, and with the settings of this numerical example, T_s cannot be smaller than 0.5 seconds. In this simulation, the simulation time increment is 1 second, the cell length is 60 meters. We extend the stochastic version of inflow to outflow mapping method (with FIFO queuing principle) to calculate the travel time (Sumalee et al., 2010b). The detailed methodology will be explained in the Section 6.5.2.4.

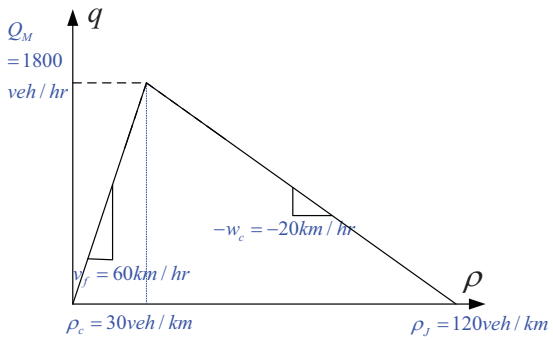


Figure 6.24: Nominal fundamental diagram of the segment

6.5.2.3 Simulation results

It has been reported in Dion et al. (2004) and illustrated in Figure 6.25 that there is a general consistency between the analytical delay models when they are applied to the analysis of signalized intersections with low saturation ratios. In addition, the agreement tends to decrease with increasing saturation ratios. The delays obtained from these formulas are compared against those of the SCTM approach and Monte Carlo Simulation (MCS) of MCTM approach with 1000 samples in Figure 6.25. As suggested by Figure 6.25, the delays obtained from the SCTM approach are close to those obtained from the MCS of the MCTM approach. The delays obtained from the SCTM are close to those obtained from the formulas considered when the degree of saturation is below 0.8. Note also that the SCTM and the MCS of the MCTM produce the lowest estimations when the degree of saturation is close to one. Similar observation can also be found in Dion et al. (2004) for the deterministic queuing and shockwave models. There is also a great difference between the probabilistic model proposed by Viti and Van Zuylen (2010) and the traditional delay formulas when the degree of saturation is close to one. A common reason for all these inconsistent results may be due to the fact that the concept of a time-dependent delay model is based on the coordinate transformation technique which transforms the equation defining a steady-state stochastic delay model such that it becomes asymptotic (tangent) to the deterministic over-saturation model as illustrated in Figure 6.26 (Dion et al., 2004). Errors and over-estimation would be introduced due to this continuous approximation. Two more reasons may be due to the CTM/SCTM: First error source, which may not be significant, would be the choice of cell length of the CTM/SCTM, which affects the transient response of the traffic dynamics. Second, the summary statistics such as the mean and standard deviation can be deceptive when applied to arbitrary signal delay distribution. We will discuss this issue in detail in the next section.

The consistency between the SCTM and these formulas appears again when the degree of saturation is above 1.2. However, compared with the delay formulas listed in Section 6.5.2.1, the SCTM approach has wider application opportunities, since it can describe the temporal evolution of the queue length distribution at signalized junctions with uncertain supply functions while assuming any arrival distribution that can be non-stationary. The applicability of the SCTM is not limited by a fixed evaluation period as we have shown in the previous simulation and will be further discussed later on.

The traffic shockwave based model estimates queue lengths at signalized junctions by tracing the trajectories of shockwaves. This kind of model successfully describes the complex queuing process in both temporal and spatial dimensions. Nevertheless, such kind of model has limitations in the practical applications since "perfect" inflow information is required as input to the model. However, vehicles arrive at the signalized junction in a stochastic dynamic manner, the above perfect information assumption cannot be satisfied for most situations. The model proposed here can be viewed as an extension of the existing shockwave models to estimate intersection queue lengths with stochastic dynamic arrivals. Figure 6.27 depicts the simulation result for the estimation of average queue length. Comparison between the result obtained by the time-dependent queue formula of Akcelik (1980) and that obtained by the SCTM is demonstrated in Figure 6.27. The result shows also good consistency between these two methods. The Akcelik's queue model gives zero queue length when $x \leq 0.67$. However, the SCTM gives a nonzero but small queue length when the degree of saturation increases to a certain level. This is because some vehicles have to wait if they happen to arrive at the junction just before or during red phase. When the degree of saturation is small enough, the average queue length is also small (close to zero) but increases with respect to the increasing inflow rate.

6.5.2.4 A discussion on different kinds of probabilistic distributions of the signal delay and the case of non-stationary inflow

In this chapter, we extend the sampling approach proposed by Sumalee et al. (2010b) to estimate the probabilistic distribution of traffic delay (or travel time) at the signalized junction. Their approach is a stochastic extension of the deterministic cumulative inflow to outflow matching method under the first in first out (FIFO) principle. The probabilistic travel time of a link is defined by the likelihood between the stochastic cumulative inflow and outflow according to the following definition:

Definition 6.1. For a vehicle enters link a at time index k ($ET = k$), the probability of k' to be the time that the vehicle exit from the link is defined as $P'_{k'|k} =$

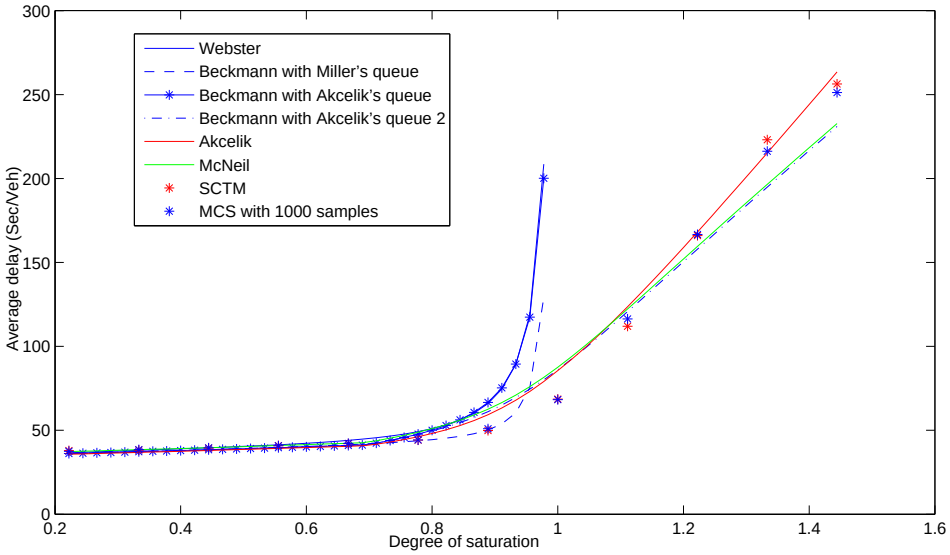


Figure 6.25: Comparison of delay estimations for different degrees of saturation

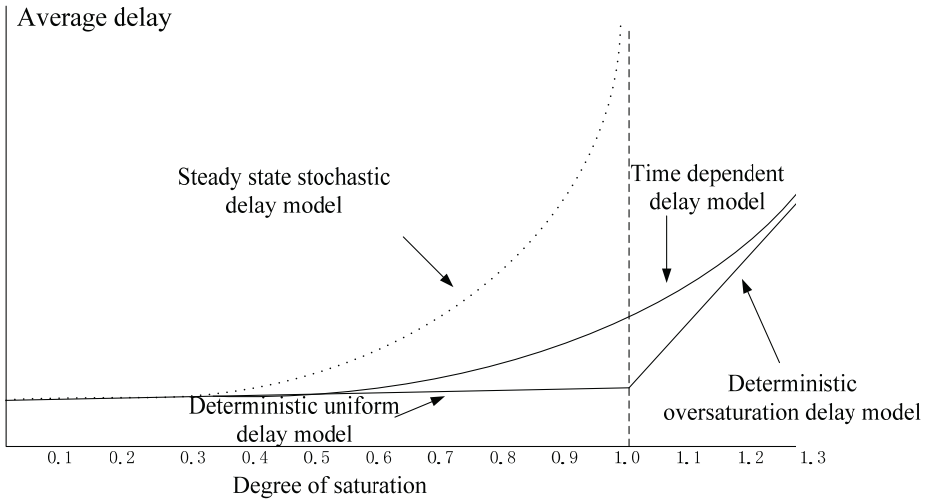


Figure 6.26: Concept of the stochastic time-dependent delay model

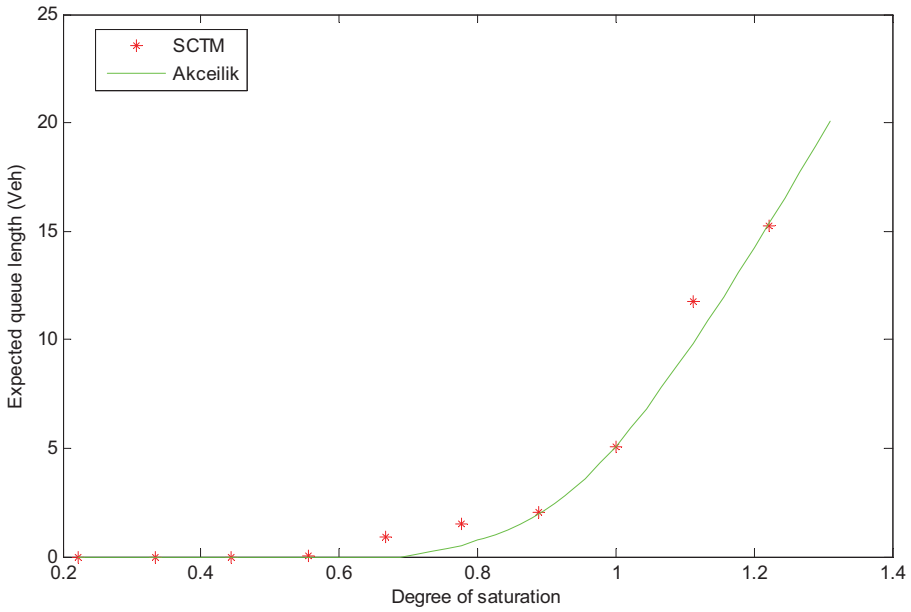


Figure 6.27: Comparison of the average queue lengths obtained by the SCTM with those of Akcelik's formula

$\Pr\left(-\varepsilon \leq C^a_{out}(k') - C^a_{in}(k) \leq \varepsilon | ET = k\right)$, with a prescribed small positive number $\varepsilon \in R^+$.

∎

The matching error $e_k(k') = C^a_{out}(k') - C^a_{in}(k)$ represents the difference between the stochastic cumulative link inflow $C^a_{in}(k)$ and outflow $C^a_{out}(k')$. The physical meaning of the $P'_{k'|k}$ is the probability that the absolute value of the mapping error to be no greater than ε vehicles. Detailed choice of mapping interval $k' \in [k_{lb}, k_{ub}]$ can be found in Sumalee et al. (2010b), we omit it here for brevity. Note that the summation of the probabilities $\sum_{k'} P'_{k'|k}$ may not equal to one, which introduces the following definition of relative frequency:

Definition 6.2. For a vehicle entering link a at time k, the relative frequency$P_{k'|k}$ is defined as:

$$P_{k'|k} = \frac{P'_{k'|k}}{\sum_{k_{lb}}^{k_{ub}} P'_{k'|k}}, \ \forall k' \in [k_{lb}, k_{ub}]. \tag{6.18}$$

∎

By definitions 6.1-6.2, we can construct the probability mass function (PMF) and the corresponding cumulative mass function (CMF) of the link travel time for traffic entering the link at time index k (strictly speaking k is an interval). To illustrate the methodology, we depict a collection of cumulative inflow and outflow distributions in Figure 6.28. As we can observe from the figure, there would be two kinds of stochastic travel time distributions for a signalized link. In the figure, the travel time distribution for the vehicle enters the

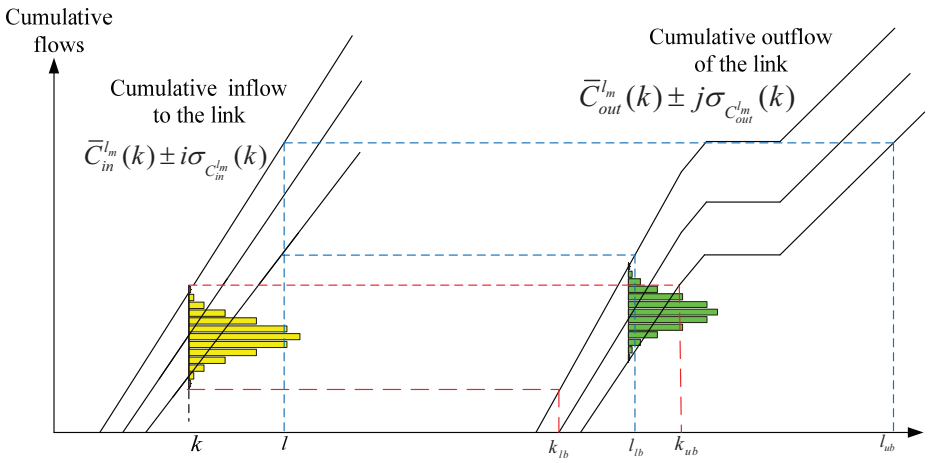


Figure 6.28: An illustration of choice of the "*sampling*" process

link at time k that would exit from the link almost before the red phase, i.e. the sample region $[k_{lb}, k_{ub}]$ falls in the green phase, is a skew normal distribution as illustrated in Figure 6.29. The other kind of travel time distribution happens when the vehicle enters the link at time l that has a chance to exit from the link before or after the red phase of the signal. As expected, this kind of distribution would be bimodal like distribution (or a skew normal distribution truncated by the red signal phase) as demonstrated in Figure 6.30.

As illustrated in Figure 6.30, the PMF of the signal delay is defined to be zero in the red phase since the traffic is not allowed to outflow in the red phase of the signal. Most of the existing methods, including what we have applied to estimate the signal delay in Section 6.5.2.3, model the dynamic signal delay by a uni-modal distribution, e.g. Figure 6.29. However, it would be somehow unjustified when a uni-modal distribution is applied to estimate the bimodal like distribution depicted in Figure 6.30[2]. For example, the distribution in Figure 6.30, the mean and median would fall in the red phase, even though the traffic cannot leave the link at that time. The standard deviation would be also unreasonably large, which does not reflect the true variance of the signal delay. Therefore, it is interesting for us to provide the travelers with the most optimistic and/or pessimistic delays that would be encountered from investigating this kind of distribution for the purpose of real time applications, e.g. route guidance and traffic control.

In the following test, we specify the non-stationary inflow to the roadway segment as

$$u_{in}(k) = A \sin\left(\frac{(k+30)\,\pi}{NT}\right),\ \forall\ k \in [1,\ NT]\,, \tag{6.19}$$

[2] Bimodal distributions are a commonly used example of how summary statistics such as the mean and standard deviation can be deceptive when applied to an arbitrary distribution.

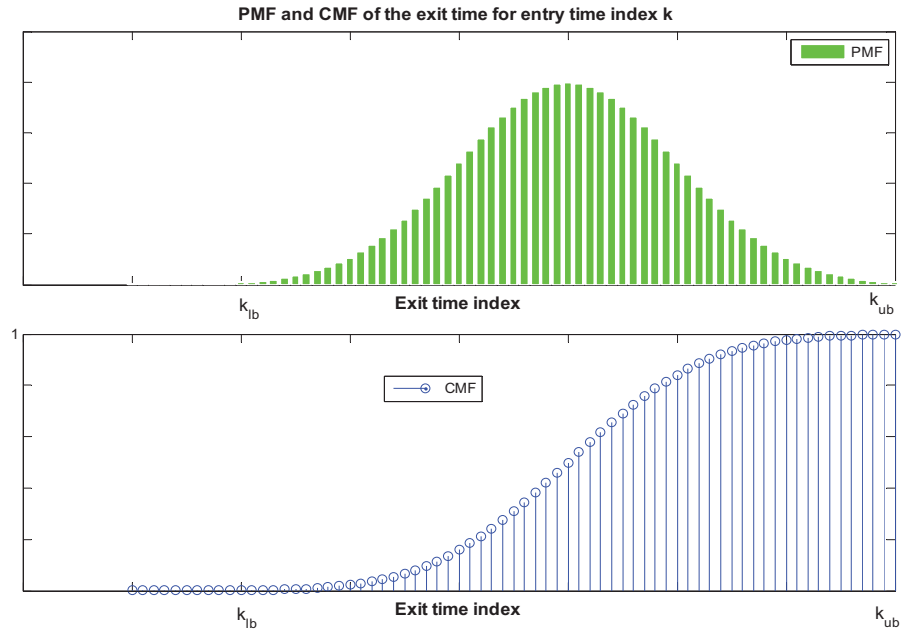


Figure 6.29: The PMF and the corresponding CMF with respect to exit time index for entry time index k

Figure 6.30: The PMF with respect to exit time index for entry time index l

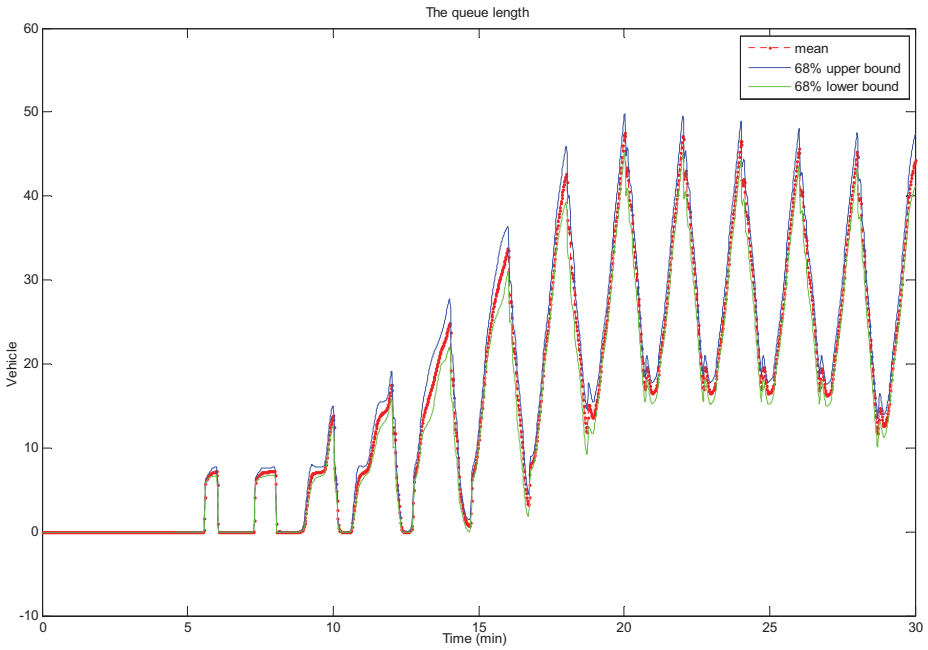


Figure 6.31: Estimated queue length and its 68% confidence interval

where A=700, and NT denotes the simulation horizon, which is chosen to be 30 min (or 1800 steps) in this example. We assume the SD of the inflow at each time is 5% of its nominal value. The estimated queue length is depicted in Figure 6.31.

We calculate the average delay for non-stationary inflow traffic volumes. We assume the inflow is given by (6.19) with different amplitudes and define the time average inflow as the average arrival rate, that is,

$$\frac{\int_0^T A \sin(\omega t + \varphi_0)\, dt}{T} = q_a.$$

The degree of saturation is then defined in line with Section 6.5.2.3. In line with the previous test, we evaluate the average delay for 15 minutes time interval. To accommodate the overflow queues at high arrival rates, we lengthen each of the cells to 100 meters long. Figure 6.32 depicts the average delays obtained by the SCTM against the approximation formulas for non-stationary inflow traffic volumes and the MCS of the MCTM with $\{50, 1000\}$ samples. As expected, when the arrival rates are low, the results are consistent. Different from the test in Section 6.5.2.3, i.e. Figure 6.25, the SCTM estimates larger delays than the coordinate transformation based methods does when degree of saturation is larger than 0.8, e.g. the Akcelik's formula. Note also that the MCS of the MCTM approach produces even larger delays than the SCTM approach. Generally, the delays obtained by these two approaches get closer as the sample size of the MCS approach becomes larger. The computational time of the SCTM approach is almost the same with

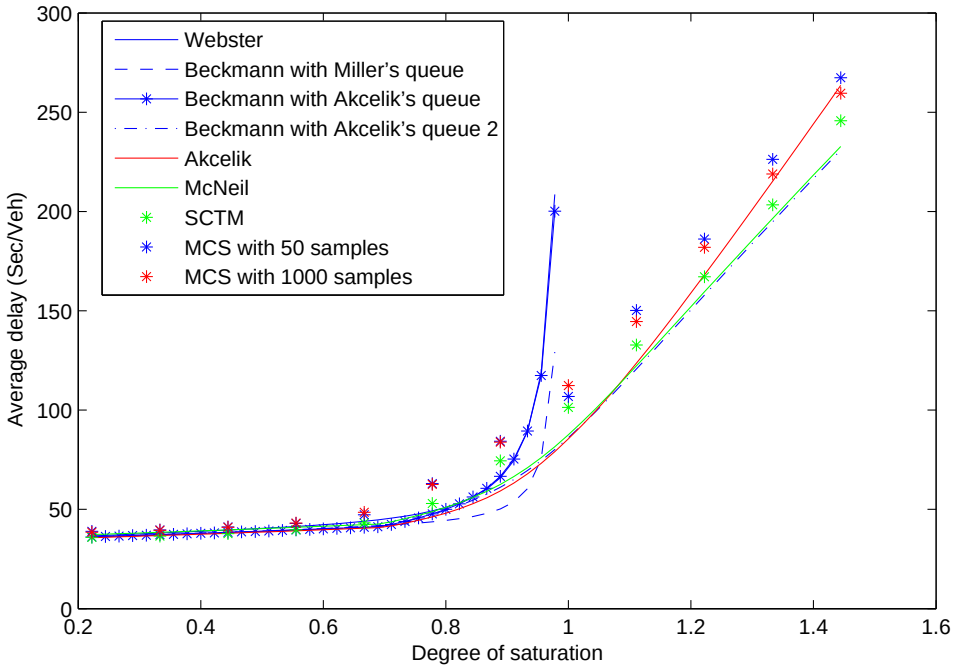


Figure 6.32: Comparison of delay estimations for different degrees of saturation (non-stationary inflows

that of the MCS of the MCTM approach with 50 samples. However, the computational time of the MCS approach does not increase linearly with respect to the sample size. The coordinate transformation based methods, e.g. the Akcelik's formula, underestimate the delays for non-stationary inflow traffic volumes when the degree of saturation is larger than 0.8 (Akcelik and Rouphail, 1993; Rouphail et al., 2000; Brilon and Wu, 1990). As explained in Figure 6.26, when the degree of saturation is high (above 1.2), the delays tend to those obtained by the deterministic queuing model. The results show good consistent again.

6.6 Conclusions

The SCTM is extended to model the stochastic traffic dynamics of a traffic network with stochastic demand and supply in this chapter. The original SCTM for one freeway segment consisting of two cells is defined as one basic subsystem of a traffic network. Four basic functional blocks of a traffic network, i.e. freeway corridor, on-/off- ramps (traffic merge/diverge), signalized junction, are identified. A long freeway corridor is represented as a system connected by several basic SCTM subsystems. For ramps with heavy traffic, we consider the ramps as SCTM subsystems. An isolated signalized junction is divided into several SCTM subsystems. Each of these subsystems consists of several phases ac-

cording to the signal phase under certain assumptions. A traffic network is modeled by all these basic functional blocks. The SCTM subsystems accept the means and variances of the stochastic travel demand and supply functions as exogenous inputs, which in turn produce cell traffic densities and outflow of the roadway segment in terms of mean and variance as well as the probabilities of occurrence of different operational modes. In the first numerical example, we demonstrate that the proposed network SCTM can be used as a stochastic dynamic traffic network model for traffic control and management. The uncertain traffic dynamics and probabilistic wave-fronts are captured. In the second test, the model is applied to estimate queues and delays at signalized intersections. Comparison with some traditional delay and queue estimation formulas is conducted. The numerical results show good consistency between the SCTM and these formulas. However, compared with these delay and queue formulas the SCTM approach has wider application opportunities, since it can describe the temporal evolution of the queue length distribution at signalized junctions with uncertain supply functions by assuming any arrival distribution, which can be non-stationary. Two kinds of dynamic delay distributions are found, i.e. skew normal distribution and bimodal like distribution. The summary statistics such as the mean and variance can be deceptive when applied to the second kind of distribution.

These are several potential applications of the SCTM. For the surveillance purpose, the SCTM can be utilized to provide a short-term prediction using the historical and on-line data of travel demand and traffic state. The prediction (in terms of travel time and traffic state) under the SCTM considers both demand and supply uncertainties in the future time-step. This allows traffic operators to monitor and devise robust control strategies for freeways. For the dynamic traffic assignment and control, we will extend the SCTM framework to model traffic flows on a general network. The key operational benefit of the SCTM for traffic assignment purpose is the potential continuity of the delay operator which is not the case for the deterministic CTM (due to the potential blocking back condition of an arterial). This is due to the introduction of the stochastic delay in the SCTM which can also be considered as a better paradigm for a long-term traffic prediction. In the next chapter, we will discuss the decision making problem for traffic management under the demand and supply uncertainties.

