Traffic modeling, estimation and control for large-scale congested urban networks

THÈSE N^{O} 6460 (2014)

PRÉSENTÉE LE 5 DÉCEMBRE 2014

À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT
LABORATOIRE DE SYSTÈMES DE TRANSPORTS URBAINS
PROGRAMME DOCTORAL EN GÉNIE CIVIL ET ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Mohsen RAMEZANI GHALENOEI

acceptée sur proposition du jury:

Prof. S. Takahama, président du jury
Prof. N. Geroliminis, directeur de thèse
Prof. C. N. Jones, rapporteur
Prof. J. Laval, rapporteur
Prof. M. Papageorgiou, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2014

Je sais que je ne sais rien.

To Mehdi, Akram, and Parastoo with all my love . . .

Acknowledgements

I am extremely indebted to Nikolas (Prof. Nikolas Geroliminis), my thesis adviser, for his invaluable guidance, scientific enthusiasm, believe in high-standard research approach, continuous support, endless encouragement, and most importantly for offering me an opportunity to carry out this thesis on original and challenging topics with creative freedom. He is a great scientist and a good friend and I am immensely grateful to know him and work with him.

I would like to thank the members of my jury, Prof. Satoshi Takahama, Prof. Colin Jones, Prof. Jorge Laval, and Prof. Markos Papageorgiou for taking the time to review this dissertation and for providing valuable comments and discussions. I would like also to acknowledge the Swiss National Science Foundation (Grant #200021-130165) for funding this dissertation.

I feel extremely lucky to be a member of LUTS, a friendly, relaxing, yet motivating working place. Many thanks to LUTS caring and helpful secretary, Christine Droguet. I would like to thank my current and former LUTS colleagues: Konstantinos Ampountolas, Burak Boyaci, Jack Haddad, Yuxuan Ji, Anastasios Kouvelas, Raphael Lamotte, Wei Liu, Mikhail Murashkin, Mohammadreza Saeedmanesh, Mehmet Yildirimoglu, and Nan Zheng. It has been a pleasure to know you and work with you. My appreciation goes to Raphael for translating the English abstract to French résumé. Moreover, I would like to express my sincere gratitude to Jack (Prof. Jack Haddad), with whom I had the opportunity to have close and fruitful collaboration on some part of this thesis.

I am grateful to my friends for making my life more fun, enjoyable, and exciting. My special thanks to Mostafa Ajallooeian, Parastoo Alizadeh, Amir Firouzeh, Asiye Meskin, and many others with whom I shared many great memories in Lausanne. I also thank my distant friends, Behzad Dehlaghi Jadid, Mahdi Izadkhast, Vahid Jabi, Sina Moharramzadeh, Mohsen Rezaei, Hosein Taghavi, and Saber Taghvaeeyan.

Finally, my deepest gratitude goes to Mehdi and Akram, my parents, for their unconditional love and endless support through all these 28 years. Thank you for everything. This work is dedicated to you. I would also like to thank Fahimeh, my sister. And most importantly, to Parastoo, my love, for all her boundless love, understanding, kindness, patience, support, and encouragement. Thank you for your being, all that we had, and we will have together. This work is dedicated to you.

Lausanne, November 12, 2014
Mohsen Ramezani

Abstract

NOWADAYS, urbanization has become a universal trend resulting in growing cities and congested transportation networks all around the world. For congestion alleviation, constructing new infrastructures is not a sustainable solution since urban development and transportation infrastructures compete for the same space resource. Alternatively, the development of Intelligent Transportation Systems (ITS) technologies with new monitoring paradigms and computational tools enables to estimate traffic states in real-time and implement traffic-responsive control schemes.

However efficient management/control of transportation systems remains a big challenge, due to the complexity and extent of transportation networks and unpredictability of travelers choices (in terms of route, departure time, and mode of transport). This dissertation proposes novel methods to ameliorate traffic congestion in large-scale urban transportation networks with focus on improvement of performance of existing transportation infrastructure.

Part I of the thesis investigates novel urban traffic state estimation methods utilizing probe vehicle data. Analysis and monitoring of operating conditions on signalized arterials and networks requires estimates of performance measures that is beneficial for both system operators and travelers. The development of new sensing hardware and integration of global positioning system (GPS) technology within the ITS framework promise a new paradigm in traffic surveillance: probe vehicles, which compared to fixed traffic sensors, e.g. loop detectors, offer richer data, e.g. vehicle trajectories. Based on probe vehicle data, Chapter 2 proposes a method to integrate the collective effect of dispersed probe data with traffic kinematic wave theory and data mining techniques to model the spatial and temporal dynamics of queue formation and dissipation in arterials. The queue estimation method captures interdependencies in queue evolutions of successive intersections, and moreover, the method is applicable in oversaturated conditions and includes a queue spillover statistical inference procedure.

Chapter 3 develops a travel time reliability model to estimate arterial route travel times distribution (TTD) considering spatial and temporal correlations between traffic states in consecutive links. The arterial route TTD contains more information regarding arterial performance measurements and travel time variability. The model uses link-level travel time data and a heuristic grid clustering method to estimate the state structure and transition probabilities of a Markov chain. By applying the Markov chain procedure, the correlation between states of successive links is integrated and the route-level TTD is estimated. The methods in Part I are tested with various probe vehicle penetration rates on case studies with field measurements and simulated data. The methods are straightforward in implementation and have demonstrated

promising performance and accuracy through numerous experiments.
Part II studies network-level modeling and control of large-scale urban networks. Over decades, control strategies based on disaggregated traffic flow models have been proposed for isolated intersections or coordinated intersections in arterials. However, microscopic modeling of traffic flow dynamics and centralized control schemes for a large-scale network are computationally intractable and unstable. On the other hand, network traffic models, e.g. the macroscopic fundamental diagram (MFD), aim at simplifying the micro-modeling of urban networks in which the collective traffic flow dynamics of subnetworks capture the main characteristics of traffic congestion. Recent analysis of empirical data from cities showed that MFD provides for homogenous urban regions a unimodal low-scatter relationship between network vehicle density and network space-mean flow. Chapter 4 is the pioneer that introduces the urban perimeter control for two-region urban cities as an elegant control strategy to improve mobility and decrease delays in large urban networks. Perimeter controllers operate on the border between the two regions, and manipulate the percentages of transfer flows between the two regions, such that the number of trips reaching their destinations is maximized. The optimal perimeter control problem is solved by the model predictive control (MPC) scheme, where the prediction model and the plant (reality) are formulated by MFDs.
Chapter 5 extends the perimeter control strategy and MFD modeling to mixed urban-freeway networks to provide a holistic approach for large-scale integrated corridor management (ICM). The network consists of two urban regions, each one with a well-defined MFD, and a freeway, modeled by the asymmetric cell transmission model, that is an alternative commuting route which has one on-ramp and one off-ramp within each urban region. Perimeter controllers on the border of the urban regions operating to manipulate the interflow between the two regions, and controllers at the on-ramps for ramp metering are considered to control the flow distribution in the mixed network. The optimal traffic control problem is solved by the MPC approach to minimize total delay in the entire network considering several control policies with different levels of urban-freeway control coordination.
Chapter 6 integrates traffic heterogeneity dynamics in large-scale urban modeling and control to develop a hierarchical control strategy for heterogeneously congested cities. Two aggregated models, region- and subregion-based MFDs, are introduced to study the effect of link density heterogeneity on the scatter and hysteresis of MFD. A hierarchical perimeter flow control problem is proposed to minimize the network delay and to homogenize the distribution of congestion. The first level of the hierarchical control problem is solved by the MPC approach, where the prediction model is the aggregated parsimonious region-based MFD and the plant is the subregion-based MFD, which is a more detailed model. At the lower level, a feedback controller tries to maximize the network outflow, by increasing regional homogeneity.

Keywords: *Cooperative decentralized control, Coordinated urban-freeway control, Heterogeneity modeling, Hierarchical control, Integrated corridor management, Macroscopic fundamental diagram, Model predictive control, Perimeter control, Probe vehicle data, Queue profile estimation, Spillover identification, Traffic hysteresis, Travel time distribution, Travel time variability*

Résumé

LE phénomène d'urbanisation observé actuellement à l'échelle mondiale a pour conséquences des villes toujours plus grandes et des réseaux de transports toujours plus encombrés. La construction de nouvelles infrastructures pour soulager celles existantes n'est pas une solution viable à long terme car l'espace urbain est limité. Cependant, le développement des Systèmes de Transports Intelligents (STI) ainsi que de nouveaux outils de surveillance et de calcul permettent d'estimer en temps réel l'état du trafic et de mettre en place des stratégies de contrôle adaptatives.

Malgré ces avancées, la gestion des systèmes de transports pose toujours de nombreux défis, du fait du caractère imprévisible des usagers (en termes d'itinéraire, d'horaire et de mode de transport) et de la complexité et de l'étendue des réseaux. Cette thèse propose de nouveaux outils pour réduire la congestion dans les grands réseaux de transports urbains et met l'accent sur une meilleure utilisation des infrastructures existantes.

La première partie de cette thèse est consacrée à des méthodes innovantes d'estimation du trafic basées sur des données provenant de véhicules en mouvement. L'analyse et la surveillance des conditions routières nécessitent l'estimation de diverses mesures de performance, à la fois du point de vue de l'opérateur du réseau et du point de vue des usagers. Le développement de nouveaux capteurs mobiles et l'intégration des outils de localisation par satellite (GPS) dans le cadre des STI ont un immense potentiel. Par rapport à des capteurs conventionnels (par exemple des boucles dans le sol), ces capteurs fournissent des donnés bien plus riches, nous informant sur la trajectoire complète des véhicules. En utilisant ces techniques, le Chapitre 2 propose une méthode basée sur la théorie de propagation du trafic et la fouille de données pour agréger l'ensemble des informations provenant des véhicules dispersés dans le réseau et modéliser les dynamiques spatiales et temporelles de formation des queues et de leur dissipation dans les artères urbaines. La méthode d'estimation des queues identifie les interdépendances dans les évolutions des queues à des intersections successives et reste applicable dans des conditions de sursaturation, grâce à une modélisation statistique des débordements de queues.

Dans le Chapitre 3, un modèle est proposé permettant d'estimer la distribution des temps de parcours dans les artères urbaines. La distribution de ces temps de parcours est particulièrement utile pour estimer la performance de ces artères ainsi que la variabilité de ces performances. Le modèle a été obtenu en considérant les corrélations spatiales et temporelles entre les états de liens consécutifs. Il utilise les temps de parcours de chaque lien et une méthode heuristique d'agrégation pour obtenir une structure d'états. Les corrélations entre

les états de liens successifs sont estimées via une chaîne de Markov, permettant ainsi d'obtenir la distribution des temps de parcours. Les méthodes de cette première partie sont testées avec différents taux de pénétration des véhicules sondes, à la fois dans des études de cas basées sur des données réelles et avec des simulations. Ces méthodes peuvent être aisément mises en place et ont obtenu d'excellents résultats dans de multiples expériences.
La deuxième partie de cette thèse est consacrée à la modélisation des réseaux dans leur ensemble et au contrôle à grande échelle des réseaux urbains. Au cours des dernières décennies, des stratégies de contrôle basées sur des modèles désagrégés du trafic ont été proposées pour les intersections isolées et pour la coordination des intersections d'artères urbaines. Cependant, la modélisation microscopique du trafic et les stratégies de contrôle qui lui sont associées deviennent extrêmement instables et incommodes à grande échelle du fait leur complexité. En revanche, des modèles spécifiques ont été développés pour les réseaux, comme le Diagramme Macroscopique Fondamental (MFD). Le but de ces modèles est de résumer l'état de l'ensemble du réseau par certaines grandeurs caractéristiques. Des analyses récentes de données empiriques provenant de différentes villes ont montré que le MFD permet d'obtenir une relation bien définie entre la densité de véhicules dans le réseau et la moyenne spatiale du flux de véhicules. Le Chapitre 4 présente une stratégie nouvelle et élégante de contrôle du périmètre urbain pour une ville séparée en deux régions, destinée à fluidifier la circulation et à réduire les délais. Les contrôleurs opèrent aux frontières de ces régions et régulent les flux d'échange entre ces régions, de manière à maximiser le nombre d'usagers servis. Ce problème d'optimisation est résolu à l'aide de la technique de commande prédictive, utilisant comme image du monde réel un MFD et comme outil de prédiction un MFD similaire mais simplifié. Le Chapitre 5 étend cette stratégie de contrôle du périmètre à un réseau mixte, constitué de deux régions urbaines et d'une autoroute, qui représente un itinéraire alternatif entre les régions. Les régions urbaines sont modélisées à l'aide de MFDs tandis que l'autoroute est modélisée par un modèle de transmission asymétrique de cellule à cellule. Aux contrôleurs situés à l'interface entre les régions s'ajoutent ici des contrôleurs sur les rampes d'accès à l'autoroute. La technique de commande prédictive est utilisée afin de minimiser la somme des délais occasionnés dans le réseau et différents niveaux de coopération entre les régions et l'autoroute sont envisagés.
Le Chapitre 6 intègre l'évolution de l'hétérogénéité dans la modélisation et le contrôle à grande échelle et développe une stratégie de contrôle hiérarchique pour les villes où la congestion est hétérogène. Deux types de MFDs -au niveau régional et au niveau infrarégional- sont présentés pour étudier l'effet de liens avec différents niveaux de congestion sur la dispersion des données et le phénomène d'hystérésis dans le MFD. Une stratégie de contrôle du périmètre hiérarchique est proposée pour minimiser la somme des délais et homogénéiser la distribution de la congestion. Le premier niveau de ce problème est résolu par la technique de commande prédictive tandis qu'un contrôleur à rétroaction est utilisé au niveau inférieur afin de maximise le flux total dans le réseau, en le rendant plus homogène.

Mots-clés : *Contrôle coopératif décentralisé, Contrôle coordonné mixte urbain/autoroutier, Modélisation de l'hétérogénéité, Contrôle hiérarchique, Gestion intégrée des couloirs routiers, Diagramme Macroscopique Fondamental, Commande Prédictive, Contrôle de périmètre, Vé-*

hicules sondes, Estimation des profils de queues, Identification des débordements de queues, Hystérésis du trafic, Distribution des temps de parcours, Variabilité des temps de parcours.

Table of Contents

List of Figures

List of Tables

1 Introduction

TRAFFIC congestion is a prevalent phenomenon in cities that has bidirectional cause and effect relation with human beings' daily social behaviour. With growing transportation demand and urbanization trend, severe traffic congestion is imminent in metropolitan areas that requires thorough and effective investigations for congestion alleviation. To this end, numerous traffic management policies and technologies have been devised and implemented to improve traffic mobility and accessibility. This dissertation proposes and develops new methods based on aggregated traffic models for traffic state estimation and control in large-scale urban networks.

Recently, cities are becoming smarter in ways that enable us to monitor, analyze, and plan to improve the quality of life. This requires a fundamental consideration in modeling the transport network and its connection with the social behaviour of travelers through a multidisciplinary combination of applied physics, engineering, and data mining researches. Moreover, recent advances in user-based data collection systems, i.e. crowdsourcing, and emergence of "big data" in transport networks offer an opportunity to advance the mobility of cities through the integration of big data, understanding of mobility patterns interactions, and coordination and optimization of urban traffic systems.

In this introductory chapter, Section 1.1 explains the thesis motivation and briefly introduces the research background in the current literature, while a corresponding detailed state-of-the-art literature review is provided later at each chapter. Section 1.2 overviews the objectives that the thesis follows to achieve. Section 1.3 elaborates the main contributions of our study corresponding to each chapter. And finally, Section 1.4 outlines the structure of the thesis.

1.1 Research motivation and background

Traffic congestion becomes apparent on road networks when the transport demand exceeds the infrastructure supply capacity, which is characterized by high vehicle density on roads, low traffic flows, and long trip travel times. Traffic congestion imposes several adverse effects

on urban communities, such as unproductive wasted time, air and sound pollution, stressful environment, etc. Due to the significant functionality of transportation networks in modern societies, the economical aspects of transport are of great importance for local and federal governments. For instance, the Texas transportation institute estimated that the cost of congestion in US was more than $120 billion in 2011 [83]. Furthermore, congestion annually costs Europe about 1% of its gross domestic product (GDP) that is projected to increase by about 50% to nearly €200 billion by 2050 [23]. Hence, the significance of transportation systems in the framework of sustainable and smart cities cannot be overestimated.

To alleviate traffic congestion, numerous potential resolutions have been taken into consideration. The following provides a few instances of such policies. (i) Increasing and modifying the usage of the road infrastructure, e.g. high occupancy vehicle (HOV) lanes [122, 114] and dynamic intermittent public transit lanes [39, 27]; (ii) traffic supply and demand management, e.g. congestion pricing [197, 174, 205], car sharing [24, 15], and parking management [167, 173]; and (iii) traffic control strategies such as coordinated signal control [124, 72], ramp metering [136, 57], route guidance [64], etc. For a detailed overview of road traffic control strategies see [135].

However due to the complexity and extent of transportation networks, efficient management and control of transportation systems remains a big challenge. Moreover, the current traffic control schemes are fragmented because (i) there are multiple jurisdictions responsible for different parts of the network and (ii) the modeling of fully centralized control systems is complex and computationally intractable. Hence, the aim of the thesis is to develop a holistic approach for large-scale urban traffic estimation and control based on three interconnected foundations: (i) understanding interactions of mobility (*modeling*), (ii) monitoring traffic states and measuring traffic data (*estimation*), and (iii) optimizing the transportation system (*control*). The thesis proposes traffic state estimation methods (Chapters 2 and 3) along with a hierarchical control structure (Chapters 4, 5, and 6) to address the complexity and extent of urban networks. Note that in the following, a brief research background related to the objectives of the thesis is presented, while a detailed state-of-the-art literature review specifically for every research problem is presented at each chapter.

The scope of Chapters 2 and 3 is traffic state estimation in urban networks. The basic ingredient for the new wave of smart cities that has emerged during the last decade is massive data sets concerning human mobility, fostered by the widespread distribution of sensors, such as Global Positioning System (GPS) devices in many modes of transport and smart phones. These tools allow for sensing and collecting substantial spatiotemporal data, such as the trajectories of many vehicles from navigation devices, which represent proxies for human mobility patterns. This “big mobility data” provides a unique social observatory that can help us understand how congestion develops and evolves, and discover hidden patterns and identify models that can contribute to efficient traffic management systems to improve cities’ mobility and accessibility.

Transportation data can be collected in real-time from various types of sensors, e.g. fixed loop detectors [169, 56, 194], video sensors [82], and automatic vehicle location (AVL) mobile sensors, which are in abundance in taxi and transit fleets [80]. To obtain precise traffic measurements and state estimations, various methodologies have been exploited, e.g. (Extended, Unscented) Kalman filter from estimation theory [186] and classification and clustering methods from data mining field [8, 9]. Recently, the emergence and steady increase of public deployment of user-based data collection systems, e.g. GPS-equipped vehicles [79, 14, 88] and vehicle re-identification using cellphones or wireless magnetic sensors [105], provide a great potential for probe vehicles in intelligent transportation systems (ITS) applications. However, utilizing probe vehicles for traffic observation and estimation raises challenges because of (i) errors in raw measurements and post-processing algorithms, e.g. map-matching [146], and (ii) the fact that the probe data are inherent samples of the true traffic state. Thus, probe data cannot be readily applied to well-established traffic estimation methods.

A crucial feature of traffic flow in urban networks is vehicle queues at intersections. An accurate and practical queue estimation method (i) should provide a better understanding of urban flow dynamics, (ii) can be utilized for traffic state estimation, and (iii) can be integrated in a traffic signal control framework. Literature of queue estimation can be categorized to two modeling classes: (i) models based on the cumulative traffic input-output [188, 6, 40, 183], with the limitation that these models are insufficient to provide the spatial distribution of queue dynamics [123] and suffer from measurement errors and bias [182], and (ii) models based on kinematic shockwave theory [169, 10, 193], which provide the temporal-spatial dynamics of the queuing process with input data from loop detectors or probe vehicles. Further, many of aforementioned works and pivotal studies of queue and delay estimation in transportation research [128, 34, 127, 121] are based (fully or partially) on rather limiting assumptions. These assumptions can be summarized as: (i) known signal settings, (ii) known initial queue size at the start of the cycle, (iii) known arrival pattern (e.g. uniform, Poisson), and (iv) undersaturated conditions that expected value of arrivals do not (constantly) exceed the signal capacity [36].

The motivation of the thesis with respect to the traffic estimation is to propose a queue shockwave estimation method that is based on probe data as the single source of information while relaxing the aforementioned common limiting assumptions in the literature. By estimating the queue shockwave profile for each cycle, one can derive performance measures including queue length, position of back and front of the queue, and delays per vehicle (average and distribution) that can be further used to estimate travel time distributions as a travel time reliability index [151].

Furthermore for traffic state estimation, another research direction of the thesis is toward the trip travel time distribution estimation. Travel time is a crucial index in assessing the operational efficiency of traffic networks. It establishes a common perception among the travelers and practitioners, while it can be an indicator of congestion level of transport network once compared to the free flow travel time. However, a single-value estimator of travel time does not convey the full picture of travel time characteristics. Thus, the aim of the dissertation is to

develop methods for travel time distribution estimation to capture the travel time variability in signalized arterials.

With respect to monitoring, reliable and efficient estimation of travel time is still not a wide spread accomplishment on arterials, since it requires extensive sensor infrastructure, normally found only on freeway networks. The issue is not only that the existing monitoring infrastructure in arterials is less dense than in freeways but also that arterial network traffic dynamics are inherently different than these of freeways and fixed sensors cannot always provide the required level of data. The main reasons are randomness in supply and demand of the dynamic urban network [38], the signaling effect (alternation of green and red phases in short time intervals), and the characteristics of route choice (vehicles in arterials can randomly turn at intersections and either begin or finish their trips along the street itself, which is not the case in freeways). Meanwhile, speed of vehicles at a given time in the network is not a deterministic quantity over space because of drivers' behaviors (conservative vs. aggressive drivers), the spatial effect of signals (near the stop line vs. further upstream) and temporal-spatial pockets, where average speed is temporarily different than the widespread average.

Moreover, reduction in travel time variability is at least as desirable as reduction in mean travel time for travelers [87], since it decreases commuting stress and uncertainty of mode- and route-choice decision making. Travel time variability designates the variation of various trip travel times over a specific path. Travel time variability can be investigated from several point of views [130]: vehicle-to-vehicle variability which corresponds to different vehicles traveling the same route at the same time, period-to-period variability corresponding to vehicles traveling the same route at different periods within a day, and day-to-day variability addressing the travel time variations of vehicles crossing the same route at the same period of time on different days. Different indexes of travel time stochasticity-reliability are presented in [93]. This dissertation progresses a further step in modeling the vehicle-to-vehicle travel time variability and analyzes the probability *distribution* of travel time for arterial routes (expressed as series of links). It is worth mentioning that, the thesis explicitly addresses the spatial and temporal correlations between traffic states of links, which has a significant impact on the estimation performance.

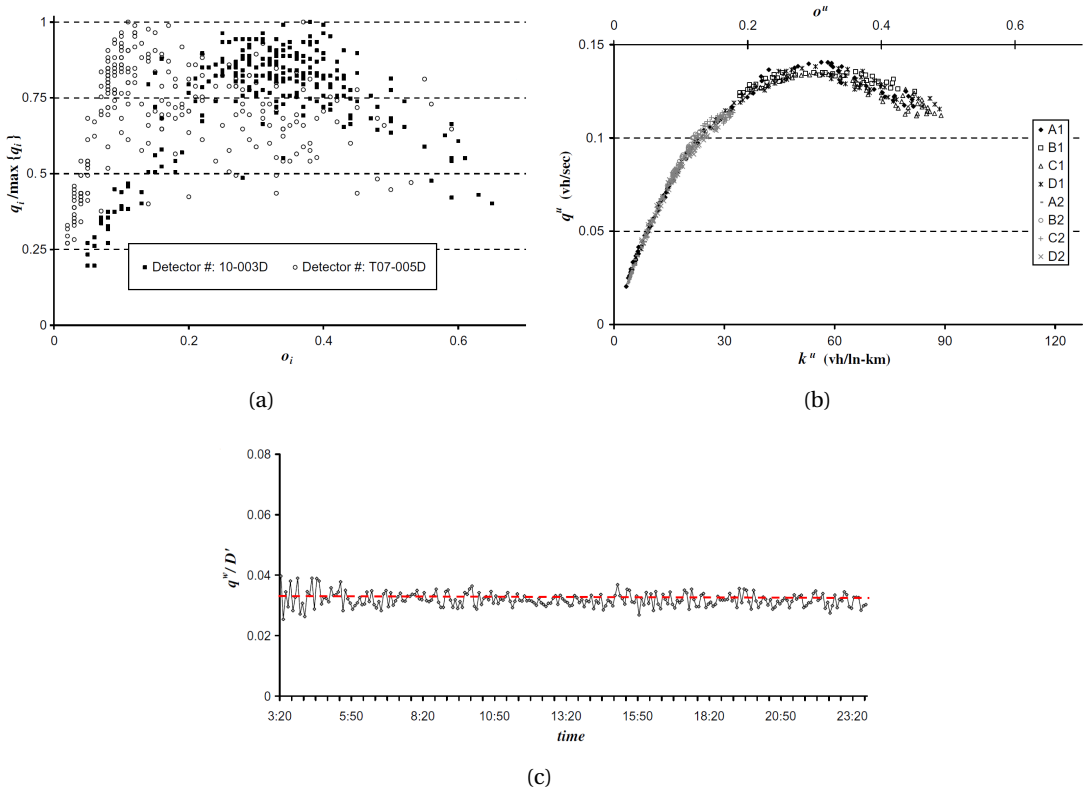
The scope of Chapters 4, 5, and 6 is large-scale traffic modeling and control. Traffic flow modeling mainly is developed in analogy to physical laws inspired from fluid mechanics [110, 163] and many-particle systems (e.g. Gas-kinetic models) [144, 76]. However, the key difference is that in vehicular traffic systems humans show different driving behaviours and make different choices in terms of routes, mode, etc. that create additional complexity. The common feature of traffic models is the fundamental diagram (FD) [62] that relates the traffic flow to traffic speed and density in a link. A short list of main traffic models is presented in the following. Car-following models that are based on the dynamics of single vehicles and their interactions with other vehicles [142, 48, 95, 190]. Later, cellular-automata models [26, 126] are developed to decrease the computational burden of car-following models by discretizing the space, where the road is divided into cells that a cell could be occupied by

a vehicle or not. The cellular-automata models comprise a set of rules to determine the movement of vehicles from upstream cells to downstream ones. The two mentioned model are regarded as microscopic models as they deal with individual vehicles. On other hand, macroscopic models aim at describing the traffic as a continuum by relating aggregated traffic state variables such as average traffic flow and density. First-order flow models such as LWR model [110, 163] are based on a partial differential equation representing the mass conservation law and the FD relation. The examples of discretized and multi-class versions of LWR model are respectively, [29, 106] and [31, 192]. To cover also velocity dynamics for non-steady state conditions along with the mass conservation law, higher-order models are proposed, e.g. see [141, 107] for second-order models and [94] for the three-phase traffic theory. All these models are suitable for link traffic flow, however, the literature on network flow modeling is limited. A comprehensive review of traffic flow models can be found in [180].

For network-level traffic modeling, recently [55] observed from empirical data in Yokohama, that by spatial and temporal aggregation of individual scattered link FDs, i.e. traffic flow vs. traffic density (see Fig. 1.1(a)), a well-defined low-scatter macroscopic fundamental diagram (MFD) exists (see Fig. 1.1(b)). The MFD provides a relation between network space-mean flow and density as opposed to link state variables. Figure 1.1(a) illustrates the scattered FD for two loop detectors while Fig. 1.1(b) shows the Yokohama region low-scatter MFD. The idea of an MFD with an optimum accumulation belongs to Godfrey [60] and similar approaches were introduced later in [78, 116, 28]. The empirical verification of MFD existence with dynamic features is recent [55], while there, it is demonstrated that (i) homogenous urban regions approximately exhibit an MFD relating region vehicle accumulation to flow (or space-mean speed), (ii) an almost linear relation between the region average flow and its outflow (i.e. the trip completion rate, the rate vehicles reach their destinations) exists, which means an almost time-invariant value for the region average trip length, see Fig. 1.1(c), and (iii) the well-defined MFD is a property of the network topology and control and is not very sensitive to the time-dependent demand.

These findings provide new opportunities for traffic network analysis. Property (i) offers a large-scale model to describe the congestion level of cities as details of individual links are not needed. Property (ii) is important as the region flow can be readily observed with different types of sensors, e.g. loop detectors, contrary to the region outflow, which is cumbersome to observe yet essential for control purposes as it designates the output rate of the system. In this thesis, we demonstrate that network average trip length might experience variability with change of congestion level, if the network is not homogenously congested. Further, property (iii) ensures that a detailed knowledge of origin-destination demand is not needed to develop efficient traffic control schemes. Overall, the MFD concept can be utilized to introduce elegant control strategies to improve mobility in cities, e.g. [28, 97, 98], where the core logic of the strategies is to decrease the traffic inflow to regions once the region is congested (represented by the decreasing part of the MFD).

Despite the findings related to the existence of low-scatter MFDs, a well-defined MFD is not



(a) (b) (c)

Figure 1.1: (a) Loop detector flow vs. occupancy data for two single detectors across a day, (b) region average flow vs. region average occupancy (density) from all the detectors across two different days, and (c) time series of ratio of region average flow and outflow. (taken from [55])

universal. In particular, [58, 120, 32] have identified that the spatial distribution of congestion in the network is a key factor related to the shape and scatter of the MFD. Briefly speaking, it is demonstrated that for the same network density, the network average flow is higher when link density variance is lower, meaning that a homogenously congested network is more desirable than the heterogenous network with the same number of vehicles. A solution for heterogeneous networks is to partition the network to form a number of homogeneous regions with smaller variances of link densities [90]. Figure 1.2 shows an application of such partitioning algorithms for a microscopic simulation of a real study site, San Francisco, with time dependent traffic conditions. Figure 1.2(a) illustrates a snapshot of link densities (the darker the more congested links), while Fig. 1.2(b) shows the result of the partitioning algorithm that results in 3 regions. Figure 1.2(c) depicts the MFD of the three regions. Notably, the time each of the regions reaches the congested regime is different. The partitioning outcome of [90] is the basis for the proposed control strategies in this thesis as the MFD can be applied to heterogeneously congested cities once they are partitioned to homogeneous regions. The proposed control strategy offers elegant tools to integrate smart traffic management schemes for large-scale congested networks. Nevertheless, significant effort is needed to apply these control concepts in field implementations.

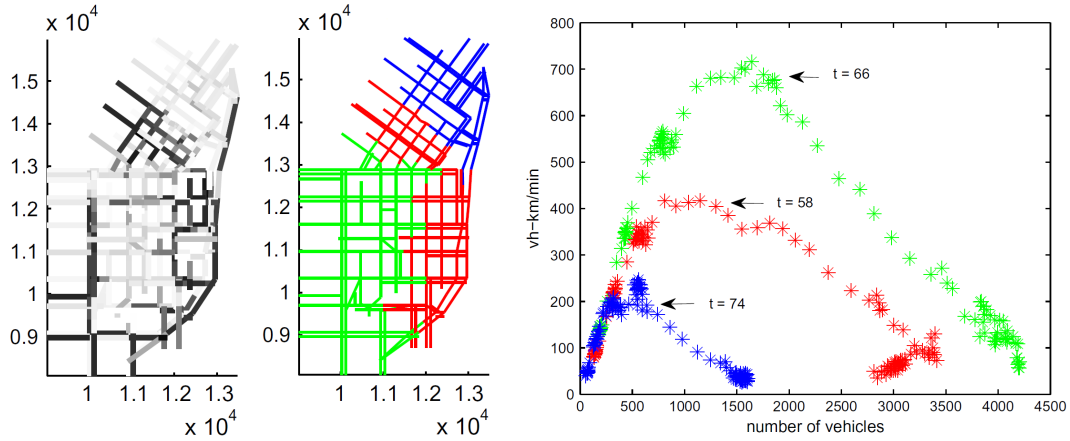


Figure 1.2: The effect of spatial heterogeneity of link density on the MFD. (a) A snapshot of the San Francisco network during the peak hour (dark colors mean congested links), (b) partitioned network into 3 regions, and (c) MFDs for the 3 regions. (taken from [90])

Furthermore, transportation systems consist of multiple interacting modes including pedestrians, non-motorized vehicles, cars, taxis, and more productive modes, such as buses or trams. Despite the different features of motorized modes in terms of passenger occupancy, driving behavior, duration of travel, and scheduled vs. non-scheduled service, their common characteristic is that they all cause delays to the transportation system as a whole. Morover, metropolitan transportation networks are a mixture of urban roads and freeways that are inherently interconnected. Nevertheless, the traffic management of urban-freeway corridors is challenging due to dissimilar urban and freeway traffic dynamics and the diversity in characteristics of urban traffic controllers, i.e. traffic signals; and freeway controllers, i.e. variable speed limits and ramp metering. Although integrated mixed control policy could enhance the network performance, lack of coordination among the urban and freeway network jurisdictions and/or limited means of traffic monitoring and data communication impede such goal. Recently, a large effort for the development of integrated corridor management (ICM) has been promoted, with many case studies around metropolitan areas. Most of the implementations and case studies mainly perform scenario analysis and considers alternative routes under extreme events, e.g. accidents, while it is expected that a more formal optimization approach could lead to a better system performance. This thesis also contributes toward this direction by (i) integrating MFD modeling with a freeway model, (ii) incorporating the novel perimeter flow controller with conventional ramp metering controllers to formulate the control problem of large-scale urban-freeway networks, and (iii) proposing cooperative decentralized control structures as an alternative to tackle possible untractable centralized control problems.

The MFD aims at simplifying the micro-modeling task of the urban network, where the collective traffic flow dynamics of subnetworks capture the main characteristics of traffic congestion, such as the evolution of space-mean flows and densities in different regions of

the city. Nevertheless, if the MFD is applied in a non-careful way, e.g. ignoring the effect of congestion heterogeneity, it can hide critical patterns of congestion and results in inefficient control methods. Hence, the thesis further investigates the relation between the heterogeneity and the MFD by scrutinizing the dynamics of heterogeneity and how it can affect accuracy and scatter of the MFD. While there is some work regarding how heterogeneity influences the shape of the MFD [120, 58, 46, 118, 101], there is no theoretical work to investigate how an asymmetric demand pattern can affect the distribution of congestion over time and space. In the current body of the literature, existing MFD-based models are hysteresis-free and as a result the developed control frameworks based on such models are not efficient when hysteresis appears. To fill the gap in the literature, the thesis integrates the dynamics of heterogeneity in the optimization framework and develops perimeter control strategies that can decrease congestion heterogeneity and consequently increase the system performance. This is a crucial step for efficient control strategies, where strong hysteresis loop in the MFD appears. Such an advanced model also allows to develop a two-level hierarchical control framework that decreases system delays and hysteresis loops, which are directly related to the capacity loss at the network level. The hierarchical structure is of great importance to practitioners and city managers to unveil practical, transparent, and robust control schemes for optimizing the network capacity and serving the maximum number of travelers.

1.2 Thesis objectives

The goal of the thesis is to develop real-time monitoring and control strategies for large-scale congested transportation networks to improve the travelers mobility and the network performance. To this end, the research approach deals with 3 interacting key research areas, large-scale modeling, estimation, and (hierarchical) control of the transport system, i.e. modeling the dynamics of urban traffic at the network level, estimating performance measures for signalized arterials, and developing coordinated control strategies for large-scale heterogeneously congested networks.

In the light of the aforementioned research background and motivation, the specific objectives of this thesis are categorized to two distinct groups, (i) to develop novel traffic state (travel times and queue shockwaves) estimation methods for congested arterials and (ii) to propose a network-level control strategy as the cornerstone of the hierarchical urban traffic control scheme. The detailed objectives of the thesis according to the structure of the chapters are listed as follows:

- Part I: Traffic state estimation
  - The main objective of Chapter 2 is to study a method to estimate queue characteristics in signalized arterials. The method should be based on probe data to exploit the massive data of crowdsourcing and floating sensors new era. The method should provide accurate estimations with sparse probe data and be robust to measurement noise and

data processing errors. The proposed method should be consistent with the physics of traffic, easy to implement, transparent, and above all, applicable to oversaturated traffic conditions when spillovers are apparent. As the estimation problem scope is arterials, the method should relax the literature common assumption of uniform arrival flow because of synchronization among the traffic signals. Furthermore, the method should be able to produce accurate estimates even when signal settings information is missing because obtaining signal settings is not cost-effective and feasible in case of actuated and adaptive signal control strategies at the network scale.

  - The key objective of Chapter 3 is to construct a travel time estimation method based on probe vehicle data for arterial routes. It is essential that the spatial domain of the estimation is *arterial routes*, because commuting trips consist of several links and not only a single link. Moreover, the purpose of estimation method is travel time *distribution* that offers full information of travel time variability instead of only a single mean or variance indexes. Further, the method should capture the correlations between traffic states of successive links to fill the gap in the research literature and shed light on the effect of traffic state correlation on the network performance.

- Part II: Network-level traffic modeling and control

  - Chapter 4 main objective is to develop a novel network-level control strategy to decrease the network total delay. The modeling basis of the method should be consistent with MFD aggregated traffic dynamics. To this end, the network should be assumed to be partitioned to two regions designating city center and suburb parts of the network. Ultimately, the method should be elegant, scalable, and tractable to fit as the upper-level controller of the hierarchical urban traffic control framework. The modeling framework should also be expandable to cities with R-regions (see Chapter 6).

  - The chief objective of Chapter 5 is to integrate the developed urban traffic control strategy in Chapter 4, i.e. perimeter flow control, with the ramp metering strategy for traffic control of large-scale mixed urban-freeway networks. The freeway and the urban network are inherently coupled in metropolitan transportation networks, yet they have dissimilar traffic flow dynamics which challenges the traffic control of mixed networks of two interconnected (urban and freeway) traffic control entities. To this end, the aggregated MFD urban model should be integrated with a freeway model plus a proper route choice procedure to develop the mixed urban-freeway traffic model. As the state and control variables of the problem becomes large, to keep the control problem tractable, a coordinated decentralized control structure should be investigated. This provides insights regarding the impact of data communication manner and level of cooperation between urban and freeway control entities. The analysis should also highlight how restrictions in full coordination can influence the overall mobility objectives.

  - The main objective of Chapter 6 is to further extend the perimeter flow control strategy developed in Chapter 4, to handle heterogenous congested networks. First, the effect of

link density heterogeneity and time-varying average trip length on the MFD modeling should be explored. Then, a lower-level controller should be proposed and augmented with the perimeter controller to decrease the network total delay by homogenizing congestion in the network. This modeling approach brings the opportunity to investigate more rigorously several assumptions in the MFD literature that have been empirically observed, e.g. trip length in a region is about constant, if and how route choice, perimeter control, and O-D affect the heterogeneity and the distribution of congestion. In addition, the perimeter control strategy in Chapter 4 is solved by the model predictive control (MPC) approach, while the optimization model and the plant are inherently the same models. The second objective of this chapter is to introduce two models with different levels of granularity aggregation to utilize them in the MPC framework.

The following section introduces contributions of each chapter according to the aforementioned objectives.

1.3 Thesis contributions

Considering the thesis motivation and objectives, the contributions of the thesis at each chapter are as follows.

- **Chapter 2: Queue profile estimation in congested urban networks with probe data**
  A queue shockwave profile estimation method is proposed to estimate a comprehensive set of urban network performance measures. The list includes delay, queue size, position of back of queue, etc. The queue shockwave profiles designate the queue formation and dissipation that capture the most dominant traffic flow dynamic in the urban networks. The estimation method utilizes probe vehicle data that are a useful and growing monitoring data source because of network-wide coverage and the rapid increase in deployment of floating sensors such as vehicle-based GPS devices and cell phones. The method incorporates the data within an integrated framework of LWR traffic theory and data mining techniques, while there are specific considerations for physics of traffic, longitudinal kinematics, and traffic state interdependencies between adjacent links. The method relaxes the literature common assumption of uniform arrival flow and does not need the explicit knowledge of signal settings for estimation. The method is based on an optimization program that is formulated as a least squares problem, with various constraints related to the physical properties of the LWR model, to provide numerical robustness to noisy measurements. Notably, the significance of the proposed method is that it is applicable to oversaturated conditions and it contains a probabilistic procedure for spillover occurrence inference. The method can be integrated to decompose the travel time between samples of probe data to link travel times, which are utilized in Chapter 3 for route travel time estimation.
- **Chapter 3: Estimation of arterial route travel time distribution with Markov chains**

A travel time distribution (TTD) estimation method based on probe data is developed. The estimation spatial scope is arterials with signalized intersections because a trip in urban networks consists of several consecutive links instead of only one link. Furthermore, a distribution is desirable for travel time estimation since each vehicle encounters a random delay due to the stochasticity of traffic flow and control. The novelty of estimation approach is to address spatial and temporal traffic state correlations in signalized arterials. The approach is to define the traffic states based on the condition of vehicles travel time in a link instead of aggregated traffic density or flow states. This is in line with the intrinsic property of probe vehicle data. Consequently, the method captures the correlation of vehicles travel time states in successive links by incorporating a Markov chain procedure to account for traffic progression and state correlation. A diagram based on the probe data is introduced to represent the joint distributions of successive link travel times that in combination with a heuristic grid clustering algorithm provides state definitions, and initial and transition probabilities of the Markov chain procedure. The method utilizes probe vehicle link travel times to infer the arterial route TTD. The route TTD provides the big picture of travel time variability in arterials that could be utilized for travel time reliability measures. The proposed method demonstrates a sound performance capturing the fundamental characteristics of travel time variability even with sparse probe data.

- **Chapter 4: Model predictive perimeter control for two urban regions with MFDs**
  The chapter proposes a novel large-scale control strategy for urban networks named as perimeter flow control. The network is assumed to be partitioned to two regions designating the city center and suburb areas while the control strategy identifies the inter-transfer flows between the two regions of the network to maximize the whole network output. The perimeter flow control can be realized with the traffic signals at the border of the regions. The underlying traffic modeling is consistent with the MFD aggregated model that provides a relation between region accumulation and outflow, which leads to the 2-region MFD model that is formulated based on the vehicle mass conservation principle. The control problem is tackled by the MPC scheme, where the open-loop optimal control problem is solved using the direct sequential (or control vector parameterization (CVP)) method that converts the optimal control problem into a finite-dimensional nonlinear problem through discretization of the control variables with piecewise constant controls, while the ODEs are embedded in the nonlinear problem. In addition, we investigate two methods based on augmenting a constraint into the optimization problem and a penalty term into the objective function to smooth the control inputs for field implementation considerations.

- **Chapter 5: Cooperative traffic control of a mixed network with two urban regions and a freeway**
  The control problem of a large-scale mixed traffic network, comprising two urban regions and one freeway route is formulated. The perimeter control on the boundary of regions manipulates the inter-transfer flows among regions, while two on-ramp controllers control the traffic flow from urban regions to the freeway. The urban region traffic modeling is based on the MFD (similar to Chapter 4 modeling) and the freeway traffic dynamics are based

on the asymmetric cell transmission model (ACTM). The two different models along with a simple ("current best") route choice procedure are integrated to form the mixed traffic network dynamics. We solve the traffic control problem using the MPC scheme. Since the size of state variables and control inputs becomes large, several control policies with different controller structures and levels of urban-freeway coordination are introduced and investigated. This shed light on the characteristics of decentralized control schemes as an alternative of the centralized control strategy in case of limited data communication or coordination between urban and freeway jurisdictions. A cooperative decentralized MPC approach is proposed that provides promising results. Moreover, the system optimum (SO) route choice is incorporated within the centralized MPC framework that leads to increase in the network performance compared to the user equilibrium condition. This work highlights the significance of coordination (and communication) between the urban and the freeway networks under oversaturated conditions. This study is among the first scientific efforts that provides a rigorous proof-of-concept toward the use of system of systems approach for ICM in network management framework.

- **Chapter 6: Dynamics of heterogeneity in urban networks, aggregated traffic modeling and hierarchical control**
  This chapter moves a further step toward the development of the hierarchical control framework for network-level urban traffic control. The hierarchical structure is of great importance because of the spatial extent and complexity of the urban network. The modeling is an extension of the MFD modeling in Chapter 4, while the network is partitioned to few regions (R-region model instead of 2-region) where each region consists of several subregions. We assume that a well-defined (without hysteresis loop and low-scatter) MFD exists for each subregion. However the superposition of well-defined subregion MFDs does not guarantee a well-defined region MFD as heterogeneity among subregions can create hysteresis. This approach, i.e. modeling the traffic dynamics in two different levels of spatial aggregation (regional and subregional), offers the opportunity to investigate more rigorously the effect of link density heterogeneity on the MFD. Moreover, the characteristics of average trip length, and effect of route choice and perimeter control on the heterogeneity and the distribution of congestion can be further explored. These are challenging research questions that have been raised by many researchers and it is not clear yet under what network conditions MFD provides an accurate representation of network performance. Further based on an analysis of a field dataset, a functional form is developed to model the effect of heterogeneity on subregion and region MFDs. Afterwards, the upper-level control of the hierarchical control structure is solved by the MPC scheme, while the subregion- and region-based models are respectively the plant (reality) and the optimization model in the MPC scheme. This is also an important control contribution. The lower-level of the hierarchical control structure consists of a feedback controller that manipulates the perimeter control inputs for each subregion in the region boundary to actively control the spatial distribution of traffic density in regions. In addition, the subregional receiving boundary capacity is augmented in the subregional MFD traffic model. The significance of heterogeneity modeling and the feedback homogeneity controller in control applications is demonstrated with comprehensive

tests while the importance of efficient control strategies to decrease the magnitude of the hysteresis loop in the MFD is highlighted.

1.4 Thesis outline

This thesis consists of 7 chapters that are briefly described in the following paragraphs. The main 5 chapters (excluding Chapter 1, Introduction, and Chapter 7, Conclusion and future research) are organized into 2 parts. Part I includes Chapters 2 and 3 that propose two novel traffic state estimation methods, respectively, a queue shockwave profile estimation method and a travel time distribution estimation method. The methods are based on probe vehicle data and are applicable to congested arterials. Part II comprises Chapters 4, 5, and 6 that covers the research toward the development of a hierarchical control strategy for large-scale transport networks, which also includes large-scale urban traffic models based on the MFD. Note that each chapter is a complete stand-alone research article including an abstract, introduction, methodology, results, and conclusions with its own (mathematical) notations.

Chapter 2 proposes a novel method to estimate queue shockwave profiles at signalized arterials. The queue shockwaves offer comprehensive dynamics of traffic in the arterial and can be used to estimate numerous traffic states such as link travel time and queue size. The estimation process utilizes the probe vehicle data without need of signal settings or any assumption on arrival traffic flow. The method significance is that it contains a statistical Bayesian inference procedure to identify the probability of spillover occurrence. Thus, the method can be applied in oversaturated regimes while the estimations are accurate even with sparse probe data. The preliminary results of this chapter are presented in:

- M. Ramezani and N. Geroliminis,"Queue profile estimation in signalized arterials by utilizing probe vehicle data," 13th Swiss Transportation Research Conference, Ascona, Switzerland, April 2013, [156].
- M. Ramezani and N. Geroliminis, "Exploiting probe data to estimate the queue profile in urban networks," *16th International IEEE Conference on Intelligent Transportation Systems*, The Hague, The Netherlands, October 2013, [152].
- M. Ramezani and N. Geroliminis, "Queue profile estimation with probe vehicle data for urban arterials," *93rd Annual Meeting of the Transportation Research Board*, Washington D.C., January 2014, [153].

Chapter 2 is a self stand-alone article published as:

- M. Ramezani and N. Geroliminis, "Queue profile estimation in congested urban networks with probe data," *Computer-Aided Civil and Infrastructure Engineering*, DOI: 10.1111/mice.12095, 2014, [157].

Chapter 3 presents a TTD estimation method based on probe data to model travel time reliability. The TTD provides traffic operators and travelers a performance measurement regarding urban network travel time variability. The probe vehicles provide the link travel times to the estimation method that utilizes the data to infer the arterial route TTD. The method captures the correlation of vehicles travel times in successive links that is crucial to improve the estimation performance, which is verified by extensive analyses. The preliminary results of the method are presented in:

- M. Ramezani and N. Geroliminis, "Arterial route travel time distribution estimation with a Markov chain procedure," 11th Swiss Transportation Research Conference, Ascona, Switzerland, May 2011, [155].
- M. Ramezani and N. Geroliminis, "Markov chain procedure for arterial route travel time distribution estimation," *2nd international conference on Models and Technologies for Intelligent Transportation Systems*, Leuven, Belgium, June 2011, [149].
- M. Ramezani and N. Geroliminis, "Estimation of arterial route travel time distribution with Markov chains," *91st Annual Meeting of the Transportation Research Board*, Washington D.C., January 2012, [150].

Chapter 3 is organized as a self stand-alone article published as:

- M. Ramezani and N. Geroliminis, "On the estimation of arterial route travel time distribution with Markov chains," *Transportation Research Part B*, vol. 46, no. 10, pp. 1576-1590, 2012, [151].

Chapter 4 is the pioneer to propose the perimeter flow control strategy as a large-scale traffic control method for urban networks based on the MFD modeling. The network is assumed that is partitioned to two regions, e.g. according to topological attributes (city center and suburb), while the perimeter flow controllers are on the boundary between the regions to manipulate the inter-transfer flows between the regions to maximize the trip completion rate for the whole network. The perimeter flow control strategy can be readily realized by traffic signals on the border of regions. The control problem is tackled with the MPC approach. Chapter 4 preliminary results are presented in:

- J. Haddad, M. Ramezani, and N. Geroliminis, "Model predictive perimeter-control for two-region urban cities," *91st Annual Meeting of the Transportation Research Board*, Washington D.C., January 2012, [67].
- J. Haddad, M. Ramezani, and N. Geroliminis, "Model predictive perimeter control for urban areas with macroscopic fundamental diagrams," *proceedings of IEEE 2012 American Control Conference*, Montreal, Canada, pp. 5757-5762, [68].

Chapter 4 is structured as a self stand-alone article published as:

- N. Geroliminis, J. Haddad, and M. Ramezani, "Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 348-359, 2013, [49].

Chapter 5 extends the perimeter control method in Chapter 4 to large-scale mixed urban-freeway networks by incorporating the ramp metering control strategy with the perimeter control. The mixed network is modeled based on the MFD and ACTM, respectively representing urban regions and the freeway. A centralized controller is achieved by the MPC scheme, while a cooperative decentralized control structure is developed to overcome the possible lack of data communication and coordination between urban and freeway management entities. Further, the study includes a SO routing within the proposed mixed urban-freeway control framework. Chapter 5 preliminary results are presented in:

- M. Ramezani, J. Haddad, and N. Geroliminis, "Macroscopic traffic control of a mixed urban and freeway network," 12th Swiss Transportation Research Conference, Ascona, Switzerland, May 2012, [158].
- M. Ramezani, J. Haddad, and N. Geroliminis, "Macroscopic traffic control of a mixed urban and freeway network," *13th IFAC Symposium on Control in Transportation Systems*, Sofia, Bulgaria, September 2012, [159].
- M. Ramezani, J. Haddad, and N. Geroliminis, "Cooperative traffic management policies for mixed freeway and urban networks," *92nd Annual Meeting of the Transportation Research Board*, Washington D.C., January 2013, [160].

Chapter 5 is a self stand-alone article published as:

- J. Haddad, M. Ramezani, and N. Geroliminis, "Cooperative traffic control of a mixed network with two urban regions and a freeway," *Transportation Research Part B*, vol. 54, pp. 17-36, 2013, [69].

Chapter 6 proposes a hierarchical and large-scale urban control framework based on the MFD modeling and the perimeter control strategy. The control structure comprises a upper-level (MPC-based) controller that based on regional traffic states identifies the inter-transfer flows between the (multiple) regions of the network, and a lower-level (feedback) controller aiming at homogenizing the congestion spatial distribution inside the regions by locally tuning the control actions. The modeling significance of the study is to investigate the effect of density heterogeneity on the MFD and also to introduce two traffic models with different levels of granularity and utilize them as the optimization model and the plant in the MPC scheme. The preliminary results of this chapter are presented in:

- M. Ramezani, J. Haddad, and N. Geroliminis, "Integrating the dynamics of heterogeneity in aggregated network modeling and control," *93rd Annual Meeting of the Transportation Research Board*, Washington D.C., January 2014, [162].
- M. Ramezani, J. Haddad, and N. Geroliminis, "Integrating heterogeneity dynamics in aggregated urban network modeling and control," 14th Swiss Transportation Research Conference, Ascona, Switzerland, May 2014, [161].

Chapter 6 is organized as a self stand-alone article submitted for publication as:

- M. Ramezani, J. Haddad, and N. Geroliminis, "Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control," *under 2nd round of review in Transportation Research Part B*, [154].

Finally, **Chapter 7** concludes the dissertation by summarizing the main contributions of each chapter, discussing the potential field applications of the proposed estimation and control methods, and presenting the future research directions.

Traffic State Estimation Part I

2 Queue profile estimation in congested urban networks with probe data

QUEUES at signalized intersections are the main cause of traffic delays and travel time variability in urban networks. In this chapter, we propose a method to estimate queue profiles that are traffic shockwave polygons in the time-space plane describing the spatiotemporal formation and dissipation of queues. The method integrates the collective effect of dispersed probe vehicle data with traffic flow shockwave analysis and data mining techniques. The proposed queue profile estimation method requires position and velocity data of probe vehicles; however, any explicit information of signal settings and arrival distribution is indispensable. Moreover, the method captures interdependencies in queue evolutions of successive intersections. The significance of the proposed method is that it is applicable in oversaturated conditions and includes queue spillover identification. Numerical results of simulation experiments and tests on NGSIM field data, with various penetration rates and sampling intervals, reveal the promising and robust performance of the proposed method compared with a uniform arrival queue estimation procedure. The method provides a thorough understanding of urban traffic flow dynamics and has direct applications for delay analysis, queue length estimation, signal settings estimation, and vehicle trajectory reconstruction.

2.1 Introduction

Vehicle queues at intersections are the crucial cause of variability in drivers' experienced travel time in urban networks. Hence, an accurate and practical queue estimation method is of great importance for intelligent transportation systems (ITS) (i) to provide a better understanding of urban flow dynamics, (ii) to be utilized for traffic state estimation, and (iii) to be integrated in a traffic signal control framework. In this chapter, we aim at estimating the evolution of queues by utilizing probe data. We introduce the concept of queue profile that describes the formation and dissipation of queues in the time-space (x-t) plane. Queue profile is a polygon in the x-t plane that its edges designate traffic shockwave boundaries (see Fig. 2.1), based on kinematic LWR theory [110, 163]. By estimating the queue profile for each cycle, one can derive performance measures including queue length, position of back and front of the queue,

and delays per vehicle (average and distribution) that can be further used to estimate travel time distributions as a travel time reliability index [151]. In addition, given the queue profile vehicle trajectories can be reconstructed (see for example [171]), which are fundamental for fuel consumption and emission estimation [170] and travel time decomposition [77].

Literature of queue estimation can be categorized to two modeling classes: (i) models based on the cumulative traffic input-output [188, 6, 40, 183], with the limitation that these models are insufficient to provide the spatial distribution of queue dynamics [123] and suffer from measurement errors, and (ii) models based on LWR shockwave theory [169, 10, 193], which provide the temporal-spatial dynamics of the queuing process with input data from loop detectors or probe vehicles. Further, many pivotal studies of queue and delay estimation in transportation research [128, 34, 127, 121] are based (fully or partially) on rather limiting assumptions to result in closed-form theoretical solutions. These assumptions can be summarized as: (i) known signal settings, (ii) known initial queue size at the start of the cycle, (iii) known arrival pattern (e.g. uniform, Poisson), and (iv) undersaturated conditions that expected value of arrivals do not (constantly) exceed the signal capacity [36].

The main contribution of this chapter is to introduce a queue profile estimation method that relaxes the aforementioned assumptions. Because signal settings, i.e. cycle length and splits, might not be readily available specifically from actuated traffic signals and network-wide systems [70]. Note that signal settings are crucial for capturing queue spatiotemporal characteristics, as the spatial extent of queue should be estimated at the correct time instance. For example, if the exact queue length is estimated with a time lag the estimation error would be significant. Whereas in the analyses of maximum queue length estimation (or studies with assumption of known signal settings), temporal term of estimation error is zero. In addition, a known arrival traffic flow distribution might be considered valid in case of an isolated intersection; whereas such an assumption is not valid in arterials, as it neglects the impact of upstream intersections that alter the arrival pattern further downstream [151]. The other significance of the proposed queue estimation method is to cope with oversaturated conditions and identify the occurrence of spillovers in arterials with probabilistic inference.

The basic ingredient for the new wave of smart cities that has emerged during the last decade is massive data sets concerning human mobility, fostered by the widespread distribution of sensors, such as Global Positioning System (GPS) devices in many modes of transport, smart phones, and traffic fixed sensors (e.g. loop detectors). These network infrastructures allow for sensing and collecting substantial spatiotemporal data, such as the trajectories of many vehicles from navigation devices, which represent proxies for human mobility patterns. This "big mobility data" provides a unique social observatory that can help us understand how congestion develops and evolves, and discover hidden patterns and identify models that can contribute to efficient traffic management systems to improve cities' mobility and accessibility.

Prevailing queue estimation methods employ several monitoring technologies, e.g. loop detectors [169, 56, 194] and event-based signal and vehicle data [195]. It has been reported in many

publications that input-output diagrams have strong difficulties in estimating queue lengths accurately under congested conditions. The main reason is that detector errors (even unbiased) accumulate over time (see for example [182]). Furthermore, in case of detectors in the beginning and end of a link, stop-and-go traffic and non-smooth flow due to upstream departures creates strong disturbances in the estimation. For mid-block detectors when queues are overpassing the detector, the accuracy of estimation degrades. The emergence and a steadily increase of public deployment of user-based data collection systems, e.g. GPS-equipped vehicles [79, 14, 88] and vehicle re-identification using cellphones or wireless magnetic sensors [105], provides a great potential for probe vehicles in ITS applications.

We introduce a method that is based on probe data as the single source of information. Utilizing probe vehicles for traffic observation and estimation raises challenges because of (i) errors in raw measurements and post-processing algorithms, e.g. map-matching, and (ii) the fact that the probe data are inherent samples of a true traffic state. Thus, probe data cannot be readily applied to well-established traffic estimation methods. In this study, the position and instantaneous speed of probe vehicles are the input to the queue profile estimation method and the robustness of the method to measurement errors is investigated.

[105] and [81] are of the few efforts to estimate performance measures without the need for signal settings information. The former requires wide deployment of sensors and the latter proposes a probabilistic framework for arterial traffic state estimation using sparse probe data that infer the average signal settings and arrival rates from historical data. Utilizing probe data, [151] develop a method based on the link travel times of probe vehicles to estimate the arterial travel time distribution by considering spatiotemporal nonlinear correlations. [22] provide an analytical model for queue length estimation of an isolated intersection with the assumption of Poisson arrival distribution and examined the effect of probe vehicles penetration rate. In addition, [84] and [19] study a queue estimation method based on the identification of shockwaves from probe vehicle trajectories. Although queue estimation is straightforward given trajectory of probe vehicles with large penetration rates, for realistic cases with low penetration rates and high sampling intervals, an approach that combines data mining, optimization techniques, and physical properties of traffic flow is needed (e.g. see [70, 81]; also see [184] for incident duration modeling, [59] for delay and queue length estimation at freeway work zones, and [92] for traffic flow forecasting). A brief description of the developed method without considerations for spillovers, capturing the interdependencies of queuing dynamics of adjacent links and less robust optimization framework to noisy measurements is presented in [152].

The remainder of the chapter is organized as follows. Section 2.2 presents the preliminaries of the traffic flow modeling utilized for the queue profile estimation. Next, we introduce the proposed method and elaborate on its details in Section 2.3. Section 2.4 presents the results for two case studies based on field and simulation data, comparisons with a queue estimation procedure based on the uniform arrival assumption, and the effect of probe data penetration rates, sampling intervals, and measurement errors on the performance of the

proposed method. The spillover probabilistic inference method is introduced in Section 2.5. Finally, Section 2.6 draws the conclusions.

2.2 Background

This chapter aims at estimating the queue shockwave profile that is a polygon in the x-t plane such that its every edge designates a traffic shockwave. The shockwaves model the formation and dissipation of the queue during signal cycles. Based on LWR theory, shockwaves are boundaries determining flow-density discontinuities in the traffic stream, i.e. points on the border between two different traffic states. Thus, crossing traffic shockwaves, the slope of vehicle trajectory in the x-t plane (i.e. speed of the vehicle) changes abruptly. In this study, for each urban network link we assume a triangular fundamental diagram (FD) characterized by, the maximum flow (capacity), q_{m}, the free flow speed, v_{ff}, and the jam density, k_{j}, see Fig. 2.1. While LWR theory can integrate more complicated shapes of FDs, we choose the widely used triangular FD as it can represent well the development and dissipation of shockwaves in arterials and allows for elegant data mining techniques consistent with the physics of traffic. In this case, the platoon dispersion effect cannot be estimated. We refer to [123] or [168] for non-triangular FDs for arterial estimation.

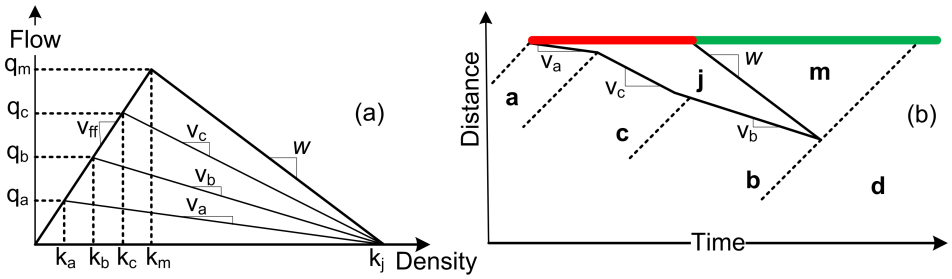


Figure 2.1: The fundamental diagram and a queue profile at a signalized intersection.