Appendix: notations adopted in this chapter

Notations	
mix	Finite mixture distribution
$\Pr(\cdot)$	Probability evaluation
$q_{u,j}$	Inflow to the upstream boundary of subsystem j, with $\rho_{u,j}$ the corresponding density
$q_{d,j}$	Outflow from the downstream boundary of subsystem j, with $\rho_{d,j}$ the corresponding density
$r_{on,j}$	Measured on-ramp flow to subsystem j
$r_{off,j}$	Measured off-ramp flow from subsystem j
s	Operational mode of the SCTM, $s = FF, CC, CF, FC1, FC2$
$\rho_{c,i}$	Critical density of cell i
$\rho_{J,i}$	Jam density of cell i
$w_{J,i}$	Backward wave speed of cell i
$v_{f,i}$	Free-flow speed of cell i
$Q_{M,i}$	Capacity of cell i
$P_s(k)$	Probability of occurrence of mode s at time k
$\bar{\rho}(k)$	Vector of joint traffic density at time k with $\mu(k)$ the corresponding mean
$\theta_s(k)$	Extended state of mode s
$S_j(k)$	Sending function of subsystem j
$R_j^a(k)$	Receiving function of subsystem j
$g_R(\cdot)$	Probability density function of receiving function
$d_j(k)$	On-ramp demand to node (or subsystem) j at time k
$f_j(k)$	Off-ramp demand from node (or subsystem) j at time k
$r_j(k)$	Actual on-ramp flow to node (or subsystem) j at time k
$f_j^a(k)$	Actual off-ramp flow from node (or subsystem) j at time k
$Q_1^{f_j}$	Capacity of the first cell of off-ramp connected to node j
$w_1^{f_j}$	Backward wave speed of the first cell of off-ramp connected to node j
$\rho_{J,1}^{f_j}$	Jam density of the first cell of off-ramp connected to node j
$\rho_1^{f_j}$	Density of the first cell of off-ramp connected to node j
$P_y^{f_j}(k)$	Probability of occurrence of off-ramp event y at time k
$f_y^{f_j}(k)$	Off-ramp flow of off-ramp event y at time k
$P_x^{r_j}(k)$	Probability of occurrence of on-ramp event x at time k
$d_x^{r_j}(k)$	On-ramp flow of on-ramp event x at time k

Chapter 7

Traffic management under demand and supply uncertainties

This chapter investigates the decision making for traffic management under demand and supply uncertainties. The problem is formulated as a stochastic dynamic programming problem. The stochastic traffic flow under demand and supply uncertainties is described by the proposed SCTM. To be more specific, we represent the SCTM as a class of discrete time stochastic bilinear systems with Markov switching. Based on this model, we investigate the optimal decision making for traffic management of a freeway segment. A closed form of optimal control law is derived in terms of a set of coupled generalized recursive Riccati equations. As the optimal control laws may be fragile with respect to the model miss-specifications, we further pursue a robust (optimal) decision making law which is aimed to act robust with respect to the parameter miss-specifications in the traffic flow model (which can be originated from the calibration process), and to attenuate the effect of disturbances in the freeway network (where demand uncertainty is usually taken as a kind of disturbance). This robust decision making problem can be also recognized as an equivalent optimal decision making problem. Finally, we list some practical issues in traffic management that can be addressed by extending the current framework.

7.1 Motivation and introduction

There are several categories of traffic management schemes to alleviate traffic congestion. Among these schemes, road pricing, ramp metering, and urban traffic signal control are frequently applied to regulate the traffic flows on urban transportation networks to improve their efficiency. Road pricing has also been shown to be an efficient approach to traffic *demand* management and control (see e.g. Lindsey (2006); Tsekeris and Voß (2009) and

the references therein). The ramp metering and urban traffic signal control schemes are, however, concentrated on regulating the *supply* side of a transportation network. Ramp metering aims to improve the freeway traffic conditions by regulating ramp flows to the freeway mainstream and flows at freeway-to-freeway intersections. As commented by Papageorgiou and Kotsialos (2002), the objective of ramp metering schemes is to operate the freeways at their capacities (capacity flows on the freeways during the peak hours) at the price of introducing short delays at the on-ramps and freeway-to-freeway intersections, which leads to substantial savings of travel time for each individual road user. Urban traffic signal control strategies intend to optimize the (network-wide) traffic signal timing of an urban arterial road network to reduce the total travel time and delay experienced by vehicles (Papageorgiou et al., 2003).

In literature, these three major categories of traffic management schemes are developed independently. However, in practice a traffic network would have all these traffic management schemes simultaneously or at least a hybrid of them. It would be interesting for us to look into the performance of hybrid combinations of these traffic management schemes. By using the cell transmission model (CTM) as the network loading model, Varaiya (2008) compared the performance of four congestion-reducing schemes for a freeway with three lanes (where the case with uncontrolled ramps and no tolls was considered as the base case): (R) ramp control only; (T) one lane is tolled and ramps are uncontrolled; (B) bottlenecks are tolled and ramps are uncontrolled; and (RB) ramps are controlled and bottlenecks are tolled. It was found that

1. Scheme (T) is inefficient and may leave all travelers worse off in the following sense. As explained in Chapter 4, the tolled travelers should experience no congestion delay when traveling on the tolled lane. Demand on the two free lanes will increase which in turn yields the traffic on the free lanes be settled to the most congested equilibrium. These lanes will become congested throughout their length. As a consequence, the tolled lane must extend all along the freeway, which is a waste of the freeway capacity. All travelers (except for those with a very high value of travel time or tolled travelers) will be worse off.

2. Schemes (R), (B) and (RB) can achieve efficient freeway use to different levels:

   Scheme (R) can achieve efficient use of the freeway by keeping traffic flow on the freeway strictly below its capacity. Further more, a ramp metering strategy can achieve an uncongested equilibrium at a small capacity loss while causing a deadweight welfare loss from queuing delays at the metered ramps.

Scheme (B) can eliminate queues, but has adverse spatial and equity side effects. When the bottlenecks are tolled, as discussed in Chapter 4, the queues will be eliminated. In this sense, the bottleneck toll achieves both efficient use of the freeway and eliminates the deadweight loss of queuing delay at ramps. However, as commented by Varaiya (2008) implementation difficulties and adverse spatial and equity side effects reduce the attractiveness of bottleneck tolls. For the implementation aspect, for example, a freeway would have several bottlenecks and imposing a toll on each of them will render the travelers try to avoid the bottleneck tolls by exiting the freeway before the tolled sections. The bottleneck could be moved upstream or a new bottleneck would be created. "Furthermore, the streets leading from these exits will carry more traffic, which may create new congestion and public opposition." as claimed by Varaiya (2008). The analysis of equity side effect is more complex, and discussing this issue is out of the scope of this chapter.

Scheme (RB) minimizes the adverse spatial and equity side effects of the pure bottleneck toll scheme and is likely to be least costly to implement and maintain. Roughly speaking, this scheme imposes ramp metering control to ensure that traffic flows on the freeways are free-flowing. The consequence of doing this is creating queues at ramps. The bottleneck tolls are implemented to eliminate these queues. As a matter of fact, the RB scheme can be viewed as a combination of the demand (i.e. tolling) and supply (i.e. ramp metering control) management schemes.

Ramp metering control strategies can be classified as fixed-time scheme or traffic-responsive scheme[1] (see Papageorgiou and Kotsialos (2002); Papageorgiou et al. (2003) for an overview of the ramp metering control strategies). Fixed-time strategies are derived off-line for particular times of the day, based on simple static models and the available historical data on the demand and supply sides. Due to the absence of real-time measurements and their static nature (derived from static models and cannot respond to the real-time traffic condition), fixed-time strategies may lead either to overload of the mainstream flow (and thus cause congestion) or under-utilization of the freeway (Papageorgiou and Kotsialos, 2002; Gomes, 2004). Owning to the drawback of fixed-time strategies,

[1]Some authors categorized ramp metering control strategies into predictive and reactive strategies. The reactive type is corresponding to the traffic-responsive or feedback scheme, while predictive type is corresponding to the fixed-time scheme. A detailed review on predictive on-ramp metering control strategies can be found in Gomes (2004).

traffic-responsive ramp metering strategies are preferred. Based on real-time measurements from sensors (or other intelligent transportation system infrastructures) installed, traffic-responsive ramp metering strategies are developed by adjusting the control to react to the real-time traffic condition. In terms of control method and information used, it can be further classified as reactive strategies or proactive strategies. The former aims at maintaining the freeway traffic conditions close to pre-specified (or desired) values by adjusting the metering rates (or control) using real-time measurements, while the later aims at specifying optimal traffic conditions for a freeway or a freeway network based on demand and model predictions over a sufficiently long time horizon (Smaragdis et al., 2004).

In terms of network topology, ramp metering strategies can be further classified as local or coordinated schemes. Local strategies make use of traffic measurements of the current ramp and its adjacent ramps to adjust the corresponding individual ramp metering rates while coordinated strategies make use of all traffic measurements from the freeway network (or part of the network to coordinate). Compared with the coordinated ramp metering strategies, local strategies are far more easy to design and implement. It has been proved that, under recurrent traffic congestion conditions, the performance of local ramp metering, e.g. the ALINEA—a local traffic-responsive ramp metering algorithm, is close to that of the coordinated approach, e.g. the METALINE—the coordinated version of ALINEA (Papageorgiou and Kotsialos, 2002; Papageorgiou et al., 2003). For the purpose of freeway management, the control efficiency of local ramp metering with unlimited (or sufficiently high) ramp storage could would be very high (Papageorgiou and Kotsialos, 2002; Papageorgiou et al., 2003). However, attentions should be paid to the concomitant side effects, e.g. unfairly long waiting times at that particular ramp. Ramp queues must also be restricted to avoid interference with adjacent street traffic when we apply local ramp metering. For these reasons, coordinated ramp metering is proposed to address these issues (Papageorgiou and Kotsialos, 2002; Papageorgiou et al., 2003; Papamichail and Papageorgiou, 2008; Papageorgiou et al., 2008b). Despite their complex structure and implementation issues, the coordinated ramp-metering strategies, e.g. the METALINE and SWARM (System Wide Adaptive Ramp Metering), are more efficient than the local ramp-metering strategies when there are multiple bottlenecks on the freeway, restricted ramp storage spaces, and non-recurrent congestions Papageorgiou and Kotsialos (2002); Papageorgiou et al. (2003); Papamichail and Papageorgiou (2008); Papageorgiou et al. (2008b). The optimal on-ramp flows is obtained through minimizing objective functions such as total time spent (TTS) under limited ramp storage space re-

striction. The problem can be converted to a model-based optimal control formulation which can be solved and implemented in the generic Advanced Motorway Optimal Control (AMOC) tool (Kotsialos et al., 2002; Kotsialos and Papageorgiou, 2004). As required by optimal control theory, this approach requires a perfect model calibration and a perfect information with respect to the future disturbances (i.e., demand and exit-rate profiles). However, freeways are exposed to various demand and supply uncertainties. Due to various inherent uncertainties (including the capacity uncertainty), the open-loop optimal solution delivered by optimal control theory becomes suboptimal when directly applied to the freeway traffic process. The efficiency of optimal control based strategies increasingly deteriorate with increasing disturbance-prediction and modeling errors. For example, in view of the uncertainty of highway capacity, any ramp metering strategy attempting to achieve a pre-specified capacity flow value, will either lead to overload and congestion or to under utilization of the infrastructure. However, as it is widely recognized in control theory and application (see e.g. Kotsialos et al. (2002)), as a control law, the applied traffic control strategy should be intelligent and robust enough to the uncertainties, and if possible, be optimal. When the network size becomes large, it is unlikely for us to consider it as a single system to optimize its performance due to the sophisticated network dynamics and interactions between the actuators, sensors, and the traffic control center and/or commercial constraints. Due to their geographically distributed and uncertainty in a dynamic environment, traffic networks are well suited for the concept of multiagent system (MAS). Applications of MAS to transportation networks were reviewed by Schleiffer (2002); Chen and Cheng (2010).

In this chapter, we will develop some policies to support decision making for traffic management under demand and supply uncertainties. The proposed policies can be ramp metering control and/or (flow-dependent) dynamic congestion pricing. The uncertain demand profiles are modeled as external disturbances. With respect to this demand uncertainty, the controller aims to achieve disturbance attenuation, that is to minimize the effect of the disturbances. With respect to the supply uncertainty, the controller aims at the robust property. The optimal decision problem is formulated as a stochastic optimal control problem based on stochastic dynamic programming problem. The stochastic traffic flow under demand and supply uncertainties is described by the SCTM proposed in Chapter 5. We investigate the optimal decision making for traffic management of a freeway segment. As the optimal control may be fragile with respect to the model miss-specifications, we further pursue a robust (optimal) decision making law which is aimed to act robust with respect to the parameter miss-specifications in the traffic flow model

(which can be originated from the calibration process), and to attenuate the effect of disturbances in the freeway network (where demand uncertainty is usually taken as a kind of disturbance). Finally, we list some practical issues in traffic management that can be addressed by extending the current framework, e.g. the MAS approach to access the traffic management for general traffic networks.

7.2 The SCTM as a Markov switching state space model

Following the previous chapters, the dynamics of the SCTM is propagated by a class of discrete time bilinear systems in conjunction with a finite mixture distribution of five Gaussian random vectors corresponding to the five events (or the five operational modes of the freeway segment). Actually, as explained in Chapter 5, we can divide one SCTM system into three functional blocks, i.e. the five operational modes represented by a class of stochastic bilinear systems, the probabilities of occurrence of the five modes, and a finite mixture distribution. The finite mixture distribution is employed based on the justification that, based on the Wiener approximation theorem, any non-Gaussian distribution can be expressed as, or approximated sufficiently well by, a finite sum of known Gaussian densities, which is summarized by Anderson and Moore (1979) in their book "Optimal Filtering" in the following lemma:

Lemma 7.1. Any probabilistic density $p(x)$ associated with an n-dimensional vector x can be approximated as closely as desired by a density of the form

$$p_A(x) = \sum_{i=1}^{k} a_i \mathcal{N}\left(\bar{x}_i, \Sigma_i\right), \tag{7.1}$$

for some integer k, and positive scalars a_i with $\sum_{i=1}^{k} a_i = 1$. ∎

The finite mixture distribution is also widely employed to approximate/estimate the "overall" effect of the switching state space models, e.g. the mixture Kalman filtering, particle filtering with finite mixture, etc, which have wide application in bioinformatics, biology, economics, finance, hydrology, marketing, medicine, and engineering (Frühwirth-Schnatter, 2006; Costa et al., 2005). Due to the Markovian property of the model, the dynamic process of the SCTM with finite mixture distribution can be classified as a finite Markov mixture distribution process which is a special case of the Markov switching state space models (Harrison and Stevens, 1976; Timmermann, 2000; Frühwirth-Schnatter, 2006; Costa et al., 2005).

Remark 7.1. In fact, such kind of models belong to the basic Markov switching state space models based on hidden Markov chains. Various terminologies have been adopted

to denote such models. The term finite Markov mixture models is preferred by biologists. Finite Markov mixture models are usually called hidden Markov models in engineering applications. The terms Markov switching models or regime-switching models are preferred by economists who used Markov switching models to analyze stock market returns, interest rates, etc. The kind of process is also known as Markov jump system (or stochastic hybrid systems, hybrid dynamic Bayesian networks) in control and telecommunication engineering. ∎

Before we classify the dynamics of SCTM into such kind of Markov switching state space models, we first give some intuitive justification on doing this. The basic idea of a Markov switching state space model is that a priori no single model is expected to hold for all time indices, rather the possibilities that different modes hold at different time points are explicitly recognized by modeling the hidden model indicator and the corresponding transition matrices as being dynamic over time. This basic idea coincides with the starting point of the SCTM, wherein the five possible operational modes with different probabilities of occurrence already written in state space form. If we identify the operational modes as different models in the Markov switching state space model and the (time-varying) probabilities of occurrence as the dynamic hidden model, we can recognize the SCTM as a special case of the Markov switching state space model. The operational modes jump between possible values following a discrete-time Markov chain with certain transition probabilities.

A switching nonlinear Gaussian state space model is based on the state space form, however, the detailed dynamics (or the system matrices of the SCTM, each mode has different system matrices) are driven by a hidden model indicator $\theta(k)$. To better understand this dynamics, we represent the SCTM in a compact form as the following discrete-time Markov jump bilinear dynamic system driven by white noises which represent the demand and supply uncertainties:

$$\begin{aligned} \rho(k+1) &= \left(A_{\theta(k)} + \sum_{s=1}^{v} A_{\theta(k),s}\omega_s(k)\right)\rho(k) + \left(\sum_{s=1}^{v} B_{\theta(k),s}\omega_s(k)\right)\lambda(k) + B_{\theta(k)}u(k), \\ &\quad \rho(0) = \rho_0,\ \theta(0) = \theta_0, \end{aligned} \tag{7.2}$$

where $u(k)$ denotes the uncertain demand and $\omega(k)$ and $\lambda(k)$ denote the uncertain supply to the freeway segment, and $\theta(k)$ indicates the five operational modes, respectively. $\{\theta(k),\ k = 1, \cdots, N\}$ is a sequence of random variables, allowed to take values in the discrete space. In the SCTM case, this space is the five operational modes, i.e. $\theta(k) = \{1, 2, 3, 4, 5\}$, where 1 represents the FF mode and 2 represents the CC mode, 3 denotes the CF mode, 4 denotes the FC1 mode, and 5 represents the FC2 mode. The

corresponding system matrices are specified in Chapter 5. To complete the model specification without introducing too many mathematical jargons, we impose some probabilistic structure on the transition probability matrix $\xi(k) = [p_{i,j}(k)]$, where $p_{i,j}(k)$ is thus defined

$$p_{i,j}(k) = \Pr\left(\theta(k) = j | \theta(k-1) = i\right). \tag{7.3}$$

To explain the physical meaning of this probability, for example, we denote i the FF mode and j the CC mode. Then $p_{i,j}(k)$ denotes the probability that the freeway segment would transfer to CC mode at the next time step k given the current free-flowing condition. By definition, all elements of ξ are nonnegative and the elements of each row sum to 1:

$$p_{i,j}(k) \geq 0, \sum_{j=1}^{5} p_{i,j}(k) = 1.$$

Note that, in Chapter 5, we only define the probabilities of occurrence of the five operational modes without specifying the transition probabilities between modes. Actually, this renders the SCTM (finite mixture distribution) as a special case of the Markov switching state space models with the following transition matrix:

$$\xi(k) = \begin{pmatrix} \Pr_{FF}(k) & \Pr_{CC}(k) & \Pr_{CF}(k) & \Pr_{FC1}(k) & \Pr_{FC2}(k) \\ \Pr_{FF}(k) & \Pr_{CC}(k) & \Pr_{CF}(k) & \Pr_{FC1}(k) & \Pr_{FC2}(k) \\ \Pr_{FF}(k) & \Pr_{CC}(k) & \Pr_{CF}(k) & \Pr_{FC1}(k) & \Pr_{FC2}(k) \\ \Pr_{FF}(k) & \Pr_{CC}(k) & \Pr_{CF}(k) & \Pr_{FC1}(k) & \Pr_{FC2}(k) \\ \Pr_{FF}(k) & \Pr_{CC}(k) & \Pr_{CF}(k) & \Pr_{FC1}(k) & \Pr_{FC2}(k) \end{pmatrix}. \tag{7.4}$$

7.2.1 A model reduction of the SCTM for control and filtering

As shown in the previous section, the whole transition matrix is not well defined in the SCTM when it is taken as a Markov switching state space model. One way to approach the control problem of the SCTM is to consider it as a Markov jump system with partly unknown transition probabilities, e.g. rather than consider it as $\Pr_{FF}(k)$, we take the transition probability from CC mode to FF mode as an unknown transition probability. The transition matrix is then

$$\xi(k) = \begin{pmatrix} \Pr_{FF}(k) & ? & ? & ? & ? \\ ? & \Pr_{CC}(k) & ? & ? & ? \\ ? & ? & \Pr_{CF}(k) & ? & ? \\ ? & ? & ? & \Pr_{FC1}(k) & ? \\ ? & ? & ? & ? & \Pr_{FC2}(k) \end{pmatrix}. \tag{7.5}$$

The elements of each row sum to 1. However, the control problem would be too complicated and computationally expensive. We refer the readers to some pioneer works, e.g. Zhang and Boukas (2009a,b), on this subject.

Another way to access the control and filtering problems is to reduce the modes of the SCTM similar to Sun et al. (2003); Sun (2005), where only the two-steady modes, i.e. FF and CC modes, are retained while the other three transient modes are ignored. This approach has been validated by some empirical studies and interfaced with various data sources and traffic simulators (Sun et al., 2003; Sun, 2005). Under such circumstance, the transition matrix can be defined as

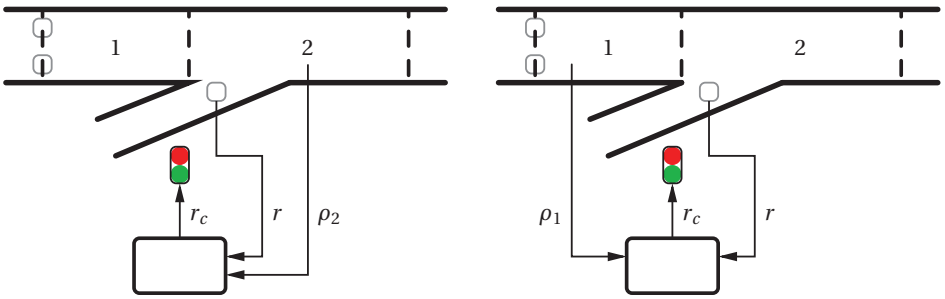
$$\xi(k) = \begin{pmatrix} \Pr_{FF}(k) & 1 - \Pr_{FF}(k) \\ 1 - \Pr_{CC}(k) & \Pr_{CC}(k) \end{pmatrix}. \tag{7.6}$$

7.2.2 A refinement of the control variables

As for decision making under different purposes, we have different control variables. For example, in ramp metering control, we usually do not control the mainstream flows of freeways, instead, we control the on-/off- ramp flows[2] to achieve the prescribed objective. However, in the derivation of the SCTM, we include the mainstream flows in the control vector. When road pricing scheme is used to optimize the performance of a freeway corridor, we can charge the drivers at the mainstream and on-ramps of the freeway but not at the off-ramps. Therefore, it is necessary for us to refine the control variables for different purposes.

For the freeway ramp metering control, the mainstream flows are now taken as disturbance $\omega(k) = \begin{pmatrix} q_u(k) \\ q_d(k) \end{pmatrix}$. The weighting matrices for this disturbance vector are $C_1(k) = \begin{pmatrix} \frac{T_s}{l_1} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}_{p\times 2}$ for the FF mode and $C_2(k) = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & -\frac{T_s}{l_p} \end{pmatrix}_{p\times 2}$ for the CC mode. The control variables are the on-/off- ramp flows $u(k) = \begin{pmatrix} r_e(k) \\ f_e(k) \end{pmatrix}$. The weighting

[2]Most of the time only the on-ramps are metered. Recently, Li et al. (2009) proposed a mixed integer model for an integrated control between off-ramp and arterial traffic flows. The off-ramps are controlled to minimize the queue spillback from off-ramps to the freeway mainline that may significantly degrade the performance quality of the entire freeway system. Here we choose off-ramps as control variables to optimize the performance of freeway-to-arterial intersections. This could be important when we go to the multiagent settings as depicted in Figure 7.3, i.e. to coordinate the performance of freeway systems and urban arterials.



(a) *Free-flow* mode. (b) *Congested* mode.

Figure 7.1: Different on-ramp metering control structures for different congestion modes of the SMM. Source: Sun (2005)

matrices for this control vector are $B_1(k) = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ \frac{T_s}{l_b} & 0 \\ \vdots & \vdots \\ 0 & -\frac{T_s}{l_{e+1}} \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}_{p\times 2}$ for the FF mode and

$B_2(k) = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ \frac{T_s}{l_{b-1}} & 0 \\ \vdots & \vdots \\ 0 & -\frac{T_s}{l_e} \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}_{p\times 2}$ for the CC mode.

To demonstrate the application of SCTM to ramp metering control for a freeway segment, let's consider a freeway segment consisting of two cells and one on-ramp. As depicted in Figure 7.1, in the SMM different on-ramp metering control structures are applied for different congestion modes (Sun, 2005). For the FF mode, the downstream traffic density is required for feedback on-ramp metering control design while the upstream traffic density is required for the CC mode. In the SCTM, due to the the Markov switching model, these two modes are possible but with different probabilities of occurrence. Both the upstream and downstream traffic densities are required for the on-ramp metering control design as shown in Figure 7.2. As it will be shown in the forthcoming section, the control actually depends on these probabilities of occurrence.