Figure 2.1 also depicts a queue profile at a signalized intersection. The queue discharging shockwave represents the front of queue and a 3-segment piecewise linear shockwave represents the back of queue. It is apparent that the queue discharging shockwave separates the traffic state at capacity (**m**) from the jammed state (**j**). Likewise, the 3-segment shockwave of the back of queue separates the jammed state from 3 different traffic states (q_i, k_i), in the flow-density plane with $i \in \{a,b,c\}$. Note that traffic state (**d**) may represent a different state than state (**b**). The slope of the shockwave between traffic states s and s' and the slope of the queue-discharging shockwave (w) are, respectively:

$$v_{ss'} = \frac{q_s - q_{s'}}{k_s - k_{s'}}, w = \frac{q_{\mathrm{m}}}{\frac{q_{\mathrm{m}}}{v_{\mathrm{ff}}} - k_{\mathrm{j}}}. \tag{2.1}$$

It is straightforward to estimate the queue shockwave profile (the extent and slope of shockwaves) given the arrival traffic flow and the signal settings (which both impose measurement difficulties). However, the proposed queue profile estimation method does not require either

arrival traffic distribution or signal settings. Note that one application of queue profile is that by reverse modeling, one can estimate the attributes of shockwaves, and consequently the characteristics of arrival traffic state can be determined. In addition, the estimated queue profiles approximate the signal setting, i.e. the start and end time of signal phases. A recent study [70], estimates signal settings utilizing vehicle travel time observations from upstream and downstream of an intersection.

In this chapter, we try to estimate the queue profile polygon from probe vehicles. Probe vehicles provide samples of their individual traffic state and we aim at leveraging the collective information of temporally and spatially dispersed probe data. As we stated earlier, traffic shockwaves designate the points in the x-t diagram associated with the sudden change of vehicles speeds. In urban networks and particularly at signalized intersections, this associates with how probe vehicles join and leave the queue, which has to be estimated. The proposed method utilizes kinematic analysis to approximate the time and the position that a probe vehicle joins and leaves the queue (joining and leaving points) from its reported data by assuming constant acceleration, a_{acc}, and deceleration, a_{dec}. This assumption is not expected to influence the accuracy of the model [168].

2.3 Methodology

In this section, first a brief overview of the proposed queue profile estimation method is given, and then the principal steps of the method are elaborated in detail. Let us assume that probe vehicle i reports its position (the distance from upstream of the link), x_i^k, and velocity, v_i^k, along with the time stamp, t_i^k, at time step k. In the sequel, the link index is omitted for simplicity, and we assume periodic reporting of the above information with sampling interval T (s). However the method can be easily applied with non-periodic probe data collected from various types of probe vehicles, e.g. taxis, transit fleet, and cars. Note that the queue profile estimation is a link-wise procedure and in case of possible spillovers (or spillbacks), the queue profile should be estimated from downstream to upstream links to be able to capture the spillover development and propagation. Note that the occurrence of spillbacks is not considered in this section while in Section 2.5, we relax this assumption and develop a spillover identification procedure based on Bayesian inference.

An example of the queue profile estimation with $T = 30$ (s) and penetration rate (ϕ) of 40% is depicted in Fig. 2.2, where ground truth and estimated queue profiles are respectively denoted by black and red polygons. Even if this penetration rate is not realistic, it is chosen for the explanation of the method. Later, smaller values of ϕ are analyzed. Note that probe data points are the input to the method and trajectories in Fig. 2.2 are shown for illustration purposes of the queue dynamics. The details shown in Fig. 2.2 will be fully elaborated in the following subsections. In addition, effects of measurements or map-matching errors on the performance of the proposed estimation method are scrutinized in Section 2.4.

A conceptualization of the queue profile estimation method is as follows. Since signal settings

are assumed to be unavailable, we need to identify stopped data points and assign them to various groups that each group represents a cycle. Therefore in subsection 2.3.1, we introduce a classifier based on the velocity attribute to classify all probe data to two groups, stopped and moving. In Fig. 2.2, red squares denote the stopped data points while black diamonds denote the moving data points. Afterwards in subsection 2.3.2, we cluster the stopped data points into different signal cycles based on the projection profile algorithm. Subsequently in subsection 2.3.3, the moving data is assigned to different cycles. The magenta lines provide linear boundaries that group the moving data in cycles. Fig. 2.2 depicts the cycle number of each stopped and moving data point. As queued vehicles discharge at the capacity flow, based on LWR theory, the queue discharging shockwave is a line with slope w in the x-t plane. In addition, the leaving points of vehicles from the queue (filled blue points in Fig. 2.2) are on the queue discharging shockwave. Hence, the estimation of front of queue for each cycle can be formulated as a constrained least squares problem, which is presented in subsection 2.3.4. The proposed method does not require the knowledge of arrival patterns to estimate queue profiles. Thus to model the back of queue, we consider a piecewise linear function that fits the joining points of vehicles to the queue (filled green, cyan, and yellow points in Fig. 2.2). In subsection 2.3.5, a curve-fitting nonlinear optimization method is introduced to identify the number and attributes of the piecewise linear function [63].

2.3.1 Classification of probe data to moving and stopped classes

The first step of the procedure is to classify the probe data into two classes of moving and stopped vehicles $\{\mathrm{m},\mathrm{s}\}$. This can be done with a simple threshold-based classifier:

$$c_i^k = \begin{cases} \mathrm{m} & \text{if } v_i^k > v_{\mathrm{th}} \\ \mathrm{s} & \text{if } v_i^k \leq v_{\mathrm{th}} \end{cases} \tag{2.2}$$

where v_{th} (m/s) is a predefined threshold parameter to designate the vehicles with velocity close to zero as stopped.

The correct classification rate of the proposed classifier is 100% in case of exact measurements. Though errors in velocity measurements degrade the classifier performance, measurement errors do not significantly affect the method (The robustness of method to measurement errors is examined later in Section 2.4). However, a more precise velocity measurement, e.g. utilizing differential GPS or fusion of GPS speed data with vehicle speedometer seems essential, given also that signal settings are considered unknown. The classification of probe data points to stopped (red square) and moving (black diamond) is illustrated in Fig. 2.2.

2.3.2 Clustering the stopped vehicles to cycles

Our motivation in this study is to relax the assumption of known signal settings. With no information of signal settings, the number and duration of cycles are unknown, especially for

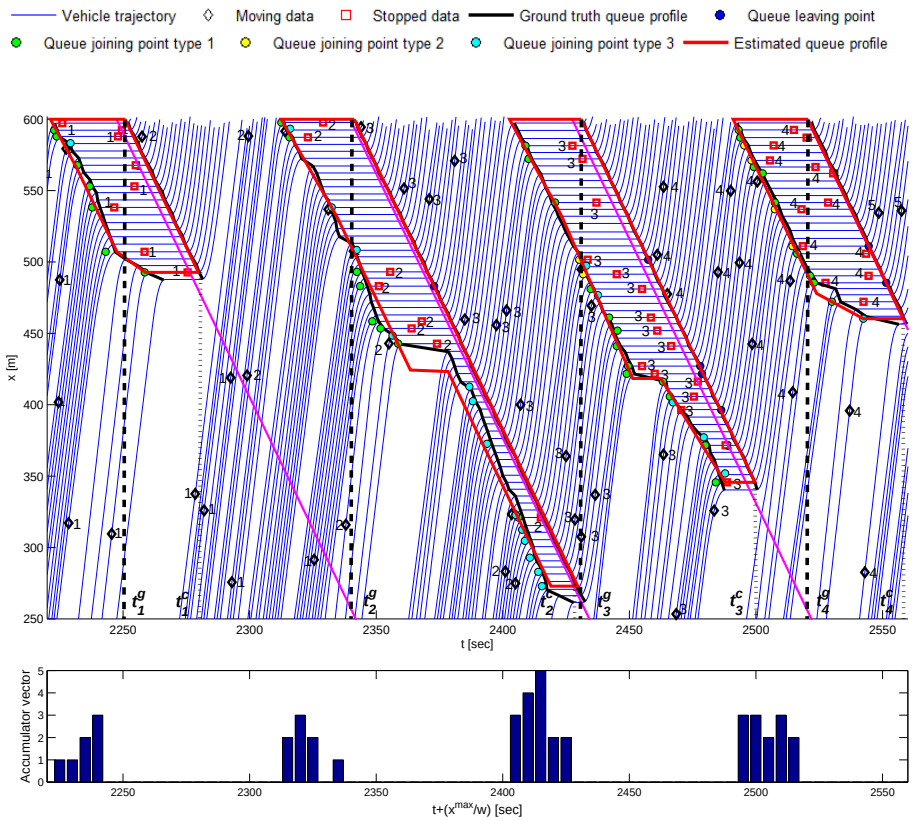


Figure 2.2: An illustrative example of the queue profile estimation method. The black diamonds and red squares respectively show input probe data classified to moving and stopped points (subsection 2.3.1). Their associated numbers are the corresponding cycle index (subsections 2.3.2 and 2.3.3), where magenta lines group the moving points to cycles. The filled blue circles are the queue leaving points of probe vehicles that designate the front of the queue profile (subsection 3.4). The front of queue shockwave determines the estimated start of green phase, t^g (subsection 2.3.4). The filled green, yellow, and cyan circles are different types of queue joining points of probe vehicles that designate the back of queue (subsection 2.3.5). t^c denotes the time that the queue clears. The bottom plot shows the corresponding accumulator projection vector (subsection 2.3.2) depicting four groups of stopped data associated with four cycles. All the details are elaborated in corresponding subsections.

actuated signals. As queues are associated with cycles, in the second step of the procedure a clustering technique is needed to cluster the stopped data into various groups that represent cycles (of the same or different duration). By considering the temporal and spatial distribution of stopped data in the x-t plane, the stopped data of two consecutive cycles can be linearly separated with a straight line with slope w that lies properly between the two cycles. Therefore, we propose a clustering technique based on the projection profile method [86] to cluster the stopped data into cycles. The purpose of projection profile method is to extract a scalar feature from the (2D) dispersed data of stopped vehicles in the x-t plane. Accordingly, we introduce a projection function, $\mathscr{P}$, which projects every stopped point along a line with slope w into an accumulator projection vector over the line $x = 0$ (the upstream of link). At this location, time is divided to equal intervals (bins) such that each bin is associated with the number of

projected points lying inside the bin, see Fig. 2.2.

$$\text{if } c_i^k = \text{s} \Rightarrow \mathscr{P}(t_i^k, x_i^k) = t_i^k - \frac{x_i^k}{w}. \tag{2.3}$$

It is expected that the projection profile method with projection function $\mathscr{P}$ results in an accumulator vector that contains (i) clusters of adjacent bins with positive values, representing the red interval of a cycle, and (ii) clusters of adjacent bins with zero values, representing the green duration of a cycle (see values of the accumulator projection vector in Fig. 2.2). Consequently, the stopped data associated with every cluster of stopped bins (adjacent bins with positive value) correspond to a common cycle. It is remarkable that no prior information on number of clusters, i.e. number of cycles, is needed for the clustering procedure. This property is important for actuated traffic signals and for spillback identification.

The accuracy of projection profile algorithm depends on the bin size (s). To determine the size of each bin, one should consider that large values cannot discriminate between cycles; while small values may decrease the chance that stopped data points of the same cycle are projected into adjacent bins. (Note: In case of small-sized bins, the number of bins with zero value during the red interval increases. This might create group(s) of zero-valued bins during the red interval. So, the clustering procedure should disregard small number of zero-valued bins inside a cluster of stopped bins; see for example cycle 2 in Fig. 2.2.) We test the sensitivity of the projection profile algorithm to various bin sizes and values between 4 and 10 (s) show satisfactory outcomes. In this study, we set the bin size equal to 5 (s). Result of this step is apparent in Fig. 2.2, where the cycle number of each stopped data point is depicted. Note that this method is still valid in case of active spillovers that block departures during green times of upstream intersections. An enhancement to the clustering procedure concerning occurrence of spillovers will be discussed in Section . In case of very low penetration rates (< 5%) this method can be improved by also considering the moving points during green durations, which can be considered as a future work.

2.3.3 Associating the moving vehicles to cycles

The previous step identifies the number of cycles (not signal settings) and assigns stopped data to their corresponding cycle. Likewise, the proposed queue profile estimation method needs the moving data ($c_i^k = \text{m}$) to be associated to cycles to formulate the estimation problem of the front and back of queue. Let S_j denotes the set of stopped data assigned to cycle j. One can fit a line with slope w to every data point that belongs to S_j. The line with the maximum intercept (traffic progresses in direction of x-axis) can be regarded as a lower envelope of the"true" discharging shockwave, see magenta lines in Fig. 2.2. This line can be considered as a linear boundary to distinguish between moving data of every cycle. In other words, the moving data of cycle j are on the left side of this line for cycle j, and right side of the corresponding line for cycle $j-1$. We denote the set of moving data associated with cycle j as M_j. Fig. 2.2 depicts the number of the cycle that each moving point is assigned to. Note that

the error between the magenta lines and the"true" discharging shockwaves only influences assigning cycles to moving points and does not intervene on cycle identification, signal setting estimation, and queue profile estimation.

2.3.4 Estimation of the front of queue

The queuing process can be divided to two distinct fundamental processes, the formation and dissipation of queues, which in urban networks is associated with the traffic signal alternation. This step of the queue profile estimation is to determine the queue dissipation for every cycle (j). Based on LWR theory, the queued vehicles discharging at the capacity flow result in a shockwave designated by a line with slope w in the x-t diagram. Hence, there is only one parameter to estimate the front of queue shockwave (discharging line), i.e. the line intercept. Initially, we need to estimate the leaving point of probe vehicles, i.e. at which time and position probe vehicles leave the queue. The estimation of leaving points of probe vehicles is important since these points are ideally supposed to be on the discharging shockwave line. From the formulas of kinematics on one dimension and based on two consecutive data points of the same probe vehicle, where the first one is stopped and the other one is moving (i.e. c_i^k = s, c_i^{k+1} = m, $(t_i^k, x_i^k) \in S_j$), the time that probe vehicle i leaves the queue can be estimated as

$$\tau_i^{\mathrm{l}} = \begin{cases} t_i^{k+1} - \dfrac{v_i^{k+1}}{2a_{\mathrm{acc}}} - \dfrac{x_i^{k+1} - x_i^k}{v_i^{k+1}} & \text{if } v_i^{k+1} > \eta \cdot v_{\mathrm{ff}} \quad (2.4a) \\ t_i^{k+1} - \dfrac{2(x_i^{k+1} - x_i^k)}{v_i^{k+1}} & \text{if } v_i^{k+1} \leq \eta \cdot v_{\mathrm{ff}}. \quad (2.4b) \end{cases}$$

Equation (2.4a) corresponds to a vehicle that reaches its desired speed, which is in a vicinity of the free flow speed, specified by $\eta \in (0,1)$. η is the parameter to designate whether a probe vehicle is in acceleration (or deceleration) mode or it has reached to the free flow speed. Likewise, (2.4b) corresponds to a vehicle that its instantaneous speed at time step $k+1$ is not its desired speed, since the vehicle is still accelerating. Note that, there is another possibility that the vehicle has a speed below its desired speed ($v_i^{k+1} \leq \eta \cdot v_{\mathrm{ff}}$) and be in the deceleration stage. In this case, (2.4b) gives a value less than t_i^k which is incorrect because, $\tau_i^{\mathrm{l}} \in (t_i^k, t_i^{k+1})$. Therefore, we need to modify (2.4b) in order to correctly estimate the leaving time of probe vehicle i that has a speed below the desired speed and is in the deceleration stage. So, if $v_i^{k+1} \leq \eta \cdot v_{\mathrm{ff}}$ and τ_i^{l} computed by (2.4b) is smaller than t_i^k, the time that probe vehicle i leaves the queue is

$$\tau_i^{\mathrm{l}} = t_i^{k+1} - \frac{v_{\mathrm{ff}}}{2a_{\mathrm{acc}}} - \frac{x_i^{k+1} - x_i^k}{v_{\mathrm{ff}}} - \frac{(v_i^{k+1} - v_{\mathrm{ff}})^2}{2v_{\mathrm{ff}} \cdot a_{\mathrm{dec}}}. \quad (2.5)$$

Equation (2.5) implies that the vehicle accelerates with acceleration a_{acc}, reaches its desired speed which is equal to v_{ff}, and then decelerates with deceleration a_{dec} to its speed at time

step $k+1$, v_i^{k+1}.

Ultimately, the leaving point is estimated as $(\hat{t}_i, \hat{x}_i) = (\tau_i^{\mathrm{l}}, x_i^k)$. In Fig. 2.2, filled blue points represent leaving points. While time spent in acceleration and deceleration mode is a minor component of the delay under congested conditions and can be ignored, it is a necessary step in our approach to classify the different types of vehicles (joining or leaving the queue).

The discharging line has slope w such that all the moving data points of the next cycle, M_{j+1}, and stopped data points of the corresponding cycle, S_j, are on its right and left side, respectively. Hence, the estimation of front of queue is formulated as a constrained least squares problem:

$$\min_{B_j} \sum_{l=1}^{N_{\mathrm{L}}} (\hat{x}_l - w\hat{t}_l - B_j)^2 \tag{2.6}$$

$$\text{subject to} \begin{cases} x - wt - B_j \le \epsilon_1 & \forall (t,x) \in S_j \\ x - wt - B_j \ge \epsilon_1 & \forall (t,x) \in M_{j+1} \end{cases} \tag{2.7}$$

where N_{L} denotes the total number of estimated leaving points and B_j is the intercept of the discharging line shockwave of cycle j. Because problem (2.6) is convex, the solution is global although not necessarily unique. Note that given B_j, the start time of the current cycle green phase, denoted by t_j^{g} can be estimated as:

$$t_j^{\mathrm{g}} = \frac{x^{\max} - B_j}{w} \tag{2.8}$$

where $x^{\max}$ is the position of intersection stop line from the entrance (upstream) of the link.

The objective of (2.6) is to minimize the sum of squared errors between the leaving points and the estimated discharging shockwave line. Regarding constraints (2.7), introduction of ϵ_1 provides a tuning parameter to regulate the extent of hardness of the constraints, e.g. $\epsilon_1 = 0$ represents hard-constraints that obliges the solution of (2.6) to fully separate data points of S_j and M_{j+1}, whereas a positive value of ϵ_1 relaxes the constraints, which is necessary in case of noisy input data, because there might be no line with slope w that fully separates noisy data of S_j and M_{j+1}. It is worth mentioning that, there is no restriction that two consecutive stopped and moving data points (c_i^k = s and c_i^{k+1} = m) are on the same link, i.e. the moving point can belong to the downstream link. Nevertheless, the chance of finding two such data points in the most downstream link is less compared to the other links. Thus, in case there is no estimated leaving point (see the first cycle in Fig. 2.2), the discharging shockwave line is estimated based on the best generalization concept which provides the maximum margin from both stopped (S_j) and moving vehicles (M_{j+1}). In other words, the discharging line should have an equal distance from the nearest data points of both S_j and M_{j+1}. This concept is widely utilized in

support vector machine (SVM) in computer science literature [25] and reads:

$$B_j = \frac{\max_{S_j} B + \min_{M_{j+1}} B}{2} \tag{2.9}$$

where $\max_{S_j} B (\min_{M_{j+1}} B)$ denotes the maximum (minimum) intercept of lines with slope w that are fitted to the data points of set $S_j (M_{j+1})$.

2.3.5 Estimation of the back of queue

The goal of this step is to estimate the queue formation for each cycle (j). The back of queue in the x-t diagram can be modeled by a piecewise linear function of several segments so that each one represents a shockwave. Without any explicit information of arrival patterns, the back of queue estimation procedure should determine the number and extent of segments of the piecewise linear function. Initially, similar to the previous step, we need to estimate the joining point of probe vehicles, i.e. at which time and position probe vehicles join the queue, because the joining points ideally belong to the piecewise linear shockwave. We consider the following types of two consecutive data points for probe vehicle i that contribute to determining the joining points:
(*Type 1*) the first point is moving and the other one is stopped, i.e. $(t_i^{k-1}, x_i^{k-1}) \in M_j$ and $(t_i^k, x_i^k) \in S_j$, (filled green points in Fig. 2.2)
(*Type 2*) both points are stopped, i.e. $(t_i^{k-1}, x_i^{k-1}) \in S_{j-1}$ and $(t_i^k, x_i^k) \in S_j$, (filled yellow points in Fig. 2.2)
(*Type 3*) both points are moving, i.e. $(t_i^{k-1}, x_i^{k-1}) \in M_j$ and $(t_i^k, x_i^k) \in \{M_{j+1} \cup ND\}$, (filled cyan points in Fig. 2.2)
(*Type 4*) first point is stopped and the second one is moving, i.e. $(t_i^{k-1}, x_i^{k-1}) \in S_{j-1}$ and $(t_i^k, x_i^k) \in \{M_{j+1} \cup ND\}$.
ND denotes the set of moving points in the downstream link that are not delayed, i.e. the points that do not belong to the queue discharged traffic state (state **m** in Fig. 2.1). Note that since the queue profile estimation method starts from the downstream link, ND is already available. By considering moving points of ND, the methodology implicitly correlates the queuing state of successive links. For types 3 and 4, first we should assess if probe vehicle i encounters any significant delay between the two consecutive observations. This is straightforward by comparing the average speed between two successive observations and the free flow speed. Afterwards, if the probe vehicle is delayed (as it joins and leaves a queue), the data at time step k belonging to M_{j+1} or ND can be converted to a queue leaving point according to (2.4a), (2.4b), and (2.5), since the discharging line of the cycle j is already estimated in the fourth step. Therefore, data points of type 3 and 4 can be readily converted to type 1 and 2, respectively.

For type 1, the time that probe vehicle i joins the queue can be calculated as

$$\tau_i^{\mathrm{j}} = \begin{cases} t_i^{k-1} - \dfrac{v_i^{k-1}}{2a_{\mathrm{dec}}} - \dfrac{x_i^k - x_i^{k-1}}{v_i^{k-1}} & \text{if } v_i^{k-1} > \eta \cdot v_{\mathrm{ff}} \quad (2.10a) \\ t_i^{k-1} - \dfrac{2(x_i^k - x_i^{k-1})}{v_i^{k-1}} & \text{if } v_i^{k-1} \le \eta \cdot v_{\mathrm{ff}}. \quad (2.10b) \end{cases}$$

Equations (2.10a, 2.10b) have the same logic as (2.4a, 2.4b), so that equation (2.10a) corresponds to a vehicle that reaches to its desired speed in the vicinity of free flow speed, and (2.10b) corresponds to a vehicle that its instantaneous speed at time step $k-1$ is not its desired speed, since the vehicle is decelerating. Similarly, the third possibility is that the vehicle has a speed below its desired speed ($v_i^{k-1} \le \eta \cdot v_{\mathrm{ff}}$) and is in the acceleration stage. In this case, (2.10b) gives a value greater than t_i^k which is incorrect because, $\tau_i^{\mathrm{j}} \in (t_i^{k-1}, t_i^k)$. Therefore, we need to modify (2.10b) in order to correctly estimate the joining time of probe vehicle i. So, if $v_i^{k-1} \le \eta \cdot v_{\mathrm{ff}}$ and τ_i^{j} computed by (2.10b) is greater than t_i^k, the time that probe vehicle i joins the queue is

$$\tau_i^{\mathrm{j}} = t_i^{k-1} - \frac{v_{\mathrm{ff}}}{2a_{\mathrm{dec}}} - \frac{x_i^k - x_i^{k-1}}{v_{\mathrm{ff}}} - \frac{(v_i^{k-1} - v_{\mathrm{ff}})^2}{2v_{\mathrm{ff}} \cdot a_{\mathrm{acc}}}. \quad (2.11)$$

Equation (2.11) implies that the vehicle with speed v_i^{k-1} accelerates with acceleration a_{acc}, reaches its desired speed which is equal to v_{ff}, and then decelerates with deceleration a_{dec} to the stopping state. Ultimately, the joining point is estimated as $(\check{t}_i, \check{x}_i) = (\tau_i^{\mathrm{j}}, x_i^k)$.

Note that for data type 2, the leaving point of probe vehicle i at cycle $j-1$ can be estimated by utilizing the stopped data point (t_i^{k-1}, x_i^{k-1}) and the discharging shockwave of the previous cycle, $j-1$. (The discharging shockwave of the previous cycle is already known, because queue profile estimation advances cycle by cycle in time.) Thus with the estimated leaving point that is certainly at the acceleration stage and by applying (2.11) the corresponding joining point can be estimated. Remarkably, because the data point at time step $k-1$ of type 2 is discharging from the queue of the previous cycle, it arrives at capacity state to the queue of the current cycle. Hence, the slope of the corresponding segment of back of queue shockwave should be w.

Given the estimated joining points, J_j, the back of queue piecewise linear shockwave, $f_{\mathrm{b}}(x)$, should be estimated according to the following objectives: (i) to be as close as possible to the joining points, and (ii) to separate all the moving data points associated with cycle j, M_j, and stopped ones of cycle j, S_j, on its left and right side, respectively (see the definition of function $\mathscr{F}$). Further, based on LWR theory, the slope of each segment of the piecewise linear function should be between zero and w (see 2.15). Note that, if a leaving point of type 2 or 4 is associated with a segment of $f_{\mathrm{b}}(x)$, the segment slope should be equal to w. To incorporate all the aforementioned physical properties, the estimation of the back of queue by an N-segment

piecewise linear function is formulated as the following nonlinear program:

$$\min_{\alpha_1,\dots,\alpha_{N-1},\beta_1,\cdots,\beta_{N+1}} \frac{\sum_{l=1}^{N_\mathrm{J}} (\mathscr{D}\langle f_\mathrm{b}(.),(\check{t}_l,\check{x}_l)\rangle)^2}{N_\mathrm{J}} + \lambda N + C\Big[\sum_{S_j}\mathscr{F}(t,x)+\sum_{M_j}\mathscr{F}(t,x)\Big] \tag{2.12}$$

$$f_\mathrm{b}(x)=\begin{cases} \beta_1+\frac{x-x^{\max}}{\alpha_1-x^{\max}}(\beta_2-\beta_1) & \alpha_1 < x \le x^{\max} \\ \vdots & \vdots \\ \beta_n+\frac{x-\alpha_{n-1}}{\alpha_n-\alpha_{n-1}}(\beta_{n+1}-\beta_n) & \alpha_n < x \le \alpha_{n-1} \\ \vdots & \vdots \\ \beta_N+\frac{x-\alpha_{N-1}}{x_j^{\min}-\alpha_{N-1}}(\beta_{N+1}-\beta_N) & x_j^{\min} \le x \le \alpha_{N-1} \end{cases} \tag{2.13}$$

$$\begin{cases} x_j^{\min} \triangleq \min_x(S_j,J_j) < \alpha_{N-1} < \cdots < \alpha_1 < x^{\max} & \text{(2.14a)} \\ t_{j-1}^{\mathrm{g}} < \beta_1 < \cdots < \beta_{N+1} \le t_j^{\mathrm{c}} = \dfrac{x_j^{\min}-B_j}{w} & \text{(2.14b)} \end{cases}$$

$$\frac{\beta_{n+1}-\beta_n}{\alpha_n-\alpha_{n-1}} \le \frac{1}{w}; \quad n=1,\cdots,N \tag{2.15}$$

$$\mathscr{F}(t,x)=\begin{cases} \mathscr{D}\langle f_\mathrm{b}(.),(t,x)\rangle & \text{if } (t,x)\in S_j \wedge f_\mathrm{b}(x)-t>0 \\ \mathscr{D}\langle f_\mathrm{b}(.),(t,x)\rangle & \text{if } (t,x)\in M_j \wedge f_\mathrm{b}(x)-t<0 \end{cases} \tag{2.16}$$

where N_J denotes the total number of estimated joining points, and function $\mathscr{D}$ determines the distance of a point from a piecewise line. Hence, the first term of the objective function (2.12) is to minimize the sum of squared distance between the joining points and the desired piecewise linear function. Given that $f_\mathrm{b}(x)$ comprises N segments, (2.12) optimizes $2N$ parameters defining slope and extent of segments, i.e. $\alpha_1,\dots,\alpha_{N-1}$ that are associated with the spatial extent of segments and $\beta_1,\cdots,\beta_{N+1}$ that are associated with the temporal extent of segments, see Fig. 2.3. It is apparent that as N increases, $f_\mathrm{b}(x)$ becomes more complex and can attain better optimization results as oppose to higher chance of overfitting. To prevent overfitting, we need to balance the number of parameters, as an indicator of complexity, and the value of the objective function. To this end, we add a complexity penalty term to the objective function, λN, where λ is the regularization parameter. In addition, function $\mathscr{F}$ calculates the distance of misclassified moving data of M_j and stopped data of S_j and penalizes them with a relatively large constant, C. Finally to find the optimum solution of (2.12), we restrict $N \le \sqrt{N_\mathrm{j}}$ and execute the optimization algorithm for several random initial points in order to prevent reaching local optima.

Furthermore, constraints (2.14a) and (2.14b) ensure that $f_\mathrm{b}(x)$ has to be connected and to be in proper intervals, i.e. $x_j^{\min} \le x \le x^{\max}$ and $t_{j-1}^{\mathrm{g}} < t \le t_j^{\mathrm{c}}$. In other words, 2.14a ensures that

the spatial extent of the back of queue shockwave is between the position of intersection stop line, $x^{\max}$, and the position of the furthest queued vehicle from the intersection for cycle j, $x_j^{\min}$. To estimate$x_j^{\min}$, we consider the furthest position of available probe information, as the most conservative estimator. Thus, $x_j^{\min}$ is the minimum between the furthest position from stop line among stopped data S_j (see cycles 1 and 3 in Fig. 2.2), and the furthest position among the estimated joining points J_j (see cycles 2 and 4 in Fig. 2.2). Note that, $x_j^{\min}$ is the position where the estimated discharging shockwave and the piecewise shockwave of back of queue are connected to each other with a horizontal line (slope equals to zero) to connect the queue profile polygon (see Figures 2.2 and 2.3). Also (2.14b) implies that the estimation of the start time of red phase of current cycle, β_1, should be after the start time of green phase of the previous cycle denoted by t_{j-1}^{g}. (If there is any information regarding a minimum duration of red phases, this can be easily incorporated in the optimization framework.) In addition, the back of queue piecewise shockwave should be always (with respect to time) before the front of queue discharging shockwave in the x-t plane, i.e. the utmost time extent of the piecewise shockwave, β_{N+1}, must be less than the estimated time of queue full clearance, which is denoted by, t_j^{c}.

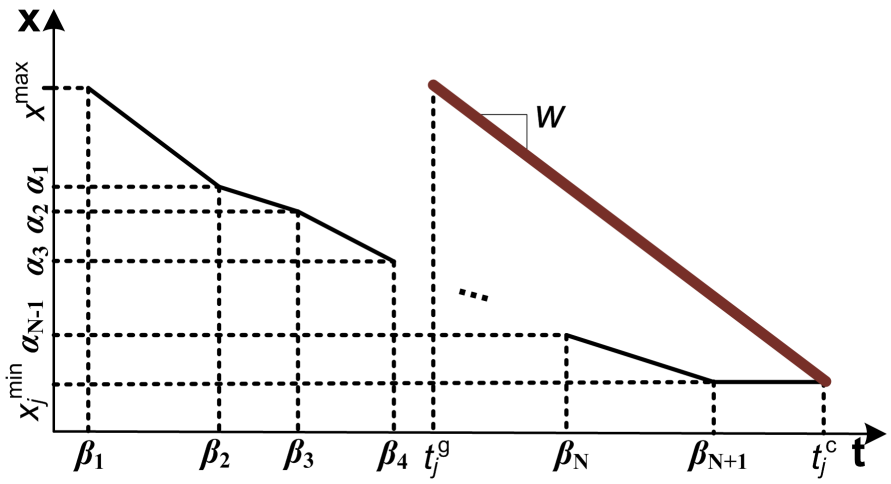


Figure 2.3: A schematic of N-segment piecewise linear back of queue at cycle j, annotated with parameters of optimization (2.12).