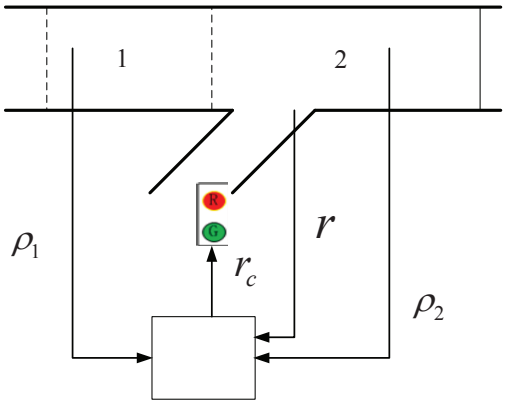


Figure 7.2: A unified on-ramp metering control structure for the SCTM

7.3 A stochastic optimal control framework for the SCTM

7.3.1 The problem formulation and basic assumptions

The optimal control problem for such kind of discrete-time stochastic bilinear system with Markov switching has been studied in several articles in control theory recently (Costa and Wanderlei, 2007; Hou et al., 2010). However, due to the fact that the SCTM is not in the same form of systems studied in the cited papers, we cannot apply their results directly. Instead, we have to extend their results for the SCTM. In this chapter, we will develop an optimal control framework for the SCTM based on the stochastic dynamic programming[3], i.e. (7.62). In most situations, a quadratic objective function is chosen for control engineering practice and traffic flow control practice (Papageorgiou and Kotsialos, 2002; Papageorgiou et al., 2003; Sun, 2005; de Oliveira and Camponogara, 2010). The objective function aims to minimize the total time spent and the risk of over saturation and the spillback of link queues. This approach is referred to as the LQ[4] (Linear Quadratic control) in control theory and was used to design a coordinated ramp-metering control (Papageorgiou and Kotsialos, 2002; Papageorgiou et al., 2003). Therefore, in this chapter, we formulate the optimal control problem for the SCTM as

$$\begin{aligned} \min_{\mathbf{u}} V(\rho_0, \theta_0, u_0) &= \sum_{k=0}^{N} E\left(\rho^T(k) Q_{\theta(k)}(k)\rho(k) + L_{\theta(k)}(k)\rho(k)\right) \\ &+ \sum_{k=0}^{N-1} E\left(u^T(k) M_{\theta(k)}(k)u(k) + H_{\theta(k)}(k)u(k)\right), \end{aligned} \tag{7.7}$$

[3] A brief introduction to the stochastic dynamic programming is given in the appendix of this chapter.

[4] The linear objective function case which aims to minimize the total travel time (TTT) or total time spent (TTS) can be regarded as a special case.

subject to

$$\begin{aligned}\rho(k+1) &= \left(A_{\theta(k)}+\sum_{s=1}^{v}A_{\theta(k),s}\omega_s(k)\right)\rho(k)+\left(\sum_{s=1}^{v}B_{\theta(k),s}\omega_s(k)\right)\lambda(k)\\ &+ C_{\theta(k)}\omega(k)+B_{\theta(k)}u(k),\ \rho(0)=\rho_0,\ \theta(0)=\theta_0,\end{aligned}\tag{7.8}$$

where $Q_{\theta(k)},\ L_{\theta(k)}(k),\ M_{\theta(k)}(k),\ H_{\theta(k)}(k)$ are weighting matrices to be specified according to different design purposes, see e.g. Papageorgiou and Kotsialos (2002); Kotsialos et al. (2002,b); Kotsialos and Papageorgiou (2004); Papageorgiou et al. (2003); Sun (2005); Papamichail and Papageorgiou (2008); Carlson et al. (2010) and the references therein.

To apply the stochastic dynamic programming framework (or Bellman's optimality condition), we define the following intermediate problem

$$\begin{aligned}\min_{\mathbf{u}} V\left(\rho(\tau),\theta(\tau),\tau\right) &= \sum_{k=\tau}^{N}E\left\{\left(\rho^T(k)Q_{\theta(k)}(k)\rho(k)+L_{\theta(k)}(k)\rho(k)\right)|\mathfrak{B}_\tau\right\}\\ &+ \sum_{k=\tau}^{N-1}E\left\{\left(u^T(k)M_{\theta(k)}(k)u(k)+H_{\theta(k)}(k)\rho(k)\right)|\mathfrak{B}_\tau\right\},\end{aligned}\tag{7.9}$$

subject to (7.8), where $\mathfrak{B}_\tau$ is the σ-field [5] generated by the random variables $\{\rho(t),\theta(t);t=0,\cdots,\tau\}$, so that $\mathfrak{B}_k\subset\mathfrak{B}_{k+1}\subset\mathfrak{F}$ with $\mathfrak{F}$ denotes the set of $\mathfrak{F}$-measurable random variables. The problem (7.9)-(7.8) need to be well posed, which is defined as follows (Ait Rami et al., 2002):

Definition 7.1. The problem (7.9)-(7.8) is well posed if

$$V\left(\rho_0\right)=\inf_{\mathbf{u}}V\left(\rho_0,u_0,\cdots,u_{N-1}\right)>-\infty,$$

for any random variable ρ_0 which is independent of the noises $\omega_s(k),\omega(k),\lambda(k),k=0,1,\cdots,N-1$. ∎

The well-posedness of the intermediate problem can be similarly defined, see e.g. Costa and Wanderlei (2007).

To simplify the analysis, we impose the following assumption (Costa et al., 2005):

Assumption 7.1. Assume that for any measurable functions f and g,

$$E\left(f\left(\nu(k)\right)g\left(\theta(k+1)\right)|\mathfrak{B}_k\right)=E\left(f\left(\nu(k)|\mathfrak{B}_k\right)\right)\sum_{j=1}^{r}p_{\theta(k),j}(k)g(j).\tag{7.10}$$

Assumption 7.1 is somehow "abstract", to explain its mathematical and physical meanings we provide an intuitive example as:

[5] Several rigorous definitions and basic results from probability theory which are used in the developments of this chapter can be found in Costa et al. (2005); Dragan et al. (2010).

Example 7.1. Consider a simple discrete-time bilinear stochastic system with Markov switching

$$x(k+1) = A_{\theta(k)}x(k) + B_{\theta(k)}x(k)\omega_e(k), \tag{7.11}$$

where $\omega_e(k)$ is a scalar white noise sequence satisfying $E(\omega_e(k)) = 0$, and $E\left(\omega_e^2(k)\right) = 1$. Let $V_e\left(x(k), \theta(k), k\right) = x^T(k)P_{\theta(k)}x(k)$ be an scalar objective function with $P_{\theta(k)} = P_i$ when $\theta(k) = i,\ \forall i = 1, 2, \cdots, r$, and r is the number of possible modes. Given a realization of the state x_k at time k and possible mode $\theta(k) = i$, we have

$$\begin{aligned}
& E\left(V_e\left(x(k+1), \theta(k+1), k+1|\mathfrak{B}_k\right)\right) \\
\triangleq\ & E\left(V_e\left(x(k+1), \theta(k+1), k+1|x_k, \theta(k) = i\right)\right) \\
=\ & \sum_{j=1}^{r} \Pr\{\theta(k+1) = j|\theta(k) = i\} x_k^T \left(A_i^T P_j A_i + B_i^T P_j B_i\right) x_k \\
=\ & \sum_{j=1}^{r} P_{i,j}(k) x_k^T \left(A_i^T P_j A_i + B_i^T P_j B_i\right) x_k \\
=\ & x_k^T \left(A_i^T \mathcal{E}_i(P,k) A_i + B_i^T \mathcal{E}_i(P,k) B_i\right) x_k,
\end{aligned} \tag{7.12}$$

where $\mathcal{E}_i(P,k)$ is defined by (7.13). ∎

7.3.2 Definitions of operators

In this chapter, we define the following notations

$$E(\omega_s(k)) = \bar{\omega}_s(k),\ E(\omega_s(k)\omega_s(k)) = \sigma_s(k),\ E(\omega_{s_1}(k)\omega_{s_2}(k)) = \sigma_{s_1,s_2}(k),$$

for the noise sequences. Instead of dealing directly with the system state, we express the system in the Markovian framework via the augmented state $(\rho(k), \theta(k))$, and define the following operators:

$$\mathcal{E}_i(X,k) = \sum_{j=1}^{r} p_{i,j}(k) X_j, \tag{7.13}$$

$$\begin{aligned}
\mathcal{A}_i(X,k) =\ & Q_i(k) + A_i^T \mathcal{E}_i(X,k) A_i + 2A_i^T \mathcal{E}_i(X,k) \sum_{s=1}^{v} \bar{\omega}_s(k) A_{i,s} \\
& + \sum_{s_1=1}^{v} \sum_{s_2=1}^{v} \sigma_{s_1,s_2}(k) A_{i,s_1}^T \mathcal{E}_i(X,k) A_{i,s_2},
\end{aligned} \tag{7.14}$$

$$\mathcal{G}_i^1(X,k) = A_i^T \mathcal{E}_i(X,k) \sum_{s=1}^{v} \bar{\omega}_s(k) B_{i,s} + \sum_{s_1=1}^{v} \sum_{s_2=1}^{v} \sigma_{s_1,s_2}(k) A_{i,s_1}^T \mathcal{E}_i(X,k) B_{i,s_2}, \tag{7.15}$$

$$\mathcal{G}_i^2(X,k) = A_i^T \mathcal{E}_i(X,k) C_i + \sum_{s=1}^{v} \bar{\omega}_s(k) A_{i,s_1}^T \mathcal{E}_i(X,k) C_i, \tag{7.16}$$

$$\mathcal{G}_i^3(X,k) = \left(A_i^T \mathcal{E}_i(X,k) B_i + \sum_{s=1}^{v} \bar{\omega}_s(k) A_{i,s}^T \mathcal{E}_i(X,k) B_i\right)^T, \tag{7.17}$$

$$\mathcal{R}_i(X,k) = B_i^T \mathcal{E}_i(X,k) B_i + M_i(k), \tag{7.18}$$

$$\mathcal{Z}_i(X,k) = \sum_{s_1=1}^{v} \sum_{s_2=1}^{v} \sigma_{s_1,s_2}(k) B_{i,s_1}^T \mathcal{E}_i(X,k) B_{i,s_2}, \tag{7.19}$$

$$\mathcal{T}_i(J,k) = \left(L_i(k) + \mathcal{E}_i(J,k) \left(A_i + \sum_{s=1}^{v} \bar{\omega}_s(k) A_{i,s} \right) \right)^T, \tag{7.20}$$

$$\mathcal{H}_i^1(J,k) = \left(H_i(k) + \mathcal{E}_i(J,k) B_i \right)^T, \tag{7.21}$$

$$\mathcal{H}_i^2(J,k) = \mathcal{E}_i(J,k) \left(\sum_{s=1}^{v} \bar{\omega}_s(k) B_{i,s} \right), \tag{7.22}$$

$$\mathcal{H}_i^3(J,k) = \mathcal{E}_i(J,k) C_i. \tag{7.23}$$

7.3.3 Derivation of an optimal strategy

For the case in which the state $\rho(k)$ is available to the controller, the solution of the quadratic optimal control problem has been solved in the literature. In terms of Bellman's optimality principle, the optimal control is given by solving a set of recursive coupled Riccati difference equations. Before we prove the Bellman's optimality condition for this optimal control problem, we first prove the following proposition in line with Costa and Wanderlei (2007); Costa and Okimura (2009).

Proposition 7.1. Let $P = (P_1, \cdots, P_r)$, $J = (J_1, \cdots, J_r)$ be matrix functions of appropriate dimension whose elements are real valued function of $\theta(k)$ and γ is a scalar real valued function of $\theta(k)$. For any admissible control, $u(k) = u$[6], $\rho(k) = \rho$, and $\theta(k) = i$, we have that

$$\begin{aligned}
& \rho^T Q_i(k) \rho + L_i(k) \rho + u^T M_i(k) u + H_i(k) u \\
+ \; & E\left(\rho^T(k+1) P_{\theta(k+1)} \rho(k+1) + J_{\theta(k+1)} \rho(k+1) + \gamma_{\theta(k+1)} | \mathfrak{B}_k \right) \\
= \; & \rho^T \mathcal{A}_i(P,k) \rho + u^T \mathcal{R}_i(P,k) u + \lambda^T \mathcal{Z}_i(P,k) \lambda + \sigma(k) \cdot tr\left(C_i^T C_i \mathcal{E}_i(P,k) \right) \\
+ \; & 2\rho^T \mathcal{G}_i^1(P,k) \lambda + 2\rho^T \mathcal{G}_i^2(P,k) \bar{\omega}(k) + 2\rho^T \left(\mathcal{G}_i^3(P,k) \right)^T u + \mathcal{T}_i^T(J,k) \rho \\
+ \; & \left(\mathcal{H}_i^1(J,k) \right)^T u + \mathcal{H}_i^2(J,k) \lambda + \mathcal{H}_i^3(J,k) \bar{\omega}(k) + \mathcal{E}_i(\gamma,k).
\end{aligned} \tag{7.24}$$

∎

Proof. By Assumption 7.1 and proceeding the calculation as shown in Example 7.1, we have that

$$\begin{aligned}
& E\left(\rho^T(k+1) P_{\theta(k+1)} \rho(k+1) + J_{\theta(k+1)} \rho(k+1) + \gamma_{\theta(k+1)} | \mathfrak{B}_k \right) \\
= \; & E\left(\rho^T(k+1) P_{\theta(k+1)} \rho(k+1) | \mathfrak{B}_k \right) + E\left(J_{\theta(k+1)} \rho(k+1) | \mathfrak{B}_k \right) \\
+ \; & E\left(\gamma_{\theta(k+1)} | \mathfrak{B}_k \right).
\end{aligned} \tag{7.25}$$

[6] This means that $u(k)$ has a mean value of u or a realization of u.

$$
\begin{aligned}
& E\left(\rho^T(k+1)P_{\theta(k+1)}\rho(k+1)|\mathfrak{B}_k\right) \\
= \;& \rho^T\left(A_i^T\mathcal{E}_i(P,k)A_i + 2A_i^T\mathcal{E}_i(P,k)\sum_{s=1}^{v}\bar{\omega}_s(k)A_{i,s}\right. \\
& \left. + \sum_{s_1=1}^{v}\sum_{s_2=1}^{v}\sigma_{s_1,s_2}(k)A_{i,s_1}^T\mathcal{E}_i(P,k)A_{i,s_2}\right)\rho \\
+ \;& u^T\left(B_i^T\mathcal{E}_i(P,k)B_i\right)u + \lambda^T\left(\sum_{s_1=1}^{v}\sum_{s_2=1}^{v}\sigma_{s_1,s_2}(k)B_{i,s_1}^T\mathcal{E}_i(P,k)B_{i,s_2}\right)\lambda \\
+ \;& 2\rho^T\left(A_i^T\mathcal{E}_i(P,k)\sum_{s=1}^{v}\bar{\omega}(k)_s(k)B_{i,s} + \sum_{s_1=1}^{v}\sum_{s_2=1}^{v}\sigma_{s_1,s_2}(k)A_{i,s_1}^T\mathcal{E}_i(P,k)B_{i,s_2}\right)\lambda \\
+ \;& 2\rho^T\left(A_i^T\mathcal{E}_i(P,k)C_i + \sum_{s=1}^{v}\bar{\omega}_s(k)A_{i,s_1}^T\mathcal{E}_i(P,k)C_i\right)\bar{\omega}(k) \\
+ \;& 2\rho^T\left(A_i^T\mathcal{E}_i(P,k)B_i + \sum_{s=1}^{v}\bar{\omega}_s(k)A_{i,s}^T\mathcal{E}_i(P,k)B_i\right)u \\
+ \;& \sigma(k)\cdot tr\left(C_i^TC_i\mathcal{E}_i(P,k)\right); \qquad (7.26)
\end{aligned}
$$

$$
\begin{aligned}
& E\left(J_{\theta(k+1)}\rho(k+1)|\mathfrak{B}_k\right) \\
= \;& \mathcal{E}_i(J,k)\left(\left(A_i+\textstyle\sum_{s=1}^{v}\bar{\omega}_s(k)A_{i,s}\right)\rho + \left(\textstyle\sum_{s=1}^{v}\bar{\omega}_s(k)B_{i,s}\right)\lambda + C_i\bar{\omega}(k) + B_iu\right); \qquad (7.27)
\end{aligned}
$$

$$
E\left(\gamma_{\theta(k+1)}|\mathfrak{B}_k\right) = \mathcal{E}_i(\gamma,k). \qquad (7.28)
$$

Adding (7.26)-(7.28) to $\rho^TQ_i(k)\rho + L_i(k)\rho + u^TM_i(k)u + H_i(k)u$ and using the operators defined in (7.13), we can obtain (7.24) by rearranging the terms. □

We cannot use the noise sequences to design the control. Based on our assumption that the state $\rho(k)$ is available for control, the control input has the form of $u(k) = \kappa(\rho(k))$. For our purpose, we are designing a control law to minimize the objective function (7.9) in terms of Bellman's optimality condition (7.62). As the objective function is in a quadratic form, we would like to achieve a minimum by considering the cross terms of the state and control, i.e. $u^T\mathcal{R}_i(P,k)u + 2\rho^T\left(\mathcal{G}_i^3(P,k)\right)^Tu + \left(\mathcal{H}_i^1(J,k)\right)^Tu$. To make this quadratic function more clear, we rewrite this term as

$$
\begin{aligned}
& u^T\mathcal{R}_i(P,k)u + 2\rho^T\left(\mathcal{G}_i^3(P,k)\right)^Tu + \left(\mathcal{H}_i^1(J,k)\right)^Tu \\
= \;& u^T\mathcal{R}_i(P,k)u + 2\left(\rho^T\left(\mathcal{G}_i^3(P,k)\right)^T + \frac{1}{2}\left(\mathcal{H}_i^1(J,k)\right)^T\right)u \\
= \;& u^T\mathcal{R}_i(P,k)u + 2\left(\rho^T\left(\mathcal{G}_i^3(P,k)\right)^T + \frac{1}{2}\left(\mathcal{H}_i^1(J,k)\right)^T\right)\mathcal{R}_i^\dagger(P,k)\mathcal{R}_i(P,k)u. \qquad (7.29)
\end{aligned}
$$

Define the control function as

$$
\kappa\left(\rho(k)\right) = \mathcal{R}_i^\dagger(P,k)\left(\mathcal{G}_i^3(P,k)\rho + \frac{1}{2}\mathcal{H}_i^1(J,k)\right), \qquad (7.30)
$$

then

$$\begin{aligned} & u^T\mathcal{R}_i(P,k)u + 2\rho^T\left(\mathcal{G}_i^3(P,k)\right)^T u + \left(\mathcal{H}_i^1(J,k)\right)^T u \\ = \;& u^T\mathcal{R}_i(P,k)u + 2\kappa^T\left(\rho(k)\right)\mathcal{R}_i(P,k)u \\ = \;& \left(u+\kappa\left(\rho(k)\right)\right)^T\mathcal{R}_i(P,k)\left(u+\kappa\left(\rho(k)\right)\right) - \kappa^T\left(\rho(k)\right)\mathcal{R}_i(P,k)\kappa\left(\rho(k)\right). \end{aligned} \tag{7.31}$$

By utilizing the property of Moore-Penrose inverse, we have that

$$\begin{aligned} \kappa^T\left(\rho(k)\right)\mathcal{R}_i(P,k)\kappa\left(\rho(k)\right) &= \left(\mathcal{G}_i^3(P,k)\rho\right)^T\mathcal{R}_i^\dagger(P,k)\mathcal{G}_i^3(P,k)\rho \\ &+ \left(\mathcal{H}_i^1(J,k)\right)^T\mathcal{R}_i^\dagger(P,k)\mathcal{G}_i^3(P,k)\rho \\ &+ \frac{1}{4}\left(\mathcal{H}_i^1(J,k)\right)^T\mathcal{R}_i^\dagger(P,k)\mathcal{H}_i^1(J,k). \end{aligned} \tag{7.32}$$

Define

$$\begin{aligned} \mathcal{P}_i(P,k) &= \mathcal{A}_i(P,k) - \left(\mathcal{G}_i^3(P,k)\right)^T\mathcal{R}_i^\dagger(P,k)\mathcal{G}_i^3(P,k), \\ \mathcal{D}_i(J,P,k) &= \mathcal{T}_i^T(J,k) - \left(\mathcal{H}_i^1(J,k)\right)^T\mathcal{R}_i^\dagger(P,k)\mathcal{G}_i^3(P,k), \\ \mathcal{K}_i(J,P,\gamma,k) &= \lambda^T\mathcal{Z}_i(P,k)\lambda + 2\rho^T\mathcal{G}_i^1(P,k)\lambda + 2\rho^T\mathcal{G}_i^2(P,k)\bar{\omega}(k) \\ &+ \mathcal{H}_i^2(J,k)\lambda + \mathcal{H}_i^3(J,k)\bar{\omega}(k) + \mathcal{E}_i(\gamma,k) \\ &- \frac{1}{4}\left(\mathcal{H}_i^1(J,k)\right)^T\mathcal{R}_i^\dagger(P,k)\mathcal{H}_i^1(J,k) + \sigma(k)\cdot tr\left(C_i^TC_i\mathcal{E}_i(P,k)\right). \end{aligned}$$

$$\begin{aligned} & \rho^TQ_i(k)\rho + L_i(k)\rho + u^TM_i(k)u + H_i(k)u \\ + \;& E\left(\rho^T(k+1)P_{\theta(k+1)}\rho(k+1) + J_{\theta(k+1)}\rho(k+1) + \gamma_{\theta(k+1)}|\mathfrak{B}_k\right) \\ = \;& \rho^T\mathcal{P}_i(P,k)\rho + \left(u+\kappa\left(\rho(k)\right)\right)^T\mathcal{R}_i(P,k)\left(u+\kappa\left(\rho(k)\right)\right) \\ + \;& \mathcal{D}_i^T(J,P,k)\rho + \mathcal{K}_i(J,P,\gamma,k). \end{aligned} \tag{7.33}$$

Now we are ready to state the optimal control law for problem (7.7)-(7.8). The Bellman's optimality condition and the optimal control law of the problem can be written in terms of a set of coupled generalized algebraic Riccati difference equations in conjunction with a set of coupled linear recursive equations, which is similar to the LQG case.