In case there is no estimated joining point, the piecewise back of queue shockwave can be estimated similarly based on the best generalization concept which has the maximum margin from both stopped (S_j) and moving vehicles (M_j). In other words, the discharging piecewise shockwave should have the maximum distance from the nearest data points of both S_j and M_{j+1}. This reads:

$$\max_{\substack{\alpha_1,\dots,\alpha_{N-1}\\ \beta_1,\cdots,\beta_{N+1}}} \left(\min_{S_j} \mathscr{D}\langle f_{\mathrm{b}}(.),(t,x)\rangle + \min_{M_j} \mathscr{D}\langle f_{\mathrm{b}}(.),(t,x)\rangle - \lambda N - C\Big[\sum_{S_j} \mathscr{F}(t,x) + \sum_{M_j} \mathscr{F}(t,x)\Big]\right) \tag{2.17}$$

where $\min_{S_j} \mathscr{D}\langle f_{\mathrm{b}}(.),(t,x)\rangle$ or $(\min_{M_j} \mathscr{D}\langle f_{\mathrm{b}}(.),(t,x)\rangle)$ denotes the distance of the nearest data point of S_j or (M_j) to f_{b}. To solve the nonlinear problem (2.17), the conditions (2.13)-(2.16) should be also satisfied.

2.4 Results

In this section, we test and discuss the proposed queue profile estimation method using (i) NGSIM field data and (ii) a more congested micro-simulated arterial with four signalized intersections. We also investigate the effect of probe data sampling interval, penetration rates, and measurement noise on the performance of method. Moreover, for further comparison the proposed method is compared with a uniform arrival queue estimation procedure, which requires signal settings data and aggregated data of an upstream (entrance) loop detector. For uncongested conditions, the uniform arrival assumption leads to a triangular queue profile for each cycle (see Fig. 2.4(a)), such that the red interval is the one edge, the second edge is the discharging shockwave starting at the end of red phase with slope w, and the third edge starts from the beginning of red phase with a slope based on (2.1), where s' is the jammed traffic state and state s is:

$$q_s = \frac{\mathrm{count}_j}{C_j}, k_s = \frac{q_s}{v_{\mathrm{ff}}}. \tag{2.18}$$

C_j denotes the duration of cycle j and count_j is number of vehicles that enter the link and with the free flow speed would reach to the intersection stop line during cycle j. In case of oversaturated conditions and residual queues at the end of the green time, the queue profile has a trapezoidal form (see for example [169]). An approach where the uniform arrival is applied without considering residual queues (i.e. a triangular queue profile) would result in much worse performance.

The method is tested on NGSIM data set. NGSIM program [129] was managed by the Federal Highway Administration to provide a data set of vehicle trajectories for traffic behavioral analyses. The data set contains comprehensive individual vehicle information such as instantaneous position and velocity along with the time, link, and direction stamps at the resolution of 0.1 second. We use NGSIM dataset of the through lane of southbound link between 11th Street and 10th Street on Peachtree Street, Atlanta, from 16:00 to 16:15. This link is approximately 130 (m) long and the speed limit is 56 (km/h). The FD parameters are set to $v_{\mathrm{ff}} = 15.65$ (m/s), $q_{\mathrm{m}} = 2100$ (veh/h), and $w = -3.7$ (m/s), and the parameters of the proposed method are chosen as: $v_{\mathrm{th}} = 1$ (m/s), $\eta = 0.8$, $a_{\mathrm{acc}} = 2$ (m/s^2), $a_{\mathrm{dec}} = -3.5$ (m/s^2), $\epsilon_1 = 0$, $C = 10^4$, and $\lambda = 1$.

Figure 2.4(a) illustrates the estimated queue profiles for 10 cycles based on the proposed method (red polygons, ($\phi = 30\%, T = 10$s)) and the uniform arrival assumption procedure (blue polygons) along with the ground truth queue profiles (black polygons). Given the estimated queue profiles, signal settings, queue length, average delay, etc. can be readily

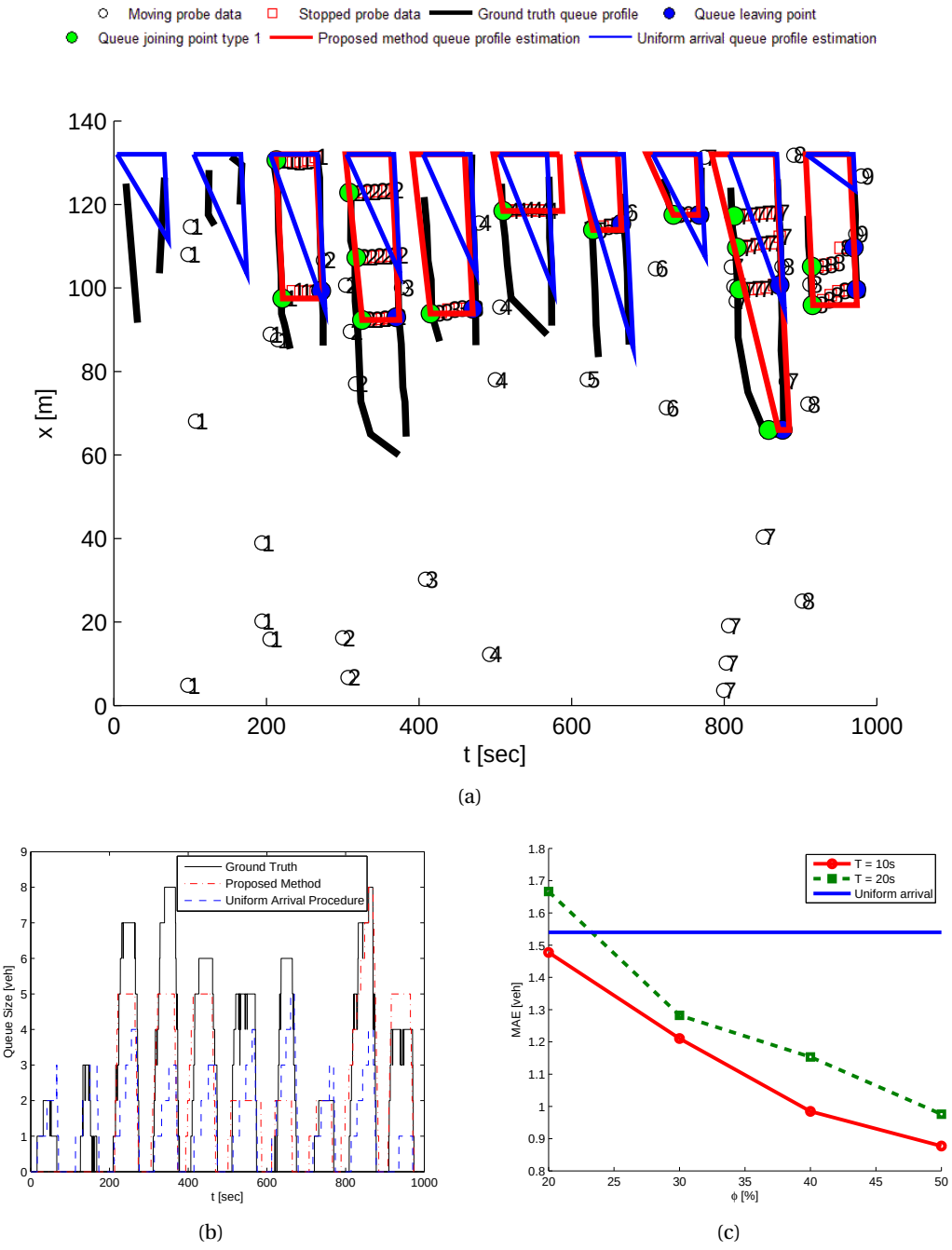


Figure 2.4: The NGSIM ground truth (black) and estimated ($\phi = 30\%$, $T = 10$s) queue profiles based on the proposed method (red) and based on the uniform arrival assumption (blue). The probe input data consist of time, position, and velocity of probe vehicles (a). The estimated and ground truth queue length (b). The MAE between the ground truth and estimated queue length with different penetration rates and sampling intervals (c).

estimated. For quantitative comparison, the queue size is derived based on the estimated queue profiles and depicted in Fig. 2.4(b). Furthermore, Fig. 2.4(c) shows the mean absolute error (MAE) between the ground truth queue size and the estimated queue size averaged over 10 runs for various penetration rates ($\phi = 20, 30, 40, 50\%$) and sampling intervals ($T = 10, 20$s). Note that during the data collection interval, there are only 80 vehicles crossing the link. Hence with a low penetration rate, the number of probe vehicles and consequently probe data would be insufficient to have an accurate estimation. It is apparent in Fig. 2.4(a) where the first two cycles are missed with the proposed method. Likewise, sampling intervals greater than 20s yields inadequate number of probe data since the link free flow travel time is below 10s. However, it is apparent in Fig. 2.4(c) that the proposed queue profile estimation method without any information of arrival distribution and signal settings is more accurate than the uniform arrival estimation procedure. Evidently with higher penetration rates or lower sampling intervals of probe vehicles, the outcomes of the proposed queue profile estimation method would be more precise.

A fundamental property of probe data is that as congestion increases, which is the case when the traffic estimation and control are crucial, the number of probe data increases (consider that congestion increases the total vehicle hours traveled by all vehicles, which is directly related to the number of available measurements). Hence in case of congestion, the proposed method performance is significantly better compared to light conditions. The NGSIM data does not include oversaturated conditions (note maximum queue size is 8 vehicles in Fig. 2.4(b)). The results are not surprising as at undersaturated conditions, simple and elegant methods can provide decent quality of results. Travel time estimation methods have reached similar conclusions about the accuracy of a simple model of uniform arrivals. Nevertheless, these models fail when congestion and long queues is present (see for example [169]).

To test the developed methodological framework on more challenging scenarios, we analyze a micro-simulated case study with relatively high traffic volumes and long queues that do not exist in the NGSIM data. This also enables us to investigate the effect of low penetration rates, high sampling intervals, and measurement noise. The simulated case study is an arterial with four signalized intersections where length of links are respectively 350, 500, 400, and 450 (m), with time-varying demand for 20 minutes. The signals are traffic-actuated each with different cycle length. Note that no spillovers occur in this case study, while in the following section an extension of the proposed methodology handling spillovers is presented.

Figure 2.5(a) illustrates the estimated ($\phi = 20\%$, $T = 20$s) and ground truth queue profiles for all intersections and demonstrates that the performance of proposed queue profile estimation method is promising and can capture the fundamental characteristics of queue profiles without any information of arrival distribution and signal settings. Further, to scrutinize the proposed method robustness to measurement noise and map-matching errors, we add an error term to position and velocity measurements. The position and velocity error terms are assumed as a normal random variable with zero mean and standard deviations equal to 2.5 (m) and 1 (m/s), respectively. (These values are relatively pessimistic considering technological

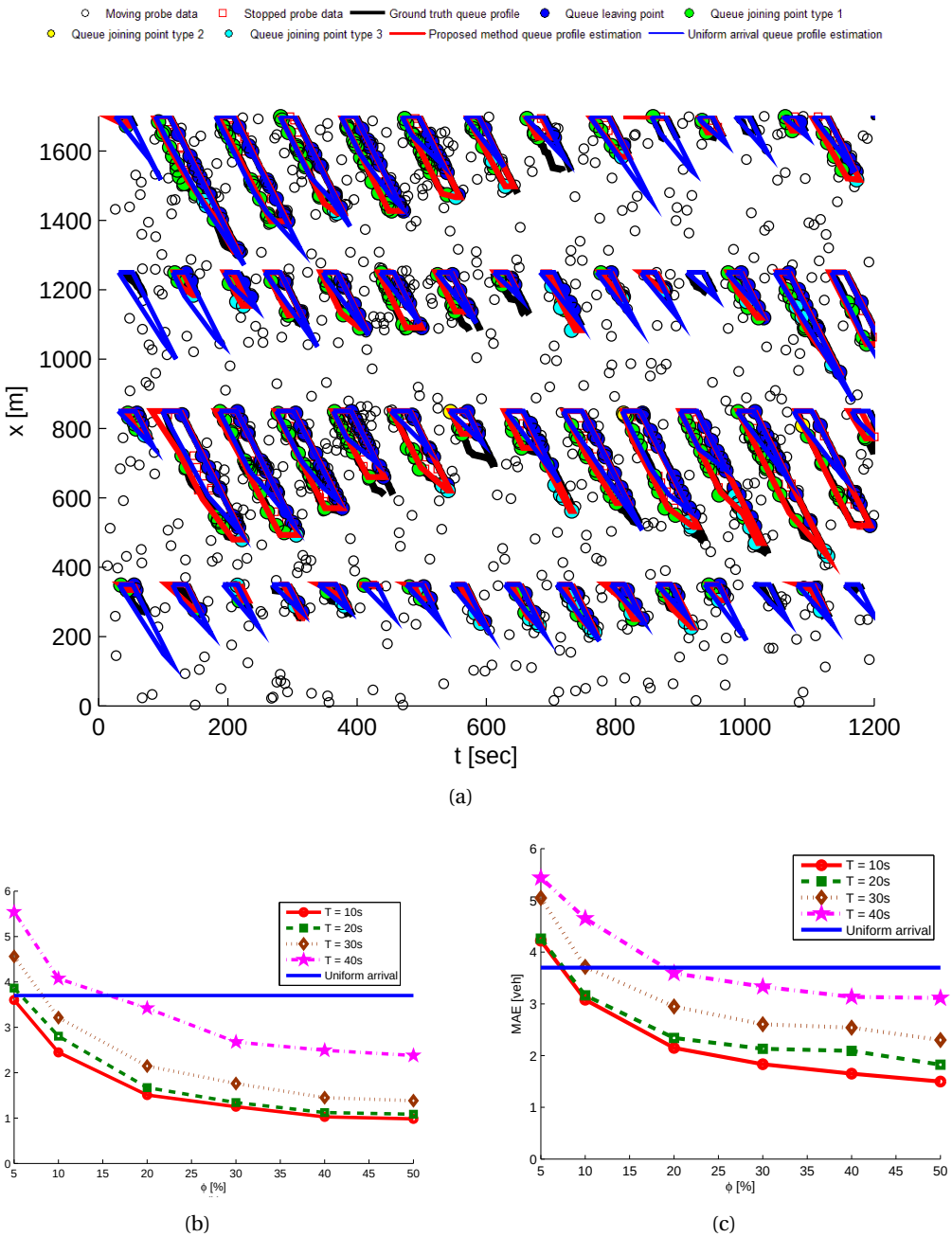


Figure 2.5: The ground truth (black) and estimated ($\phi = 20\%$, $T = 20$s) queue profiles for all 4 intersections based on the proposed method (red) and based on the uniform arrival assumption (blue) (a). The MAE between the ground truth and estimated queue length with different penetration rates and sampling intervals in case of no noise (b). The MAE between the ground truth and estimated queue length with different penetration rates and sampling intervals in case of noisy measurements (c).

advancements and developments of filtering, estimation and fusion algorithms, see [196] and [191].) Figures 2.5(b) and 2.5(c) depict the MAE between the ground truth queue size and the estimated queue size averaged over 10 runs for penetration rates, $\phi = 5, 10, 20, 30, 40, 50\%$, and sampling intervals, $T = 10, 20, 30, 40$s for two cases without and with measurement errors. It is apparent in Fig. 2.5(b) that even in case of sparse probe data the proposed method results are more accurate than the uniform arrival estimation (with the integration of residual queues modeling of [169]). It is also clear that higher penetration rates and lower sampling intervals are not necessary as the error is close to one vehicle for $T = 10$s and $\phi = 30\%$. Moreover, Fig. 2.5(c) reveals that the proposed method is robust to measurement errors. We also assess robustness of the proposed method to measurement errors closer to the state-of-the-art with standard deviations equal to 1 (m), and 0.2 (m/s) and the outcomes are statistically similar to no error case. Note that the proposed method parameters are similar in both case studies, which demonstrates the method is insensitive to a sensible range of parameters.

To scrutinize the method performance with very sparse probe data (e.g. 5% penetration rate), Fig. 2.6 shows the percentage of identified cycles, as the number of identified cycles in step 2 of the method divided by total number of cycles (a similar performance measure is also discussed in [10]). For $T = 20$s and $\phi = 20\%$, the percentage of correctly identified cycles is above 90%, while evidently cycles with lower queue size and associated with less congested traffic states have higher probability to be missed in case of very sparse probe data (see 5 missed cycles in Fig. 2.5(a)). In this example, the maximum queue size is 29 (veh) and the mean queue size is 7.9 (veh) (that is the MAE of a zero queue estimation, equivalent to no estimated queue profile or missed cycles). In case prior signal setting information is available (e.g. lower and upper bounds of red and green phases), the performance of the method can be improved. In Section 2.6, a method based on historical data is discussed to tackle missed cycles.

2.5 Spillover Bayesian Inference

In this section, we explore the effect of spillovers on the proposed queue profile estimation method, subsequently introduce an elegant probabilistic spillover identification procedure based on Bayesian inference and probe data observations, and finally incorporate the spillover identification within the queue profile estimation method.

2.5.1 Spillover inference

Spillovers occur during oversaturated conditions when exceeding queue at the downstream link impedes the arrivals from the upstream link, while the signal phase is green. This creates a *de facto* red phase that decreases the outflow of the link and consequently the network outflow. Recent studies have addressed the problem of spillover identification using conventional loop detectors ([56, 195]) and control strategies to avoid spillovers ([20]). However, the literature of methods identifying queue spillovers in urban networks with probe data is limited.

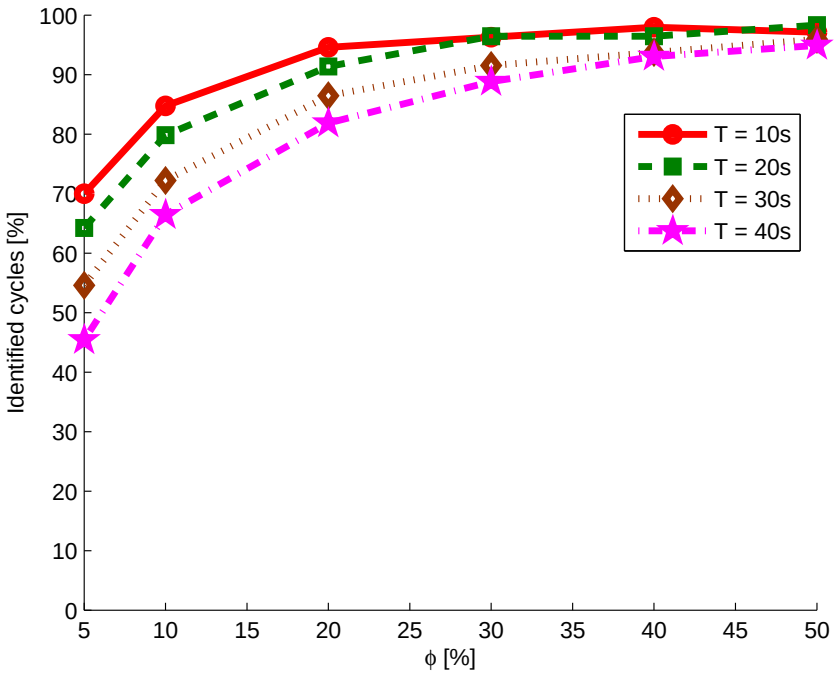


Figure 2.6: The percentage of identified cycles in step 2 of the method with different penetration rates and sampling intervals.

Given that in this chapter signal settings are unknown, occurrence of spillovers (and hence the *de facto* red) causes difficulties for step 2 of the proposed method (Section 2.3.2). In step 2, the stopped data points are clustered to cycles based on the projection profile method, while due to randomness of probe data, the clustering procedure disregards small number of zero-valued bins inside a cluster of positive-valued bins. Note that zero-valued bins of the accumulator vector represent intervals (the time between two stripes in x-t plane with slope w) without any stopped data points. So, if the *de facto* red occurs, step 2 might consider this red as a part of the nearest signal red phase (see Fig. 2.7(a)), or even as a whole with both adjacent red phases (see Fig. 2.7(b)) depending on the *de facto* red spatiotemporal extent. However in both cases, step 2 results in a longer red phase with a long interval (inside the long red phase) without any stopped data point observation, and also a systematic error in the estimation of leaving and joining points. The key idea of the proposed spillover identification method is to integrate these two indications from the upstream link along with the length of the queue into a probabilistic inference system to identify the occurrence of spillovers.

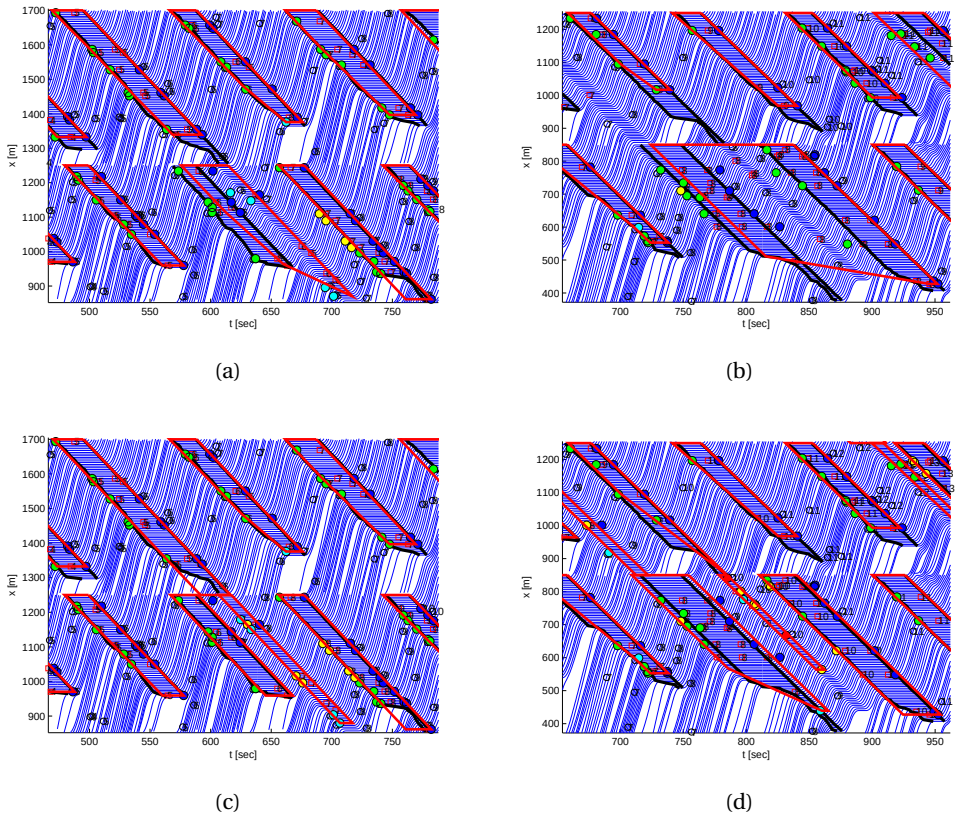
Bayesian inference method provides a probabilistic framework to update the prior probability, i.e. pr(H), of a set of hypotheses, $\{H_1, H_2, \ldots, H_h\}$, while further evidence regarding the hypotheses is acquired. Bayesian inference method estimates the posterior probability of the

hypotheses given the evidences, i.e. $\text{pr}(H|E)$, as

$$\text{pr}(H_i|E) = \frac{\text{pr}(E|H_i)}{\sum_h \text{pr}(E|H_h)\text{pr}(H_h)}\text{pr}(H_i) \tag{2.19}$$

where for a set of independent and identically distributed evidences, $\{e_1, e_2, \ldots, e_k\}$,

$$\text{pr}(E|H_i) = \prod_k \text{pr}(e_k|H_i). \tag{2.20}$$



(a) (b)

(c) (d)

Figure 2.7: The ground truth (black) and estimated ($\phi = 10\%$, $T = 20$s) queue profiles based on the proposed method (red). The adverse effect of spillover and de facto red results in longer than usual red phases, R_{r_j}, with long Z_{r_j}, and also a systematic error in leaving and joining point estimation (a). The adverse effect of the de facto red when it combines with two adjacent red phases (b). The two modifications of the spillover inference module on $x_j^{\min}$ and r_j (c). The adjustments of the spillover inference module when the de facto red combines with two adjacent red phases (d). Note that blue lines are vehicle trajectories and are depicted for illustration purposes.

To infer the probability of a spillover occurrence at cycle j of the downstream link, we need to pre-process steps 1 and 2 for the upstream link to identify corresponding red phases. Note that by definition, spillovers occur during green phases, whereas this duration will be

accounted as a (*de facto*) red phase in the proposed queue profile estimation method. So, if queue profile j spills over, it must reach to the upstream intersection during a *de facto* red phase, r_j. As mentioned above, duration of r_j, R_{r_j}, is longer than other (normal) red phases. Likewise, the maximum interval (the time between two stripes in x-t plane with slope w) within r_j without any stopped point, Z_{r_j}, is also longer than the corresponding value of other red phases. Therefore, the magnitude that R_{r_j} and Z_{r_j} are outliers and greater than the rest of samples in their corresponding set (respectively $\mathscr{R}$ and $\mathscr{Z}$ values for red phases of the upstream link) can be regarded as evidences for spillover occurrence. An intuitive equation to designate the extent of being outlier is as follows:

$$Q_a = \frac{a - \bar{\mathscr{A}}}{\sigma(\mathscr{A})}, \qquad a \in \mathscr{A} \tag{2.21}$$

where a is the sample and $\mathscr{A}$ denotes the set with mean and range (difference between the largest and smallest values), respectively as $\bar{\mathscr{A}}$ and $\sigma(\mathscr{A})$.

Lastly, we consider the evidences in Bayesian inference as, e_1: the position of the furthest queued vehicle from the intersection, $x_j^{\min}$; e_2: $Q_{R_{r_j}}$; e_3: $Q_{Z_{r_j}}$. The two complementary possible hypotheses are H_1: occurrence of spillover and H_2: no occurrence of spillover. Furthermore, the likelihood functions, $\text{pr}(e_k|H_i); i = 1,2; k = 1,2,3$, are

$$\text{pr}(e_1|H_1) = \frac{1 - e^{-(\frac{x^{\max} - x_j^{\min}}{x^{\max}})}}{1 - e^{-1}} \tag{2.22}$$

$$\text{pr}(e_2|H_1) = \max(\delta, Q_{R_{r_j}}) \tag{2.23}$$

$$\text{pr}(e_3|H_1) = \max(\delta, Q_{Z_{r_j}}) \tag{2.24}$$

$$\text{pr}(e_k|H_2) = 1 - \text{pr}(e_k|H_1) \tag{2.25}$$

where δ is a small predefined positive value near zero to avoid negative and zero values. More sophisticated inference methods or optimizing the likelihood functions are a future research direction.

2.5.2 Integration of spillover identification into queue profile estimation

The Bayesian spillover inference is an intermediate step between steps 2 and 3 of the queue profile estimation method that assigns a spillover probability to each queue profile. With this inference method, queue profiles are grouped into two distinct sets, ones with spillover probability close to zero and ones close to 1, see Fig. 2.8. Thus, a prescribed threshold distinguishes the two sets. Two modifications should be applied to queue profile j that is labeled as spillover, i.e. its associated spillover probability is higher than the predefined threshold. First, $x_j^{\min}$ should be set equal to the position of the link upstream (instead of (2.14a)). Second, r_j should be divided into the *de facto* red and the normal signal red phase during Step 2 for the upstream link. The division should be placed at the interval corresponding

to Z_{r_j}. Figure 2.7(c) illustrates the two modifications that have been applied to the example in Fig. 2.7(a). In case the *de facto* red was considered with the two adjacent red phases (e.g. Fig. 2.7(b)), r_j should be split into two intervals accordingly. A simple procedure is to check if the second largest interval without a stopped point observation is close enough to Z_{r_j}. Figure 2.7(d) illustrates the same example in Fig. 2.7(b) with modification on $x_j^{\min}$ and dividing r_j into two signal red phases and a *de facto* red. The spillover inference procedure acquires the correlation of queue spatiotemporal extent between two successive links and can identify the spillbacks that propagate to more than one upstream link.

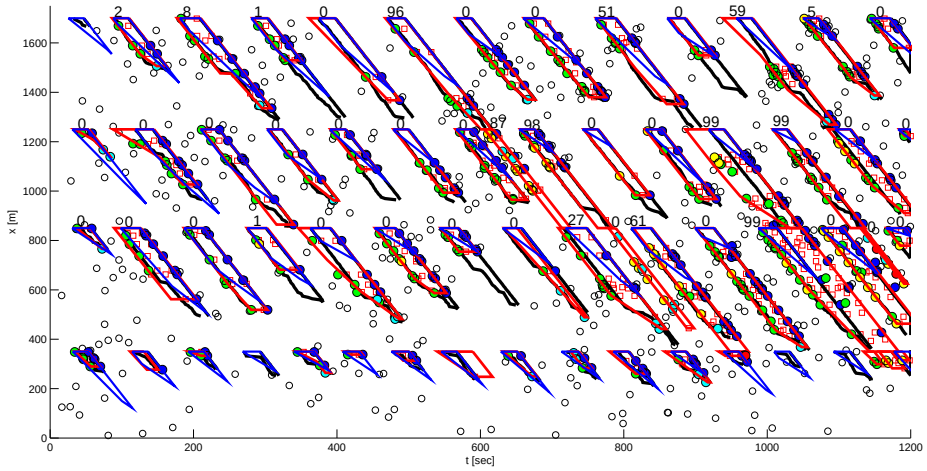


Figure 2.8: The ground truth (black) and estimated ($\phi = 10\%$, $T = 20$s) queue profiles for all 4 intersections based on the proposed method (red), and based on the uniform arrival assumption (blue). The values (in %) are the spillover probability associated for each queue profile.

Figure 2.8 illustrates the queue profiles of a case study with the same topology but higher demand with oversaturated conditions and spillovers. It is apparent that the proposed queue profile estimation method with sparse probe data captures the evolution and fundamental attributes of queues and spillovers in the signalized arterial. Furthermore, Fig. 2.9 shows the MAE between the ground truth queue size and the estimated queue size averaged over 5 runs for various penetration rates and sampling intervals. It is clear that the outcome of the proposed method is more precise than the estimation of uniform arrival procedure. Even if the effect of oversaturated condition has been integrated in the uniform arrival estimation (see for example the blue trapezium in Fig. 2.8), the developed method is superior even with 5% penetration rates. If oversaturated conditions are not embedded in the uniform arrival estimation and only triangular queue profiles are applied, the errors would be much larger.

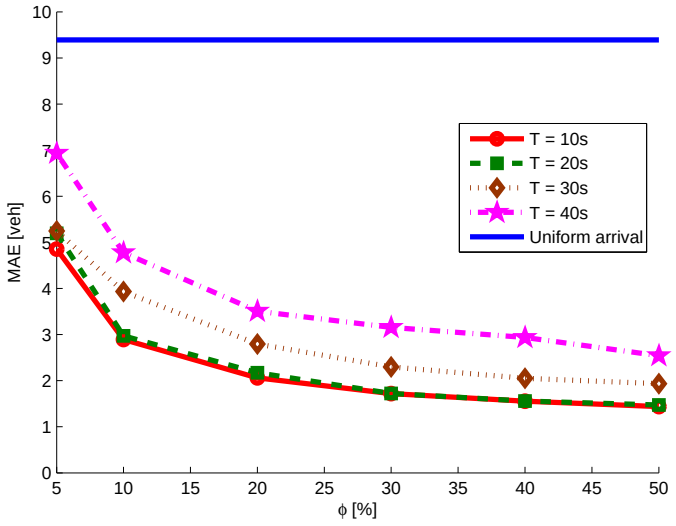


Figure 2.9: The MAE between the ground truth and estimated queue length with different penetration rates and sampling intervals in case of spillovers.