Proposition 7.2. For each time index $N-1,\cdots,0$, an optimal control law for problem (7.7)-(7.8) can be obtained by

$$u(k) = -\mathcal{R}_{\theta(k)}^\dagger\left(P(k+1),k\right)\left(\mathcal{G}_{\theta(k)}^3\left(P(k+1),k\right)\rho(k) + \frac{1}{2}\mathcal{H}_{\theta(k)}^1\left(J(k+1),k\right)\right), \tag{7.34}$$

where for $\theta(k)=i$, we have

$$\mathcal{R}_i\left(P(k+1),k\right) \succeq 0, \tag{7.35}$$

$$\left(\mathcal{G}_i^3\left(P(k+1),k\right)\right)^T = \left(\mathcal{G}_i^3\left(P(k+1),k\right)\right)^T \mathcal{R}_i^\dagger\left(P(k+1),k\right)\mathcal{R}_i\left(P(k+1),k\right), \quad (7.36)$$

$$\left(\mathcal{H}_i^1\left(J(k+1),k\right)\right)^T = \left(\mathcal{H}_i^1\left(J(k+1),k\right)\right)^T \mathcal{R}_i^\dagger\left(P(k+1),k\right)\mathcal{R}_i\left(P(k+1),k\right), \quad (7.37)$$

and

$$\begin{aligned} &P(N+1)=0, J(N+1)=0, \mathcal{P}_{\theta(N)}\left(P(N+1),N\right)=Q_{\theta(N)}(N), \\ &\mathcal{T}_{\theta(N)}(J(N+1),N)=L_{\theta(N)}(N), \gamma(N)=0. \end{aligned}$$

For each time step k the payoff function is given by

$$\begin{aligned} V\left(\rho(k),\theta(k),k\right) &= E\left(\rho(k)^T\mathcal{P}_{\theta(k)}\left(P(k+1),k\right)\rho(k)+\mathcal{D}_{\theta(k)}^T\left(J(k+1),P(k+1),k\right)\rho(k)\right) \\ &\quad +E\left(\mathcal{K}_{\theta(k)}\left(J(k+1),P(k+1),\gamma(k+1),k\right)\right). \end{aligned} \quad (7.38)$$

∎

Proof. We will prove this proposition by induction on time step k based on the fact that the stochastic dynamic programming is established in a recursive (backward) manner. At the terminal time, i.e. $k=N$, there is no control input which implies that

$$V\left(\rho(N),\theta(N),N\right) = E\left(\rho^T(N)Q_{\theta(N)}(N)\rho(N)+L_{\theta(N)}(N)\rho(N)\right). \quad (7.39)$$

By definition $P(N+1)=0$, $J(N+1)=0$, and $\gamma(N+1)=0$, we have that

$$\mathcal{P}_{\theta(N)}\left(P(N+1),N\right)=Q_{\theta(N)}(N),\ \mathcal{T}_{\theta(N)}(J(N+1),N)=L_{\theta(N)}(N),$$

and

$$\gamma(N)=E\left(\mathcal{K}_{\theta(N)}\left(J(N+1),P(N+1),\gamma(N+1),k\right)\right)=0,$$

which is consistent with (7.38). Suppose from the induction hypothesis that (7.34)-(7.38) hold for time step $k+1$. By the recursive equation for stochastic dynamic programming and Proposition 7.1, we have that for $\rho(k)=\rho, \theta(k)=i$,

$$\begin{aligned} V\left(\rho,i,k\right) &= \min_u\left(\rho^TQ_i(k)\rho+L_i(k)\rho+u^TM_i(k)u+H_i(k)u\right. \\ &\quad \left.+E\left(V\left(\rho(k+1),\theta(k+1),k+1|\mathfrak{B}_k\right)\right)\right), \\ &= \min_u\left(\rho^TQ_i(k)\rho+L_i(k)\rho+u^TM_i(k)u+H_i(k)u\right. \\ &\quad \left.+E\left(\rho^T(k+1)P_{\theta(k+1)}\rho(k+1)+J_{\theta(k+1)}\rho(k+1)+\gamma_{\theta(k+1)}|\mathfrak{B}_k\right)\right). \end{aligned} \quad (7.40)$$

$V(\rho, i, k)$ can be evaluated according to Proposition 7.1,

$$\begin{aligned} V(\rho,i,k) &= \min_{u} \left(\rho^T \mathcal{A}_i\left(P(k+1),k\right)\rho + u^T \mathcal{R}_i\left(P(k+1),k\right)u \right. \\ &+ \lambda^T \mathcal{Z}_i\left(P(k+1),k\right)\lambda + \sigma(k)\cdot tr\left(C_i^T C_i \mathcal{E}_i\left(P(k+1),k\right)\right) \\ &+ 2\rho^T \mathcal{G}_i^1\left(P(k+1),k\right)\lambda + 2\rho^T \mathcal{G}_i^2\left(P(k+1),k\right)\bar{\omega}(k) \\ &+ 2\rho^T \left(\mathcal{G}_i^3\left(P(k+1),k\right)\right)^T u + \mathcal{T}_i^T\left(J(k+1),k\right)\rho + \left(\mathcal{H}_i^1\left(J(k+1),k\right)\right)^T u \\ &+ \mathcal{H}_i^2\left(P(k+1),k\right)\lambda + \mathcal{H}_i^3\left(P(k+1),k\right)\bar{\omega}(k) + \left.\mathcal{E}_i\left(\gamma(k+1),k\right)\right) \\ &= \min_{u} \Big(\rho^T \mathcal{P}_i(P(k+1),k)\rho + \left(u+\kappa\left(\rho(k)\right)\right)^T \mathcal{R}_i(P(k+1),k)\left(u+\kappa\left(\rho(k)\right)\right) \\ &+ \mathcal{D}_i^T(J(k+1),P(k+1),k)\rho + \mathcal{K}_i(J(k+1),P(k+1),\gamma(k+1),k)\Big). \quad (7.41) \end{aligned}$$

(7.41) achieves its minimum when $u = -\kappa\left(\rho(k)\right)$ with the minimum of the payoff function given by (7.38), which completes the proof of the proposition. □

Remark 7.2. We give an intuitive interpretation to the generalized recursive Riccati equations (7.36)-(7.37) in this remark. To begin with, we first introduce the following lemma:

Lemma 7.2. (Ait Rami et al., 2002) Let be given matrices $G = G^T$ and H with appropriate sizes. Then the following conditions are equivalent

(a). $H(I - GG^{\dagger}) = 0$

(b). $Ker(G) \subseteq Ker(H)$,

where $Ker(G)$ denotes the null space of matrix G[7]. ∎

By this lemma, we have that (7.36) is equivalent to

$$Ker\left(\mathcal{R}_i\left(P(k+1),k\right)\right) \subseteq ker\left(\left(\mathcal{G}_i^3\left(P(k+1),k\right)\right)^T\right)$$

and similarly, (7.37) is equivalent to $Ker\left(\mathcal{R}_i\left(P(k+1),k\right)\right) \subseteq ker\left(\left(\mathcal{H}_i^1\left(J(k+1),k\right)\right)^T\right)$. These conditions ensure the optimal control is well-posed for all time step k (and hence admits solution). A rigorous mathematical proof can be found in Ait Rami et al. (2002); Costa and Wanderlei (2007). As for another aspect, these two conditions implies that multiplying (7.29) with $\mathcal{R}_i^{\dagger}(P,k)\mathcal{R}_i(P,k)$ will not change the solution space of the original optimal control problem. ∎

Under some circumstance, the feasibility of linear matrix inequality (LMI) and the solvability of the Riccati equation are equivalent (Boyd et al., 1994; Ait Rami and Zhou,

[7] In linear algebra, the null (or kernel) space of a matrix G is the set of all vectors x for which $Gx = 0$, i.e. $Ker(G) = \{x : Gx = 0\}$.

2000) and the appendix of this chapter. LMI conditions can be also established as alternative sufficient conditions for the optimal control problem to solve the generalized recursive Riccati equations. Since the LMIs can be solved numerically very efficiently using the interior-point methods, the LMIs can be viewed as an efficient numerical algorithm to the symmetric solutions of the ARE. Examples can be found in Ait Rami and Zhou (2000); Ait Rami et al. (2002).

Remark 7.3. Note that, we do not require the weighting matrices $Q_{\theta(k)}(k)$ and $M_{\theta(k)}(k)$ to be positive (or semi-positive) definite, while the assumption that the control weighting matrix, i.e. $M_{\theta(k)}(k)$, must be positive definite and/or the state weighting matrix, i.e. $Q_{\theta(k)}(k)$, must at least be semi-positive definite (for almost all the time), inherited from the deterministic case, have been taken for granted in both control and traffic engineering literature. To be precise, if $M_{\theta(k)}(k)$ is negative (which means a benefit rather than a cost), then the optimal control $\mathbf{u}$ can be shown to be such that $\|u(k)\| = +\infty$, namely, "the larger the better". However, there are some cases that the control weight $M_{\theta(k)}(k)$ is negative definite and the "the-larger-the-better" policy no longer applies. Let us take a more concrete example to illustrate the above argument.

An example. Many highways in China are built and operated by private companies. A company may wish more people to use their highways to maximize their benefit which is proportional to $\mathbf{u}$, i.e. the price level and the traffic volume entering the network. However, the impacts are supervised and monitored by the government so that the price can not be set to too expensive, the pollution level and LOS cannot deviate too much from an allowable level. The objective of the company is on one hand to maximize the total expected return, and on the other hand to minimize the expected negative impacts. This multi-objective optimization problem may be converted into a single-objective problem by putting weights on the different objectives (i.e. a trade-off between these two objectives). Thus the following function is to be minimized:

$$\min J = E\sum_{k=0}^{N}\left(\lambda_1||\rho(k)-\hat{\rho}||^2 - \lambda_2\hbar||u(k)||^2\right),$$

where $\lambda_1, \lambda_2 \in (0,1), \lambda_1+\lambda_2 = 1$, represent the weights. Generally, as commented by Chen and Zhou (1998, 2000), such kind of phenomenon can happen in the following situation: "Suppose, in a deterministic (minimizing) optimization problem, that the cost decreases as the level of activity the decision maker carries out increases (a typical example of such situations is an investment that would be "guaranteed" to be profitable if the risk were to be excluded from consideration). Then it is not really an optimization problem because there is no trade-off in it, and the optimal decision is simply to take the maximum possible

activity level. So the problem is trivial or ill-posed. However, in a stochastic environment, suppose that the uncertainty increases with increasing magnitude of the activity level and that the uncertainty results in certain additional cost (called risk adjustment in the terminology of economics); then there is a trade-off between the activity level and the uncertainty, and the decision maker has to carefully balance the two to achieve an optimal solution." The indefinite terms can be compensated by taking advantage of the underlying uncertainties (e.g. risk adjustment (economics); Congestion derivatives). ∎

7.4 A robust consideration

7.4.1 An introduction to robust control

In single agent problems with incomplete information, optimal decision rules depend on a decision maker's posterior distribution over hidden state[8] variables, e.g. the distribution of possible traffic operational modes in our setting (which includes the dynamics of modes and their transition probabilities), and an objective function that summarizes the pertinent history of observations, e.g. the detected traffic flow information. In control theory and economics, a decision maker expresses faith in his model when he uses Bayes' rule to deduce the transition law, e.g. the generalized algebraic Reccati equations in the previous section. As shown in the derivation of the optimal control law, the optimal decision actually depends on the transition probabilities and the calibration of the model. When a decision maker doubts his/her model and wants a decision rule that is robust to the mis-specifications of the model, how should he/she proceed? Robust control provides us a way to address this kind of problems. Robust control considers the design of decision or control rules that explicitly acknowledges decision makers' fear of model mis-specification. Robust control is inherently about model uncertainty, to be more specific, focusing on the implications of model uncertainty for decisions. It has been verified that robust control can be formulated as a two-person zero-sum game wherein a minimizing player helps a maximizing player to design decision rules that satisfy bounds on the value of an objective function over a set of stochastic models (Basar and Bernhard, 1995; Hansen and Sargent, 2008). Robust control has widely applied in control theory Basar and Bernhard (1995) and economics (Hansen and Sargent, 2001, 2008; Hansen et al., 2010).

Example 7.2. (Robust control as a two-person zero-sum game–a linear system example): The following two-person (quadratic) dynamic game is proposed to represent a preference

[8] Hidden states can include unknown parameters, dummy variables indexing different models, e.g. $\theta(k)$, hidden information variables, capital stocks, and effort levels (Hansen et al., 2010).

for robustness:

$$\begin{aligned} & -\frac{1}{2}x(k)^T Q x(k) \\ = \quad & \max_{u(k)} \min_{\omega_e(k)} -\frac{1}{2}z(k)^T z(k) + \frac{\alpha_n}{2}\omega_e(k)^T \omega_e(k) - \frac{\alpha_s}{2}x(k+1)^T Q x(k+1), \end{aligned} \tag{7.42}$$

subject to

$$x(k+1) = Ax(k) + Bu(k) + C\omega_e(k), \tag{7.43}$$

where ω_e has zero mean and covariance matrix I, $\alpha_n > 0$ is a parameter measuring a preference for robustness, and z is a controlled output to be specified. Compared with the optimal control law, wherein only u is the decision maker, we note that another person aiming to minimize the payoff function has entered the game. He/she achieves his/her objective by penalizing a term on the noise sequence $\omega_e(k)$ which is added to the payoff function as $\frac{\alpha_n}{2}\omega_e(k)^T\omega_e(k)$. Thus, the theory of dynamic games (e.g. the dynamic programming routine in the previous section) can be applied to study robust decision-making (Basar and Bernhard, 1995). As explained by Hansen and Sargent (2008), the fictitious person, i.e. ω_e, puts context specific pessimism into the control law. The robustness parameter or multiplier α_n restrains the magnitude of the pessimistic distortion. Large values of α_n keep the degree of pessimism (the magnitude of ω_e) small. By making α_n arbitrarily large, we approximate the certainty-equivalent solution to the single-agent decision problem, i.e. the original optimal control problem. ∎

7.4.2 A robust control formulation

From previous sections, we can see that the uncertainties for decision making mainly come from calibration and the definition of transition probability matrix rather than the parameters of the system matrices (as the system matrices depend on the sample time T_s and cell length l_i only, which are certain). Recall the specification of noise sequences in the SCTM that an uncertain supply function is calibrated as a nominal function plus a Gaussian white noise term, e.g. $\omega_s(k) = \omega_{s,0}(k) + \omega_s^n(k)$, where $\omega_{s,0}(k)$ is the nominal function (which is usually regarded as the mean) of $\omega_s(k)$, and $\omega_s^n(k)$ is a Gaussian white noise term from the calibration. To this end, we expand $A_{\theta(k),s}\omega_s(k)$ as

$$A_{\theta(k),s}\omega_s(k) = A_{\theta(k),s}\left(\omega_{s,0}(k) + \omega_s^n(k)\right) \triangleq A_s\left(\theta(k), k\right)\omega_s^n(k). \tag{7.44}$$

It is easy for us to mis-specify the nominal supply function, e.g. the calibration of freeway capacity in Figure 5.1, due to the fact that it is generally very difficult to calibrate a nominal freeway capacity in practice. In this example, what is the nominal freeway capacity,

7500 veh/h or 8000 veh/h? It is hope that our decision can act robust to this kind of mis-specifications, e.g. the designed control law is robust to the nominal freeway capacity within the interval $[7000, 8500]$ veh/h. Therefore, we decompose $A_s(\theta(k), k)$ into

$$A_s(\theta(k), k) = A_s(\theta(k)) + \Delta A_s(\theta(k)).$$

Similar reasoning is applied to the pure noise terms to obtain $B_s(\theta(k), s)$. We rewrite the SCTM as

$$\begin{aligned}\rho(k+1) &= \left(A_{\theta(k)} + \sum_{s=1}^{v} A_s(\theta(k), k)\,\omega_s^n(k)\right)\rho(k) + \left(\sum_{s=1}^{v} B_s(\theta(k), s)\,\omega_s^n(k)\right)\lambda^n(k) \\ &\quad + C_{\theta(k)}\omega(k) + B_{\theta(k)}u(k),\ \rho(0) = \rho_0,\ \theta(0) = \theta_0, \end{aligned} \tag{7.45}$$

where $\omega_s^n(k)$ and $\lambda^n(k)$ are now Gaussian white noise sequences. A controlled output is defined to be a linear function of the state and control as

$$z(k) = S_{\theta(k)}\rho(k) + R_{\theta(k)}u(k). \tag{7.46}$$

In this section, we define the following objective function for the robust control problem as

$$V_0(\rho_0, \theta_0, u_0, \omega) = \max_u \min_\omega \sum_{k=0}^{N} E\left(\alpha^2\|\omega(k)\|^2 - \|z(k)\|^2\right), \tag{7.47}$$

which is equivalent to

$$V_r(\rho_0, \theta_0, u_0, \omega) = \min_u \max_\omega \sum_{k=0}^{N} E\left(\|z(k)\|^2 - \alpha^2\|\omega(k)\|^2\right). \tag{7.48}$$

Note that the uncertain supply functions $\omega_s^n(k)$, $\lambda^n(k)$ are not explicitly included in the objective function while the uncertain demand $\omega(k)$ is explicitly considered. The objective function (7.48) is chosen in this form for the following two reasons:

1. The uncertain supply functions $\omega_s^n(k)$, $\lambda^n(k)$ are regarded as model mis-specifications. The design purpose to these terms is to enable the control law to act robust to these uncertainties. This is implicitly included in the term involving the output function, i.e. $\|z(k)\|^2$.

2. As explained, the uncertain demand $\omega(k)$ is always taken as disturbance in traffic control literature. The design purpose to this term is to attenuate the effect of this disturbance to the output.

Compared with (7.47), (7.48) is more consistent with the context of Section 7.3 wherein the control aims to minimize the cost function. If the signal ω maximize $V_r(0, \theta_0, u_0, \omega)$,

we call it the worst-case disturbance. From the engineering view of point, the worst-case disturbance ω^* achieves the maximum possible energy gain from the disturbance ω to the output z. This worst-case paradigm underlies a number of robust control design methodologies. In terms of a two-person dynamic zero-sum game, the uncertainty is regarded as the "maximizing" player who attempts to impair the performance of the closed-loop system. In contrast, the designer is trying his/her best to achieve a best guaranteed level of the closed-loop performance in the face of uncertainties. In terms of control methodology, this kind of robust control is also referred to as robust H_∞ control or minimax optimal control. The robust control problem (7.48)-(7.45) can be solved by the stochastic optimal control routine by converting the problem into an equivalent optimal control formulation and following the development in Section 7.3. One can refer to Yoon et al. (2005); Haddad and Chellaboina (2008); Hou et al. (2010) on how to convert the problem into an equivalent optimal control formulation. An LMI based solution can be obtained by extending the results by Xu and Chen (2005).

7.5 Multiple agent settings

There are several reasons for us to consider a traffic network as a multi-agent system. First, in traffic control engineering, freeway control schemes and dynamic urban traffic signal control strategies are developed independently. This may be due to the differences in the objectives and controls. Ramp metering and variable speed limit (VSL) are two major control strategies for freeway systems, while tuning the green times of traffic signals is the major control strategy for urban arterials. The objective function is always chosen as the total time spent (TTS) by all vehicles in the network (including the waiting time experienced in the ramp queues) (linear terms) plus the penalty terms to suppress high-frequency oscillations of the optimal control trajectories and to enforce the maximum ramp queue constraints (quadratic terms). For dynamic urban traffic signal control with store-and-forward modeling of traffic dynamics (e.g. TUC), the objective is always chosen as an energy function of the queue lengths and controls. To this end, we should model these two different components of traffic networks by different agents (with different objectives and control structures). Second, when the network size becomes large, it is unlikely for us to consider it as a single system to optimize its performance due to the sophisticated network dynamics and interactions between the actuators and sensors. Centralized control of such kind of systems from a single control agent is often not possible due to technical or commercial constraints. Communication delays, coding and transmission

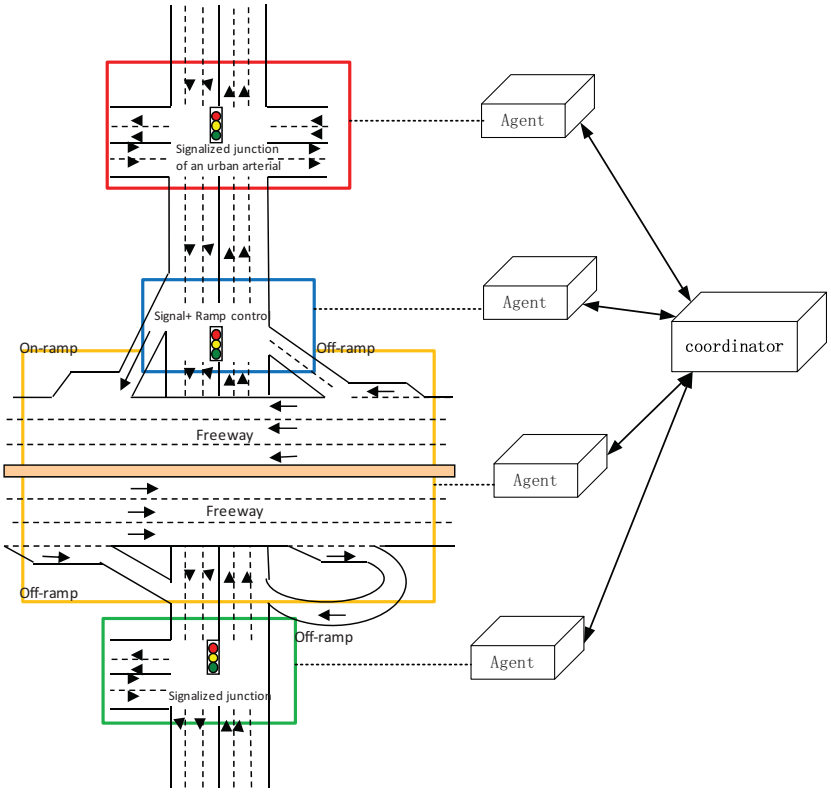


Figure 7.3: A network with a hierarchy of agents and their responsibilities

errors from the sensors (or detectors) to the control center, from the control center to actuators, extremely large amount of data, and expensively computational cost are major technical constraints. Some commercial issues may arise from privacy issues, information is not shared by different network operators, and budget constraints, etc. Moreover, the robustness and reliability may be fragile for single-agent (or centralized) control. Third, due to their geographically distributed and uncertainty in a dynamic environment, traffic networks are well suited for the concept of multiagent system (MAS), i.e. distributed structure and parallel computation (Schleiffer, 2002; Chen and Cheng, 2010). For these reasons, transportation networks typically have to be operated using a multi-agent (or distributed) system (MAS) approach.

Multi-agent techniques have been used in several stages of transportation systems which can be classified into three levels (Schleiffer, 2002): integration of traffic management systems (Adler et al., 2005; Chen et al., 2009), dynamic urban traffic signal control (see e.g. de Oliveira and Camponogara (2010), and the references therein), and traffic guidance (Adler et al., 2005). The papers on traffic management and guidance mainly concentrate on deterministic static environment while the papers on traffic signal control consider the traffic network operating under deterministic dynamic environment with simplified traffic

dynamics. These approaches cannot be applied to our situation. An example of a traffic network with mixed freeways and urban arterials represented by a multi-agent system is depicted in Figure 7.3. The traffic network (system) to be controlled is divided into subsystems, and that each subsystem has been assigned an agent. Each agent detects the current state of its subsystem, and receives information from other agents. Each of the agents determines its actions (local traffic management policy). On the above level, a coordination of these agents is preferred to optimize the performance of the network. How to coordinate and thus optimize the performance of a multi-agent system is a difficult question in general which has attracted attention from various research communities with a rapidly expanding literature. Research on this field has employed techniques from game theory (GT) (especially evolutionary GT), artificial intelligence (AI) and MAS by trying to synergizing them (see e.g. Busoniu et al. (2008); Shoham and Leyton-Brown (2009), and the references therein). We will investigate this problem in the future work.

7.6 Conclusion

We investigate traffic management schemes for freeway networks with demand and supply uncertainties modeled by the SCTM proposed in Chapters 5-6. As ramp metering control can only be applied to not too dense traffic conditions and may cause some side effects if the induced queues are not properly addressed, the traffic management schemes can be ramp metering control and/or (flow-dependent) dynamic road pricing. The road pricing aims at regulating travel demand. The optimal policy is established utilizing the stochastic dynamic programming.

To develop a stochastic optimal control framework, we reformulate the SCTM as a class of discrete time stochastic bilinear systems with Markov switching and further simplify the model. Based on this reformulation and simplification, we investigate the optimal decision making problem for traffic management of a freeway segment. A closed form of optimal policy is derived in terms of a set of coupled generalized recursive Riccati equations. As the optimal control may be fragile with respect to the model miss-specifications, we further pursued the robust (optimal) decision policy, which would act robust with respect to the parameter miss-specifications in the traffic flow model (which can be originated from the calibration process), and to attenuate the effect of disturbances in the freeway network (where demand uncertainty is usually taken as a kind of disturbance). The robust policy would further release the best performance of the traffic network under control. For the network traffic case, we propose a multiagent based approach to address the problem.

Another implication of the proposed methodology is to make benefit from the inherent uncertainties, which is achieved by extending the conventional LQ optimal control theory to consider the indefinite terms of the state and input weighting matrices. The indefinite terms can be compensated by taking advantage of the underlying uncertainties, e.g. risk adjustment (i.e. the decision maker has to balance the activity level against the uncertainty to achieve an optimal solution).

Appendix

Preliminaries

Definition 7.2. For a matrix A of appropriate dimension (no need to be invertible), the generalized inverse (or Moore-Penrose inverse) of A is defined to be the unique matrix $A^\dagger$ of appropriate dimension such that

1. $AA^\dagger A = A$
2. $A^\dagger AA^\dagger = A^\dagger$
3. $\left(AA^\dagger\right)^T = AA^\dagger$
4. $\left(A^\dagger A\right)^T = A^\dagger A$.

∎

Definition 7.3. (Linear Matrix Inequality) A linear matrix inequality (LMI) is an inequality

$$F(x) \triangleq F_0 + \sum_{i=1}^{m} x_i F_i \prec 0, \tag{7.49}$$

where $x_i \in R^m$ is the variable, $F_i,\ \forall i = 1, 2, \cdots, m$ are given symmetric matrices, and the inequality $\prec 0$ means "negative definite". ∎

Definition 7.4. (System of LMIs) A system of linear matrix inequalities is a finite set of linear matrix inequalities (LMIs)

$$F^{(1)}(x) \prec 0, \cdots, F^{(p)}(x) \prec 0. \tag{7.50}$$

∎

For a system of LMIs, we have the fact that the intersection of the feasible sets of each of the inequalities is convex. In other words, the set of all x that satisfy (7.50) is convex.

Thus the system of LMIs (7.50) can be expressed as a single LMI

$$\begin{pmatrix} F^{(1)}(x) & 0 & \cdots & 0 \\ 0 & F^{(1)}(x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & F^{(p)}(x) \end{pmatrix} \prec 0. \tag{7.51}$$

The above definitions can be also extended to the positive definite case $F(x) \succ 0$, semi-definite case (or non-strict LMI), e.g. $F(x) \preceq 0$ or $F(x) \succeq 0$. It turns out that any feasible non-strict LMI can be reduced to an equivalent LMI that is strictly feasible, by eliminating implicit equality constraints and then reducing the resulting LMI by removing any constant nulls pace.