2.6 Summary

This chapter has presented a method to estimate queue shockwave profiles at signalized intersections in urban networks, based on LWR traffic theory and dispersed data of position and velocity from probe vehicles. The significance of the proposed method is the estimation of shockwaves without any explicit information of signal settings and arrival traffic flow patterns. The method incorporates the physics of traffic, longitudinal kinematics, and interdependencies between adjacent links into an optimization framework. The outcome of the proposed method for two case studies with field and simulated data seems promising, especially for congested conditions. The method is robust to noisy measurements and more accurate once compared to a uniform arrival queue profile estimation procedure. The effect of penetration rates and sampling interval of probe data on the performance of the method is also investigated. In addition, the method is applicable to oversaturated conditions and with integration of Bayesian inference; the method infers the probability of spillover occurrence.

The queue profile estimation method also provides signal settings, i.e. phase durations, at each intersection. However, the method cannot identify a cycle without any probe data. So the proposed method needs to be further explored to tackle missing cycles, which might occur in case of low penetration rates and high sampling intervals. To this end, prior assumptions on minimum or maximum phase duration can enhance the method regarding missing cycles, while the remainder of the methodology remains the same. Also, given the estimated queue shockwaves, one can calculate a historical value for phase durations based on the average length of red and green intervals (except the outliers). Then long (outlier) green intervals can be detected, which probably include missed cycles, and be decomposed to several green and red phases, based on the historical phase durations. This provides a proxy for number

of missed cycles and, assuming uniform arrival rate, an estimator of back of queue length. However, this is not accurate for queue length estimation as biased information of signal settings (e.g. several seconds lag) leads to significant queue measurement error.

The unique feature of queue shockwave profile estimation is that it provides the comprehensive dynamics of the urban traffic flow, which can be further applied for performance measurement and control applications. For instance, the estimated queue shockwave profile can be readily utilized for delay analysis (mean and distribution) and queue length estimation. Moreover taking into account that the arrival traffic flow distribution can be estimated from the estimated queue profile, a noteworthy application of the proposed method is vehicle trajectory reconstruction, which is not limited to probe vehicles and particularly can be applied for all vehicles. In other words, the proposed method estimates trajectories of all vehicles given sample data of only sensor-equipped probe vehicles. Estimation of detailed trajectory of vehicles can be further applied for (rough) emission or fuel consumption estimation models.

Estimation of vehicles trajectories also enables the method to tackle the travel time decomposition problem [77], which is about how to decompose the travel time between two successive probe data with high sampling interval (e.g. 3 min) to individual link travel time. Although the proposed method requires a sampling interval that produces enough probe data to estimate the queue profiles, notably it can provide the link travel time for all the vehicles and not only for probe vehicles. Chapter 3 will introduce a method to estimate arterial route travel time distribution based on the link travel times of probe vehicles. Consequently, the estimated link travel times from the proposed queue shockwave estimation method provides the input to the route travel time distribution estimation method in Chapter 3. Link and route travel time distribution is a crucial performance index for travel time reliability models.

3 Estimation of arterial route travel time distribution with Markov chains

RECENT advances in the probe vehicle deployment offer an innovative prospect for research in arterial travel time estimation. Specifically, our focus is on the estimation of probability distribution of arterial route travel time, which contains more information regarding arterial performance measurements and travel time reliability. One of the fundamental contributions of this chapter is the integration of travel time correlation of route's successive links within the methodology. In the proposed technique, given probe vehicles travel times of the traversing links, a two-dimensional (2D) diagram is established with data points representing travel times of a probe vehicle crossing two consecutive links. A heuristic grid clustering method is developed to cluster each 2D diagram to rectangular sub spaces (states) with regard to travel time homogeneity. By applying a Markov chain procedure, we integrate the correlation between states of 2D diagrams for successive links. We then compute the transition probabilities and link partial travel time distributions to obtain the arterial route travel time distribution. The procedure with various probe vehicle sample sizes is tested on two study sites with time dependent conditions, with field measurements and simulated data. The results are very close to the Markov chain procedure and more accurate once compared to the convolution of links travel time distributions for different levels of congestion, even for small penetration rates of probe vehicles.

3.1 Introduction

Nowadays, traffic congestion is a widespread time-consuming phenomenon in urban areas and the primary step for improving conditions is traffic observation and data collection. Hence, network monitoring is a crucial component in management of transportation systems for traffic control and guidance purposes. The introduction of Intelligent Transportation System (ITS) technologies and new sensing hardware promise significant progress in reducing the congestion level in cities. The integration of Global Positioning System (GPS) technology within the ITS framework introduces a new paradigm in traffic surveillance: probe vehicles. Compared to fixed traffic sensors (e.g. inductive loop detectors), probe vehicles offer further data like vehicle trajectory in a more convenient manner. In principle, a steadily incremental

public deployment rate, low maintenance cost, and inherent distributed characteristics lead to tackling GPS-equipped vehicle challenges in traffic monitoring research. Nevertheless, the estimation methodology should not be constrained to GPS information from cars, but may be applied to any type of Automatic Vehicle Location (AVL) mobile sensors found in abundance in the form of commercial fleets like UPS, FedEx, taxis and transit vehicles, given the broad existence of filters which distinguish between actual congestion and a stopping delay (see for example [12] for transit vehicles as probes or [55, 79] for other vehicles). The chief struggle with utilizing probe data is that travel time of a probe vehicle is principally a sample of a random variable, i.e. travel time. This raises inquiries about the probe vehicles penetration rate required to have a proper sample set for travel time estimation.

Travel time is a crucial index in assessing the operational efficiency of traffic networks. It establishes a common perception among all the perspectives of individual travelers and practitioners. In addition, it can be an indicator of congestion level of transport network once compared to the free flow travel time. With respect to monitoring, reliable and efficient estimation of travel time and other performance measures is still not a wide spread accomplishment on arterials, since it requires extensive sensor infrastructure, normally found only on freeway networks. The issue is not only that the existing monitoring infrastructure in arterials is less dense than in freeways but also that arterial network traffic dynamics are inherently different than these of freeways and fixed sensors cannot always provide the required level of data. The main reasons are randomness in supply and demand of the dynamic urban network (see for example [38]), the signaling effect (alternation of green and red phases in short time intervals), and the characteristics of route choice (vehicles in arterials can randomly turn at intersections and either begin or finish their trips along the street itself, which is not the case in freeways). Meanwhile, speed of vehicles at a given time in the network is not a deterministic quantity over space because of drivers' behaviors (conservative vs. aggressive drivers), the spatial effect of signals (near the stop line vs. further upstream) and temporal-spatial pockets, where average speed is temporarily different than the widespread average, e.g. point bottleneck in a freeway system. To simplify matters, if we track a vehicle on a freeway for 15secs we can estimate with high level of confidence the mobility level in the time-space proximity of the vehicle. Instead, this is not the case in arterials as there are variations in travel times even for vehicles traveling in the same link during the same cycle length.

Reduction in travel time variability is at least as desirable as reduction in mean travel time [87], since it decreases commuting stress and uncertainty of mode- and route-choice decision making. Travel time variability designates the variation of various trip travel times over a specific path. Travel time variability can be investigated from several point of views [130]: vehicle-to-vehicle variability which corresponds to different vehicles traveling the same route at the same time, period-to-period variability corresponding to vehicles traveling the same route at different periods within a day, and day-to-day variability addressing the travel time variations of vehicles crossing the same route at the same period of time on different days. Different indexes of travel time stochasticity-reliability are presented in [93]. In this chapter, we model the vehicle-to-vehicle variability and analyze the probability distribution of travel

time for arterial routes (expressed as series of links). The input to the proposed model is probe vehicles travel times of all links that are traversed and belong to the route; and the output is the Travel Time Distribution (TTD) over the study time horizon. The developed model is based on Markov chain to address both the traffic progression and correlation between links.

Note that the urban TTD should not be estimated for periods less than one cycle, because this might over- or under-estimate travel times depending on the period of observation (red or green period). Also the analysis of the travel time characteristics for a short period, e.g. 2 or 3 times the cycle length, can significantly be influenced by the size of study period. For example, choosing the start or end of the study period at the beginning or end of the cycle could cause drastic changes in the results [204]. However, by analysing longer study periods; we can smooth out the traffic variations caused by time-dependent signal capacity. Nevertheless, we stress that the period of analysis should not be too large, as in this case the spatial correlations between links may be influenced by variant traffic regimes. In the latter case, classifying the period of analysis to traffic regimes is required.

This chapter is organized as follows. Section 3.2 provides the state of the art of travel time estimation. In Section 3.3, we briefly discuss the Markov chain procedure, its application in traffic engineering, and we illustrate our motivation via an example. Then, we introduce our proposed method in Section 3.4. The study site, data, and simulation details are presented in Section 3.5, while results and discussion are described in Section 3.6. Finally in Section 3.7, conclusions are drawn.

3.2 State-of-the-art of travel time estimation

There is a vast literature addressing different travel time estimation approaches for diverse applications and terms. Initial approaches for travel time estimation on signalized links include point delay models [188, 7], speed vs. volume to capacity ratio relations, and procedures based on the Highway Capacity Manual [71]. The latter calculates average travel time as the sum of the running time and the intersection delay, based on a deterministic point delay model plus some stochastic components. Such approaches are not well-suited for real-time applications, especially for time-dependent congested conditions. The results from the application of several conventional approaches on two arterial sites indicate that these methods produce large differences against field measured and predicted travel times [176]. The same reference presents a detailed review of arterial travel time models.

Recently, an analytical model was developed to estimate the travel times on arterial streets based on 15-30 seconds flow and occupancy data provided by loop detectors and the signal settings at each traffic signal [168]. The model considers the spatial and temporal queuing characteristics at the traffic signals and the signal coordination to estimate travel time in arterials. Several extensions and enhancements to the analytical model were developed and implemented by the same authors which explicitly address the issues of long queues and spillovers that frequently occur on arterials in urban areas [169, 56]. The model has also been

integrated into a pilot arterial performance measurement system in California (APemS). A similar model was also developed in [112], which utilizes the exact times that the vehicles cross the upstream detector (from individual vehicle detector actuations).

The aforementioned approaches require precise signal phase times and instrumentation at each intersection, which might be expensive or non-existing. Also, these studies focused on estimation of *average travel time* of vehicles travelling in one cycle time. Although statistical scalar indexes (e.g. mean, variance, percentiles, etc.) might characterize distributions to some extent, they are not fully enlightening about travel time variability, once compared to TTD. An alternative to characterize the TTD could be to introduce an appropriate number of percentiles and approximate the TTD, by choosing these percentiles when sharp changes occur in the slope of cumulative probability distribution. In this chapter, we utilize the full TTD approach.

To address the uncertainty issue, a new trend is seen in recent travel time research. Kwong et al. [105] developed a methodology for vehicle re-identification using wireless magnetic sensors in arterial routes, without the need of signal settings. The high accuracy of matching vehicle signatures between different locations provided accurate estimates of empirical travel times (mean and distribution) for different locations. In [65], a multi-state model is employed to fit a mixture of Gaussian distributions into travel time observations of an expressway corridor. Each normal distribution is associated with an underlying traffic state providing quantitative uncertainty evaluation. The multi-state mixture model results in better fitting, revealing that TTD usually has more than one mode which is entirely dependent on time horizon of study, demand, topology, etc. A recent work of the same group [140] tries to quantify the impact of traffic incidents on TTD using multi-state (3 states) models demonstrating that incidents increase the travel time variability. The results indicate that the state corresponding to the congested regime become more dominant, yet there is no need to add a new component to the multi-state model. A similar approach is investigated, utilizing mixtures of normal distributions to estimate mean travel times for arterial routes with Next Generation Simulation (NGSIM) [129] data [41].

Uno et al. [177] discuss route travel time variability using bus probe data. In pre-processing stage of their method, a map matching procedure and a data filtering for dwell time elimination are performed. Afterwards, they decompose a route to sections with lognormal TTD to estimate the bus route TTD. In [42] a statistical evaluation is investigated, which tries to assess the feasibility of probe vehicles employment for collecting traffic information. The authors provide analytical solutions about data sampling, reporting rates, and probe vehicles penetration level in a single road link without any validation. They assume a binomial distribution for number of probe vehicles and a Poisson distribution for reporting rate, deriving formulas for mean and variance of reports number and speed estimation, and confident reporting intervals. Delay at signalized intersections is the main source of uncertainty in urban TTD which is tackled in [204]. The authors propose an analytical method for estimation of an urban link delay distribution. The results indicate a correlation between arrival time and link

travel time under different degree of congestion. Their model demonstrates evolutions of delay distribution as well, so that both average and variance of delay increase cycle by cycle. Non-recurrent and peculiar events, such as incidents, lane closures, and sport events can also cause significant deviations from recurrent conditions [104].

Sparse probe vehicles is discussed in [80] for arterial traffic estimation and short term prediction of travel time. The authors propose a statistical modeling framework that captures the evolution of traffic flow as a Coupled Hidden Markov Model (CHMM). The authors assume that each link in the network has one state evolving over time based on a time-invariant state transition matrix. Given the states of the spatial neighbors of the link, independence of state transitions from all other current and past link states is assumed. They also consider a time-invariant travel time distribution (Gaussian) for each state of each link, representing independence of link travel time from other traffic variables, given the link state. This work also decomposes path travel times from probe data to individual link travel times. The evaluation is done using dataset from a taxis fleet in San Francisco, CA, as a part of Mobile Millennium project. To model the correlation and dependency in transport networks, Rakha et al. [148] have tried to estimate the variance of freeway route travel time by modeling correlation between segment travel time variances. Geroliminis and Skabardonis [50] also estimated the variance of urban route travel time by assuming linear correlations between successive links travel times. To the best of our knowledge, correlation between travel times in urban links has not been explicitly addressed in the recent literature.

3.3 Motivation

Given all individual link TTDs, the simplest model for route TTD estimation is to aggregate those independently. Assume a route consisted of K links with signalized intersections. The route TTD is then computed according to:

$$TTD_K = TTD_1 * TTD_2 * \cdots * TTD_k \tag{3.1}$$

$$(TTD_i * TTD_j)(t) \stackrel{\Delta}{=} \int_{-\infty}^{\infty} TTD_i(\tau) TTD_j(t-\tau)\,\mathrm{d}\tau, \tag{3.2}$$

where the $(*)$ mathematical operator expresses convolution [99] and the left term in 3.2 is the probability density over time t for two links $(i, j = 1, 2, \cdots, k)$ given both TTDs a priori. Evidently, the above method considers independence between link TTDs, and consequently, any spatiotemporal correlation information is neglected.

To show the impact of correlation in estimation of TTD, we demonstrate a hypothetical "toy example" in which there is a strong correlation between travel time data. Imagine a route, which consists of two serial signalized links. Figure 3.1a depicts a 2D diagram in order that each point denotes links 1 and 2 travel times of one probe vehicle. In this case, we can infer there are 4 types of vehicles; fast in both links (left-lower part), fast in link 1 and slow in link 2 (left-upper part), slow in link 1 and fast in link 2 (right-lower part), and slow in both links

(right-upper part). Given the link TTDs in Fig. 3.1b, the outcome of the convolution method (convolved route TTD) is far from the real TTD.

To capture correlation patterns between link travel times, one can cluster travel times of one link to different states, e.g. in this example two states (slow and fast) are defined for each link. The initial probability, π, of each of the states in link 1 and the transition probabilities, P, between states of link 1 and states of link 2 as traffic progresses are: (numbers represent the toy example)

$$\pi = \begin{bmatrix} \pi_{\text{fast}} \\ \pi_{\text{slow}} \end{bmatrix} = \begin{bmatrix} \frac{N(1)+N(2)}{N(1)+N(2)+N(3)+N(4)} \\ \frac{N(3)+N(4)}{N(1)+N(2)+N(3)+N(4)} \end{bmatrix} = \begin{bmatrix} 0.58 \\ 0.42 \end{bmatrix}, \tag{3.3}$$

$$P = \begin{bmatrix} P_{\text{fast,fast}} & P_{\text{fast,slow}} \\ P_{\text{slow,fast}} & P_{\text{slow,slow}} \end{bmatrix} = \begin{bmatrix} \frac{N(1)}{N(1)+N(2)} & \frac{N(2)}{N(1)+N(2)} \\ \frac{N(3)}{N(3)+N(4)} & \frac{N(4)}{N(3)+N(4)} \end{bmatrix} = \begin{bmatrix} 0.84 & 0.16 \\ 0.29 & 0.71 \end{bmatrix}, \tag{3.4}$$

where $N(i), i = 1, \cdots, 4$ is the number of points inside each cluster (as shown in Fig. 3.1a), $\pi_a, a \in \{\text{slow}, \text{fast}\}$ is the initial probability of vehicles being slow or fast in link 1, and $P_{a,b}, a, b \in \{\text{slow}, \text{fast}\}$ is the corresponding probability of being slow and/or fast in links 1 and 2, respectively. Given 2 states for each link, there are 4 combinations of states, hereafter named as Markov paths. Every one of 4 Markov paths has a probability of occurrence and a corresponding TTD. The TTD of each Markov path refers to the convolution of partial TTD of link states where the TTD of a Markov path is conditioned on the states of the links in the corresponding Markov path. The partial TTD of link states is the TTD of a link conditioned on the state of that link (The complete definition of partial TTD is given in Section 3.4.1). For the sake of brevity, only the TTDs for Markov paths of fast-fast and slow-slow states are depicted in Fig. 3.1c and 3.1d, respectively. At the end of procedure, the TTD of a Markov path is multiplied by the path initial and occurrence probabilities and the mixture is calculated to find the route TTD. The final result of estimation method is closely matched with the real TTD as it is apparent in Fig. 3.1e.

Note that in the above example, the structure of correlation is known a priori which makes the state definition straightforward. Nevertheless, improper state definition (erroneous clusters) will degrade the significance of the results. Further, travel times of consecutive links in reality might not be well-ordered as in Fig. 3.1a. Instead, they are more similar to Fig. 3.2 which makes the state boundaries less clear and the privilege effect of partitioning less straightforward. This contamination of data leads to a grid clustering problem, yet the whole concept remains the same. The above example can be applied to several links using Markov chain procedure which is briefly reviewed in next subsection. The complete version of the method is presented in Section 3.4.

3.3.1 Markov chain procedure

Markov chain is a technique for statistical modeling of a random process in which the state of system changes through progression. A Markov chain is entirely demonstrated with the set of

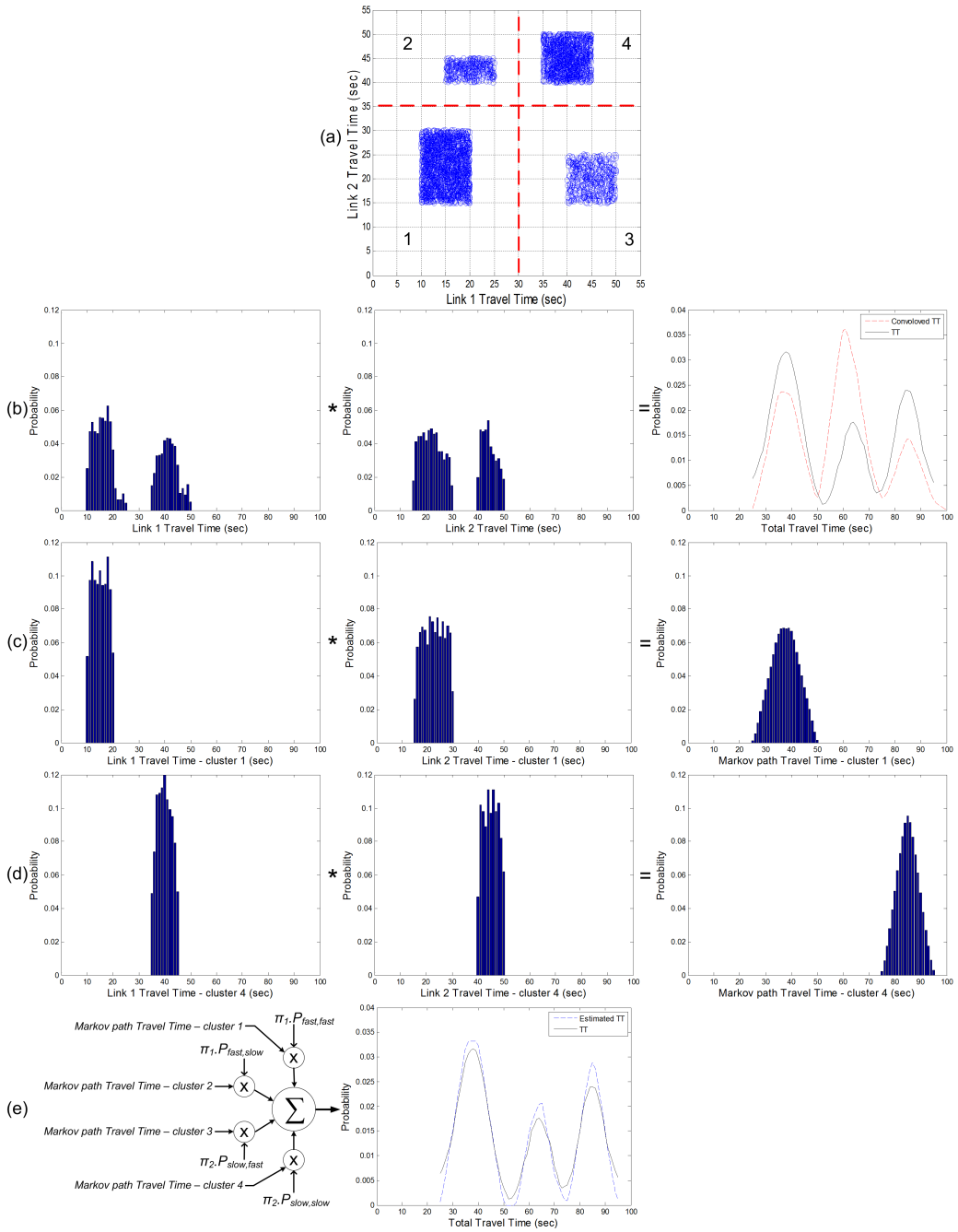


Figure 3.1: Explanatory example about motivation of Markov chain integration into TTD estimation: (a) 2D diagram representing four groups of vehicles, (b) convolution of mapping of all vehicles link 1 travel time and mapping of all vehicles link 2 travel time, (c) convolution of fast vehicles link 1 and 2 travel time distributions, (d) convolution of slow vehicles link 1 and 2 travel time distributions, and (e) mixture of TTD of all groups of vehicles.

state definition, initial probabilities and transition probabilities. The transition probabilities are associated with the manner of state progression during the system evolution. A system which has the Markov property satisfies the following: the conditional probability of the system being at the next state, s_{t+1}, given the current state, s_t, depends only on the current state and not on the previous states of the system.

$$\text{pr}\{s_{t+1} = s' | s_t, s_{t-1}, \cdots, s_1\} = \text{pr}\{s_{t+1} = s' | s_t\}. \tag{3.5}$$

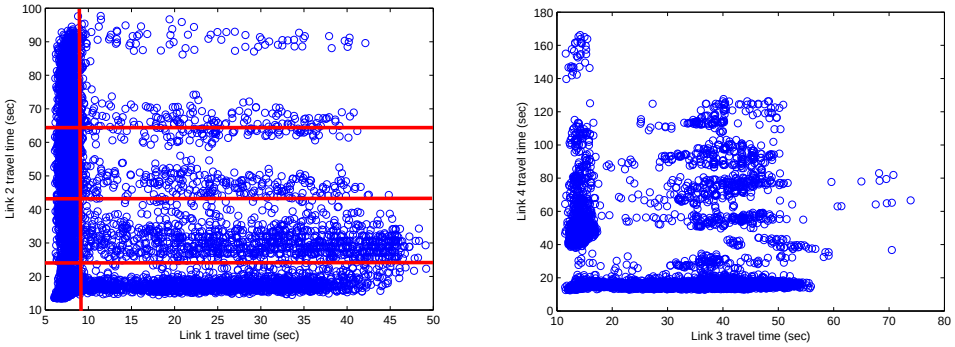
The Markov property empowers Markov chain to capture both probabilistic nature of travel time and the fundamental correlated feature of successive links travel times. In other words, traffic spatial progression in arterials is similar to a Markov chain where the current link travel time of a vehicle depends only on the travel time of immediate upstream link, equivalent to the right term in 3.5, which is well-matched with physics of traffic. Note that, the traffic spatial progression designates how the moving vehicles in the arterials encounter travel time states while traveling along links; and this definition is different from the evolution of link congestion levels (congestion propagation).

Markov chain is utilized in vast fields of transportation research. Discrete time Markov chain for estimation of expected freeway travel time is investigated in [198] where states correspond to congestion level of links. The authors find the average travel time of each link both in non-congested and congested states using field data and with consideration of transition probabilities between different states, the route mean travel time is estimated. In [37] a Markov chain is developed to model the effect of freeway flow breakdown and recovery in travel time reliability. Geroliminis and Skabardonis [168] also proposed an analytical model using Markov chain for prediction of platoon arrival profiles and queue length considering platoon dispersion in arterials. The CHMM have been also applied in [80] for traffic estimation and prediction.

3.4 Methodology

In the proposed model, the raw measurements are experienced individual link travel times traversed by a set of probe vehicles. A probe vehicle path may consist of one to as many links as the study route has, which makes the number of link travel times reported by every probe vehicle different. With high resolution GPS data (position and time stamp), finding the trajectory of a moving vehicle and link travel times is not a complicated task. Note that errors in trajectories, map matching, or low resolution data are not addressed in this work. Afterwards, travel times of all probe vehicles crossing two successive links during data collection period are used to construct a 2D diagram. The 2D diagram is a graphical representation of vehicles travel times joint distributions. Given a route consisting of K links, $K-1$ 2D diagrams are established in order to identify the Markov chain, i.e. defining its states and determining the initial and transition probabilities. Figures 3.2(a) and 3.2(b) illustrate such a diagram for links 1-2 and links 3-4 of a study site that will be described later. Each dot represents the travel

time of one probe vehicle in each of the two links. Figure 3.2(a) depicts a significant fraction of vehicles crosses link 1 or 2 without any delay, but also a large variation of travel times. Note that there is some correlation between travel times in successive links, which cannot be expressed in a linear way (linear correlation is almost zero).



(a) (b)

Figure 3.2: 2D diagrams showing joint distributions of successive link travel times. (a) links 1-2, and (b) links 3-4.

3.4.1 Markov chain identification

As 2D diagrams are constructed, states, transition probabilities, and initial state probabilities of Markov chain should be identified. We define a state of a link as travel times in a certain interval between two values. Let $\mathscr{X}_l = \{x_1^l, \cdots, x_{m_l-1}^l\}$ and $\mathscr{Y}_l = \{y_1^l, \cdots, y_{n_l-1}^l\}$ denote sets of boundaries in 2D diagram l (correspond to links l and $l+1$) producing m_l and n_l states for link l and $l+1$, respectively. In this manner, the first state of link l indicates travel times in $[\min \tau_l, x_1^l)$ and the last state represents travel times within $[x_{m_l-1}^l, \max \tau_l]$, where τ_l denotes the set of probe vehicles travel time measurements in link l. This kind of state definition yields to rectangular clusters in 2D diagrams which is used to define initial state and transition probabilities. A possible clustering of a 2D diagram is depicted in Fig. 3.2(b) where there are 2 states in link 1 and 4 states in link 2. The clustering procedure will be more elaborated in Section 3.4.2.

Initial state probabilities in Markov chain are the probabilities of link 1 states which are as follows:

$$\pi = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_{m_1} \end{bmatrix} = \begin{bmatrix} \frac{N(1)}{\sum_{i=1}^{m_1} N(i)} \\ \vdots \\ \frac{N(m_1)}{\sum_{i=1}^{m_1} N(i)} \end{bmatrix}, \tag{3.6}$$

where $N(i)$ denotes number of data in state i of link 1 (i.e. link 1 travel times in $[x_{i-1}^1, x_i^1)$). To

identify transition matrix between two successive links, transition probabilities should be defined. The generic transition matrix between each pair of successive links is as following:

$$P = \begin{bmatrix} p_{1,1} & \cdots & p_{1,n_l} \\ \vdots & \ddots & \vdots \\ p_{m_l,1} & \cdots & p_{m_l,n_l} \end{bmatrix} = \begin{bmatrix} \frac{N(1,1)}{\sum_{i=1}^{n_1} N(1,i)} & \cdots & \frac{N(1,n_l)}{\sum_{i=1}^{n_1} N(1,i)} \\ \vdots & \ddots & \vdots \\ \frac{N(m_l,1)}{\sum_{i=1}^{n_1} N(m_l,i)} & \cdots & \frac{N(m_l,n_l)}{\sum_{i=1}^{n_1} N(m_l,i)} \end{bmatrix}, \tag{3.7}$$

$$p_{i,j} = \text{pr}\{S_{l+1} = j | S_l = i\}, \tag{3.8}$$

where S_l indicates the state of travel time in link l, and $N(i,j)$ is the number of data points within each rectangular cluster confined by state i in link l and state j in link $l+1$.

By definition of 2D diagrams, travel time observation of each link (except the first and the last links) are used in two 2D diagrams, one as upstream link in x-axis and the other as downstream link in y-axis. If the states for link l are identical in both 2D diagrams $(l-1)$ and l, i.e. $\mathscr{Y}_{l-1} = \mathscr{X}_l$ and $m_l = n_l, l = 1, \cdots, k$, there will be $Q = \prod_{l=1}^{k} m_l$ state combinations from origin to the destination of the route and each one of them is named as a Markov path (violation of this condition is discussed in Section 3.4.2). For a given Markov path, all of the transition probabilities between states of all links are multiplied to compute Markov path occurrence probability (note that Markov property makes joint probability of each cluster of 2D diagrams independent),

$$\text{pr}\{S_1 = i_1, S_2 = i_2, \cdots, S_k = i_k\} = \pi_{i_1} \cdot p_{i_1,i_2} \cdot p_{i_2,i_3} \cdots p_{i_{k-1},i_k}. \tag{3.9}$$

Then we obtain a Markov path TTD using convolution of partial TTD of each link state, assuming a conditional independence between partial TTDs (hence state definition is of great importance to gain the conditional independence),

$$TTD\{S_1 = i_1, S_2 = i_2, \cdots, S_k = i_k\} = TTD(i_1) * TTD(i_2) * \cdots TTDi_k. \tag{3.10}$$

$TTD(i_l)$, partial TTD of link l, is the empirical distribution (histogram) of data points conditioned on the state of links $l-1$, l, and $l+1$, i.e. i_{l-1}, i_l, and i_{l+1}. In the same way, for the first and last links this conditioning is only valid for the following and preceding links, respectively. Finally, the route TTD is computed as a mixture of distributions such that each primary distribution component is the TTD of a Markov path (Eq. 3.10) and the corresponding weight of each distribution component is the path occurrence probability (Eq. 3.9),

$$TTD_K = \sum_{q=1}^{Q} \text{pr}(\text{Markovpath}_q) \cdot TTD(\text{Markovpath}_q). \tag{3.11}$$

3.4.2 Clustering

The chief challenge in our method is to identify rectangular clusters properly (i.e. $\mathscr{X}_l$ and $\mathscr{Y}_l$) so states exhibit homogenous travel time characteristics. In other words, any type of correlation or non-random pattern within a cluster is undesirable. Intuitive thumb rules like travel time of free flow, near capacity, and oversaturated conditions may be utilized to address the state identification problem. Furthermore, a heuristic grid clustering method is introduced to determine boundaries based on the structure of 2D diagrams in order to have more homogenous states. Although the rectangular clusters do not have the maximum flexibility to capture the entire correlation patterns in a 2D diagram, their introduction is a matter of traffic physics such that link states are expressed as time intervals. For example with this approach, the probability of a driver to be fast in the next link, given that its current link speed is slow can be directly estimated. This approach makes the proposed method to be direct and intuitively applicable, even for transportation practitioners. Note that, utilizing alternative methods to define the states of the Markov chain or different models for traffic progression, e.g. Hidden Markov Models (see [125, 80]) can also be considered.