Proposition 7.3. (Schur's Complement) Let $F(x)$ be an affine function which can be partitioned as

$$F(x) = \begin{pmatrix} F_{11}(x) & F_{12}(x) \\ F_{12}^T(x) & F_{22}(x) \end{pmatrix}. \tag{7.52}$$

The following affirmatives are equivalent:

Then $F(x) \preceq 0$.

$F_{11}(x) \preceq 0$, $F_{12}(x) = F_{11}(x)F_{11}^{\dagger}(x)F_{12}(x)$, and $F_{22}(x) - F_{12}^T(x)F_{11}^{\dagger}(x)F_{12}(x) \preceq 0$.

$F_{22}(x) \preceq 0$, $F_{12}(x) = F_{12}(x)F_{22}^{\dagger}(x)F_{22}(x)$, and $F_{11}(x) - F_{12}(x)F_{22}^{\dagger}(x)F_{12}^T(x) \preceq 0$.

∎

A very wide variety of problems arising in system and control theory, e.g. the stability in sense of Lyapunov, LQR, LQG, optimal and robust control problems, can be reduced to a few standard convex or quasi-convex optimization problems involving LMIs. The LMIs and the algebraic Riccati equation (ARE) has a closed relationship. In particular, the LMI can be viewed as an efficient numerical algorithm to the symmetric solutions of the ARE, since the LMIs can be solved numerically very efficiently using the interior-point methods. An example is shown in the following lemma.

Lemma 7.3. (Boyd et al., 1994) The symmetric solutions of the following ARE

$$S(k) = A^T P + PA - \left(PB + C^T\right) R^{-1} \left(B^T P + C\right) + Q = 0, \tag{7.53}$$

can be obtained by solving the following LMI

$$\begin{pmatrix} A^T P + PA + Q & PB + C^T \\ B^T P + C & R \end{pmatrix} \succeq 0. \tag{7.54}$$

∎

A more detailed discussion on the materials presented in this section can be found in Boyd et al. (1994).

A brief introduction to stochastic dynamic programming

The equation of motion for a nonlinear control system with a disturbance acting with the dynamics is given by

$$\rho_{k+1} = f(\rho_k, u_k, w_k, k), \tag{7.55}$$

where ρ_k is the system state, w_k is the disturbance of the model u_k is a control input. It is commonly assumed that the probability density function of the initial state ρ_0 is given, w_k is independent of $\{\rho_0, \cdots, \rho_k, w_0, \cdots, w_{k-1}\}$. For such kind of optimal control action in presence of stochastic elements, an objective function associated with the control problem becomes a random variable. Conventionally, the quantity to be minimized is the expected value of the objective function. An appropriate problem formulation can be described as

$$\min_{\mathbf{u}} E\left(G(\rho_N) + \sum_{k=0}^{N-1} g(\rho_k, u_k, k)\right), \tag{7.56}$$

subject to

$$\rho_{k+1} = f(\rho_k, u_k, w_k, k), \rho(0) = \rho_0, u_k \in A(\rho_k, k), \tag{7.57}$$

where the function $g(\cdot)$ is called the running cost function and $G(\cdot)$ is the terminal cost function, $\mathbf{u} = \{u_0, u_1, \cdots, u_{N-1}\}$, and $A(\cdot)$ represents a set of admissible controls depending on the system state and time. To solve the above stochastic optimal control problem by stochastic dynamic programming, we further define a value function known as cost-to-go, payoff function, or optimal value function by the notation $V(\rho_k, k)$, which is defined by

$$V(\rho_k, k) = \min_{u_k, \cdots, u_{N-1}} E\left(G(\rho_N) + \sum_{j=k}^{N-1} g(\rho_j, u_j, j)\right). \tag{7.58}$$

With this terminology, the principle of optimality can be stated as

Lemma 7.4. Principle of Optimality (Bertsekas, 2007) Let $\mathbf{u}^* = \{u_0^*, u_1^*, \cdots, u_{N-1}^*\}$ be an optimal control policy for problem (7.56)-(7.57). Assume that when using the optimal control policy $\mathbf{u}^*$, a given state $\rho(k)$ occurs with positive probability. Consider the subproblem where the environment is in state ρ_k at time k, and suppose we wish to minimize the cost-to-go function (7.58) for $k = 0, 1, \cdots, N-1$. Then the truncated control policy $\{u_k^*, \cdots, u_{N-1}^*\}$ is optimal for the subproblem. ∎

When we derive the stochastic dynamic programming optimality condition in terms of Bellman's backward recursion rule, we need to interchange the expectation and the minimization operations. The foundation for this exchange is stated in the following lemma.

Lemma 7.5. (Fundamental Lemma for Stochastic Control) (Speyer and Chung, 2008) Suppose that the minimum to $\min_{u\in\mathcal{U}} g(\rho, u)$ exists and $\mathcal{U}$ is a class of functions for which $E\left(g(\rho, u)\right)$ exists. Then,

$$\min_{u\in\mathcal{U}} E\left(g\left(\rho, u(\rho)\right)\right) = E\left(\min_{u\in\mathcal{U}} g\left(\rho, u(\rho)\right)\right). \tag{7.59}$$

∎

To release the iterative representation for dynamic programming, we expand the payoff function as

$$V(\rho_k, k) = \min_{u_k,\cdots,u_{N-1}} E\left(g(\rho_k, u_k, k) + \left(G(\rho_N) + \sum_{j=k+1}^{N-1} g(\rho_j, u_j, j)\right)\right). \tag{7.60}$$

By causality, the control and state at time k do not depend on the control and disturbance at future time indices, i.e. from $k+1, \cdots, N-1$. Therefore, the payoff function can be further written as

$$V(\rho_k, k) = \min_{u_k} E\left(g(\rho_k, u_k, k) + \min_{u_{k+1},\cdots,u_{N-1}} \left(G(\rho_N) + \sum_{j=k+1}^{N-1} g(\rho_j, u_j, j)\right)\right). \tag{7.61}$$

Let $V^*(\rho_k^*, k)$ denotes the optimality of the payoff function, and note that

$$V^*(\rho_{k+1}^*, k+1) = \min_{u_{k+1},\cdots,u_{N-1}} \left(G(\rho_N) + \sum_{j=k+1}^{N-1} g(\rho_j, u_j, j)\right).$$

The Bellman's optimality principle can be recognized as

$$V^*(\rho_k^*, k) = \min_{u_k} E\left(g(\rho_k, u_k, k) + V^*(\rho_{k+1}^*, k+1)\right). \tag{7.62}$$

Chapter 8

Summary of the thesis and future research topics

8.1 Summary of thesis

The main objectives of this dissertation were to develop three key components of ITS, i.e. dynamic traffic assignment with traffic volume (queue) control, real-time traffic surveillance, and decision making for traffic management under demand and supply uncertainties.

The traffic volume (or queue) control scheme is widely used in traffic control practice and has been proven to be effective to prevent the traffic network from spillback effect of congestion or gridlock. Theoretically, in the context of static traffic assignment, side constraints (or capacity constraints) are imposed to model the effects of traffic volume control schemes and link flow capacity constraints. Practically, the optimal state(s) of a traffic network operating under queue control strategies would provide us some guidelines on queue control design, which in turn yields better implementation of these queue control strategies. The traffic volume control can also be used to prevent the traffic network from spillback effect of congestion or gridlock. Controlling the traffic volumes on critical infrastructures is also an easy and efficient approach to increase the safety of such facilities. However, dynamic traffic assignment (DTA) considering the effects of traffic volume control schemes has been missing from literature. Toward these ends, we extended the dynamic user equilibrium (DUE) and dynamic system optimal (DSO) concepts to investigate the optimal design of dynamic traffic volume control scheme.

The static user equilibrium (UE) traffic assignment with side constraints was extended to the dynamic case in Chapter 3 to investigate dynamic user equilibrium under traffic volume control scheme with both route and departure time choices. The side constraints

are related to the desired temporal traffic volumes on certain links, which can be set according to the congestion levels, safety or environmental requirements. The dynamic user equilibrium problem with side constraints and departure time choice was formulated as an infinite-dimensional variational inequality (VI) problem in line with the DUE problem formulated by Friesz et al. (1993). We proved the existence of the DUE with simultaneous route and departure time choices by extending the results of Zhu and Marcotte (2000) based on the VI formulation of the DUE problem. Based on this existence result of the DUE problem and the VI formulation of the DUE-SC, we proved the existence of equilibrium to the DUE-SC problem under certain assumptions. To analyze the necessary condition, we represented the DUE-SC as an optimal control problem for a class of dynamical systems with input and state constraints. The necessary condition of the DUE-SC was shown to satisfy a generalized DUE condition wherein the equilibrium dynamic travel cost of a given OD pair equals to the effective path delay plus a term of additional travel cost induced by the side constraints. The additional travel cost term is governed by the accumulation of the Lagrange multipliers associated with the side constraints over time (unlike the static case). This additional travel cost term represents the control to be imposed to achieve the link traffic volume restrictions. The similarity between the additional delay terms from the static and dynamic cases was highlighted in the chapter.

As a Nash non-cooperative differential game, dynamic user equilibrium is used to represent the distribution of traffic that arises when travelers do not have knowledge about other travelers' strategies and compete with each other to minimize their own travel cost. In some situations, such as traffic diversion under incidents (wherein queue control is always necessary to prevent the spillback effect), it is necessary for the system manager to look for the best usage of the network under queue control. The dynamic system optimal assignment, as a Monopoly game in which the system manager can control all decision variables and thus achieves the system optimum, may be not a realistic representation of traffic. However, it provides a bound on the best performance of a traffic network, which makes it as a benchmark for evaluating various transport policy measures, e.g. time-dependent pricing (Yang and Meng, 1998; Chow, 2009a), network access control (Smith and Ghali, 1990; Lovell and Daganzo, 2000; Shen and Zhang, 2009; Zhang and Shen, 2010), and road capacity allocation (Ghali and Smith, 1995). Therefore, we further explored the dynamic system optimal traffic assignment with access control (DSO-AC) problem for general networks modeled by the two vertical queue models, i.e. the whole link model (WLM) and the deterministic queueing model (DQM) in Chapter 4. We derived dynamic marginal costs for paths and users' external costs for controlled links under the DSO-AC.

Road pricing, as a traffic management and bottleneck capacity allocation scheme, is frequently applied to regulate the traffic flows on urban transportation networks. In the static road pricing framework, it has been shown to be an efficient approach to traffic *demand* management and control (see e.g. Lindsey (2006); Tsekeris and Voß (2009) and the references therein). However, only few studies have been conducted on its dynamic counterpart. Chapter 4 provided a sound theoretical discussion to fill this gap. As a strategy to implement the access control, the dynamic access pricing problem was formulated as a DSO-AC problem, wherein the access constraints represent the restrictions on the traffic volumes and/or environmental constraints (e.g. vehicle emission). For the WLM case, we derived and discussed the necessary condition for operating the transportation system with capacity/environmental constraints optimally, i.e. the total system travel cost is minimized. For the DQM case, we utilized the formulation adopted by Akamatsu (2007) wherein the inflow to a bottleneck is saturated such that no queue would be formed. The access price is then given by the penalty associated with this constraint. Similar to the telecommunication network bandwidth allocation scheme, the DSO-AC analysis reveals the variety of economic effect of a certain amount of road capacity with respect to its spatial and temporal allocation, e.g. decide which links can be used and how to use their available capacities as "holding" capacities for queues. By comparing the dynamic externalities for the two vertical queue models, we showed that their different externality structures result in different tolling structures to achieve DSO. Based on this access pricing analysis and an "equivalent" environmental capacity that converts the environmental constraint into traffic volume restriction, we studied the traffic induced air pollution pricing scheme for networks modeled by the two point queue models. It was found that the traffic capacity based access price and traffic induced air pollution price would not become effective simultaneously for the DQM case. The boundary to determine the dominant price is defined by the minimum of the two capacities. However, for the WLM case, there is a circumstance that both prices would be effective simultaneously.

The analysis on dynamic traffic assignment with traffic volume control problems contributes to DTA literature also in the following prospects:

1. By restricting the link traffic volumes to be equal to or less than the link storage capacities on the network such that no spillback would happen, we can capture the spillback effect of traffic congestion while avoiding the drawbacks, e.g. discontinuous and non-differentiable traffic dynamics and travel cost functions, of physical queue models, see e.g. Szeto and Lo (2006); Lo and Szeto (2002); Gentile et al. (2005); Nie and Zhang (2010), and the restricted piecewise linear link exit-flow function

approach, see e.g. Nie (in press), which enables us to solve the DTA problems analytically. This was motivated by some recent comparison studies which have revealed that the macroscopic link models (including vertical queue models, e.g. WLM and bottleneck model; and physical queue models, e.g. CTM) would produce almost the same traffic assignment result unless there is a spillback (shockwave) Mun (2007); Nie and Zhang (2005, 2010). The spillback effect is captured in terms of the effect needed to prevent the network from spillback. Similar to the bottleneck case, we can regard this effect as additional travel cost imposed on travelers for using the controlled links.

2. The separable link travel time function adopted in DTA literature, i.e. $D_a(x_a(t))$, would be more meaningful under this circumstance since if there is spillback the link travel time function would not be separable, e.g. depends on the downstream traffic conditions like the case of physical queue model (Szeto and Lo, 2006).

Traffic networks are exposed to extensive demand and supply uncertainties, especially under incident and adverse weather circumstances. For instance, the rainfall on 17 September 2010 in Beijing caused a network-wide traffic congestion. Some other important scenarios would also introduce inherent uncertainties. For instance, as explained in Chapter 2, traffic information is provided to drivers by the ATMIS through various information media to support traveler's decision making which in turn influences their travel choices and consequently reduce the (total/individual) travel time and improve efficiency of the traffic network. However, divers behave in a very different way in face of the information provided to them due to their different degrees of risk aversion and perception errors on travel times, which in turn affects their routing decisions and the travel demand. Because of these unexpected events, traffic flows and travel times on the roadways are uncertain. Stochastic traffic flow models are urgently needed to estimate/predict the traffic flows and travels time under these uncertain demand and supply conditions. Therefore, in the surveillance part, we proposed the stochastic cell transmission model (SCTM) to capture traffic density and possible wavefronts on a freeway segment under supply and demand uncertainties in Chapter 5. In the SCTM, demand and supply functions are assumed to be perturbed by some wide sense stationary, second-order processes consisting of uncorrelated random vectors with known means and variances. The stochasticities of the sending and receiving functions in the SCTM are governed by the random parameters of the fundamental flow-density diagrams, including the capacities, backward wave speeds, and the free-flow speeds. Operational modes and the corresponding probabilities of occur-

rence have been introduced to replace the wavefront concept in the original CTM and to avoid the dis-continuity introduced by the "min" operator and the wavefront. The SCTM was formulated as a class of discrete time stochastic bilinear systems with finite mixture which can be recognized as a finite Markov mixture model. We derived the analytical approximations of the means and SDs of the traffic densities. Numerical examples and an empirical study were then carried out to illustrate the advantages of the SCTM over the Monte Carlo Simulation approach in terms of computation time and memory requirement. The propagation of uncertainties of traffic states over time and space was illustrated. The empirical study confirmed that the SCTM performs well for all traffic conditions ranging from light to very dense traffic conditions. This can be regarded as an advantage of the proposed model over the previous proposed macroscopic stochastic dynamic traffic models, (*e.g.* Boel and Mihaylova (2006); Kim and Zhang (2008)). The empirical study also revealed that the SCTM outperforms the MCTM, and the MCS of MCTM.

The SCTM was extended to model the stochastic traffic dynamics on a network with uncertain demand and supply in Chapter 6. We identified four basic functional blocks for modeling a traffic network, *i.e.* freeway corridor, on-/off- ramps (traffic merge/diverge), signalized junction. To increase the accuracy, the original SCTM for one freeway segment consisting of two cells was defined as one basic subsystem. A long freeway corridor was represented as a stretch system connected by several basic SCTM subsystems. Link-node formulation of CTM has been utilized to represent traffic merge, diverge, and signalized junctions. We considered the ramps as SCTM subsystems for ramps with heavy traffic. The flow propagation for uninterrupted facilities was defined as a finite mixture distribution of the four probabilistic events governed by the sending and receiving flows based on a link-node formulation. Similar logic was applied to model an isolated signalized junction. The junction was represented by several SCTM subsystems. Each of these subsystems consists of several phases according to the signal phase under certain assumptions. A traffic network was then modeled by all these basic functional blocks. The SCTM subsystems accept the means and variances of the stochastic travel demand and supply functions as exogenous inputs, which in turn produce cell traffic densities and outflow of the roadway segment in terms of mean and variance as well as the probabilities of occurrence of different operational modes. We demonstrated the proposed network SCTM as a stochastic dynamic traffic network model for traffic control and management in the numerical examples. As shown in the simulation, the uncertain traffic dynamics and probabilistic wave-fronts could be captured by the proposed model. In the second test, the proposed network SCTM was applied to estimate queues and delays at signalized intersections. Comparison with some

traditional delay and queue estimation formulas was conducted. The numerical results showed good consistency between the SCTM and these formulas. However, compared with these delay and queue formulas the SCTM approach has wider application opportunities, since it can describe the temporal evolution of the queue length distribution at signalized junctions with uncertain supply functions by assuming any arrival distribution, which can be non-stationary. Two kinds of dynamic travel time distributions were found, i.e. skew normal distribution and bimodal like distribution. The summary statistics such as the mean and variance can be deceptive when applied to the second kind of distribution.

For the surveillance purpose, the SCTM can be utilized to provide a short-term prediction using the historical and on-line data of travel demand and traffic state. The prediction (in terms of travel time and traffic state) under the SCTM considers both demand and supply uncertainties in the future time-step. This allows traffic operators to monitor and devise robust control strategies for freeways. For the dynamic traffic assignment and control, the key operational benefit of the SCTM for traffic assignment purpose is the potential continuity of the delay operator which is not the case for the deterministic CTM (due to the potential blocking back condition of an arterial). This is due to the introduction of the stochastic delay in the SCTM which can also be considered as a better paradigm for a long-term traffic prediction.

In Chapter 7, we investigated traffic management schemes for freeway networks influenced by demand and supply uncertainties. The traffic management schemes could be ramp metering control and/or (flow-dependent) dynamic road pricing. The problem was established utilizing the stochastic dynamic programming. The stochastic traffic flow under demand and supply uncertainties is described by the proposed SCTM. To be more specific, we represented the SCTM as a class of discrete time stochastic bilinear systems with Markov switching. Based on this reformulation, we investigated the optimal decision making for traffic management of a freeway segment. A closed form of optimal policy was derived in terms of a set of coupled generalized recursive Riccati equations. As the optimal control might be fragile with respect to the model miss-specifications, we further pursued the robust (optimal) decision policy, which would act robust with respect to the parameter miss-specifications in the traffic flow model (which can be originated from the calibration process and some other exogenous conditions), and to attenuate the effect of disturbances in the freeway network (where demand uncertainty is usually taken as a kind of disturbance). The robust policy would be useful under incident conditions in the following sense: As mentioned in Wang et al. (2008), Wang et al. (2009a), and Wang and Papageorgiou (2009b), in case of incidents, the traffic flow characteristics along the concerned freeway

stretch may change substantially, this may also be reflected in correspondingly drastic changes of some model parameter values. If these changed model parameter values still falls in the admissible uncertainty intervals, the robust policy can be applied even under the presence of incidents. The robust policy would further release the best performance of the traffic network under control. For the network traffic case, a multiagent system (MAS) based approach was proposed to address the problem. Another implication of the proposed methodology is to make benefit from the inherent uncertainties. This is due to the essential difference between the optimal control formulation applied in Chapter 7 of this dissertation and the one utilized in the conventional LQ optimal control theory, i.e. the indefinite terms of the state and input weighting matrices. The indefinite terms can be compensated by taking advantage of the underlying uncertainties, e.g. risk adjustment (i.e. the decision maker has to balance the activity level against the uncertainty to achieve an optimal solution).

8.2 Future works

The future works will concentrate on the following aspects with special attention be paid to incident management applications.

8.2.1 On the DTA aspect

With respect to the DTA aspect, the following problems should be addressed in the future works.

1. To apply the proposed DUE-SC scheme, we will develop a more efficient numerical algorithm to solve it. However, developing such an algorithm for the DUE-SC or even for the standard DUE is still an open problem. A scheme to use a system of ordinary differential equations (ODEs) to approximate the differential algebraic equation (DAE) system for network loading based on the whole link model is proposed by Friesz et al. (2011). It is verified that this approximation scheme increase the computational efficiency of the traditional numerical methods for solving DUE problems. Our next step is to develop an efficient solution algorithm for DUE-SC but with a trade-off between the tractability and theoretical property of the algorithm, e.g. extending the heuristic solution algorithms of DUE to solve the DUE-SC (Tong and Wong, 2009). Possibly, by applying the approximate network loading developed by Friesz et al. (2011) and the cell-based discretized model proposed by Nie and Zhang (2010).

2. By Proposition 4.2, the dynamic traffic assignment with access constraints can be applied as an alterative way to obtain the optimal toll (or permit price) for networks modeled by deterministic queuing models with time-varying bottleneck capacities. The optimal toll (or permit price) can be obtained by solving the penalty associated with the access constraint. The extension to the heterogeneous user case and elastic demand is also worth while to look into. The network auction approach to access the capacity allocation would be also interesting. In the future work, we will formulate this problem in detail, and discuss the existence, uniqueness, convergence, and stability issues for such kind of problem.

3. Revisit the stability and uniqueness analysis of the equilibriums of DTA in sense of Lyapunov theory, passivity/dissipativity, and input-to-output and/or input-to-state stability (ISS): Uniqueness and stability of the equilibriums of DTA are essential for the applications of DTA. However, the works on the stability and uniqueness analysis of the equilibriums of DTA by different researchers tell very different stories. For example, the works by Mounce and Smith, see Mounce (2006) and Mounce and Smith (2007), proposed that the uniqueness and stability of DUE can be guaranteed only for network with single bottleneck per route case under various assumptions, while the works by Peeta and Yang (Peeta and Yang, 2003) and Iryo (Iryo, 2008) proposed that the DUE and DSO are unique and stable for general network under certain assumptions. The stability analysis in Peeta and Yang (2003) was for a time dependent network rather than for a dynamic traffic network since the requirements for a DTA model listed in Chapter 2 were barely fulfilled. However, it worth while to mention that the Lyapunov functions $V(x)$ proposed in Peeta and Yang (2003), where x represents the vector of cumulative link traffic volumes, for the SO and UE objectives are their corresponding objective functions in DTA problems. This overcomes the key difficulty of constructing a physically meaningful Lyapunov function for traffic systems. While the Lyapunov function lacks intuitive physical meaning in Smith (1982), Mounce (2006), Mounce and Smith (2007), and Iryo (2008). Also, it is wise for us to choose the vector of cumulative link volumes as state vector rather than reconstructing a dynamic system with route-flow vector as state vector.

4. Introducing control and/or closed-loop structure to DTA problems: The analysis in Chapter 7, to be more specific the dynamic programming approach, provides a potential way to introduce feedback (closed-loop) control (e.g. road pricing) structure to the dynamic traffic assignment problems. Other approaches are also possible, e.g.

nonlinear H_∞ feedback control theoretic approach is applied to real-time user equilibrium and system optimal dynamic traffic routing problems (Kachroo and Özbay, 2005, 2006). Note that the nonlinear H_∞ feedback control theoretic approach is also applied in Chapter 7 to design the robust traffic management schemes. The closed-loop feedback structure of DTA problems may be enabled by combining these two approaches.

5. Travelers behavior: During an incident with lane blockage, congestion forms when the time-varying travel demand exceeds the reduced roadway capacity. In the meanwhile, the growing incident induced lane changes and queue spillbacks significant interrupt the traffic flows among the adjacent lanes and exacerbate the incident induced congestion. Usually, traffic information is provided to drivers by Advanced Traveler Information Systems (ATIS) through various information media. The information are broadcasted in order to support traveler's decision making which in turn influence their travel choices and consequently reduce the (total/individual) travel time and improve efficiency of the traffic network. However, divers behave in a very different way in face of the information provided to them due to their different degrees of risk aversion and perception errors on travel times, which in turn affects their routing decisions and the travel demand. Drivers' decision on route choice is a major determinant of network performance. Traffic management strategies developed without considering the driver behavior (or being very behaviorally restrictive) can result in misleading control strategies, and thus potentially deteriorate network performance. Therefore, in the development of traffic management schemes, the driver behavior should be explicitly considered (or reasonable driver behavior should be assumed).

8.2.2 On the traffic surveillance and control aspects

Note that there are several assumptions which were made to simplify the construction and analysis of the SCTM. Several key future research issues are envisaged including:

1. investigation of theoretical relationship between the SCTM and the LWR model with stochastic components;

2. the study of the existence and property of the dynamic user equilibrium (DUE) solution based on the SCTM framework should also be carried out;

3. Enable the SCTM to estimate/predict stochastic travel time distribution. Travel time is one important element in dynamic traffic assignment. It is also one important

factor for the travelers to make their route choices. This function would provide a model for stochastic dynamic traffic assignment. In the traffic control aspect, travel time, such as Total Time Spent (TTS), is also an important performance index for optimal design. This work would contribute to the real time traffic control design under uncertainties and disturbances as well.