Traffic engineering guidelines

Intuitive traffic engineering rules to estimate travel time for conditions with different level of congestion are employed to define state boundaries. For instance in uncongested regime, we consider free flow travel time boundary (TT_{ff}) as:

$$TT_{\text{ff}} = \frac{\text{linklength}}{V_{\text{ff}}}(1+\epsilon) \tag{3.12}$$

where V_{ff} denotes free flow speed and ϵ is a coefficient, to account for the variability of drivers' behavior (e.g. $\epsilon = 0.1$). Furthermore, the second boundary which estimates the maximum travel time in near capacity regime is $R+TT_{\text{ff}}$, where R is the red interval for through movement. The next boundary for oversaturated conditions where vehicles are delayed for more than one cycle time can be set as $C+TT_{\text{ff}}$, where C is the cycle time [169]. The major drawback of these guidelines is that they are practical when all of the route intersections are controlled by traffic signals. They also ignore the effect of offsets and the fact that within these boundaries some correlation might be present. It might also be difficult to be estimated in case of actuated traffic signals. Offsets and turning movements create sharp functions of arrival profiles and travel time characteristics and thus lead to different boundaries in the 2D diagrams. Note that in this case, the numbers of states are the same for all the links which makes the number of Markov paths (Q) equal to $\prod_{l=1}^{k} m_l, m_l = 2,3,4$.

Grid clustering

Due to the implication of travel time states as time intervals, grid clustering seems a reasonable clustering approach. The chief trait of grid clustering method is that it uses a multi-

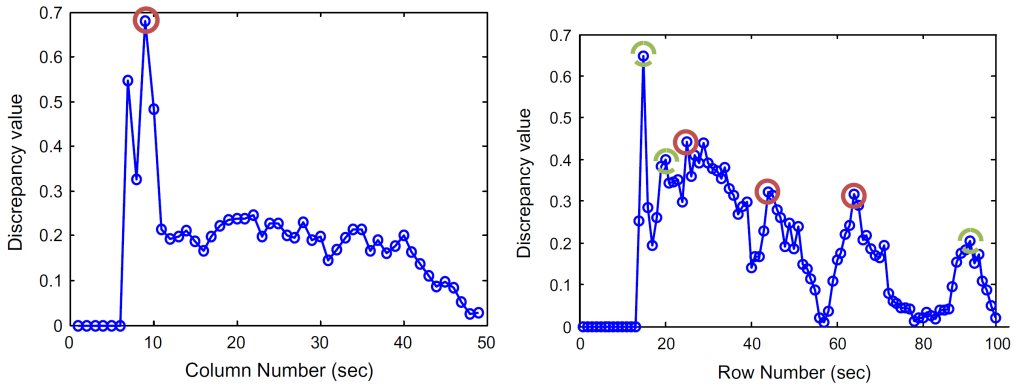
dimensional imaginary grid structure which partitions the data space to hyperrectangles (rectangles in 2D space) in order to find hidden patterns in data. Then as the next step, the hyperrectangles are grouped with respect to a topological-neighboring criterion and the problem attributes to produce clusters. The grid clustering has shown to be very effective for analyzing enormous datasets and demonstrates superior performance over fuzzy k-means and Radial Basis Function (RBF) methods [201].

In the proposed model, we discretize the time space and draw a 1 sec × 1 sec 2D Cartesian grid over each 2D diagram and our goal is to find a subset of boundaries in both axes to have homogenous rectangular clusters. In other words, our objective is to identify clusters with negligible correlation between successive links travel times, where convolution is a proper estimator of sum of distributions. We introduce a metric to measure the discrepancy between data of two consecutive columns (or rows) of Cartesian grid as follows (link index is omitted for simplicity):

$$\alpha_c = \sum_{u=1}^{n_r} \left| \frac{N(c,u) - N(c+1,u)}{N(c,u) + N(c+1,u)} \right| \tag{3.13}$$

where n_r (or n_c) is the number of rows (or columns) of the Cartesian grid over 2D diagram l, which represents the range of travel time (maximum minus minimum value). Metric α_c is the measured discrepancy between columns c and $c+1$; $N(c,u)$ and $N(c+1,u)$ are the number of data points in small squares in column c and $c+1$. A similar procedure should be done for the rows of 2D diagram (α_r instead of α_c). The discrepancy metric α_c represents how data in two successive columns of grid (travel time of c and $c+1$ for a link) are similar. The less α_c means that the distributions of travel times at columns c and $c+1$ of the 2D diagram projected on the other axis (next link) is more similar. Thus, points with high α_c value are more likely to represent a drastic change in the distribution of travel time, which can be used as an identification mark for a state boundary. Thus according to discrepancy values (α_r and α_c), we should choose a few high peak value points as state boundaries. Figure 3.3 illustrates the discrepancy values for columns and rows of 2D diagram of Fig. 3.2(a).

The objective of the heuristic grid clustering algorithm is to generate (i) a few numbers of states in order to have reasonable computational complexity, (ii) large enough states that can characterize a traffic condition (uncongested, semi-congested, etc.), and (iii) negligible correlation of any type within each cluster. We introduce two mechanisms to control these settings. First, two threshold levels are defined so that the selected points (boundaries) should have a discrepancy value greater than that; one static and one dynamic. The former is a multiplication of a static factor ($\mu > 1$) by the average discrepancy value of columns (rows), i.e. $\bar{\alpha} = \sum_{u=1}^{n_c} \alpha_u / n_c$. This static threshold ensures that the selected point is a peak that is greater than a factor of the average value, $\bar{\alpha}$. The latter is the product of an increasing parameter ($\lambda < 1$) by the latest selected α value, where λ is initiated near zero and increases at each iteration. This dynamic threshold guarantees that the boundary point candidate has a sensible discrepancy difference with previous selected boundary point. Hence using both thresholds, a large number of



(a) (b)

Figure 3.3: Discrepancy values for (a) columns and (b) rows. The red circles show the selected data points as boundaries and the green circles show peaks with high discrepancy values that are rejected as boundaries (see explanation in the text).

boundaries is avoided. For example, point 90sec in Fig. 3.3(b) is rejected because its α value is less then $\mu\bar{\alpha}$. Second, the discrepancy values of β data points within the neighborhood area are forced to zero to prevent a small size state whenever a new boundary is selected. For example, point 21sec in Fig. 3.3(b) is rejected since it is in the vicinity of already chosen point 25sec. In addition, a few initial and final data points should be forced to zero to avoid small states at the beginning and end of travel time interval. For instance, point 15sec in Fig. 3.3(b) is rejected since it would create a small size state in the beginning. Note that in Fig. 3.3(a), only one boundary is selected, as the discrepancy values do not have any further peak greater than $\mu\bar{\alpha}$ after point 10sec. In this manner the procedure achieves the predefined three objectives. We have also noticed that the effectiveness of two aforementioned mechanisms is robust to random fluctuations of α and not very sensitive to parameter calibration. Values of the parameters after a fine-tuning procedure are selected as $\mu = 1.6$ and $\beta = 5$sec. The proposed heuristic algorithm for grid clustering performs iteratively and independently for columns and rows to identify the state boundaries set. The pseudo code of the heuristic algorithm is as follows:

1. Compute the discrepancy values (α)
2. Loop for $n = 1, 2, \cdots$
3. Select the highest α, i.e. $\alpha^n_{\max}$
4. If ($\alpha^n_{\max} > \alpha^{n-1}_{\max} \cdot \lambda$) and $\alpha^n_{\max} > \mu\bar{\alpha}$) then
   (a) Append the corresponding number of $\alpha^n_{\max}$ to the boundaries set
   (b) Make α of β neighbors $\alpha^n_{\max}$ equal to zero
   (c) Increase λ
5. Else end
6. Return

Since grid clustering algorithm operates based on each 2D diagram structural data, it is apparent that the states of link l in the 2D diagrams $l-1$ and l might not be consistent,

i.e. $\mathscr{Y}_{l-1} \neq \mathscr{X}_l$. Therefore some modifications of the Markov chain formulations, i.e. (3.6)–(3.11), are needed. We consider a new Markov chain with intermediate stages such that at the beginning of Markov chain there are link 1 states (from 2D diagram 1) then link l states ($l = 2,3,\cdots,k-1$) (as downstream link from 2D diagram of links $l-1$ and l), and again link l states (as upstream link from 2D diagram of links l and $l+1$), and finally link k states (from 2D diagram $k-1$). In this manner, computing the initial state probabilities and the transition probabilities between link l and $l+1$ remain the same as in (3.6)–(3.8). Note that there is no concern about the first and the last links since they only belong in one 2D diagram. The concern is about the intermediate links (2nd to $(k-1)$th), which have two different sets of states. Let us denote them $\mathscr{Y}_{l-1}$ and $\mathscr{X}_l$ for link l, where $\mathscr{Y}_{l-1} = \{y_1^{l-1},\cdots,y_{n_{l-1}-1}^{l-1}\}, \mathscr{X}_l = \{x_1^l,\cdots,x_{m_l-1}^l\}$ and τ_l is the set of all link l reported travel times. Figure 3.4 illustrates such intermediate stage.

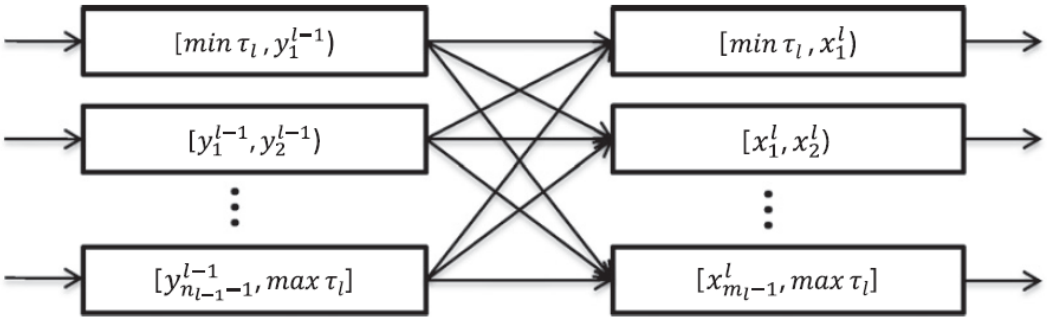


Figure 3.4: Schematic of the Markov chain intermediate stage for link l.

Computing the intermediate transition probabilities for link l is modified as: (similar to interlink transition probabilities (3.7)–(3.8))

$$\acute{P}_l = \begin{bmatrix} \acute{p}_{1,1} & \cdots & \acute{p}_{1,m_l} \\ \vdots & \ddots & \vdots \\ \acute{p}_{n_{l-1},1} & \cdots & \acute{p}_{n_{l-1},m_l} \end{bmatrix} \tag{3.14}$$

$$\acute{p}_{i,j} = \text{pr}\{tt_l \in [x_{j-1}^l, x_j^l) | tt_l \in [y_{i-1}^{l-1}, y_i^{l-1})\} = \frac{N\Big(\{tt_l \in [x_{j-1}^l, x_j^l) \cap [y_{i-1}^{l-1}, y_i^{l-1})\}\Big)}{N\Big(\{tt_l \in [y_{i-1}^{l-1}, y_i^{l-1})\}\Big)};$$

$$i = 1,\cdots,n_{l-1}; j = 1,\cdots,m_l; l = 2,\cdots,k-1 \tag{3.15}$$

Consequently, there will be $Q = \prod_{l=1}^{k-1} m_l \cdot n_1$ Markov paths. For a given Markov path, the path occurrence probability is computed similarly to (3.9):

$$\text{pr}\{S_1 = i_1, S_2 = i_2, \acute{S}_2 = \acute{i}_2, S_3 = i_3, \cdots, \acute{S}_{k-1} = \acute{i}_{k-1}, S_k = i_k\} = \pi_{i_1} p_{i_1,i_2} \acute{p}_{i_2,\acute{i}_2} p_{\acute{i}_2,i_3} \cdots p_{\acute{i}_{k-1},i_k}. \tag{3.16}$$

And the Markov path TTD is computed similarly to (3.10):

$$TTD\{S_1 = i_1, S_2 = i_2, \acute{S}_2 = \acute{i}_2, \cdots, \acute{S}_{k-1} = \acute{i}_{k-1}, S_k = i_k\} = TTD(i_1) * TTD(i_2 \cap \acute{i}_2) * \cdots * TTD(i_k). \tag{3.17}$$

Finally, equation 3.11 remains valid to compute the route TTD. In the following sections, the study sites are presented and the results are elaborated.

3.5 Study sites

3.5.1 Peachtree Street

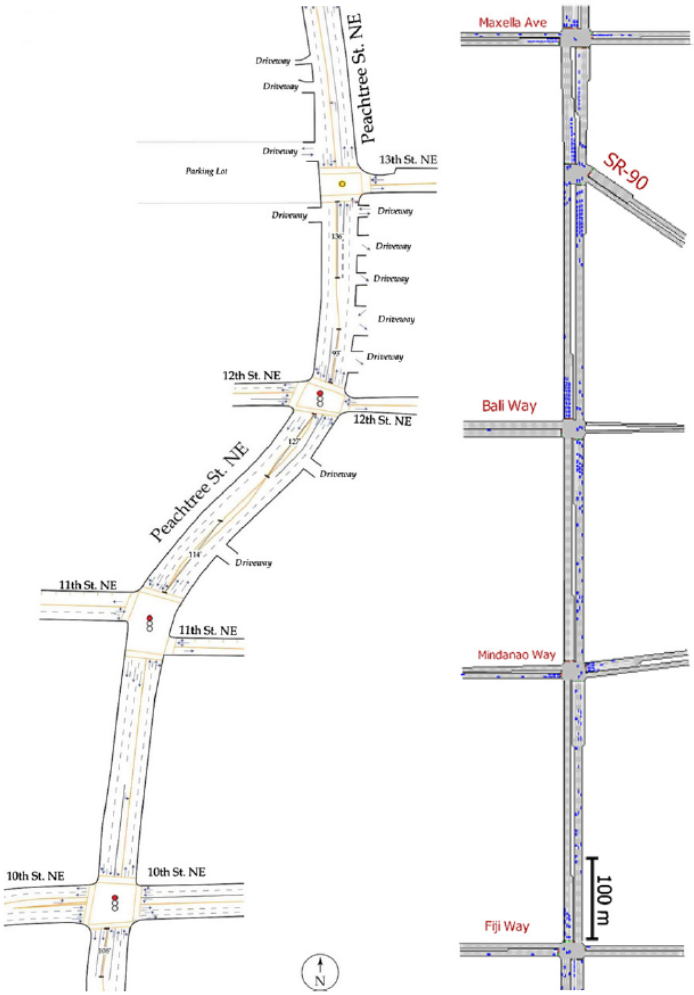
The first selected site is Peachtree Street, an arterial in Atlanta, approximately 640 meters in length, with five intersections, six links and two to three through lanes in each direction. The intersections are signalized except the most north one. The speed limit on this north-south arterial is 35 (mph) and we only study the southbound direction of traffic, which is the most congested. The Peachtree Street dataset is a part of NGSIM program [129] which managed by Federal Highway Administration, intended to provide a dataset of arterial vehicle trajectories for traffic behavioral analyses. Detailed traffic data (vehicle trajectories) were collected using video camera between 12:45 p.m. and 1:00 p.m. at a resolution of 10 frames per second on November 8, 2006. The dataset contained comprehensive individual vehicle trajectories with time, link, and direction stamps, from which the link travel times of vehicles are calculated. Figure 3.5(a) shows the study area schematic.

3.5.2 Lincoln Boulevard

This study site is 1.1 (km) long stretch of a major urban arterial with speed limit of 35 (mph), north of the Los Angeles International Airport, between Fiji Way and Maxella Ave. in the cities of Los Angeles and Santa Monica. The study section includes five signalized intersections with link lengths varying from 150 to 300 meters. The number of lanes for through traffic per link is three and additional lanes for turning movements are provided at intersection approaches. Traffic signals are all multiphase operating as coordinated under traffic responsive control as part of the Los Angeles central traffic control system. Loop detectors are located on each lane approximately 90 meters upstream of the intersection stop line. Detectors are also placed on the major cross street approaches.

A field study was undertaken to obtain a comprehensive database (loop detector data, manual turning movement counts, probe vehicles, etc.) of operating conditions. The study period enabled us to attain data for a wide range of traffic conditions: from low volume off-peak conditions, peak period conditions, and post-peak mid-day flow conditions. Traffic demand is high especially during the peak hour and heavily directional with the higher through and turning volumes in the northbound direction. For a more detailed analysis of the study

network the reader can refer to [169]. The vehicle data (demand and turning movements every 15min) and signal timing data were incorporated into the AIMSUN microscopic simulator. The proposed model was then applied to estimate the TTD on northbound travel direction in a 4 hours simulation. The data sampling rate from probe vehicles is every one second and the travel time of each vehicle at each link is recorded and treated as GPS data. The simulation output was first compared with field data (delays and travel times) to verify that the model reasonably replicates field conditions at the test sites. Figure 3.5(b) shows a snapshot of the study site.



(a) (b)

Figure 3.5: Study sites, (a) Peachtree Street schematic and (b) Lincoln Blvd. snapshot.

3.6 Results and discussion

We evaluate the proposed methodology for the through movement of both study sites. Results for Peachtree Street are shown in Fig. 3.6. It is evident that the results of proposed method are promising and can capture the fundamental and the most details of TTD profile. Note that since there is no traffic signal at the first intersection, it is not possible to utilize traffic guidelines as Section 3.4.2. For more comprehensive comparison, the Mean Absolute Error (MAE) between the ground truth TTD and the estimated TTD is calculated:

$$MAE = \frac{\sum_{u=\min(TTD)}^{\max(TTD)} |estimated_u - real_u|}{Range}, \tag{3.18}$$

where $Range$ is the longest TTDs time range (for ground truth and all estimated TTDs, max minus min value). (It should be noted that, in this article all the calculations are discretized with the accuracy of 1 sec. Thus, division by Range means to normalize the absolute error between the true TTD probability mass function and the different estimated TTD probability mass functions with various range of travel time. With different accuracy of discretization, different definition of Range is needed. Also note that, the largest possible value of MAE is $2/Range$.) The MAE of grid clustering TTD estimation is $13.59 \cdot 10^{-4}$ whereas the MAE of convolution method is $29.32 \cdot 10^{-4}$ which shows a significant (54%) improvement. Note that the real TTD is too spiky and the results (only of the real TTD) have been smoothed in the graphs with a moving average of 9secs every 1sec. Another issue about TTD being spiky is that convolution operator has the disadvantage that generally tends to produce smooth output [43] contrary to our method which can create sharp results.

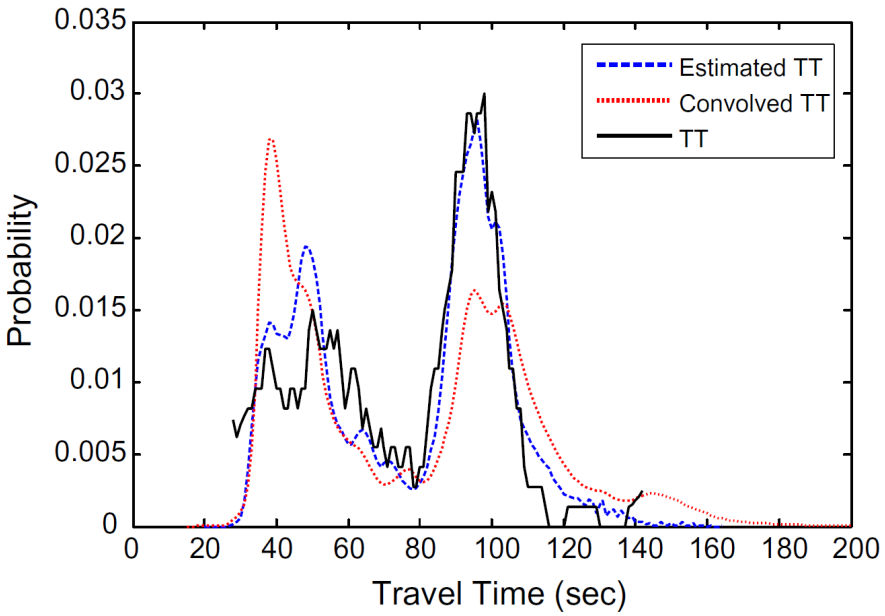
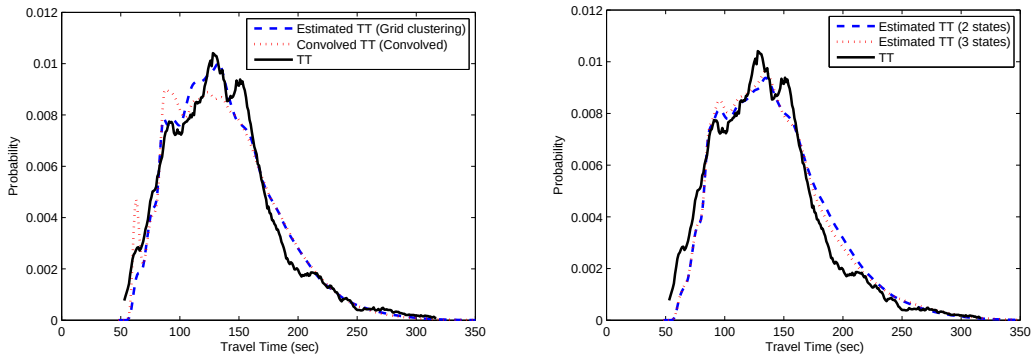


Figure 3.6: Estimation and ground truth of TTD for Peachtree Street.

The ground truth TTD for Lincoln Blvd. case study, convolved estimation and result of our methodology with grid clustering approach are depicted in Fig. 3.7(a). The outputs of the proposed method using traffic guidelines with 2 and 3 states in each link are also illustrated in Fig. 3.7(b). It is apparent that TTD is very close to the outcomes of proposed methodology and more accurate once compared to the convolution estimation. It is worth mentioning that the period of analysis should not be too large as it might contain variant traffic conditions. In this case, the spatial correlations between links may be diminished by different characteristics of traffic regimes, i.e. the grid clustering might result in different boundaries (states) for congested and uncongested conditions. This can be inferred by comparing Fig. 3.6 and 3.7. The superiority of grid clustering over convolution is more significant in Peachtree Street than Lincoln Blvd., because Peachtree street data covers only 15 min of mostly invariant traffic regime while the 4hr data of Lincoln Blvd. includes more variations.



(a) (b)

Figure 3.7: Estimation of TTD, (a) convolution and grid clustering methods and (b) traffic guidelines with 2 and 3 states.

Furthermore, in reality not all vehicles are equipped with GPS devices and only a subset of information might be available. To evaluate the robustness of the methodology, different deployment rates of probe vehicles are studied and the results are given in Tables 3.1–3.3. The values in parentheses in the rightmost column show the average number of states in grid clustering method. Hence, grid clustering results to a better estimation than traffic guidelines even with less number of states (less computation burden and complexity). Note that grid clustering algorithm detects the states deterministically based on the spatial structure of 2D diagrams. So for different sample sizes, different number of states is selected for every link (usually less than 3 states). The effect of probe vehicles penetration rates on evolution of TTD profile is shown in Fig. 3.8(a). The TTD estimation with 2% data are also depicted in Fig. 3.8(b) demonstrating that our proposed algorithm still performs very well under sparse probe vehicles condition and low number of states.

The same experiments are also done with less demand (75% and 60%) to investigate the effect of demand level on outcome of TTD estimation. In all cases, the proposed methodology

Table 3.1: MAE ($\times 10^{-4}$) of described methods for 100% demand level in Lincoln study site, Range: 440 sec (value in parenthesis of grid clustering indicates the average number of clusters).

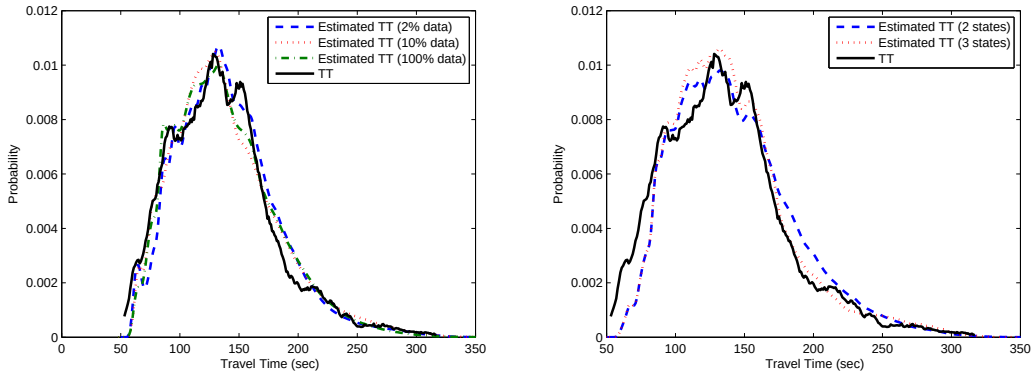
Probe Vehicles Sample Size	Convolution	2 States Markov chain	3 States Markov chain	Grid clustering
100%	3.27	3.34	3.09	2.67 (2.875)
50%	3.32	3.45	3.18	3.00 (2.75)
20%	3.48	3.75	3.31	3.12 (2.625)
10%	3.53	3.72	3.40	2.98 (2.625)
5%	3.75	3.95	3.47	3.01 (2.875)
2%	4.09	4.89	4.44	3.63 (2.625)
1%	4.49	4.49	-	3.95 (3.125)

Table 3.2: MAE ($\times 10^{-4}$) of described methods for 75% demand level in Lincoln study site, Range: 310 sec (value in parenthesis of grid clustering indicates the average number of clusters).

Probe Vehicles Sample Size	Convolution	2 States Markov chain	3 States Markov chain	Grid clustering
100%	6.90	6.48	6.47	5.23 (2.5)
50%	7.06	6.55	6.60	5.22 (2.25)
20%	7.11	6.76	6.76	5.30 (2.375)
10%	6.90	6.82	6.78	5.88 (2.125)
5%	7.32	6.82	6.75	5.65 (2.75)
2%	7.89	7.35	-	8.05 (3)
1%	9.65	8.98	-	8.61 (2.875)

Table 3.3: MAE ($\times 10^{-4}$) of described methods for 60% demand level in Lincoln study site, Range: 305 sec (value in parenthesis of grid clustering indicates the average number of clusters).

Probe Vehicles Sample Size	Convolution	2 States Markov chain	3 States Markov chain	Grid clustering
100%	9.36	7.61	7.35	6.09 (2.5)
50%	9.36	7.84	7.60	7.22 (2.25)
20%	10.24	7.68	-	7.00 (2.625)
10%	9.39	8.06	-	7.88 (2.75)
5%	10.54	8.30	-	7.52 (2.625)
2%	10.55	10.78	-	9.54 (2.375)
1%	11.78	12.64	-	11.48 (3.5)

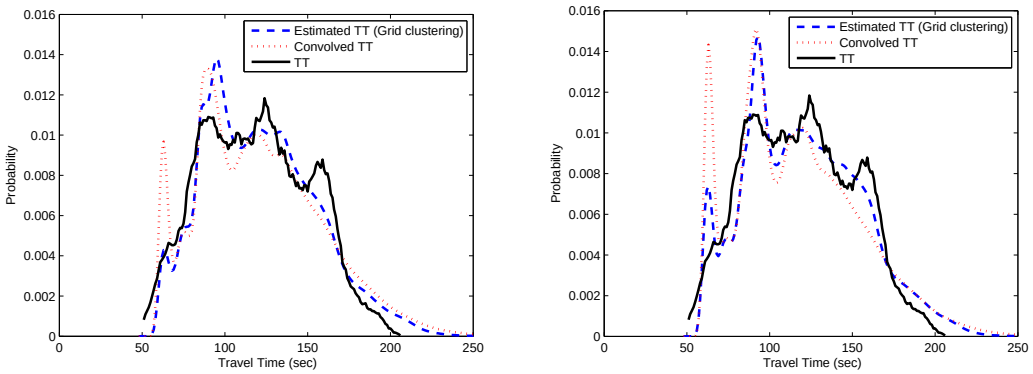


(a) (b)

Figure 3.8: TTD Estimation with, (a) different probe vehicles penetration rates and (b) 2% probe vehicles penetration rate.

utilizing only 5% of the data, gives a better estimation than convolution (neglecting the correlation between links) with 100% data. By comparing results of 2- and 3-states Markov chain in Tables 3.2 and 3.3, we notice less good outcomes for 3-states compare to 2-states contrary to results in Table 3.1. The reason is that less demand makes less congestion and thus introducing the third state regarding to the near capacity condition is not sensible and causes difficulties for the procedure. In addition being in the free flow regime, the third state boundary may be higher than the maximum travel time, which makes the procedure to fail (see last rows of Table 3.3). The results in Tables 3.1–3.3 also reveal that, with very low penetration rates the performances of proposed grid clustering and convolution are declining. It can be justified such that with very low penetration rate the sample size is too small such that to identify the correlation within the data is not sensible and assuming the total independence (convolution method) seems reasonable. In cases with so low penetration rate, a proper (Bayesian) fusion of a priori knowledge based on the historical data and online information based on the real-time data is a crucial future work. The ground truth, convolved, and estimated TTD of 75% and 60% demands are depicted in Fig. 3.9(a) and 3.9(b), respectively. It is evident that convolution method always overestimates the probability of fast and very slow moving vehicles by producing a high peak at the beginning and a long tail.

To further investigate the proposed method and given literature attempts to approximate the TTD with common single-state and multi-state continuous probability distributions, the TTD of both case studies are fitted with Normal, Gamma, Log-normal, and Gaussian Mixture Models (GMM) (with 2 components; 5 degrees of freedom) and the MAEs of these approximations are given in Table 3.4. The advantage of proposed grid clustering estimation over the fitting methods seems promising, given that the grid clustering utilizes the link-level data to estimate the unknown route-level information, while the fitting methods try to approximate the route-level TTD, which is considered known. (In addition, the results verify the visual inspections that Lincoln Blvd. TTD resembles more the unimodal distributions



(a) (b)

Figure 3.9: Estimation of TTD with, (a) 75% demand and (b) 60% demand.

whereas Peachtree TTD is a multi-modal distribution.)

Table 3.4: MAE ($\times 10^{-4}$) of estimation and fitting methods for both case studies.

Study Site	Convolution	Grid clustering	Gamma	Log normal	Normal	GMM
Lincoln	3.27	2.67	2.79	3.11	4.15	2.88
Peachtree	29.32	13.59	32.04	45.32	29.57	14.14

To provide more comprehensive comparisons, we do additional statistical investigation such that the link travel time data set is split to two parts, one for estimation methods and the other one for computing the ground truth TTD and subsequently the error metric. In this manner, the data used for the estimation is entirely different from the base ground truth. The Lincoln case study with 100% demand is selected and for various fractions of data we run the grid clustering and convolution estimation methods and compare those with the TTD computed from remaining part of data set. The average results for 5 runs are presented in Table 3.5 validating the statistical superiority of the proposed grid clustering. Considering all the facts, incorporation of Markov chain into TTD estimation in a synergy produces promising results which can capture the fundamental characteristic of field TTD.

Table 3.5: MAE ($\times 10^{-4}$) of statistical tests for Lincoln case study .

Fraction of data for estimation	Convolution	Grid Clustering
50%	3.52	3.08
20%	3.56	3.29
10%	3.47	3.17
5%	3.79	3.29
2%	4.35	3.67

3.7 Summary

In this chapter, we introduce an approach to address traffic progression and correlation in arterials for travel time distribution estimation. In this method, probe vehicles provide travel time of links of the arterial route. For each pair of consecutive links, a 2D diagram is established to graphically represent the joint distributions of successive link travel times. Then, using these 2D diagrams, we incorporate a Markov chain procedure into the model and identify its initial and transition probabilities from the observed data. For Markov chain state identification, a heuristic grid clustering algorithm is also developed. The procedure is tested with various deployment rates of probe vehicles to tackle the problem of probe vehicles sample size. Our proposed method shows a coherent performance capturing the fundamental characteristics of field measurements even under condition of sparse probe vehicles. The inputs to the proposed procedure are link travel times and since they are not directly reported by probe vehicles, the estimation of link travel times from GPS data is of interest to integrate with our method. The method presented in Chapter 2 provides the link travel time estimations.

Part II Network-Level Traffic Modeling and Control

4 Model predictive perimeter control for two urban regions with MFDs

RECENT analysis of empirical data from cities showed that the macroscopic fundamental diagram (MFD) of urban traffic provides for homogenous network regions a unimodal, low-scatter relationship between network vehicle density and network space-mean flow. In this chapter, the optimal perimeter control for two-region urban cities is formulated with the tool of MFDs. The controllers operate on the border between the two regions, and manipulate the percentages of flows that transfer between the two regions, such that the number of trips reaching their destinations is maximized. The optimal perimeter control problem is solved by model predictive control, where the prediction model and the plant (reality) are formulated by MFDs. Examples are presented for different levels of congestion in the regions of the city and the robustness of the controller is tested for different sizes of error in the MFDs and different levels of noise in the traffic demand. Moreover, two methods for smoothing the control sequences are presented. Comparison results show that the performances of the model predictive control are significantly better than a “greedy" feedback control. The results of this chapter can be extended to develop efficient hierarchical control strategies for heterogeneously congested cities.

4.1 Introduction

Efficient monitoring and traffic management of large-scale urban networks still remain a challenge both for traffic researchers and practitioners. A large urban network consists mainly of two elements, urban links and signalized intersections. Modeling the traffic flow dynamics of each element in a large urban network with large number of links and intersections is a complex task. One has to model the evolution of queues at each signalized intersection, and to account for the queue dynamic interactions between adjacent intersections, i.e. capturing the dynamics of propagation and spillback of queues due to high demand. Even if this task is completed, a centralized control approach would be a very challenging task, not only because of the computational complexity, but also because users might change their travel patterns (e.g. time of departure, route choice, mode choice, etc). Hence, instead of this micromodeling approach, the macroscopic fundamental diagram (MFD) aims at simplifying

the micromodeling task of the urban network where the collective traffic flow dynamics of subnetworks capture the main characteristics of traffic congestion, such as the evolution of space-mean flows and densities in different regions of the city. The MFD can be utilized to introduce elegant control strategies to improve mobility and decrease delays in large urban networks, that local ones are unable to succeed.

The MFD provides for different network regions a unimodal, low-scatter relationship between network vehicle density (veh/km) and network space-mean flow or outflow (veh/h) if congestion is roughly homogeneous in the region. Alternatively, the MFD links *accumulation* defined as the number of vehicles in the region and *trip completion flow* defined as the output flow of the region. Network flow or trip completion flow increases with density or accumulation up to a critical point, while additional vehicles in the network cause strong reductions in flow. The first theoretical proposition of such a physical model was developed by [60], while similar approaches were also initiated by [78] and [28]. The physical model of MFD was observed with dynamic features in congested urban networks in Yokohama by [55]. This work showed two important properties of MFD that can be utilized for management and control purposes: (i) some urban regions approximately exhibit an MFD, (ii) the shape of the MFD is not very sensitive to different demand patterns. Property (i) is important for monitoring purposes as flow can be easily observed with different types of sensors while outflow is more difficult. Property (ii) is important for control purposes as efficient active traffic management schemes can be developed without a detailed knowledge of origin–destination (OD) tables. Other investigations of MFD using empirical or simulated data can be found in [16], [89], [164], [120] and others, while routing strategies based on the MFD can be found in [100].

Recent studies [51], [120], [32] have shown that networks with heterogeneous distribution of density exhibit network flows smaller than those that approximately meet homogeneity conditions (low spatial variance of link density), especially for high network densities. Moreover, networks with small variance of link densities have *a well-defined* MFD, i.e. low scatter of flows for the same densities. A possible solution for heterogeneous networks is that they might be partitioned to a number of more homogeneous regions with small variances of link densities as each region will have a well-defined MFD. For more information regarding network partitioning, the reader can refer to [90]. Studies [51] and [164] showed strong hysteresis phenomena in freeways that might not disappear after partitioning. Nevertheless, the work of [90] showed that urban networks can be partitioned in a way to decrease the degree of heterogeneity within clusters. Partitions should not have very small size because the law of large numbers will not apply and high scatter might exist in the MFDs. In addition, a large number of partitions will not allow the development of simple control strategies, because control might change the route choices and detailed ODs might be needed.

Management and control of multi-region MFDs systems can improve urban mobility, prevent overcrowding, and relieve congestion in cities. The optimal control policy was derived for a single MFD system in [28]. The main logic of the strategies is that they aim to decrease inflows in regions with high densities of destinations and points in the decreased part of an MFD, and

manage the accumulation to maintain the flow in the city on its maximum. However, in case of multi-region cities with multiple centers of congestion and/or attraction, control policies are more complicated and not well-understood. For stability analysis of controlling two urban regions see [66].

Due to the scatter in the MFD, mainly in the congested regime, errors are expected between the MFD model and the plant (reality). Therefore, an optimal open-loop control for the multi-region MFDs system would be a sub-optimal solution compared with optimal closed-loop control. The closed-loop control takes into account the errors between the model and the plant by utilizing a feedback monitored-information. Furthermore, the closed-loop control can tackle disturbances that the model was not designed for, e.g. noises in the traffic demand. The optimal closed-loop control is obtained by implementing the model predictive control (MPC) framework. A historical survey for industrial applications of MPC can be found in [145], while theoretical issues of MPC can be found in [44], [17], [18], [119], [115].

The model predictive control is a receding horizon scheme, where at each time step an optimal open-loop of the problem with finite horizon is optimized, then only the first controller is applied to the plant and the procedure is carried out again. A receding horizon framework has been used for optimization in different traffic control problems, e.g. ramp metering of freeway networks in [11], [139], variable speed limits and route guidance for freeway networks in [74], [73], signal control for large-scale urban networks in [45], [3], [111], and mixed urban and freeway networks in [178]. The open-loop optimization in the traffic MPC models of e.g. [74], [73] uses *direct simultaneous method* to transcript it into a finite-dimensional nonlinear programming through discretization of both control and state variables, while in [139], [3] a feasible direction method is utilized to solve the open-loop optimization problem. Overviews of different control applications in transportation problems can be found for example in [35], [185].

In this chapter, the optimal perimeter control problem for two-region urban cities is formulated where the dynamic equations are modeled according to their MFDs. Moreover, the optimal control solution is obtained by applying the MPC framework. The open-loop optimal control problem is solved using *direct sequential method* that discretizes *only* the control variables with piecewise constant controls, while the state variables are “continuous" and integrated using the state-of-the-art methods for ODE solvers.

This chapter is organized as follows: the control problem for two-region MFDs system is presented in Section 4.2. Then in Section 4.3, the model predictive control is formulated, the parameters are tuned, and the control laws of a greedy controller are presented. Comparison results of case study examples are presented in Section 4.4 showing the performances difference between the model predictive control and the greedy control. Finally, two different methods are introduced to smooth the control sequences in Section 4.5. Section 4.6 summarizes the chapter.

4.2 Two-region MFDs system

In this chapter, a heterogeneous traffic network that can be partitioned in two homogeneous regions is considered. A traffic network for two-region system is schematically shown in Fig. 4.1. There are two regions R_i, $i = 1,2$, where each region has a well-defined MFD, the

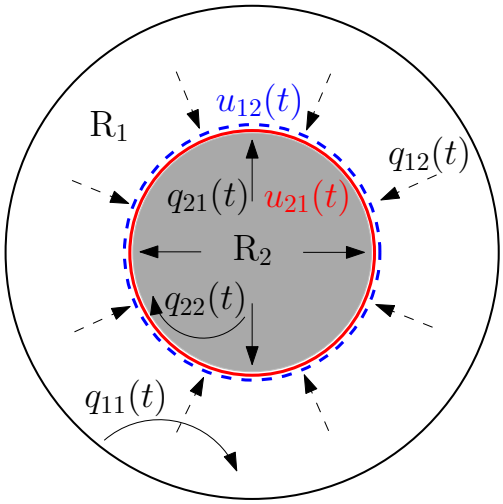


Figure 4.1: Two-region MFDs system: two regions R_1 and R_2 with four traffic demand $q_{11}(t)$, $q_{12}(t)$, $q_{21}(t)$, $q_{22}(t)$ and two perimeter controllers $u_{12}(t)$ and $u_{21}(t)$.

periphery of the center R_1 and the city center R_2. Note that the geographical relative position of these regions does not affect the dynamics of the problem, e.g. it can be two neighbors next to each other. An endogenous traffic demand is defined as a flow that its origin and destination are the same region, while the origin and destination of an exogenous traffic demand are not the same. For the two-region system, there are two endogenous traffic demand in R_1, denoted by $q_{11}(t)$ (veh/s), and in R_2, denoted by $q_{22}(t)$ (veh/s), and two exogenous traffic demand generated in R_1 and R_2 with destination to R_2 and R_1 denoted by $q_{12}(t)$ and $q_{21}(t)$ (veh/s), respectively. Corresponding to the endogenous and exogenous traffic demand, four accumulation states are used to model the dynamic equations: $n_{ij}(t)$ (veh), $i, j = 1,2$, where $n_{ij}(t)$ is the total number of vehicles in R_i with destination to R_j at time t. Let us denote $n_i(t)$ (veh) as the accumulation or the total number of vehicles in R_i at time t, i.e. $n_i(t) = \sum_j n_{ij}(t)$.

The MFD is defined by $G_i(n_i(t))$ (veh/s) which is the trip completion flow for region i at $n_i(t)$. The trip completion flow for region i is the sum of transfer flows, i.e. trips from i with destination j, $i \neq j$, plus the internal flow, i.e. trips from i with destination i. The transfer flow from i with destination to j is calculated corresponding to the ratio between accumulations, i.e. $M_{ij} = (n_{ij}/n_i) \cdot G_i(n_i(t))$, $i \neq j$, while the internal flow from i with destination to i is calculated by $M_{ii} = (n_{ii}/n_i) \cdot G_i(n_i(t))$. These relationships assume that trip length for all trips within a region (internal or external) are similar, i.e. the distance traveled per vehicle inside a region is independent of the origin and destination of the trip. For a description of different cases the reader can refer to [52], which will not alter the methodology. Simulation and

empirical results [55], show that the shape of MFD can be approximated by a non-symmetric unimodal curve skewed to the right, i.e. critical density that maximizes network flow is smaller than half the jammed density. Thus, we utilize a 3rd-order function of $n_i(t)$, e.g. $G_i(n_i(t)) = a_i \cdot n_i^3 + b_i \cdot n_i^2 + c_i \cdot n_i$, where a_i, b_i, c_i are estimated parameters.

In our formulated problem, the perimeter controllers denoted by $u_{12}(t)$ and $u_{21}(t)$ (-) are introduced on the border between the two regions as shown in Fig. 4.1, where the purpose is to control the transfer flows between the two regions such that the total number of vehicles that complete their trips and reach their destinations in the two-region MFDs system is maximized. Since the perimeter controllers exist only on the border between the two regions, the internal flows cannot be controlled or restricted, while the transfer flows are controlled by the controllers such that only a ratio transfers at time t. The perimeter controllers $u_{12}(t)$ and $u_{21}(t)$, where $0 \leq u_{12}(t) \leq 1$ and $0 \leq u_{21}(t) \leq 1$, control are the ratio of the transfer flow that transfers from R_1 to R_2 and R_2 to R_1 at time t, respectively. It is also assumed that these controllers will not change the shape of the MFDs. Implementations of the controllers in real networks are discussed in the summary Section.

The criterion is to maximize the output of the traffic network, i.e. the number of vehicles that complete their trips and reach their destinations. Therefore, the two-region MFDs control problem with four state variables is formulated as follows (similarly to [68]):

$$J = \max_{u_{12}(t), u_{21}(t)} \int_{t_0}^{t_\mathrm{f}} \Big[M_{11}(t) + M_{22}(t) \Big] \mathrm{d}t \tag{4.1}$$

subject to

$$\frac{\mathrm{d}n_{11}(t)}{\mathrm{d}t} = q_{11}(t) + u_{21}(t) \cdot M_{21}(t) - M_{11}(t) \tag{4.2}$$

$$\frac{\mathrm{d}n_{12}(t)}{\mathrm{d}t} = q_{12}(t) - u_{12}(t) \cdot M_{12}(t) \tag{4.3}$$

$$\frac{\mathrm{d}n_{21}(t)}{\mathrm{d}t} = q_{21}(t) - u_{21}(t) \cdot M_{21}(t) \tag{4.4}$$

$$\frac{\mathrm{d}n_{22}(t)}{\mathrm{d}t} = q_{22}(t) + u_{12}(t) \cdot M_{12}(t) - M_{22}(t) \tag{4.5}$$

$$0 \leq n_{11}(t) + n_{12}(t) \tag{4.6}$$

$$0 \leq n_{21}(t) + n_{22}(t) \tag{4.7}$$

$$n_{11}(t) + n_{12}(t) \leq n_{1,\mathrm{jam}} \tag{4.8}$$

$$n_{21}(t) + n_{22}(t) \leq n_{2,\mathrm{jam}} \tag{4.9}$$

$$u_{\min} \leq u_{12}(t) \leq u_{\max} \tag{4.10}$$

$$u_{\min} \leq u_{21}(t) \leq u_{\max} \tag{4.11}$$

$$\begin{aligned} &n_{11}(t_0) = n_{11,0}\,;\; n_{12}(t_0) = n_{12,0} \\ &n_{21}(t_0) = n_{21,0}\,;\; n_{22}(t_0) = n_{22,0} \end{aligned} \tag{4.12}$$

where t_f (s) is the final time, $n_{ij,0}$, $i,j = 1,2$ are the initial accumulations at t_0, $n_{1,\text{jam}}$ and $n_{2,\text{jam}}$ (veh) are the accumulations at the jammed density in R_1 and R_2, respectively, $u_{\min}$ and $u_{\max}$ are the lower and upper bounds for $u_{12}(t)$, $u_{21}(t)$, respectively. Recall that $M_{ij} = (n_{ij}/n_i)\cdot G_i(n_i(t))$, $i,j = 1,2$. The equations (4.2)–(4.5) are the conservation of mass equations for $n_{ij}(t)$, while the equations (4.6), (4.7), (4.8), and (4.9) are the lower and upper bound constraints on accumulations in R_1, R_2, respectively.

4.3 Model Predictive Control for two-region MFDs problem

The two-region MFDs problem (4.1)–(4.12) aims at findng the perimeter controllers, i.e. ratios of transfer flows of R_1 and R_2, that maximize the number of vehicles completing their trips (reach their destinations). This problem is an optimal control problem with a nonlinear objective function (4.1) and dynamic equations (4.2)–(4.5), inequality state constraints (4.6)–(4.9), and control constraints (4.10)–(4.11). Moreover, errors are expected in the modeling due to the scatter in the MFDs, mainly in the congested regime and of the demand profile. Therefore, the optimal control problem is solved by applying the model predictive control (MPC) approach which has the ability to handle the state and control constraints, and the errors in the MFDs modeling. Furthermore, the MPC is a real-time implementable solution that can be utilized for real-time urban traffic applications.

The MPC is a form of rolling horizon control in which the current control variables are obtained by solving a finite horizon open-loop optimal control problem at each time step with a feedback current state from the plant as the initial state of the model, see Fig. 4.2. The open-loop optimization problem yields a sequence of optimal control variables after several iterations of solving nonlinear programming, and the first control in this sequence is applied to the plant, then the procedure is carried out again. This scheme of feedback control, i.e. the feedback loop of states from the plant to the model as initial states for the optimization, can handle the errors between the prediction model and the plant.

4.3.1 Two-region MFDs prediction model and optimization problem

The MPC controller obtains the optimal control sequence for the current horizon by solving an optimization problem formulated with prediction model, see bottom of Fig. 4.2.

The prediction model used in the MPC scheme is formulated with equations (4.2)–(4.5). The dynamic equations predict the evolution of accumulations for the two regions with MFDs given the initial accumulations and future values of perimeter controllers and demand.

In this chapter, we follow the direct methods to solve the optimization problem (other solution methods include dynamic programming and indirect methods). The direct methods are most commonly used methods due to their applicability and robustness, where their basic principle is “first discretize and then optimize”. These methods can handle inequality constraints and

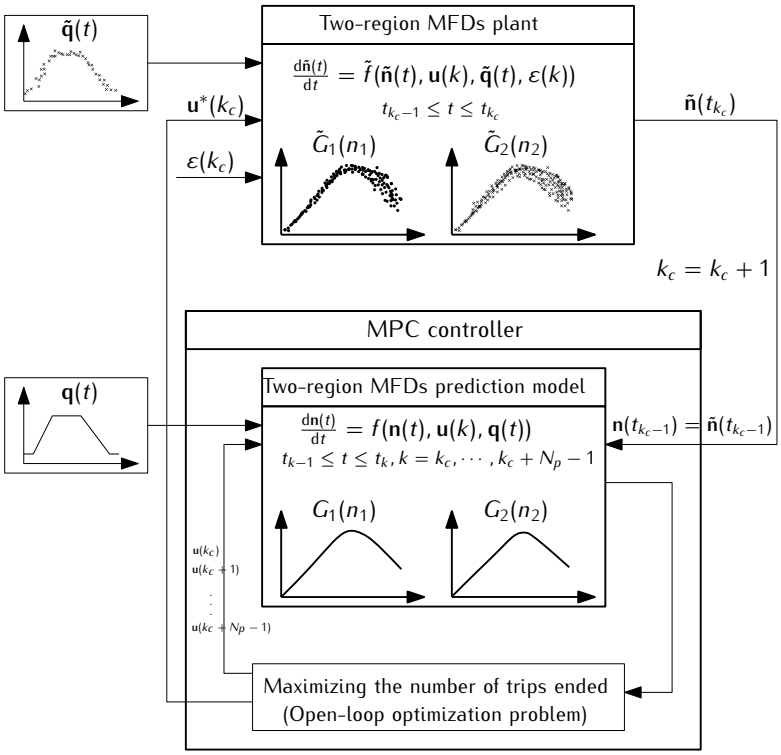


Figure 4.2: Model predictive control scheme for two-region MFDs system.

use the state-of-the-art methods for nonlinear problem solvers.

The open-loop optimal control problem is solved using the direct sequential method, also referred to as single-shooting or control vector parameterization (CVP) in the literature, e.g. [172], [13]. The direct sequential method transcripts the open-loop optimal control problem into a finite-dimensional nonlinear problem through discretization of the control variables *only* with piecewise constant controls, while the ODEs are embedded in the nonlinear problem, i.e. numerical integration is used between the time steps. A schematic description of the direct sequential method is shown in Fig. 4.3. Note the continuous dynamics of the state variables $n_{ij}(t)$, $i, j = 1, 2$. Let N_p (-) be the finite-dimensional horizon which starts from the current control step k_c. At each discrete time step k, $k_c \le k \le k_c + N_\mathrm{p} - 1$, there are two perimeter control variables: $u_{12}(k)$ and $u_{21}(k)$ which are assumed to be constant during the time period $t_{k-1} \le t \le t_k$. For online computational complexity, the number of control variables that should be optimized are reduced to a horizon smaller than N_p, called the control horizon N_c where $N_\mathrm{c} \le N_\mathrm{p}$. The rest of the control variables, i.e. $u_{12}(k)$ and $u_{21}(k)$ for $k_c + N_\mathrm{c} \le k \le k_c + N_\mathrm{p} - 1$ are assumed to be equal to the control variables at the end of the control horizon.

Following the direct sequential method, the control vector is discretized and the two-region MFDs optimal control problem (4.1)–(4.12) is approximated by a finite-dimensional nonlinear programming problem in the piecewise constants controls variables. First, the equations

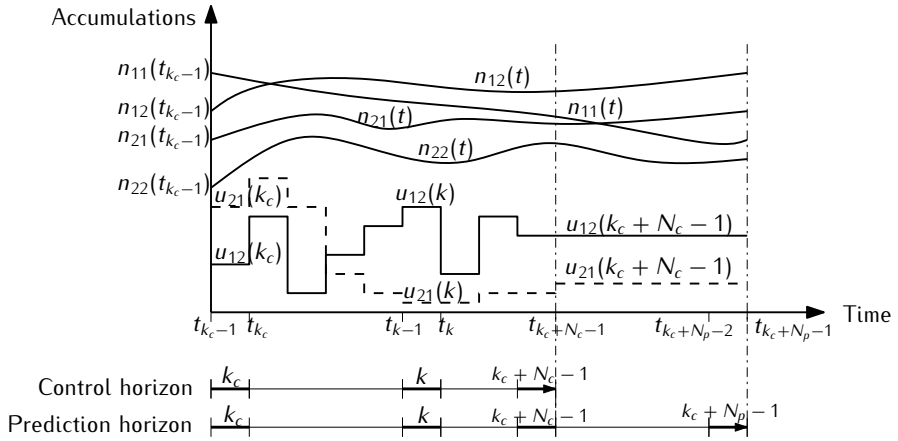


Figure 4.3: Direct sequential method for solving the open-loop optimization problem.

of the prediction model (4.2)–(4.5) are re-written in a compact form with *discrete control variables*, at time step k_c with finite-dimensional N_p as follows:

$$\frac{\mathrm{d}\mathbf{n}(t)}{\mathrm{d}t} = f(\mathbf{n}(t), \mathbf{u}(k), \mathbf{q}(t)) \tag{4.13}$$
$$t_{k-1} \le t \le t_k, \quad k = k_c, k_c+1, \cdots, k_c+N_p-1$$

where $\mathbf{n}(t) = [n_{11}(t), n_{12}(t), n_{21}(t), n_{22}(t)]^\mathrm{T}$, $\mathbf{q}(t) = [q_{11}(t), q_{12}(t), q_{21}(t), q_{22}(t)]^\mathrm{T}$, $\mathbf{u}(k) = [u_{12}(k), u_{21}(k)]^\mathrm{T}$. Then, the Lagrange form (4.1) is transferred into Mayer form by introducing an additional state variable $z(t)$, and an additional differential equation $\mathrm{d}z(t)/\mathrm{d}t$. Moreover, the path constraints (4.6), (4.7), (4.8), and (4.9) must hold for all t (continuous variable) where $t_0 \le t \le t_f$, hence the number of constraints would be infinite. However, several methods are efficient to deal with path constraints in the sequential method, e.g. transcription as integral constraints. The optimization problem is now formulated as follows:

$$\min_{\mathbf{u}(k_c), \mathbf{u}(k_c+1), \cdots, \mathbf{u}(k_c+N_p-1)} -z(t_{k_c+N_p-1}) \tag{4.14}$$

subject to

$$\frac{\mathrm{d}\mathbf{n}(t)}{\mathrm{d}t} = f(\mathbf{n}(t), \mathbf{u}(k), \mathbf{q}(t)), \tag{4.15}$$
$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = M_{11}(t) + M_{22}(t), \tag{4.16}$$
$$\mathbf{u}_\mathrm{min} \le \mathbf{u}(k) \le \mathbf{u}_\mathrm{max}, \tag{4.17}$$
$$\text{where } t_{k-1} \le t \le t_k, \quad k = k_c, k_c+1, \cdots, k_c+N_p-1,$$
$$\mathbf{u}(k) = \mathbf{u}(k_c+N_c-1) \quad k = k_c+N_c, \cdots, k_c+N_p-1, \tag{4.18}$$

$$\sum_{k=k_c}^{k_c+N_p-1} \int_{t_{k-1}}^{t_k} \max\Big\{0; -n_{11}(t) - n_{12}(t)\Big\}^2 \,\mathrm{d}t \le \epsilon, \tag{4.19}$$

$$\sum_{k=k_c}^{k_c+N_p-1} \int_{t_{k-1}}^{t_k} \max\Big\{0; -n_{21}(t) - n_{22}(t)\Big\}^2 \,\mathrm{d}t \le \epsilon, \tag{4.20}$$

$$\sum_{k=k_c}^{k_c+N_p-1} \int_{t_{k-1}}^{t_k} \max\Big\{0; n_{11}(t) + n_{12}(t) - n_{1,\mathrm{jam}}\Big\}^2 \,\mathrm{d}t \le \epsilon, \tag{4.21}$$

$$\sum_{k=k_c}^{k_c+N_p-1} \int_{t_{k-1}}^{t_k} \max\Big\{0; n_{21}(t) + n_{22}(t) - n_{2,\mathrm{jam}}\Big\}^2 \,\mathrm{d}t \le \epsilon, \tag{4.22}$$

where $\mathbf{u}_{\min} = [u_{\min}, u_{\min}]^{\mathrm{T}}$, and $\mathbf{u}_{\max} = [u_{\max}, u_{\max}]^{\mathrm{T}}$. Note that the path constraints (4.6), (4.7), (4.8), and (4.9) are reformulated as integral constraints (4.19), (4.20), (4.21), and (4.22), respectively, with relaxation where $\epsilon > 0$ a small nonnegative constant.

4.3.2 Two-region MFDs plant

The dynamic equations of the two-region MFDs plant differ from the prediction model (4.2)–(4.5) as they include *errors* in the MFDs for both regions and *noises* in the traffic demand; see top of Fig. 4.2 for an illustration. Hence, the evolutions of the accumulations over time are not the same for the prediction model and the plant, which are considered to have different magnitudes and profile (biased and unbiased). Thus, the plant and the model accumulation, MFD and demand can significantly differ.

Errors in MFDs

Let us denote the macroscopic fundamental diagrams with errors for R_1 and R_2 by $\tilde{G}_1$ and $\tilde{G}_2$, respectively. The errors in the MFDs result errors in accumulations of the plant which are denoted by $\tilde{n}_{ij}(t)$, $i, j = 1,2$ to distinguish them from the accumulations of the prediction model $n_{ij}(t)$. Note that $\tilde{n}_i(t) = \sum_j \tilde{n}_{ij}(t)$. The variance of the MFD increases with the accumulations in the region, as described in [58], see top of Fig. 4.2. Reasons for this variance are asymmetric O-D and route choices, which increase the heterogeneity in the distribution of congestion within a region. It is assumed that the variance of the macroscopic fundamental diagram is uniformly distributed where the error at step k is calculated at the time instant t_{k-1}

$$\varepsilon(\tilde{n}_1(t_{k-1})) \sim U\big(-\alpha_1 \cdot \tilde{n}_1(t_{k-1}), \alpha_1 \cdot \tilde{n}_1(t_{k-1})\big) \tag{4.23}$$

$$\varepsilon(\tilde{n}_2(t_{k-1})) \sim U\big(-\alpha_2 \cdot \tilde{n}_2(t_{k-1}), \alpha_2 \cdot \tilde{n}_2(t_{k-1})\big) \tag{4.24}$$

where α_1 and α_2 (1/s) are given parameters. It is assumed that the errors $\varepsilon(\tilde{n}_1(t_{k-1}))$ and $\varepsilon(\tilde{n}_2(t_{k-1}))$ are constant during the time step $t_{k-1} \le t \le t_k$, therefore the MFDs of the plant $\tilde{G}_1$

and $\tilde{G}_2$ for $t_{k-1} \le t \le t_k$ are:

$$\tilde{G}_1(\tilde{n}_1(t)) = G_1(\tilde{n}_1(t)) + \varepsilon(\tilde{n}_1(t_{k-1})) \tag{4.25}$$

$$\tilde{G}_2(\tilde{n}_2(t)) = G_2(\tilde{n}_2(t)) + \varepsilon(\tilde{n}_2(t_{k-1})) \tag{4.26}$$

Unbiased and biased noises in demand

Two different types of noise in demand are considered: (i) unbiased noise with Gaussian distribution, and (ii) biased noise with a sudden jump in the demand profile for a time period. In both cases, let us denote the traffic demand $\mathbf{q}(t)$ with noises as $\tilde{\mathbf{q}}(t) = [\tilde{q}_{11}(t), \tilde{q}_{12}(t), \tilde{q}_{21}(t), \tilde{q}_{22}(t)]^{\mathrm{T}}$. Unbiased demand noises represent random and recurrent variation of demand from day to day, because travel patterns, while biassed demand noises might represent cases of non-recurrent events (special events, accidents, etc).

The unbiased noises in demand are assumed to have Gaussian distribution as follows:

$$\tilde{q}_{ij}(t) = \max\left(q_{ij}(t) + \mathcal{N}(0, \sigma_{ij}^2), 0\right) \tag{4.27}$$

where $i, j = 1, 2$ and σ_{ij}^2 (veh/s)^2 is the variance for the traffic demand $q_{ij}(t)$.

Substituting the MFDs with errors (4.25) and (4.26) and the demand with noises $\tilde{\mathbf{q}}(t)$ in the dynamic equations (4.2), (4.3), (4.4), and (4.5), one gets the two-region MFDs plant in a compact form, see also Fig. 4.2,

$$\frac{\mathrm{d}\tilde{\mathbf{n}}(t)}{\mathrm{d}t} = \tilde{f}(\tilde{\mathbf{n}}(t), \mathbf{u}(k), \tilde{\mathbf{q}}(t), \boldsymbol{\varepsilon}(k)) \tag{4.28}$$

where $\boldsymbol{\varepsilon}(k) = [\varepsilon(\tilde{n}_1(t_{k-1})), \varepsilon(\tilde{n}_2(t_{k-1}))]^{\mathrm{T}}$.