4. Enable the surveillance tool the function of incident alarms: As explained in Chapter 2 and Özbay and Kachroo (1999), timely incident detection and incident alarm is essential to prevent secondary incidents. As previously mentioned, in case of incidents, the traffic flow characteristics along the related freeway stretch may change substantially. By the online estimating/predicting the supply functions, such abrupt changes may be identified in real time, and hence the incident occurrence may be recognized, leading to corresponding incident alarms (Wang et al., 2008, 2009a; Wang and Papageorgiou, 2009b). By identifying the abrupt changes of supply functions may be not sufficient for some scenarios especially for the urban arterials. Another potential approach would be inspired by the incident detection algorithms which use different combinations of the traffic measurements: measured traffic speed, measured traffic volume (or density), and measured occupancy (Adeli and Jiang, 2009) and the travel time (Lam et al., 2008). Occupancy can be always obtained from the detectors, which can be applied directly. Other indexes can be estimated/predicted by different components of the SCTM. For instance, traffic speed as one of the supply functions can be estimated/predicted by adding a best linear predictor to the SCTM. Traffic density (or volume) and travel time can be estimated/predicted by the SCTM. Criteria combining these traffic states for incident detection algorithms will be defined in this research and incident alarms will be enabled to the SCTM framework.

5. Enable the function of traffic state prediction under abnormal scenarios such as incidents, adverse weather conditions: Due to high traffic density and congestion (especially under abnormal circumstances) in the network as well as the interaction of the demand and supply uncertainties along with the dynamic nature of traffic flow, the demand and supply uncertainties are correlated in both space and time domains. For example, the free-flow speeds are spatial correlated (cell-to-cell, lane-to-lane correlated); the demand profiles are temporal correlated. By considering these spatial and temporal correlations along with the traffic dynamics bring significant potential advantages for development of efficient traffic state estimation/prediction for the

SCTM paradigm.

Ramp metering is often insufficient for traffic management under incidents during the lane blockages and over-saturated traffic conditions. Implementing the incident pricing is also not practical. To this end, we need to extend our control strategies to consider possible scenarios.

1. Enable the traffic controller(s) to control the queue lengths: As explained, queue control is essential to curve the spillback effect or even the girdlock of congestion. It is important for us to explicitly incorporate the queue constraints when developing traffic management policies especially under incident scenarios with saturated traffic conditions.

2. Other traffic management strategies for incident management: Refer to a typical fundamental diagram of a freeway lane, the region can be covered by ramp metering is on the stable (left) side of the fundamental diagram, see Figure 8.1, and close to the top where a breakdown can happen Hegyi et al. (2005a) and Hegyi et al. (2005b). Ramp metering is only useful when traffic is not too light (otherwise ramp metering is not needed) and not too dense (otherwise traffic breakdown will happen anyway). In this sense, ramp metering control is only applicable to not severe incidents (including lane blocking incidents, lane closures for maintenance) under the assumption that the traffic is not too dense.

   However, in the case of saturated traffic conditions, what can we do in case of these not serve incidents? Obviously by ramp metering only is no enough for our purpose. Detouring (or equally rerouting) the traffic is one choice for us. But this approach may involve some issues, one key issue among them is the time-varying OD matrices. How can we obtain this time-varying OD matrices? How can we reroute the traffic back to its destination based paths? What is the criteria for this rerouting? As mentioned by Özbay and Kachroo (1999), only 1 percent to 1.5 percent of the recorded incidents are categorized as major (or severe) accidents and only few of those major accidents will cause major traffic disruptions both locally and regionally. Most of the time we do not need to do such complex rerouting (or detouring) works in case of incidents. Most of the time, proper control of the traffic flows is enough especially in the case that the traffic is not too dense, in which case the proposed robust ramp metering plus queue control is enough. To handle the high density traffic flow, a potential approach is to impose speed limit control. In Hegyi et al. (2005a,b), and Carlson et al. (2010) speed limits were imposed on the main-

stream traffic while the on-ramps were metered. The main idea of the limited speed in conjunction with ramp metering approach is that when ramp metering is unable to prevent congestion, the application of variable speed limits (VSL) upstream of a (temporary) bottleneck, that is close to become active, could prevent a breakdown by limiting the inflow into the area where the traffic breakdown starts (or decrease the mainstream flow arriving the bottleneck area, thus retarding the bottleneck activation and the resulting congestion). By referring to the fundamental diagram, see e.g. Figure 8.2, the effect of the speed limit is to change the shape of the fundamental diagram and reduce the outflow of the controlled segment. Suppose the traffic state on the freeway is A. When a speed limit is applied, the speed drops and the density increases, so the traffic state will be somewhere between B and C. However, because of the high traffic demand the traffic state will approach to state C, the capacity of the new fundamental diagram. The critical density (or occupancy) under VSL control is higher than the original uncontrolled one. Since this flow is lower than the capacity of the freeway without speed limit, there will be some space left to accommodate the traffic from the on-ramp and a breakdown is prevented. As stated in Hegyi et al. (2005a): "*This effect can be explained by the fact that the number of vehicles in the network is equal to the accumulated net inflow of the network (where the net inflow is the difference between the inflow and the outflow). However, the outflow is lower when there is congestion, so the queue grows faster, and consequently congestion will last longer, and the outflow will be low for a longer time (the time that the queue needs to dissolve). This is why one should try to prevent or postpone a breakdown as much as possible.*" Roughly speaking, we would benefit from VSL in terms of:

(a) increase of throughput, and

(b) retarding of congestion at overcritical occupancies.

Besides the reduction of mean speed at under-critical occupancies, another criticism of the VSL scheme could be that the approach keeps the controlled network congestion free, but at the cost of creating congestion at the entrances of the controlled network (or causing the side effects discussion in Chapter 7). Similar to ramp metering control, which induces delays to vehicles queuing on the ramp, the VSL introduces some delays to vehicles traveling on the mainstream. However, delays caused by bottleneck congestion may be much more than the vehicle delays induced

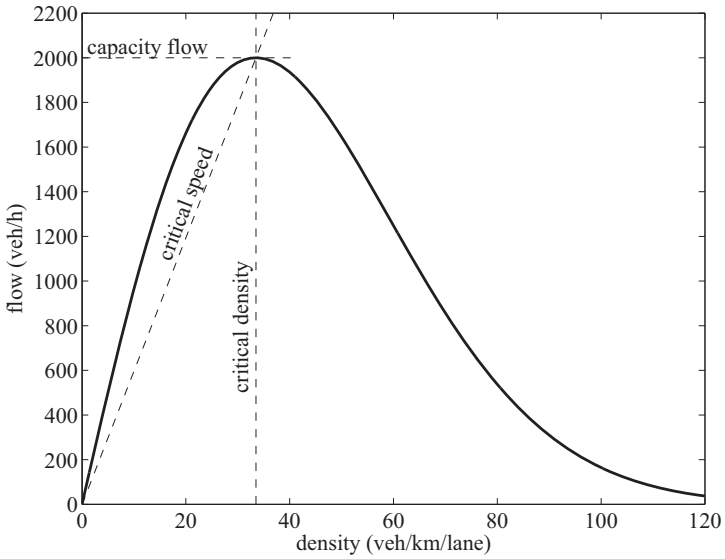


Figure 8.1: A typical fundamental diagram of a freeway lane (Source: Hegyi (2004))

by VSL Hegyi et al. (2005a); Carlson et al. (2010)[1]. To solve the side effect caused by the VSL, Hegyi et al. (2005a) proposed the following methodology: *"A remedy could be to extend size of the network with as many (uncontrolled) upstream sections as necessary to cover the congested area. In this way the congestion caused by the speed limits will not spill back to the mainstream origin queue and the congestion dynamics can be taken into account by the controller. Second, the network that is considered (i.e., evaluated and controlled) can be chosen larger, because the traffic is apparently so dense that the effects of the control reach beyond the bounds of the actual network."* As we can see that the side (spillback) effects caused by the VSL control were not properly addressed. The approaches recommended by Hegyi et al. (2005a) do not benefit the problem but somehow introduce the problem of the curses of dimensionality of the optimal control problem by expending a local problem to a regional or global problem. However, thanks to the queue control and access pricing schemes proposed in this study, we can properly address this problem as discussed in Chapter 7. To sum up, the control we proposed in this study, will be a scheme combines the proposed robust ramp metering with queue control and the speed limit control and access pricing to reduce travel demand if necessary.

3. To control the traffic flow in case of (severe) incidents: As mentioned previously, only tiny part of the recorded incidents are major accidents will cause major traffic disruptions both locally and regionally Özbay and Kachroo (1999). In the case of

[1]More detailed discussion on the ramp metering and VSL control can be found in Carlson et al. (2010).

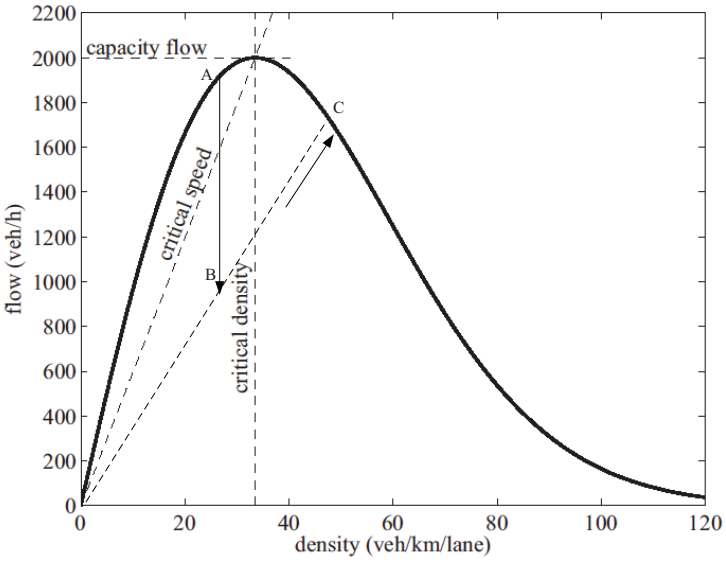


Figure 8.2: A typical fundamental diagram of a freeway lane with speed limit control

severe incidents, the freeway maybe completely blocked. The microscopic based stochastic optimal control approach developed by Sheu (2007), Sheu and Chang (2007) *etc.*, and the above proposed macroscopic based robust ramp metering and the speed limit control schemes which are developed for lane blocking incidents, may be no longer feasible. New traffic management schemes are needed. It is found that a large proportion (about 30 percent) of recorded incidents are secondary incidents caused by the primary incidents, this happens especially when the primary incident is a severe one and lasts long; at the mean while, the duration of primary incident would be longer if a secondary incident occurs (Khattak et al., 2009). Enabling the traffic incident alarm in time is essential to prevent a secondary incident and helps the local authority to response quickly to clear the incident (Özbay and Kachroo, 1999), which in turn reduce the severity of the incident. In this sense, timely incident alarm can be viewed as a kind of efficient control in the case of (severe) incident.

Besides to the incident alarm, and the above proposed robust ramp metering with queue control in conjunction with speed limit control, diversion of the traffic through adjunct and parallel arterials is essential in this situation. A heuristic optimal freeway traffic diversion control has been proposed in Liu et al. (2009) in the case of incidents. As to the traffic diversion control, we introduce similar idea to Liu et al. (2009) with extensions and combinations of the methodology we proposed to handle not severe traffic incidents.

(a) Traffic is diverted to the arterial through the off-ramp just upstream to the

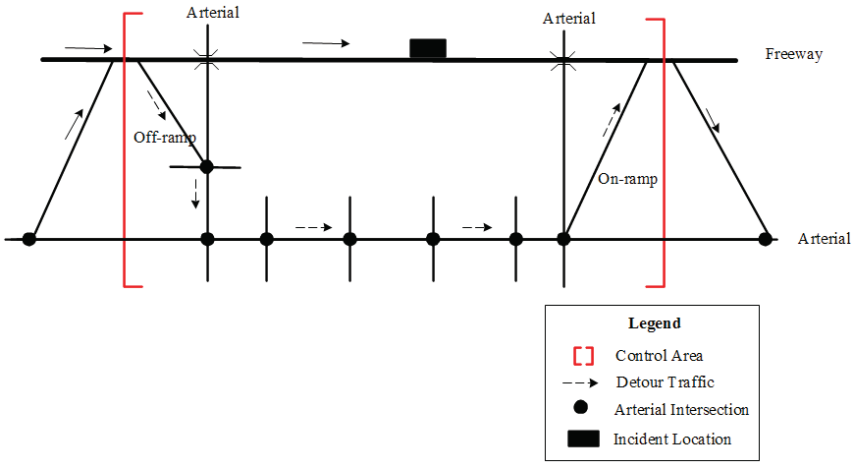


Figure 8.3: The traffic diversion control in case of incident (Source: Liu et al. (2009))

incident section, and guided back to the on-ramp right after the incident section. This is presumed to avoid the technical difficulties on OD matrices. A constant compliance rate for drivers is assumed for drivers in the control area. This assumption would be accomplished by the on-line calibration of behavior parameters for behavior-consistent route guidance proposed in Paz and Peeta (2009a). This assumption can be further relaxed by the framework proposed in Paz (2007) and Paz and Peeta (2009b). Methods proposed in Wang et al. (2003) and Karimi et al. (2004), which integrate the predicted travel times based dynamic route guidance and the advantages of feedback based freeway ramp metering approach (relatively simple, robust, fast) would be helpful to this study too.

(b) Normal traffic patterns, including off-ramp exit rates and arterial intersection turning proportions, are assumed to be stable and not impacted by the diverse traffic. This assumption is proposed to simplify the problem.

(c) The ramp metering rates (or traffic signals) are activated by the queue lengths on the corresponding signalized junctions. This is proposed to enable queue control. The robust ramp metering and speed limit control will be adopted.

More detailed discussion on the traffic diversion strategies can be found in Özbay and Kachroo (1999). To sum up, the overall picture of the incident detection and traffic control framework is depicted in Figure 8.4.

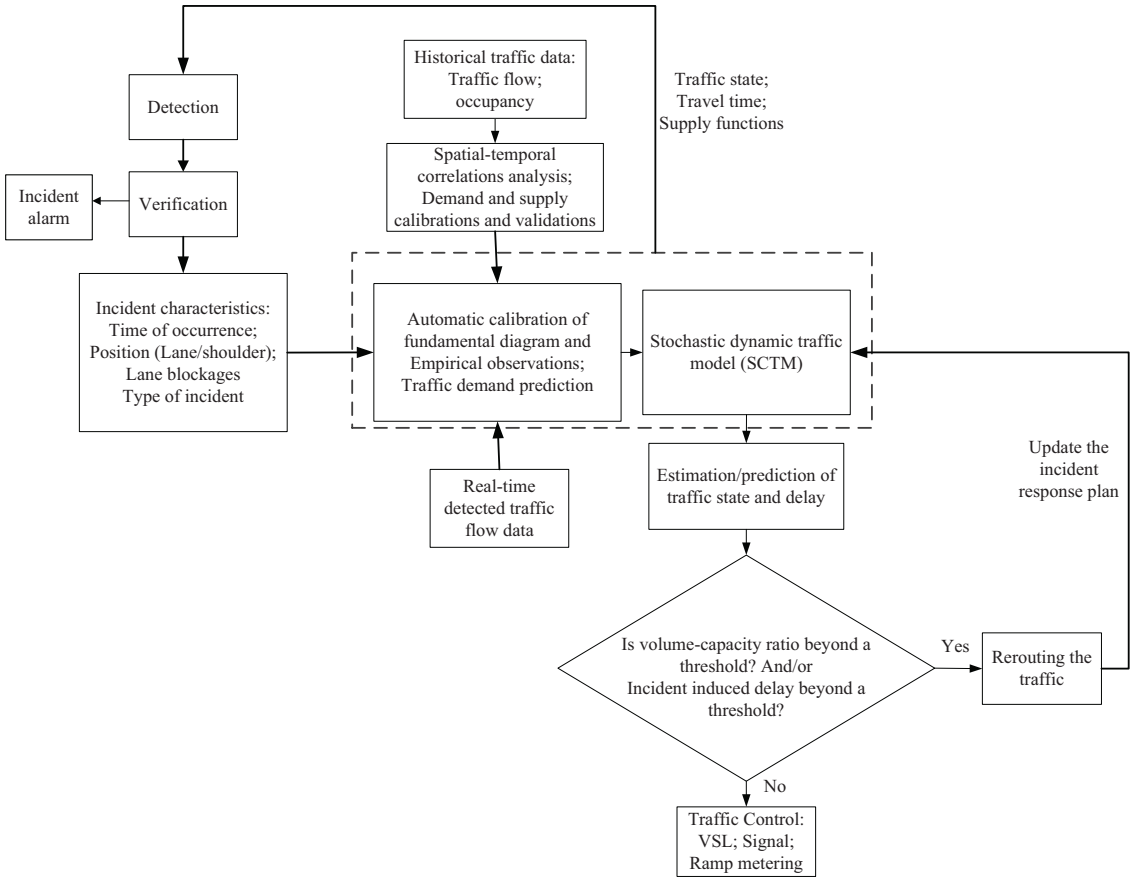


Figure 8.4: Sequence of the incident detection and traffic management

8.2.3 An approximate dynamic programming (ADP) approach to overcome the curse of dimensionality of dynamic programming

Traffic management aims to optimize the performance of the related traffic network under incident circumstances. As incidents introduce significant uncertainties to the related traffic network, the underlying optimization problem belongs to stochastic (dynamic) optimization. The challenge arises in stochastic (dynamic) optimization is that decisions are made sequentially. A decision is first made, and then information that we did not know when we made the first decision is observed. We then proceed to make another decision, after which we can observe more information. The decisions are made over time so as to minimize the objective function. Conventionally, the total travel time (TTT) or the total time spent (TTS) (Papageorgiou et al., 2003) is chosen as the objective function in traffic engineering. Several ways to model these problems as proposed by different communities have evolved modeling and algorithmic strategies to deal with specific problem classes. The most popular way to establish the optimal policy (analytically) would be the (stochastic) dynamic programming (DP) approach, wherein the optimal policy is obtained by solving Bellman's optimality equation (which requires stepping backward through time and can be shown as a fixed point equation). We adopted this approach to investigate the optimal and robust traffic management schemes for traffic networks influenced by demand and supply uncertainties which are modeled by the SCTM in Chapter 7. The DP approach is theoretically elegant and admits several powerful solution algorithms. However, they require enumerating the set of potential states. The method breaks down when the state space consists of a vector of elements (the number of states to be enumerated grows exponentially with the number of dimensions) which is common in traffic engineering involving a network more than one link. This phenomenon is the well-known curse of dimensionality of (stochastic) DP. This phenomenon renders dynamic programming intractable when the scale of problem is large.

The approximate dynamic programming ADP overcomes the above problem and provides an extremely flexible framework for modeling and solving stochastic optimization problems by combining the strengths of simulation with the intelligence of optimization. A series of stories of empirical success of ADP in applications of practical scale proved the ADP as a powerful tool for solving large-scale stochastic optimization problems (Powell, 2007; Bertsekas, 2007). Three perspectives on the ADP were argued in Powell (2007):

1. a method for large-scale optimization;

2. a method to solve complex (stochastic) dynamic programs;

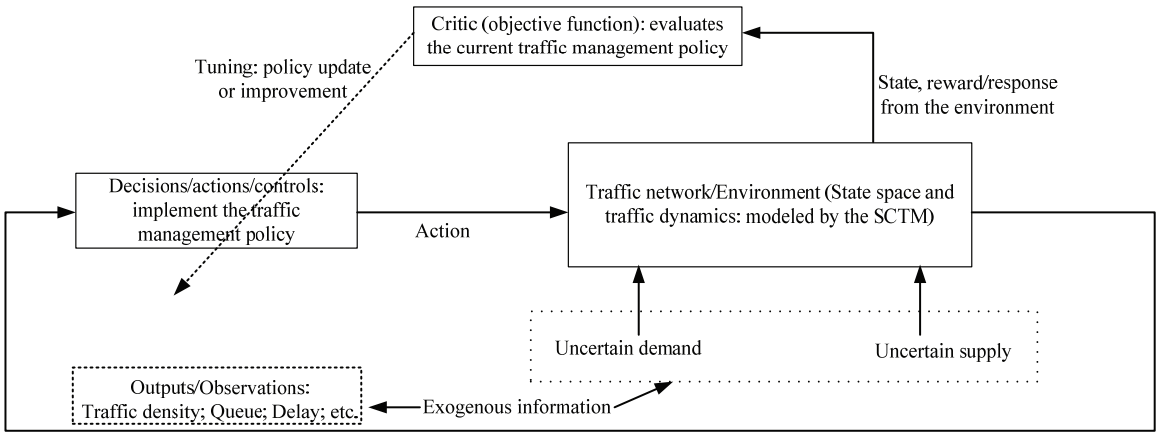


Figure 8.5: An illustration of the five fundamental dimensions of a stochastic (dynamic) optimization and the basic idea of implementation of ADP to incident management

3. a method for making simulations intelligent.

There are several distinctions between the ADP and DP. First, rather than evaluating the true value of the objective function, the ADP replaces it with some sort of statistical (or stochastic) approximation. Second, instead of conducting the optimization backward through time, the ADP steps forward in time. The other difference would be the assumptions on traveler behavior: the DP would enforce full rationality while the ADP assumes bounded rationality.

The structure of the ADP approach to determine the (optimal) traffic management policy is depicted in Figure 8.5. To begin with, we identify the following five fundamental spaces for a dynamic stochastic optimization problem i.e.

1. states (which are the traffic states of the network that can be simulated by the SCTM),
2. decisions/actions/controls (which are the traffic management policies),
3. exogenous information/random processes (which can be the demand and supply uncertainties as well as the field observations),
4. transient function/dynamics (which is the traffic dynamics modeled by the SCTM),
5. objective function/critic.

ADP offers a powerful framework for calculating the impact of a decision on the future, and using this measurement to make better decisions. A traffic management policy is designed based on the perceived traffic states of the network (or environment) and implemented to control the traffic dynamics on the network, which causes the environment to

transit into a new state. After the implementation, we observe the resulting behavior and reward from the traffic network by real-time observations and evaluation of the objective function, which reflects the quality of the applied traffic management policy and can be viewed as feedback. We then utilize the difference between the predicted performance and the observed reward (real performance) and the (short-term) prediction of traffic states based on the current observation to adjust the policy for the next time step. The adjustment is designed to make the difference between the predicted performance and the observed reward smaller, which can be captured formally in the Bellman Equation.

Several solution routines for ADP can be found in encyclopedic references on the topic (Powell, 2007; Bertsekas, 2007). The central challenge with any ADP algorithm is to find a value functional approximation which can be represented using the fewest possible number of parameters. The problem of finding the best value function approximation is closest to value iteration of DP. Detailed discussion on finding a functional approximation to the original objective function can be found in Powell (2007) and Cao (2007), which would fall into the category of stochastic approximation. It should be pointed out that before designing the value function approximation, it is extremely important for us to understand the properties of our problem, and the behavior we expect to achieve with the approximation. Then we design an approximation which captures the shape of the original function, and which will give us the desired behavior. For example, if only the TTS is taken us the objective function, then the widely applied linearly parameterized function class would be enough (Cao, 2007; de Fariasand and Roy, 2003; Powell, 2007).

The other important issue of the ADP is to find the approximate policy optimization. The problem of finding the best approximation of a policy is closest to policy iteration of DP. Several methods to address this problem are proposed by the artificial intelligence (AI) community. It is out of the scope of this chapter to review and discuss these methods. We refer the readers to the encyclopedic references on the topic (Powell, 2007; Bertsekas, 2007; Si et al., 2004). We emphasize here that different learning algorithms should be used for different purposes based on different data sets. For instance, the calibration of SCTM and its application to traffic state prediction require sufficient large amount of data (historical traffic flow data and incident records). In the case when we are sort of historical traffic flow data and incident records especially at the initial stage of the project, we cannot rely on the traffic states estimated/predicted by the SCTM. Under such circumstances, model-free ADP approach (e.g. Q-learning) is preferred to develop the incident management strategies. When the historical data and incident records are cumulated to a rather sufficient level, we would like to apply the model-based ADP approach for de-

veloping incident management strategies to reduce the learning time and thus to increase the efficiency of the incident management policy by making full usage of the traffic state prediction from the SCTM.

8.2.4 Multiagent reinforcement learning to coordinate the performance of agents

In Chapter 7, we applied the multiagent system (MAS) approach to solve the traffic management problem for a traffic network. But how to coordinate and thus optimize the performance of a multi-agent system is a difficult question. Multi-agent reinforcement learning (MARL) may be the most popular method to address this problem. In the future work, we will extend methods in multiagent reinforcement learning (MARL) to address this problem. We choose MARL for it well fits the distributed structure of MAS and the parallel computation. Information and experience sharing in the MARL also promote agents with similar tasks to learn faster and better. A MAS with MARL is robust in sense that when one or more agents fail, other agents can take over some of their tasks. Meanwhile, several challenges arise in the MARL. Generally difficult to define a good common learning goal for the multiple agents may be the foremost challenge. Others include the nonstationarity of the learning problem, the need for coordination and those inherited from single-agent reinforcement learning, including the curse of dimensionality and the exploration-exploitation tradeoff.

The curse of dimensionality could be addressed by the ADP approach. The exploration-exploitation tradeoff would not be a significant issue in our case, as the SCTM has been validated as a powerful tool to estimate/predict traffic states which can be used to support the decision making. The need for coordination is obviously. Coordination is typically required in cooperative settings. Due to the fact that the agents' returns are correlated and thus cannot be maximized independently, it is difficult to specify a good MARL goal in the general stochastic game. Nonstationarity arises because all the agents are learning simultaneously due to the distributed structure and possible parallel computation. The best policy of an agent changes as the other agents' policies change. The above two questions are related, i.e. we cannot specify a MARL goal without considering the nonstationarity and vice versa. Two typical goals are desired in the literature, i.e. stability and the adaptation. Different from the concept in control literature, stability essentially means the convergence to a stationary policy in MARL, whereas adaptation aims to maintain/improve the performance as the other agents change their policies. These two concepts fit the concepts of stability and rationality in evolutionary game theory (EGT)

(or learning in games) pretty well. Methodologies on convergence usually under some rationality assumptions of learning in games can be used to address the underlying problems. Regarding to the EGT, the ADP or RL is also used to explain how equilibrium may arise under bounded rationality.

Bibliography

Aboudolas, K., Papageorgiou, M., Kosmatopoulos, E., 2009. Store-and-forward based methods for the signal control problem in large-scale congested urban road networks. Transportation Research Part C, 17 (2), 163-174.

Aboudolas, K., Papageorgiou, M., Kouvelas, A., 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, 18 (5), 680-694.

Adeli, H., Jiang, X., 2009. Intelligent infrastructure: neural networks, wavelets, and chaos theory for intelligent transportation systems and smart structures, Boca Raton: CRC Press.

Adida, E., Perakis, G., 2007. A nonlinear continuous time optimal control model of dynamic pricing and inventory control with no backorders. Naval Research Logistics, 54, 767-795.

Adler, J., Satapathy, G., Manikonda, V., Bowles, B., Blue, V., 2005. A multi-agent approach to cooperative traffic management and route guidance. Transportation Research Part B, 39 (4), 297-318.

Ait Rami, M., Zhou, X., 2000. Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. IEEE Transactions on Automatic Control, 45 (6), 1131-1143.

Ait Rami, M., Chen, X., Zhou, X., 2002. Discrete-time indefinite LQ control with state and control dependent noises. Journal of Global Optimization, 23, 245-265.

Akamatsu, T., 2007. Tradable network permits: A new scheme for the most efficient use of network capacity. submitted to Transportation Science.

Akcelik, R., 1980. Time-dependent expressions for delay, stop rate and queue length at traffic signals. Australian Road Research Board. Internal Report AIR 367-1.

Akcelik, R., Rouphail, N., 1993. Estimation of delays at traffic signals for variable demand conditions. Transportation Research Part B, 27(2), 109-131.

Anderson B., Moore, J., 1979. Optimal Filtering. Englewood Cliffs, NJ: Prentice-Hall.

Armstrong, M., Doyle, C., Vickers, J., 1996. The access pricing problem: A synthesis. The Journal of Industrial Economics, 44 (2), 131-150.

Arnott, R., Palma, A., Lindsey, R., 1993. A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. The American Economic Review, 83 (1), 161-179.

Arnott, R., de Palma, A., Lindsey, R., 1995. Recent developments in the bottleneck model. Research paper No. 95-11, Department of Economics, University of Alberta, Canada.

Arutyunov, A., Aseev, S., 1997. Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints. SIAM Journal on Control and Optimization, 35 (3), 930-952.

Astarita, V., 1996. A continuous time link model for dynamic network loading based on travel time function. In Proceedings of 13th International Symposium on Transportation and Traffic Theory, 79-102.

Balakrishna, R., 2006. Off-line calibration of dynamic traffic assignment models. PhD thesis, Massachusetts Institute of Technology.

Balijepalli, N.C., Watling, D.P., 2005. Doubly dynamic equilibrium distribution approximation model for dynamic traffic assignment. In: Mahmassani, H. (Ed.), Transportation and Traffic Theory. Pergamon, Oxford, UK, 741-760.

Ban, X., Liu, X., Ferris, M., Ran, B., 2008. A link-node complementarity model and solution algorithm for dynamic user equilibria with exact flow propagations. Transportation Research Part B, 42 (9), 823-842.

Basar, T., Bernhard, P., 1995. H_∞ Optimal control and related minimax design problems, Basel: Birkhauser.

Basar, T., Olsder, G., 1999. Dynamic noncooperative game theory, SIAM.

Becerra, V.M., 2004. Solving optimal control problems with state constraints using nonlinear programming and simulation tools. IEEE Transactions on Education, 47, 377-384.

Berkowicz, R., 1998. Street Scale Models. In: Fenger, J., Hertel, O., Palmgren, F. (Eds.), Urban Air Pollution-European Aspects, Kluwer Academic Publishers. 223-251.

Berkowicz, R., 2000a. OSPM-A parameterised street pollution model. Environmental Monitoring and Assessment. 65, 323-331.

Berkowicz, R., 2000b. A simple model for urban background pollution. Environmental Monitoring and Assessment. 65, 259-267.

Berkowicz, R., Ketzel, M., Jensen, S., Hvidberg, M., Raaschou-Nielsen, O., 2008. Evaluation and application of OSPM for traffic pollution assessment for a large number of street locations. Environmental Modelling & Software, 23 (3), 296-303.

Bertsekas, D., 2007. Dynamic programming and optimal control, 3rd Edition Vol. II, Athena Scientific.

Betts, J., 2010. Practical methods for optimal control and estimation using nonlinear programming. Philadelphia, Society for Industrial and Applied Mathematics.

Boel, R., Mihaylova, L., 2006. A compositional stochastic model for real time freeway traffic simulation, Transportation Research Part B, 40, 319-334.

Boyce, D., Lee, D., Ran, B., 2001. Analytical models of the dynamic traffic assignment problem. 2001, 1 (3-4), 377-390.

Boyd, S., Ghaoui, L., Feron, E., Balakrishnan, V., 1994. Linear matrix inequalities in system and control theory. Society for Industrial and Applied Mathematics, Philadelphia.

Brilon, W., Wu, N., 1990. Delays at fixed-time traffic signals under time dependent traffic conditions. Traffic Engineering and Control, 31(12), 623-631.

Bristow, D., Tharayil, M., Alleyne, A., 2006. A survey of iterative learning control. IEEE Control System Magazine, 26 (23), 96-114.

Bulirsch, R., and Kraft, D., 1994. Computational optimal control, Basel, Boston: Birkhäuser Verlag.

Busoniu, L., Babuska, R., De Schutter, B., 2008. A comprehensive survey of multi-agent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 38 (2), 156-172.

Buskens, C., Maurer, H., 2000. SQP-methods for solving optimal control problems with control and state constraints: adjoint variables, sensitivity analysis and real-time control. Journal of Computational and Applied Mathematics, 120 (1), 85-108.

Buchanan, C., 1963. Traffic in towns, HMSO, London.

Calzada, J., 2007. Capacity-based versus time-based access charges in telecommunications. Journal of Regulatory Economics, 32 (2), 153-172.

Cao, X., 2007. Stochastic learning and optimization-A sensitivity-based approach. New York: Springer.

Carlson, R., Papamichail, I., Papageorgiou, M., Messmer, A., 2010. Optimal motorway traffic flow control involving variable speed limits and ramp metering. Transportation Science, 44 (2), 238-253.

Carey, M., 1987. Optimal time-varying flows on congested networks. Operations Research, 35, 58-69.

Carey, M., Srinivasan, M., 1993. Externalities, average and marginal costs, and tolls on congested networks with time-varying flows. Operations Research, 41(1), 217-231.

Carey, M., McCartney, M., 2002. Behaviour of a whole-link travel time model used in dynamic traffic assignment. Transportation Research Part B, 36 (1), 83-95.

Carey, M., Ge, Y., 2003. Comparing whole-link travel time models used in DTA. Transportation Research Part B, 37 (10), 905-926.

Carey, M., McCartney, M., 2004. An exit-flow model used in dynamic traffic assignment. Computers and Operations Research, 31 (10), 1583-1602.

Carey, M., 2004a. Link travel times I: desirable properties. Network and Spatial Economics, 4 (3), 257-268.

Carey, M., 2004b. Link travel times II: properties derived from traffic-flow models. Network and Spatial Economics, 4 (4), 379-402.

Carey, M., 2008. A framework for user equilibrium dynamic traffic assignment. Journal of the Operational Research Society, 60, 395-410.

Cassidy, M., Rudjanakanoknad, J., 2005. Increasing the capacity of an isolated merge by metering its on-ramp. Transportation Research, Part B, 39 (10), 896-913.

Chang, T., Nozick, L., Turnquist, M., 2005. Multiobjective path finding in stochastic dynamic networks, with application to routing hazardous materials shipments. Transportation Science, 39 (3), 383-399.

Chen, C., Jia, Z., Varaiya, P., 2001. Causes and cures of highway congestion. IEEE Control Systems Magazine, 21, 26-33.

Chen, C., 2003. Freeway performance measurement system (PeMS). PhD dissertation, University of California, Berkeley.

Chen, C., Kwon, J., Rice. J., Skabardonis, A., Varaiya, P., 2003. Detecting errors and imputing missing data for single loop surveillance systems. Transportation Research Record, No. 1855.

Chen, B., Cheng, H., Palen, J., 2009. Integrating mobile agent technology with multi-agent systems for distributed traffic detection and management systems. Transportation Research Part C, 17 (1), 1-10.

Chen, B., Cheng, H., 2010. A review of the applications of agent technology in traffic and transportation systems. IEEE Transaction on Intelligent Transportation Systems, 11 (2), 485-497.

Chen, S., Zhou, X., 1998. Stochastic linear quadratic regulators with indefinite control weight costs. SIAM Journal on Control and Optimization, 36 (5), 1685-1702.

Chen, S., Zhou, X., 2000. Stochastic linear quadratic regulators with indefinite control weight costs. II. SIAM Journal on Control and Optimization, 39 (4), 1065-1081.

Cheng, Y., Hui, H., Lo, H., 2005. Report of the task force on emergency transport coordination. Report to the Secretary for the Environment, Transport and Works, Hong Kong.

Chow, A., 2007a. Analysis of dynamic system optimum and externalities with departure time choice. In: Allsop, R.E., Bell, M.G.H., Heydecker, B.G. (Eds.), Transportation and Traffic Theory. Elsevier, Amsterdam, 301-326.

Chow, A., 2007b. System optimal traffic assignment with departure time choice. PhD dissertation, University of London.

Chow, A., 2009a. Dynamic system optimal traffic assignment-a state-dependent control theoretic approach. Transportmetrica, 5 (2), 85-106.

Chow, A., 2009b. Properties of system optimal traffic assignment with departure time choice and its solution method. Transportation Research Part B, 43 (3), 325-344.

Costa, O., Fragoso, M., and Marques, R., 2005. Discrete-time Markov jump linear systems. London, Springer.

Costa, O., Wanderlei, L., 2007. Indefinite quadratic with linear costs optimal control of Markov jump with multiplicative noise systems. Automatica, 43, 587-597.

Costa, O., Okimura, R., 2009. Discrete-time mean variance optimal control of linear systems with Markovian jumps and multiplicative noise. International Journal of Control, 82 (2), 256-267.

Daganzo, C., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B, 28, 269-287.

Daganzo, C., 1995a. The cell transmission model: network traffic. Transportation Research Part B, 29, 79-93.

Daganzo, C., 1995b. Properties of link travel time functions under dynamic loads. Transportation Research Part B, 29 (2), 95-98.

Davis, J., 2002. Foundations of deterministic and stochastic control. Birkhäuser, Boston.

de Farias, D., Roy, B., 2003. The linear programming approach to approximate dynamic programming. Operations Research, 51 (6), 850-856.

de Oliveira, L., Camponogara, E., 2010. Multi-agent model predictive control of signaling split in urban traffic networks. Transportation Research Part C, 18 (1), 120-139.

Dion, F., Rakha, H., Kang, Y., 2004. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. Transportation Research Part B, 38(2), 99-122.

Drissi-Kaitouni, O., Hameda-Benchekroun, A., 1992. A dynamic traffic assignment model and a solution algorithm. Transportation Science, 26 (2), 119-128.

Dramitinos, M., Stamoulis, G., Courcoubetis, C., 2007. An auction mechanism for allocating the bandwidth of networks to their users. Computer Networks, 51, 4979-4996.

Dragan, V., Morozan, T., Stoica, A., 2010. Mathematical methods in robust control of discrete-time linear stochastic systems. New York, Springer.

European Commission, 2001. White paper–European transport policy for 2010: time to decide. Office for official publications of the European communities, Luxembourg.

Fattorini, H., 1999. Infinite dimensional optimization and control theory. Cambridge University Press, Cambridge, New York.

Friesz, T., Luque, F., Tobin, R., Wie, B., 1989. Dynamic network assignment considered as a continuous time optimal control problem. Operations Research, 37 (6), 893-901.

Friesz, T., Bernstein, D., Smith, T., Tobin, R., Wie, B., 1993. A variational inequality formulation of the dynamic networks user equilibrium problem. Operations Research, 41 (2), 179-191.

Friesz, T., Bernstein, D., Suo, Z., Tobin, R., 2001. Dynamic network user equilibrium with state-dependent time lags. Networks and Spatial Economics, 1 (3-4), 319-347.

Friesz, T., Mookherjee, R., 2006. Solving the dynamic network user equilibrium problem with state-dependent time shifts. Transportation Research Part B, 40 (3), 207-229.

Friesz, T.L., Kwon, C., Mookherjee, R., 2007. A computable theory of dynamic congestion pricing. In: Allsop, R.E., Bell, M.G.H., Heydecker, B.G. (Eds.), Transportation and Traffic Theory. Elsevier, Amsterdam, 1-26.

Friesz, T., 2010. Dynamic optimization and differential games. New York, Springer.

Friesz, T., Mookherjee, R., Yao, T., 2008. Securitizing congestion: The congestion call option. Transportation Research Part B, 42 (5), 407-437.

Friesz, T., Kim, T., Kwon, C., Rigdon, M., 2011. Approximate network loading and dual-time-scale dynamic user equilibrium. Transportation Research Part B, 45 (1), 176-207.

Frühwirth-Schnatter, S., 2006. Finite mixture and Markov switching models, New York, Springer.

Geistefeldt, J., Brilon, W., 2009. A comparative assessment of stochastic capacity estimation methods, in Transportation and Traffic Theory 2009, Springer.

Gentile, G., Meschini, L., Papola, N., 2005. Macroscopic arc performance models with capacity constraints for within-day dynamic traffic assignment. Transportation Research Part B, 39, 319-338.

Ghali, M., Smith, M., 1995. A model for the dynamic system optimum traffic assignment problem. Transportation Research Part B, 29(3), 155-170.

Gomes, G., 2004. Optimization and microsimulation of on-ramp metering for congested freeways. Ph.D. dissertation, University of California, Berkeley.

Gomes, G., and Horowitz, R., 2006. Optimal freeway ramp metering using the asymmetric cell transmission model. Transportation Research Part C, 14, (4), 244-262.

Gomes, G., Horowitz, R., Kurzhanskiy, A., Varaiya, R., and Kwon, J., 2008. Behavior of the cell transmission model and effectiveness of ramp metering, Transportation Research Part C, 16 (4), 485-513.

Goodwin, G., Seron, M., Dona, J., 2005. Constrained control and estimation: an optimisation approach, London, Springer.

Guberinić, S., Šenborn, G., and Bratislav Lazić, 2008. Optimal traffic control: urban intersections. Boca Raton, CRC Press.

Haberman, R., 1977. Mathematical models, mechanical vibrations, population dynamics and traffic flow. Prentice-Hall, Englewood Cliffs, NJ.

Harrison, P., Stevens, C., 1976. Bayesian forecasting (with discussion). Journal of the Royal Statistical Society, Ser. B, 38, 205-247.

Haddad, W., Chellaboina, V., 2008. Nonlinear dynamical systems and control : a Lyapunov-based approach, Princeton University Press.

Hansen, L.P., Sargent, T.J., 2001. Robust control and model uncertainty. American Economic Review, 91(2), 60-66.

Hansen, L.P., Sargent, T.J., 2005. Robust estimation and control under commitment. Journal of Economic Theory, 124(2), 258-301.

Hansen, L.P., Sargent, T.J., 2007. Recursive robust estimation and control without commitment. Journal of Economic Theory, 136, 1-27.

Hansen, L.P., Sargent, T.J., 2008. Robustness. Princeton University Press, Princeton, NJ.

Hansen, L.P., Mayer, R., Sargent, T.J., 2010. Robust hidden Markov LQG problems. Journal of Economic Dynamics and Control, doi:10.1016/j.jedc.2010.05.004.

Han, D., Yang, H., Wang, X., 2010. Efficiency of the plate-number-based traffic rationing in general networks. Transportation Research Part E, 46 (6), 1095-1110.

Hartl, R., Sethi, S., Vickson, R., 1995. A Survey of the Maximum Principles for Optimal Control Problems with State Constraints, SIAM Review, 37 (2), 181-218.

Transportation Research Board, 2000. Highway capacity manual 2000. National Research Council, Washington, D.C.

Hegyi, A., 2004. Model predictive control for integrating traffic control measures. PhD disertation, TRAIL Thesis Series, The Netherlands.

Hegyi, A., Schutter, B., Hellendoorn, H., 2005. Model predictive control for optimal coordination of ramp metering and variable speed limits, Transportation Research Part C, 13, 185-209.

Hegyi, A., Schutter, B., Hellendoorn, H., 2005. Optimal coordination of variable speed limits to suppress shock waves. IEEE Transactions on Intelligent Transportation Systems, 6, 102-112.

Heydecker, B., Addison, J., 2005. Analysis of Dynamic Traffic Equilibrium with Departure Time Choice. Transportation Science, 39 (1), 39-57.

Holdsworth, J., and Singleton, D., 1979. Environmental capacity of roads. Proceedings of the 5th Australian Transport Research Forum, Australian Government Pub. Service, Canberra, Australia, 219-238.

Holdsworth, J., Singleton, D., 1980. Environmental capacity as a basis for traffic management at local government level. Proceedings of the 10th ARRB Conference, Melbourne, Australia, 165-173.

Hou, T., Zhang, W., Ma, H., 2010. Finite horizon control for discrete-time stochastic systems with markovian jumps and multiplicative noise. IEEE Transaction on Automatic Control, 55 (5), 1185-1191.

Huang, H., Lam, W., 2002. Modeling and solving the dynamic user equilibrium route and departure time choice problem in network with queues. Transportation Research Part B, 36, 253-273.

Iryo, T., 2008. An analysis of instability in a departure time choice problem, Journal of Advanced Transportation, 42, 333-356.

Johnston, R., Lund, J., Craig, P., 1995. Capacity-Allocation Methods for Reducing Urban Traffic Congestion. Journal of Transportation Engineering, 121 (1), 27-39.

Kachroo, P., Özbay, K., 1999. Feedback control theory for dynamic traffic assignment. Springer-Verlag Series Advances in Industrial Control, Springer-Verlag.

Kachroo, P., Özbay, K., 2003. Feedback ramp metering in intelligent transportation systems New York: Kluwer Academic/Plenum.

Kachroo, P., Özbay, K., 2005. Feedback control solutions to network level user-equilibrium real-time dynamic traffic assignment problems. Networks and Spatial Economics, 5 (3), 243-260.

Kachroo, P., Özbay, K., 2006. Feedback control solutions to network level system optimal real-time dynamic traffic assignment problems. Journal of Intelligent Transportation Systems, 10 (4), 159-171.

Karimi, A., Hegyi, A., Schutter, B., Hellendoorn, H., and Middelham, F., 2004. Integration of dynamic route guidance and freeway ramp metering using model predictive control, Proceedings of the 2004 American Control Conference, Boston, Massachusetts, 5533-5538.

Katwijk, R., 2008. Multi-agent look-ahead traffic adaptive control. PhD dissertation, Delft University of Technology.

Khalil, H., 2002. Nonlinear systems, Prentice Hall.

Khattak, A., Wang, X., Zhang, H., 2009. Are incident durations and secondary incidents interdependent? Pre-print CD-ROM, the 88th Transportation Research Board (TRB) Annual Meeting, Washington, D.C.

Kuwahara, M., 2007. A theory and implications on dynamic marginal cost. Transportation Research Part A, 41 (7), 627-643.

Kim, T., Zhang, H., 2008. A stochastic wave propagation model, Transportation Research Part B, 42, 619-634.

Kosmatopoulos, E., Papageorgiou, M., Bielefeldt, C., Dinopoulou, V., Morris, R., Mueck, J., Richards, A., and Weichenmeier, F., 2006. International comparative field evaluation of a traffic-responsive signal control strategy in three cities. Transportation Research Part A, 40(5), 399-413.

Kosmatopoulos, E., Papageorgiou, M., Manolis, D., Hayden, J., Higginson, R., McCabe, K., Rayman, N., 2006. Real-time estimation of critical occupancy for maximum motorway throughput. Transportation Research Record, No. 1959, 65-76.

Kosmatopoulos, E., Papageorgiou, M., Vakouli, A., Kouvelas, A., 2007. Adaptive fine-tuning of nonlinear control systems with application to the urban traffic control strategy TUC. IEEE Transactions on Control Systems Technology, 15 (6) , 991-1002.

Kotsialos, A., Papageorgiou, M., Mangeas, M, Haj-Salem, H., 2002. Coordinated and integrated control of motorway networks via nonlinear optimal control. Transportation Research Part C, 10(1), 65-84.

Kotsialos, A., Papageorgiou, M., Diakaki, C., Pavlis, Y., Middelham, F., 2002. Traffic flow modeling of large-scale motorway networks using the macroscopic modeling tool METANET. IEEE Transactions on Intelligent Transportation Systems, 3 (4), 282-292.

Kotsialos, A., Papageorgiou, M., 2004. Efficiency and equity properties of freeway network wide ramp metering with AMOC, Transportation Research, Part C, 12(6), 401-420.

Kurzhanskiy, A., 2007. Modeling and Software Tools for Freeway Operational Planning. PhD thesis, University of California, Berkeley.

Lam. W.H.K., Tam, M.L., Sumalee, A., Li, C.L., Chen, W., Kwok, S.K., Li, Z.L., Ngai, E.W.T., 2008. Incident detection based on short-term travel time forecasting. Proceedings of the 13th International Conference of Hong Kong Society for Transportation Studies, 13-15 December 2008, Hong Kong, 83-92.

Larsson, T., Patriksson, M., 1995. An augmented lagrangean dual algorithm for link capacity side constrained traffic assignment problems, Transportation Research Part B, 29 (6), 433-455.

Larsson, T., Patriksson, M., 1999. Side constrained traffic equilibrium models-analysis, computation and applications. Transportation Research Part B, 33 (4), 233-264.

Larsson, T., Patriksson, M., Rydergren, C., 2004. A column generation procedure for the side constrained traffic equilibrium problem, Transportation Research Part B, 38 (1), 17-38.

Lazar, A., Semret, N., 1999. Design, analysis and simulation of the progressive second price auction for network bandwidth sharing. Telecommunications Systems-Special Issue on Network Economics.

Lebacque, J. P., 1996. The Godunov Scheme and what it means for first order traffic flow models. In Lesort (Ed.), Transportation and Traffic Theory, Pergamon-Elservier, New York, 647-677.

Lee, J., Özbay, K., 2009. A new calibration methodology for microscopic traffic simulation using enhanced simultaneous perturbation stochastic approximation approach. Journal of Transportation Research Record, No. 2124, 233-240.

Leoni, G., 2009. A first course in Sobolev spaces. American Mathematical Society, Providence.

Li, T., Lin, J., Wu, M., and Wang, X., 2009, Concept and spatial analysis method of urban environmental traffic capacity, Journal of Transportation Engineering, 135 (11), 873-879.

Li, X., 2008. Large-eddy simulation of wind flow and air pollutant transport inside unban street canyons of different aspect ratios. Ph.D. dissertation, The University of Hong Kong, Hong Kong, China.

Li, J., Chen, Q., Wang, H., Ni, D., 2009. Analysis of LWR model with fundamental diagram subject to uncertainties. Pre-print CD-ROM, the 88th Transportation Research Board (TRB) Annual Meeting, Washington, D.C.

Li, Z., Chang, G., Natarajan, S., 2009. Integrated off-ramp control model for freeway traffic management. Pre-print CD-ROM, the 88th Transportation Research Board (TRB) Annual Meeting, Washington, D.C.

Lindley, J., 1986. Quantification of urban freeway congestion and analysis of remedial measures. Report RD/87-052. FHWA, U.S. Department of Transportation.

Lindley, J., 1987. Urban freeway congestion: quantification of the problem and effectiveness of potential solutions. Journal of Institute of Traffic Engineering, 57, 27-32.

Lindsey, R., 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. Transportation Science, 38 (3), 293-314.

Lindsey, R., 2006. Do economists reach a conclusion on road pricing? The intellectual history of an idea. Econ Journal Watch, 3 (2), 292-379.

van Lint, J., 2004. Reliable travel time prediction for freeways. TRAIL Thesis Series, The Netherlands.

Liu, Y., Yu, J., Chang, G., 2009. A multi-objective model for optimal diversion control of a freeway corridor under incident conditions. Pre-print CD-ROM, the 88th Transportation Research Board (TRB) Annual Meeting, Washington, D.C.

Liu, H. X., Wu. X., Ma, W., Hu, H., 2009. Real-time queue length estimation for congested signalized intersections. Transportation Research Part C, 17(4), 412-427.

Lo, H., 1999a. A dynamic traffic assignment formulation that encapsulates the cell transmission model. Transportation and Traffic Theory. Edited by A. Cedar, Elsevier Science, 327-350.

Lo, H., 1999b. A novel traffic signal control formulation. Transportation Research Part A, 33, 433-448.

Lo, H., Chang, E., Chan, Y. C., 2001. Dynamic network traffic control. Transportation Research Part A, 35, 721-744.

Lo, H.P., Szeto, W.Y., 2002. A cell-based variational inequality formulation of the dynamic user optimal assignment problem. Transportation Research Part B, 36, 421-443.

Lovell, D., Daganzo, C., 2000. Access control on networks with unique origindestination paths. Transportation Research Part B, 34 (3), 185-202.

Lu, C., Mahmassani, H., Zhou, X., 2008. Equivalent gap function-based reformulation and solution algorithm for the dynamic user equilibrium problem. Transportation Research Part B, 43 (3), 345-364.

McNeill, D.R., 1968. A solution to the fixed cycle traffic light problem with compound Poisson arrivals. Journal of Applied Probability, 5(3), 624-635.

Merchant, D.K., Nemhauser, G.L., 1978a. A model and an algorithm for the dynamic traffic assignment problem. Transportation Science 12 (3), 183-199.

Merchant, D.K., Nemhauser, G.L., 1978b. Optimality conditions for a dynamic traffic assignment model. Transportation Science 12 (3), 200-207.

Messmer, A., Papageorgiou, M., 1990. METANET: a macroscopic simulation program for motorway networks. Traffic Engineering and Control, 31, (8/9), 466-470.

Mihaylova, L., Boel, R., Hegyi, A., 2007. Freeway traffic estimation within particle filtering framework, Automatica, 43, 290-300.

Miller, A.J., 1968. The capacity of signalised intersections in Australia. Bulletin No. 3, Australian Road Research Board.

Minoux, M., 1986. Mathematical Programming: Theory and Algorithms. New York: John Wiley.

Mohler, R., 1973. Bilinear control processes: with applications to engineering, ecology, and medicine, New York, Academic Press.

Mounce, R., 2006. Convergence in a continuous dynamic queueing model for traffic networks. Transportation Research Part B, 40, 779-791.

Mounce, R., Smith, M., 2007. Uniqueness of equilibrium in steady state and dynamic traffic networks. R. Allsop, M. Bell and B. Heydecker eds, Transportation and Traffic Theory. Elsevier, Amsterdam, 281-299.

Mounce, R., 2007. Convergence to equilibrium in dynamic traffic networks when route cost is decay-monotone. Transportation Science, 41, 409-414.

Muralidharan, A., Horowitz, R., 2009. Imputation of ramp flow data for freeway traffic simulation. Transportation Research Record, No. 2099, 58-64.

Muñoz, L., Sun, X., Horowitz, R., Alvarez, L., 2003. Traffic density estimation with the cell transmission model, *Proceedings of the American Control Conference*, Denver, Colorado, June, 3750-3755.

Muñoz, L., Sun, X., Sun, D., Gomes, G., and Horowitz, R., 2004. Methodological calibration of the cell transmission model, *Proceeding of the 2004 American Control Conference*, Boston, Massachusetts, July, 798-803.

Mun, J., 2007. Traffic performance models for dynamic traffic assignment: an assessment of existing models. Transport Reviews, 27, 231-249.

Nagae, T., Sasaki, S., 2009. A mean-variance approach to mixed strategies for dispatching problems under travel time uncertainty. The 14th Hong Kong Society of Transportation Studies International Conference, 189-196.

Nagurney, A., 2000. Sustainable Transportation Networks, Edward Elgar Publishers, Cheltenham, England.

Nagurney, A., Dong, J., Mokhtarian, P., 2002. Traffic Network Equilibrium and the Environment: A Multicriteria Decision-Making Perspective. In Computational Methods in Decision-Making, Economics and Finance, E. Kontoghiorges, B. Rustem, and S. Siokos, Editors, Kluwer Academic Publishers.

Nagurney, A., Liu, Z., Cojocaru, M., Daniele, P., 2007. Dynamic electric power supply chains and transportation networks: an evolutionary variational inequality formulation. Transportation Research Part E, 43, 624-646.

Nagurney, A., Qiang, Q., Nagurney, L., 2010. Environmental impact assessment of transportation networks with degradable links in an era of climate change. International Journal of Sustainable Transportation, 4, 154-171.

Nie, X., Zhang, H.M., 2002. Delay-function-based link models: their properties and computational issues. Transportation Research Part B, 39 (8), 729-751.

Nie, X., Zhang, H., 2005. A comparative study of some macroscopic link models used in dynamic traffic assignment. Networks and Spatial Economics, 5 (1), 89-115.

Nie, Y., Zhang, H. M., 2010. Solving the dynamic user optimal assignment problem considering queue spillback. Networks and Spatial Economics, 10, 49-71.

Nie, Y., in press. A cell-based Merchant-Nemhauser model for the system optimum dynamic traffic assignment problem. Transportation Research Part B.

Ngoduy, D., 2009. Multiclass first order model using stochastic fundamental diagrams. Transportmetrica, in press.

Noiri, T., 1974. On weakly continuous mappings. Proceedings of the American Mathematical Society, 46 (1), 120-124.

Oke, T., 1988. Street design and urban canopy layer climate. Energy and Buildings, 11, 103-113.

Özbay, K., Kachroo, P., 1999. Incident management in intelligent transportation systems. Artech House, Boston.

Özbay, K., Yanmaz-Tuzel, O., 2008. Valuation of travel time and departure time choice in the existence of time-of-day pricing. Transportation Research Part A, 42, 577-590.

Özbay, K., Bartin, B., Yanmaz-Tuzel, O., Berechman, J., 2007. Alternative methods for estimating full marginal costs of highway transportation. Transportation Research Part A, 41, 768-786.

Özbay, K., Yasar, I., Kachroo, P., 2004. Modeling and PARAMICS based evaluation of new local freeway ramp metering strategy that takes into account ramp queues. Journal of Transportation Research Record, No. 1867, 89-97.

Özbay, K., Xiao, W., Jaiswal, G., Bartin, B., Kachroo, P., Baykal-Gursoy, M., 2009. Evaluation of incident management strategies and technologies using an integrated traffic/incident management simulation. World Review of Intermodal Transportation Research, 2 (2/3), 155-186.

Ozguven, E., Özbay, K., 2008. Nonparametric Bayesian estimation of freeway Capacity distribution from censored observations. Journal of Transportation Research Record, No. 2061, 20-29.

Papageorgiou, M., 1983. Application of automatic control concepts to traffic flow modeling and control. Springer- Verlag, New York.

Papageorgiou, M., Kotsialos, A., 2002. Freeway ramp metering: an overview. IEEE Transactions on Intelligent Transportation Systems, 3, 271-281.

Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., Wang, Y., 2003. Review of road traffic control strategies, Proceedings of the IEEE, 96 (12), 2043-2067.

Papageorgiou, M., Papamichail, I., Spiliopoulou, A., Lentzakis, A., 2008. Real-time merging traffic control with applications to toll plaza and work zone management, Transportation Research Part C, 16, 535-553.

Papageorgiou, M., Kosmatopoulos, E., Papamichail, I., Wang, Y., 2008. A misapplication of the local ramp metering strategy ALINEA. IEEE Transactions on Intelligent Transportation Systems, 9 (2), 360-365.

Papamichail, I., Kotsialos, A., Margonis, I., Papageorgiou, M., 2010. Coordinated ramp metering for freeway networks—a model-predictive hierarchical control approach. Transportation Research, Part C, 18 (3), 311-331.

Papamichail, I., Papageorgiou, M., 2008. Traffic-responsive linked ramp-metering control. IEEE Transactions on Intelligent Transportation Systems, 9 (1), 111-121.

Papageorgiou, M., Haj-Salem, H., Blosseville, J., 1991. ALINEA: a local feedback control law for on-ramp metering. Transportation Research Record, No. 1320, 58-64.

Peeta, S., 1994. System optimal dynamic traffic assignment in congested networks with advanced information systems, PhD dissertation, The University of Texas at Austin.

Peeta, S., Mahmassani, H.S., 1995. System optimal and user equilibrium time-dependent traffic assignment in congested networks. Annals of Operations Research, 60, 81-113.

Peeta, S., Ziliaskopoulos, A., 2001. Foundations of dynamic traffic assignment: the past, the present and the future. Networks and Spatial Economics, 1, (3/4), 233-266.

Peeta, S., Yang, T., 2003. Stability issues for dynamic traffic assignment. Automatica, 39, 21-34.

Peeta, S., Yu, J., 2006. Behavior-based consistency-seeking models as deployment alternatives to dynamic traffic assignment models. Transportation Research Part C, 14 (2), 114-138.

Peeta, S., Zhou, C., 2006. Stochastic quasi-gradient algorithm for the off-line stochastic dynamic traffic assignment problem Transportation Research Part B, 40, 179-206.

Paz, A., 2007. Behavior-consistent deployable real-time traffic routing under advance traveler information systems, PhD dissertation, Purdue University.

Paz, A., Peeta, S., 2008. Fuzzy control model optimization for behavior-consistent traffic routing under information provision. IEEE Transactions on Intelligent Transportation Ssystems, 9 (1), 27-37.

Paz, A., Peeta, S., 2009. On-line calibration of behavior parameters for behavior-consistent route guidance. Transportation Research Part B, 43 (4), 403-421.

Paz, A., Peeta, S., 2009. Information-based network control strategies consistent with estimated driver behavior, Transportation Research Part B, 43 (1), 73-96.

Freeway Performance Measurement Project. http://pems.eecs.berkeley.edu/.

Powell, W., 2007. Approximate Dynamic Programming: Solving the curses of dimensionality. John Wiley & Sons, Hoboken, NJ.

Ran, B., Boyce, D., 1996a. Modeling dynamic transportation networks: an intelligent transportation systems oriented approach, 2^{nd} revised edition. Springer-Verlag, New York.

Ran, B., Boyce, D., 1996b. A link-based variational inequality formulation of ideal dynamic user optimal route choice problem. Transportation Research Part C, 4, 1-12.

Rouphail, N., Tarko, A., Li, J., 2000. Traffic flow at signalized intersections. In: Lieu, H. (Ed.) Revised Monograph of Traffic Flow Theory, Update and Expansion of the Transportation Research Board (TRB) Special Report 165. "Traffic Flow Theory", Published in 1975 (chapter 9).

Rose, D., 1984. Weak continuity and strongly closed sets. International Journal of Mathematics and Mathematical Science. 7 (4), 809-816.

Ross, P., 1988. Traffic dynamics. Transportation Research Part B, 22 (6), 421-435.

Rudin, W., 1976. Principles of mathematical analysis, 3rd edition, McGraw-Hill.

Schrank, D., Lomax, T., 2009. The 2009 urban mobility report. Technical report, Texas Transportation Institute, http://mobility.tamu.edu.

Schönhof, M., Helbing, D., 2007. Empirical features of congested traffic states and their implications for traffic modeling. Transportation Science, 41 (2), 135-166.

Schleiffer, R., 2002. Intelligent agents in traffic and transportation, special issue. Transportation Research Part C, 10 (5-6), 325-329.

Si, J., Barto, A., Powell, W., Wunsch, D., 2004. Handbook of learning and approximate dynamic programming. New York: Wiley-IEEE Press.

Sheffi, Y., 1985. Urban transportation networks. Prentice-Hall, Englewood Cliffs.

Shen, W., Zhang, H., 2009. On the morning commute problem in a corridor network with multiple bottlenecks: Its system-optimal traffic flow patterns and the realizing tolling scheme. Transportation Research Part B, 43 (3), 267-284.

Sheu, J., 2003. Erratum: A Stochastic Modeling Approach to Real-Time Prediction of Queue Overflows. Transportation Science, 37, 230-252.

Sheu, J., 2007. Microscopic modeling and control logic for incident-responsive automatic vehicle movements in single-automated-lane highway systems. European Journal of Operational Research, 182, 640-662.

Sheu, J., Chang, M., 2007. Stochastic optimal-control approach to automatic incident-responsive coordinated ramp control, IEEE Transactions on Intelligent Transportation Systems, 8 (2), 359-367.

Shiran, G., 1997. Area-wide environmental capacity based on air pollution criteria. Ph.D. dissertation, University of New South Wales, Kensington, Australia.

Shoham, Y., Leyton-Brown, K., 2009. Multi-agent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press, New York.

Smaragdis, E., Papageorgiou, M., 2003. A series of new local ramp metering strategies. Transportation Research Record, No. 1856, 74-86.

Smaragdis, E., Papageorgiou, M., Kosmatopoulos, E., 2004. A flow maximizing adaptive local ramp metering strategy. Transportation Research, Part B, 38 (3), 251-270.

Smith, M., 1979. The existence, uniqueness and stability of traffic equilibria. Transportation Research Part B, 13 (4), 295-304.

Smith, M., 1984. The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. Transportation Science, 18, 245-252.

Smith, M., Ghali, M., 1990. Dynamic traffic assignment and dynamic traffic control. In: Koshi, M., ed., Transportation and Traffic Theory. Elsevier, New York, 223-263.

Smith, M., 1993. A new dynamic traffic model and the existence and calculation of dynamic user equilibria on congested capacity-constrained road networks. Transportation Research Part B, 27 (1), 49-63.

Subchan, S., Żbikowski, R., 2009. Computational optimal control: tools and practice. Chichester, Wiley.

Sumalee, A., Xu, W., 2010. First-best marginal cost toll for a traffic network with stochastic demand. Transportation Research Part B. In press.

Sumalee, A., Zhong, R. X., Pan, T. L., Szeto, W. Y., 2010a. Stochastic cell transmission model (SCTM): a stochastic dynamic traffic model for traffic state surveillance and assignment, Transportation Research Part B, accepted.

Sumalee, A., Pan, T., Zhong, R., Uno, N., 2010b. Dynamic stochastic journey time estimation and reliability analysis using stochastic cell transmission model: algorithm and case studies. Submitted to Transportation Research Part C.

Smith, M., 1979. The existence, uniqueness and stability of traffic equilibria. Transportation Research Part B, 13 (4), 295-304.

Speyer, J., Chung, W., 2008. Stochastic processes, estimation, and control. Society for Industrial and Applied Mathematics, Philadelphia.

Sun, X., Muñoz, L., Horowitz, R., 2003. Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control, *Proceedings of the 42nd IEEE Conference on Decision and Control*, Hawaii. December, 6333-6338.

Sun, X., 2005. Modeling, estimation, and control of freeway traffic. Ph.D dissertation, University of California, Berkeley.

Szeto, W., Lo, H., 2006. Dynamic traffic assignment: properties and extensions. Transportmetrica, 2 (1), 31-52.

Szeto, W.Y., 2008. The enhanced lagged cell transmission model for dynamic traffic assignment, Transportation Research Record, 2085, 76-85.

Tam, M., Lam, W., 2008. Using automatic vehicle identification data for travel time estimation in Hong Kong. Transportmetrica, 4 (3), 179-194.

Tampère, C., Corthout, R., Cattrysse, D., Immers, L., 2011. A generic class of first order node models for dynamic macroscopic simulation of traffic flows. Transportation Research Part B, 45 (1), 289-309.

Timmermann, A., 2000. Moments of Markov switching models. Journal of Econometrics, 96, 75-111.

Tong, C., Wong, S., 2009. Heuristic algorithms for simulation-based dynamic traffic assignment. Transportmetrica, In Press.

The Economist, 1996. Living with the car, 22, 3-18.

Tsekeris, T., Voß, S., 2009. Design and evaluation of road pricing: state-of-the-art and methodological advances. Netnomics, 10 (1), 5-52.

Tu, H., 2008. Monitoring travel time reliability on freeways. TRAIL Thesis Series, The Netherlands.

Tuan, P., 1985. Bilinear Markovian representation and bilinear models, *Stochastic Processes and Their Applications*, 20, 295-306.

Varaiya, P., 2008. Congestion, ramp metering and tolls. Philosophical Transactions of Royal Society (A), 366, 1921-1930.

Viti F., Van Zuylen, H.J., 2009. The dynamics and the uncertainty of queues at fixed and actuated controls: a probabilistic approach. Journal of Intelligent Transportation Systems, 13(1), 39-51.

Viti, F., van Zuylen, H.J., 2010. Probabilistic models for queues at fixed control signals. Transportation Research Part B, 44(1), 120-135.

Vardoulakis, S., Fisher, B., Gonzalez-Flesca, N., Pericleous, K., 2002. Model sensitivity and uncertainty analysis using roadside air quality measurements. Atmospheric Environment, 36, 2121-2134.

Vardoulakis, S., Fisher, B., Pericleous, K., Gonzalez-Flesca, N., 2003. Modelling air qualityin street canyons: a review. Atmospheric Environment, 37, 155-182.

Vardoulakis, S., Valiantis, M., Milner, J., ApSimon, H., 2007. Operational air pollution modelling in the UK–street canyon applications and challenges. Atmospheric Environment, 41, 4622-4637

Vickrey, W., 1969. Congestion theory and transport investment. American Economics Review, 59, 251-260.

Vlahogianni, E., Golias, J., Karlaftis, M., 2004. Short-term traffic forecasting: overview of objectives and methods. Transport Reviews, 24 (5),533-557.

Wang, H., Li, J., Chen, Q., Ni, D., 2009. Speed-Density Relationship: from Deterministic to Stochastic. Pre-print CD-ROM, the 88th Transportation Research Board (TRB) Annual Meeting, Washington, D.C.

Wang, Y., Lindsey, R., Yang, H., 2010. Nonlinear pricing on private roads with congestion and toll collection costs. Transportation Research Part B, In press.

Wang, Y., Papageorgiou, M., Messmer, A., 2003. A predictive feedback routing control strategy for freeway network traffic. Transportation Research Record, No. 1856, 62-73.

Wang, Y., Papageorgiou, M., 2005. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transportation Research Part B, 39, 141-167.

Wang, Y., Papageorgiou, M., 2006a. Local ramp metering in the case of distant downstream bottlenecks. In Proceedings of IEEE Intelligent Transportation Ssystems Conference, Toronto, ON, Canada, 426-431.

Wang, Y., Papageorgiou, M., 2006b. RENAISSANCE—a unified macroscopic model-based approach to real-time freeway network traffic surveillance. Transportation Research Part C, 14, 190-212.

Wang, Y., Papageorgiou, M., Messmer, A., 2007. Real-time freeway traffic state estimation based on extended Kalman filter: a case study. Transportation Science, 41, 167-181.

Wang, Y., Papageorgiou, M., Messmer, A., 2008. Real-time freeway traffic state estimation based on extended Kalman filter: adaptive capabilities and real data testing. Transportation Research Part A, 42, 1340-1358.

Wang, Y., Papageorgiou, M., Messmer, A., Coppola, P., Tzimitsi, A., Nuzzolo, A., 2009a. An adaptive freeway traffic state estimator. Automatica, 45 (1), 10-24.

Wang, Y., Papageorgiou, M., 2009b. A joint state and parameter estimation approach to freeway traffic state estimation, incident alarm, and detector fault diagnosis. Transportation Research Board 88th Annual Meeting January 11-15, Washington D. C.

Webster, F.V., 1958. Traffic signal settings. Paper No. 39, Road Research Lab., Her Majesty Stationary Office, London.

Wen, Y., 2008. Scalability of dynamic traffic assignment. PhD Thesis, Massachusetts Institute of Technology.

Wen, Y., Balakrishna, R., Ben-Akiva, M., Smith, S., 2006. Online deployment of dynamic traffic assignment: architecture and run-time management. IEE Proceedings Intelligent Transport Systems, 153 (1), 76-84.

Wie, B., 1993. A differential game model of Nash equilibrium on a congested traffic network. Networks, 23, 557-565.

Wie, B., 1995. A differential game approach to the dynamic mixed behavior traffic network equilibrium problem. European Journal of Operational Research, 83(1), 117-136.

Wie, B., Tobin, R., 1998. On the relationship between dynamic Nash and instantaneous user equilibria. Networks, 32, 141-163.

Wie, B., Tobin, R., Carey, M., 2002. The existence, uniqueness and computation of an arc-based dynamic network user equilibrium formulation. Transportation Research Part B, 36 (10), 897-918.

Xu, S., Chen, T., 2005. Robust H_∞ control for uncertain discrete-time stochastic bilinear systems with Markovian switching. International Journal of Robust and Nonlinear Control, 15, 201-217.

Xu, Y.W., Wu, J.H., Florian, M., Marcotte, P., Zhu, D.L., 1999. Advances in the continuous dynamic network loading problem. Transportation Science, 33 (4), 341C353.

Yang, H., Meng, Q., 1998. Departure time, route choice and congestion toll in a queuing network with elastic demand. Transportation Research Part B, 32 (4), 247-260.

Yang, H., Bell, M., Meng, Q., 2000. Modeling the capacity and level of service of urban transportation networks. Transportation Research Part B, 34 (4), 255-275.

Yang, H., Meng, Q., Lee, D., 2004. Trial-and-error implementation of marginal-cost pricing on networks in the absence of demand functions. Transportation Research Part B, 38 (6), 477-493.

Yang, H., Huang, H., 2005. Mathematical and economic theory of road pricing. Elsevier, Amsterdam.

Yang, H., Xu, W., He, B., Meng, Q., 2010. Road pricing for congestion control with unknown demand and cost functions. Transportation Research Part C, 18 (2), 157-175

Yao, T., Friesz, T., Wei, M, Yin, Y., 2010. Congestion derivatives for a traffic bottleneck. Transportation Research Part B, In Press.

Yanmaz-Tuzel, O., Özbay, K., 2009. Modeling Learning Impacts on Day-to-Day Travel Choice. Transportation and Traffic Theory 2009: Golden Jubilee, Eds: William H.K. Lam, S.C. Wong, 387-403, Springer.

Yazici M., Özbay, K., 2007. Impact of probabilistic road capacity constraints on the spatial distribution of hurricane evacuation shelter capacities. Journal of Transportation Research Record, No. 2022, 55-62.

Yin, Y., Lawphongpanich, S., 2006. Internalizing emission externality on road networks. Transportation Research Part D, 11, 292-301.

Yoon, M., Ugrinovskii, V., Petersen, I., 2005. On the worst-case disturbance of minimax optimal control. Automatica, 41, 847 -855.

Zhang, H., Ritchie, S., Recker, W., 1996. Some general results on the optimal ramp control Problem, Transportation Research Part C, 4 (2), 51-69.

Zhang, H., Recker, W., 1999. On optimal freeway ramp control policies for congested traffic corridors. Transportation Research Part B, 33 (6), 417-436.

Zhang, H., Nie, Y., 2005. Modeling network flow with and without link interactions: properties and implications, 2005 TRB Conference, CD-ROM.

Zhang, H., Shen, W., 2010. Access control policies without inside queues: Their properties and public policy implications Transportation Research Part B, In Press.

Zhang, L., Boukas, E., 2009a. Stability and stabilization of Markovian jump linear systems with partly unknown transition probabilities. Automatica, 45, 463-468.

Zhang, L., Boukas, E., 2009b. Mode-dependent H_∞ filtering for discrete-time Markovian jump linear systems with partly unknown transition probabilities. Automatica, 45, 1462-1467.

Zhang, Y., Lv, J., Ying, Q., 2010. Traffic assignment considering air quality. Transportation Research Part D, in press, doi:10.1016/j.trd.2010.04.011.

Zhong, R., Sumalee, A., Lam, W.H.K, 2010. Dynamic user equilibrium with side constraints for a traffic network: theoretical development and numerical solution algorithm, submitted to Transportation Research Part B.

Zhou, C., 2002. Stochastic dynamic traffic assignment for robust on-line operations under real-time information system. PhD dissertation, Purdue University.

Zhu, D., Marcotte, P., 2000. On the existence of solutions to the dynamic user equilibrium problem. Transportation Science, 34 (4), 402-414.

Ziliaskopoulos, A., 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. Transportation Science, 34 (1), 37-49.