4.3.3 "Greedy" control

In order to investigate and estimate the performance of the MPC controller, comparison results are done with a "greedy" control (GC) for different level and type of errors. The greedy control is a state feedback control that its policy is determined by the current accumulations $n_1(t)$ and $n_2(t)$. Let $n_{1,\mathrm{cr}}$ and $n_{2,\mathrm{cr}}$ (veh) be the accumulations that maximize G_1 and G_2, respectively. The greedy control is designed according to the following policy: if both regions are uncongested, i.e. $n_1(t) \le n_{1,\mathrm{cr}}$ and $n_2(t) \le n_{2,\mathrm{cr}}$, then both controllers should maximize the transfer flows, therefore $[u_{12}(t), u_{21}(t)] = [u_{\max}, u_{\max}]$. If one region is congested and the other one is uncongested, i.e. $n_1(t) \le n_{1,\mathrm{cr}}$ and $n_2(t) > n_{2,\mathrm{cr}}$, or, $n_1(t) > n_{1,\mathrm{cr}}$ and $n_2(t) \le n_{2,\mathrm{cr}}$, then the controllers should minimize the transfer flow to the congested region and maximize the transfer flow to the uncongested region. If both regions are congested, i.e. $n_1(t) > n_{1,\mathrm{cr}}$ and $n_2(t) > n_{2,\mathrm{cr}}$, then controllers should minimize the transfer flow to the "more congested" region, and maximize the transfer flow to the "less congested" region, e.g. if $n_1(t)/n_{1,\mathrm{jam}} > n_2(t)/n_{2,\mathrm{jam}}$, then R_1 is more congested than R_2, therefore $[u_{12}(t), u_{21}(t)] = [u_{\max}, u_{\min}]$. The

greedy control law is summarized in Table 4.1.

Table 4.1: Greedy controllers $u_{12}(t)$ and $u_{21}(t)$ policy.

$[u_{12}(t), u_{21}(t)]$	$n_1(t) \leq n_{1,\text{cr}}$	$n_1(t) > n_{1,\text{cr}}$
$n_2(t) \leq n_{2,\text{cr}}$	$[u_{\max}, u_{\max}]$	$[u_{\max}, u_{\min}]$
$n_2(t) > n_{2,\text{cr}}$	$[u_{\min}, u_{\max}]$	$[u_{\max}, u_{\min}]$ if $\frac{n_1(t)}{n_{1,\text{jam}}} > \frac{n_2(t)}{n_{2,\text{jam}}}$
		$[u_{\min}, u_{\max}]$ otherwise

4.3.4 Tuning the prediction and control horizons parameters

The performance of the MPC controller is affected by the prediction horizon N_p and the control horizon N_c. The prediction horizon N_p should be large enough such that the model can accurately predict the accumulations of the plant corresponding to the control variables. Increasing the prediction horizon improves the performances of the MPC controller, however, large N_p increases the optimization computing time which may add some barriers for online implementation, i.e. the control actions cannot be implemented in the current step if the computing time corresponding to large N_p is larger than the time duration of the control time step. Similar considerations regarding the tradeoff between computation complexity and results accurately should be done for the control horizon N_c.

The perimeter controllers can be actuated by signalized intersections placed in the border between the two regions of the urban network, i.e. the perimeter control sequences can be applied by choosing appropriate timing plans for the signalized intersections. The effect of perimeter control to the rest of the network and its MFDs are discussed later. Let us assume that the signalized intersections have a fixed common cycle length, e.g. equals to 60 (s). Then the time duration of the time step k_c is set to be equal to the length of the cycle, i.e. $t_k - t_{k-1} = 60$. This duration is much larger than the time-consuming for solving the open-loop optimization problem (a few seconds).

Tuning the MPC parameters N_c and N_p is done for several examples and similar results have been obtained. In the following, the tuning analysis is presented only for one of the case study examples as shown in Fig. 4.4. More information for the case study examples is presented in the next section.

The MPC parameters N_p and N_c are tuned according to the relative improvement of the trip completed corresponding to MPC controller compared with the greedy controller, i.e. improvement (%) = $(J_\text{MPC} - J_\text{GC})/J_\text{GC}$, where J is computed according to (4.1). As shown in Fig. 4.4, the improvement of the results is decreasing as the prediction horizon N_p increases, however for $N_\text{p} \geq 20$ only minor improvement is achieved. The MPC controller is less sensitive to the control horizon N_c, where $N_\text{c} \geq 2$ yields similar results for trip completion. It is shown that $N_\text{c} = 1$ yields to a low MPC performance since the open-loop optimization problem is more constrained. Therefore, the parameters are set as $N_\text{p} = 20$ and $N_\text{c} = 2$ for all subsequent

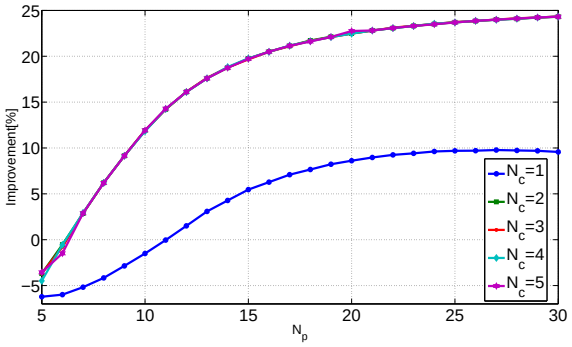


Figure 4.4: Tuning parameters N_p and N_c for MPC controller.

case study examples. Note that for small prediction horizon $N_\mathrm{p} < 6$ the MPC controller does not perform well compared with the greedy controller.

4.4 Case study examples

In this section, results of several case study examples are presented to explore the features of the MPC controller under different conditions. The examples aim at examining the efficiency of MPC controller in congested and uncongested regimes which may vary with time because of variations in demand and the MFDs. Hence, examples with different levels of demand and sizes of MFDs are presented. Furthermore, the robustness of the MPC and greedy controllers are also examined by introducing different uncertainties, i.e. different levels of error in MFD, and noises in demand. For all examples presented in this chapter, the selected MPC parameters are $N_\mathrm{p} = 20$ and $N_\mathrm{c} = 2$, the lower bound $u_\mathrm{min} = 0.1$, and the upper bound $u_\mathrm{max} = 0.9$. In examples $1-4$, the MFDs for both regions are the same where $G_i(n_i(t)) = (1.4877 \cdot 10^{-7} \cdot n_i^3 - 2.9815 \cdot 10^{-3} \cdot n_i^2 + 15.0912 \cdot n_i)/3600$, $i = 1,2$, $n_{1,\mathrm{cr}} = n_{2,\mathrm{cr}} = 3400$ (veh), $G_1(n_{1,\mathrm{cr}}) = G_2(n_{2,\mathrm{cr}}) = 6.3$ (veh/s), $n_{1,\mathrm{jam}} = n_{2,\mathrm{jam}} = 10000$ (veh). This shape is consistent with the MFD observed in Yokohama, see [55]. Note that the shape of the MFD is not predetermined in the problem formulation. In example 5, the MFD of region R_1 is increased by 25% as shown in Fig. 4.8(d) to test the control problem with two different MFDs.

In example 1, both regions R_1 and R_2 are initially congested, i.e. the initial accumulations $n_1(t_0) = 5400$ and $n_2(t_0) = 4000$ are in the decreasing part of the MFD, where $n_2(t_0)$ is 18% larger than $n_{\mathrm{cr},2}$, while $n_1(t_0)$ is 59% larger than $n_{\mathrm{cr},1}$, which means that region R_2 operates close to capacity conditions. The time varying demands shown in Fig. 4.5(d) simulate a morning peak hour with high demand $q_{12}(t)$ for trips from R_1 to R_2, i.e. from the periphery to the city center. The evolution of accumulations over time $n_{ij}(t)$, $0 \le t \le 3600$, corresponding to the MPC controller are presented in Fig. 4.5(a), while the evolutions presented in Fig. 4.5(b) are corresponding to the greedy controller, see Table (4.1). Note that at the beginning of the control process, both MPC and greedy controllers decrease the total accumulation in R_1, $n_1(t)$, and keep the total accumulation in R_2, $n_2(t)$, unchanged. Afterwards, the MPC controller tries to

decrease $n_2(t)$ by changing $u_{21}(t)$ from 0.1 to 0.55 to let more vehicles enter to R_1. In contrast, the greedy controller brings the two accumulations equal, i.e. $n_1(590) = n_2(590) = 4125$ (veh), and after that instance, both region accumulations increase together while the *chattering* behavior occurs as a result of switching control between $u_{\min}$ and $u_{\max}$ according to Table (4.1); note the saw lines of accumulations after $t = 600$ (s).

The cumulative trip completion corresponding to MPC and greedy controllers are shown in Fig. 4.5(c), while the control sequences $u_{12}(t)$ and $u_{21}(t)$ are shown in Fig. 4.5(e). The third polynomial MFDs $G_1(n_1)$ and $G_2(n_2)$ are coincided as shown in Fig. 4.5(f), while the circle points are the calculated $\tilde{G}_1$ and $\tilde{G}_2$, see (4.25) and (4.26). In Fig. 4.5(f), it is assumed that there are no errors in both MFDs where $\alpha_1 = \alpha_2 = 0$, see (4.23) and (4.24).

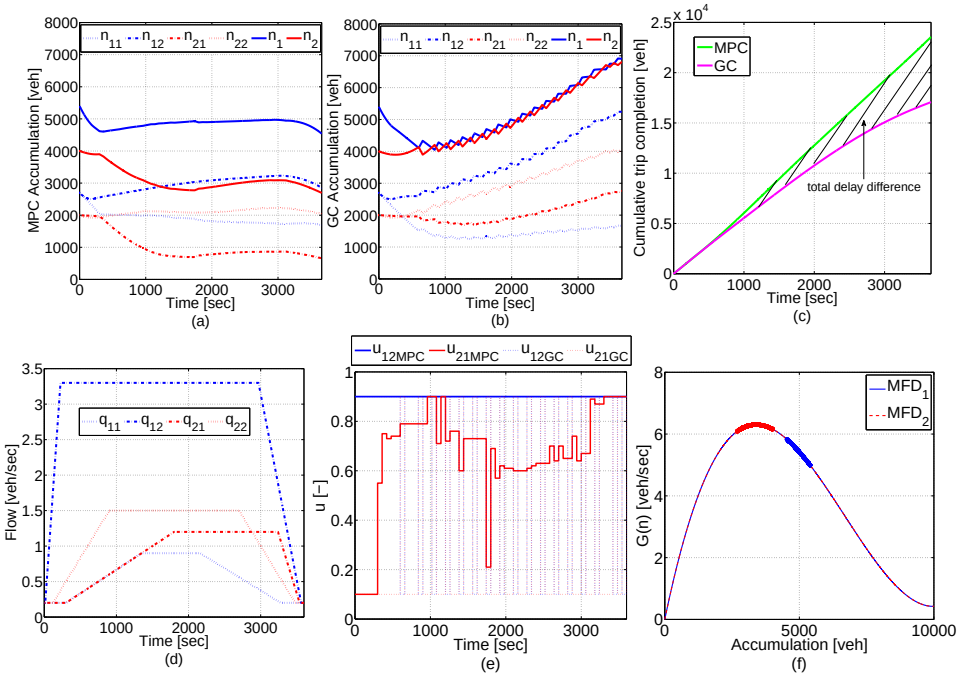


Figure 4.5: Example 1: regions R_1 and R_2 are initially congested (without errors in MFDs). (a) Region accumulations with MPC, (b) Region accumulation with GC, (c) Cumulative trip completion for MPC and GC, (d) time-varying demand, (e) MPC and GC control actions, and (f) MFD with MPC.

The MPC performances for small ($\alpha_1 = \alpha_2 = 0.2$) and large ($\alpha_1 = \alpha_2 = 1$) errors in the MFDs are shown in Fig. 4.6(a) and Fig. 4.6(b), respectively. Comparison between the three levels of error, i.e. without errors in Fig. 4.5(e), small and large errors in Fig. 4.6, shows that the controller $u_{21}(t)$ becomes less smoother when the errors in the MFD of the plant (reality) increase. Note that the performance of the MPC is not significantly affected when large errors in the MFD occur. Even if the macroscopic traffic congestion modeling is a rough approximation of urban networks this methodology can still improve traffic conditions. Nevertheless, non smooth controllers might be difficult to apply in reality and might create heterogeneities in the boundary and scatter in the MFDs. We discuss smoothing issues in Section 4.5.

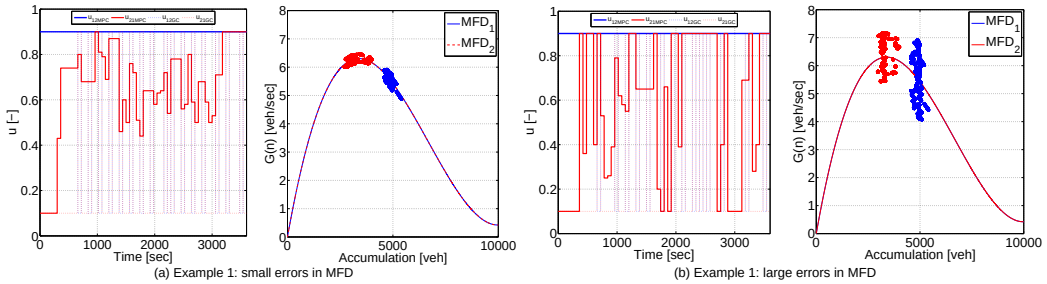


Figure 4.6: Example 1 control actions and MFDs. (a) small errors in MFD ($\alpha_1 = \alpha_2 = 0.2$) and (b) large errors in MFDs ($\alpha_1 = \alpha_2 = 1$).

In example 1 the demand is high such that at the end of the control process both regions are congested (with greedy controller regions move forward to face gridlock). The effect of demand on the MPC controller is scrutinized by examples 2 and 3. These examples have the same initial accumulations of example 1, however, the demands $q_{ij}(t)$ for examples 2 and 3 are proportionally decreased by 16% and 32%, respectively, compared with the demand for example 1 in Fig. 4.5(d). The performance of GC and MPC for different errors in demand and MFDs are summarized in Table 4.2. The results shown in the table are an average over 10 runs for small and large errors, and low and high noises. We show the cumulative trip completion by the end of the simulation and the difference in the total delays (veh·s) as this is expressed by the the area between the MPC and GC cumulative trip completion curves in the figures, e.g. Fig. 4.5(c). In example 1 there are 22 – 24% savings in total, which represent on average 5.5 (min) savings per traveler trip.

Table 4.2: The trip completion corresponding to MPC and GC, and the total delay difference.

Example 1	without errors ($\alpha_1 = \alpha_2 = 0$)			small errors ($\alpha_1 = \alpha_2 = 0.2$)			large errors ($\alpha_1 = \alpha_2 = 1$)		
	MPC (veh$\cdot 10^3$)	GC (veh$\cdot 10^3$)	MPC-GC (veh$\cdot$s$\cdot 10^3$)	MPC	GC	MPC-GC	MPC	GC	MPC-GC
without noises	23.55	17.07	7791.6 (22.5%)	23.63	17.11	7883.5 (22.7%)	24.04	17.21	8370.0 (24.1%)
low noises	23.37	16.78	7864.3 (22.9%)	23.48	16.82	7997.3 (23.3%)	23.93	17.00	8379.5 (24.3%)
high noises	22.80	16.07	7892.6 (23.5%)	23.02	16.35	7931.9 (23.4%)	23.34	15.84	9146.5 (27.6%)
Example 2	without errors ($\alpha_1 = \alpha_2 = 0$)			small errors ($\alpha_1 = \alpha_2 = 0.2$)			large errors ($\alpha_1 = \alpha_2 = 1$)		
	MPC	GC	MPC-GC	MPC	GC	MPC-GC	MPC	GC	MPC-GC
without noises	24.11	20.49	6536.8 (17.3%)	24.12	20.51	6548.8 (17.3%)	24.13	20.56	6664.4 (17.6%)
low noises	24.15	20.43	6654.2 (17.7%)	24.15	20.47	6666.2 (17.7%)	24.17	20.56	6741.1 (17.9%)
high noises	24.29	20.28	6885.2 (18.4%)	24.30	20.34	6899.7 (18.4%)	24.33	20.45	6966.9 (18.6%)
Example 3	without errors ($\alpha_1 = \alpha_2 = 0$)			small errors ($\alpha_1 = \alpha_2 = 0.2$)			large errors ($\alpha_1 = \alpha_2 = 1$)		
	MPC	GC	MPC-GC	MPC	GC	MPC-GC	MPC	GC	MPC-GC
without noises	21.70	21.63	1789.3 (4.4%)	21.70	21.64	1736.1 (4.2%)	21.70	21.64	1836.7 (4.5%)
low noises	21.78	21.69	1923.3 (4.7%)	21.78	21.69	1985.8 (4.9%)	21.78	21.70	2051.6 (5.0%)
high noises	22.11	21.94	2466.7 (6.1%)	22.11	21.94	2490.4 (6.1%)	22.11	21.95	2578.9 (6.4%)
Example 4	without errors ($\alpha_1 = \alpha_2 = 0$)			small errors ($\alpha_1 = \alpha_2 = 0.2$)			large errors ($\alpha_1 = \alpha_2 = 1$)		
	MPC	GC	MPC-GC	MPC	GC	MPC-GC	MPC	GC	MPC-GC
without noises	24.40	22.72	1143.7 (2.6%)	24.40	22.78	1135.4 (2.6%)	24.43	22.96	1120.4 (2.5%)
low noises	24.35	22.61	1200.0 (2.7%)	24.36	22.66	1204.5 (2.7%)	24.39	22.85	1192.8 (2.7%)
high noises	24.25	22.26	1439.3 (3.3%)	24.29	22.31	1435.9 (3.3%)	24.28	21.85	2032.9 (4.7%)
Example 5	without errors ($\alpha_1 = \alpha_2 = 0$)			small errors ($\alpha_1 = \alpha_2 = 0.2$)			large errors ($\alpha_1 = \alpha_2 = 1$)		
	MPC	GC	MPC-GC	MPC	GC	MPC-GC	MPC	GC	MPC-GC
without noises	27.37	21.77	9596.6 (23.3%)	27.37	21.82	9576.3 (23.2%)	27.38	21.93	9615.9 (23.2%)
low noises	27.40	21.60	9784.1 (23.8%)	27.37	21.67	9739.1 (23.7%)	27.38	21.85	9705.4 (23.5%)
high noises	27.37	21.09	10249.1 (25.3%)	27.38	21.12	10282.5 (25.4%)	27.41	21.31	10284.0 (25.3%)

Comparing examples 1, 2, and 3, see Fig. 4.5(c) and Table 4.2, we notice the advantage of MPC controller compared to greedy controller according to the total delay (veh·s) that would be obtained if the greedy controller is used instead of the MPC controller. The differences

between the total delays are proportional to the congestion level, i.e. in uncongested situation (the current accumulations are below the jam accumulations and the future demand is not going to change them) the performance of the greedy controller is almost the same as the MPC controller, and the total delay difference is almost zero.

In example 4, in contrast to other examples, both regions R_1 and R_2 are initially uncongested, i.e. the initial accumulations are in the increasing part of the MFD, while at the end of the control process both regions are congested as shown in Fig. 4.7 because of high level of time-varying demand. The accumulation profiles for both MPC and greedy controllers show the same trend in uncongested regime. Whereas, once the system reaches to the critical switching point, a significant difference is seen between MPC and greedy controllers. It can inferred that when the regions become more congested the difference between MPC and greedy controllers is more apparent.

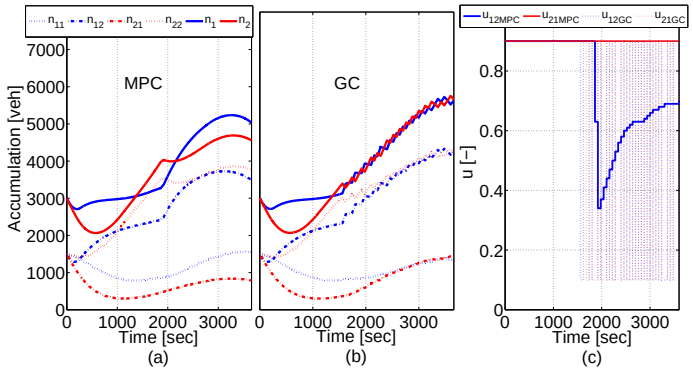


Figure 4.7: Example 4: regions R_1 and R_2 are initially uncongested and finally congested. (a) Region accumulations with MPC, (b) Region accumulations with GC, and (c) MPC and GC control actions.

In example 5 the region R_1 MFD and its corresponding internal demand $q_{11}(t)$, as shown in 4.5(d), are increased by 25%. The MFDs for this example are depicted in Fig. 4.8(d), while the MFDs for examples 2, 3, and 4 are shown in Fig. 4.8(a), (b), and (c), respectively.

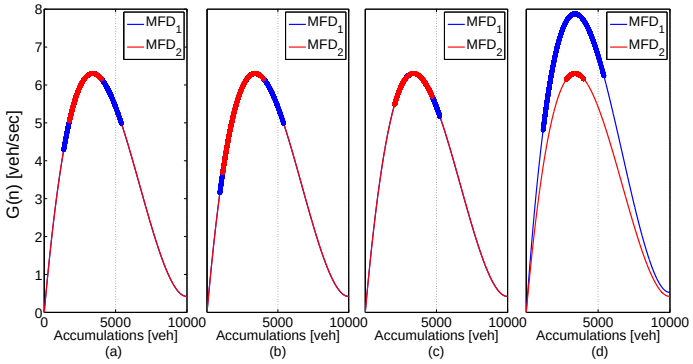


Figure 4.8: The MFDs for both regions without errors in the MFDs or noises in the demands for (a) example 2, (b) example 3, (c) example 4, and (d) example 5.

In addition, the performance of MPC controller encountering unbiased and biased noises in demand is investigated. Example 1 with high unbiased noises in demand ($\sigma_{ij} = 0.5$, $i, j = 1,2$, see (4.27)) is illustrated in Fig. 4.9. The overall results of MPC remain similar, however, the corresponding applied MPC shows more fluctuations than the base example 1. The differences between the trip completion corresponding to MPC and greedy control without errors, with small ($\alpha_1 = \alpha_2 = 0.2$), and large errors ($\alpha_1 = \alpha_2 = 1$) in the plant MFDs and without noises, low ($\sigma_{ij} = 0.25$, $i, j = 1,2$) and high ($\sigma_{ij} = 0.5$, $i, j = 1,2$) unbiased noises in demand profile are also summarized in Table 4.2.

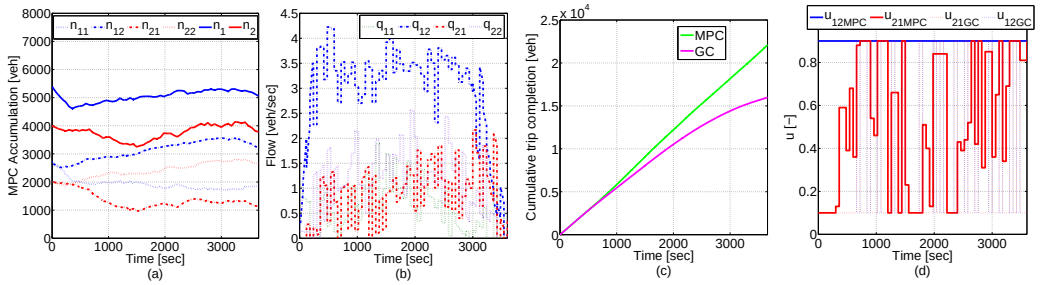


Figure 4.9: Example 1: high unbiased noise in demand. (a) Region accumulations with MPC, (b) Demand with noise, (c) Cumulative trip completion for MPC and GC, and (d) MPC and GC control actions.

In Fig. 4.10, a biased noise in demand which occurs at time instant 1200 (s) for duration of 600 (s) is added to the base setup of example 1, see Fig. 4.10(b). The MPC controller does not receive the information of the sharp change in demand. As Fig. 4.2 shows, the input to the MPC controller is the demand without any noises, which can be estimated with traditional methods. Nevertheless, it results in a similar performance, whereas greedy controller makes both regions to gridlock (it can be inferred from almost horizontal ending part of greedy controller trip completion profile, see Fig. 4.10(c)). Note that the MPC profile in Fig. 4.10(d) is identical to Fig. 4.5(e) for times before 1200 (s) and after that with decreasing $u_{12}(t)$ from $u_{\max}$, the MPC can handle the unbiased sudden augmentation in demand which has a great impact on $n_2(t)$ accumulation, see Fig. 4.10(a). The results also show that the trip completion corresponding to MPC are similar comparing between unbiased noise (23.49 (veh$\cdot 10^3$)) and without noise in demand (23.55 (veh$\cdot 10^3$), see Table 4.2) for no errors in MFDs case. However, large differences are obtained for GC as the trip completion decreased from 17.07 to 13.97 (veh$\cdot 10^3$), and the total delay difference increased from 7791.6 (veh$\cdot$s$\cdot 10^3$) (22.5%) to 10684.3 (veh$\cdot$s$\cdot 10^3$) (33.2%).

The results shown in this section imply that the MPC controller is superior for all examples with different levels of error in MFD and noise in demand, biased or unbiased.

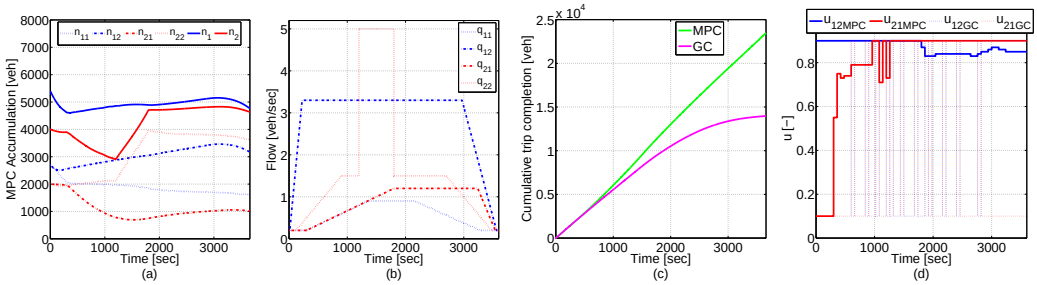


Figure 4.10: Example 1: biased noise in demand. (a) Region accumulations with MPC, (b) Demand with noise, (c) Cumulative trip completion for MPC and GC, and (d) MPC and GC control actions.

4.5 Smoothing control

In the previous section, the results of various examples have shown that the control values of two successive steps may significantly vary, e.g. see Fig. 4.9(d), especially in cases of demand variations which are more realistic. These large jumps make the implementation of control policies difficult for real cases, they might jeopardize safety and might increase the heterogeneity of congestion distribution, which can also result in highly-scattered MFDs. Thus, one would like to limit the change (jump) in the control values and smooth the control sequences over the control process. In this section, two different methods are introduced to smooth the control sequences: (i) imposing control constraints and (ii) modifying the objective function.

4.5.1 Constraints for smoothing control

One method to smooth the control sequences resulting from MPC is to impose smoothing control constraints to the optimal open-loop problem (4.14)–(4.22) over the control horizon N_c. The imposed smoothing control constraints that limit jumps in the control sequence up to u_{jump} (−) are as follows:

$$|u_{12}(k) - u_{12}(k-1)| \leq u_{\text{jump}} \tag{4.29}$$

$$|u_{21}(k) - u_{21}(k-1)| \leq u_{\text{jump}} \tag{4.30}$$

for $k = k_c, k_c+1, \ldots, k_c+N_c-1$, where u_{jump} is a given parameter, and $u_{12}(k_c-1)$ and $u_{21}(k_c-1)$ are the applied controls in the previous horizon.

In Section 4.4, we note that MPC for example 1 with high noise in demand profile is very jumpy. To smooth the control sequences of MPC, we utilize the MPC formulation with the confining constraints on control (4.29) and (4.30). The results of two values $u_{\text{jump}} = 0.1$ and $u_{\text{jump}} = 0.2$ are shown in Fig. 4.11(a) and (b), respectively. Both of them yield similar results to the unsmoothed example 1, see Fig. 4.9(a), however, control profiles are smoother and show an identical trend. This test also revels that the MPC formulation is robust to selection of u_{jump} value. It is not only a theoretical tool, but it can also have direct applications in the field.

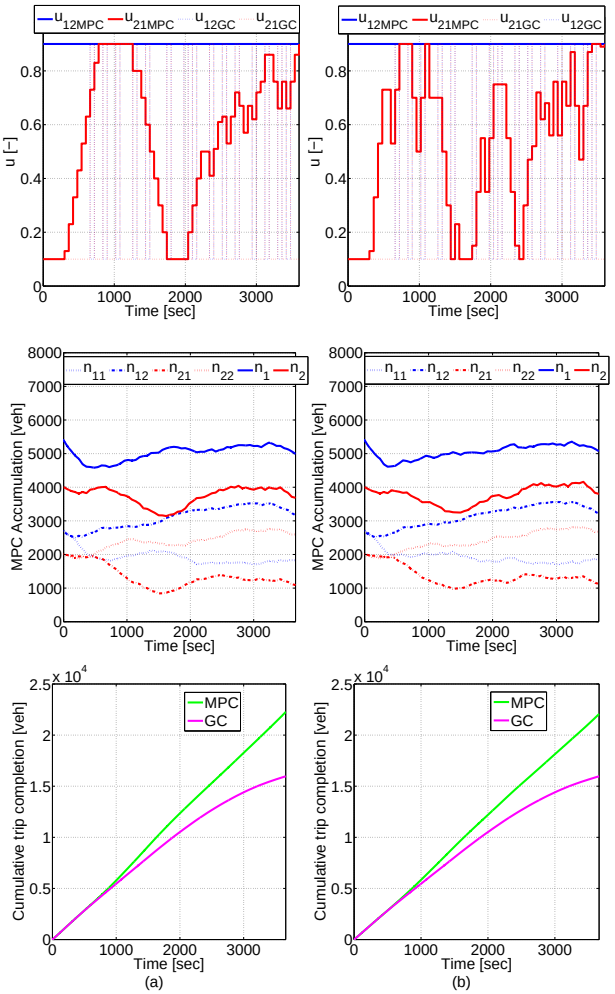


Figure 4.11: Smoothing control by imposing constraints with: (a) $u_{\text{jump}} = 0.1$, and (b) $u_{\text{jump}} = 0.2$.

4.5.2 Modified objective function

The second method to smooth the control sequences is done by introducing a tradeoff between the objective function, i.e. the maximum number of vehicles that complete their trips, and the sum of the square absolute difference between each two control sequences, see e.g. [1]. Therefore, the objective function (4.14) in the optimal open-loop problem (4.14)–(4.22) is modified as follows:

$$\min_{\mathbf{u}(k_c),\cdots,\mathbf{u}(k_c+N_p-1)} \left\{ -z(t_{k_c+N_p-1}) + \beta \sum_{k=k_c}^{k_c+N_c-1} \left(|u_{12}(k) - u_{12}(k-1)|^2 + |u_{21}(k) - u_{21}(k-1)|^2 \right) \right\} \tag{4.31}$$

where β is the weight on control sequence changes.

The role of β in (4.31) and u_{jump} in (4.29) and (4.30) are similar in the sense of making a balance

between desired smoothness and trip completion. Nevertheless, calibration of u_{jump} is easier than β since it represents a physical measure.

Similarly, the same example is chosen for smoothing via modifying the objective function. Figure 4.12 shows the results where each column corresponds to $\beta = 1, 10, 50,$ and 200. It is apparent that the higher the β the smoother is control profile and the less is objective value. For this specific example $\beta = 10$ produces the best trade-off between smoothness and optimality. Nevertheless, the value of β does not have a direct physical meaning (as u_{jump} does) and it might require additional calibration in the field. We tried to apply this approach in the other examples, and the appropriate value of β to obtain smooth control and efficient condition was varying. The first approach gave more robust results.

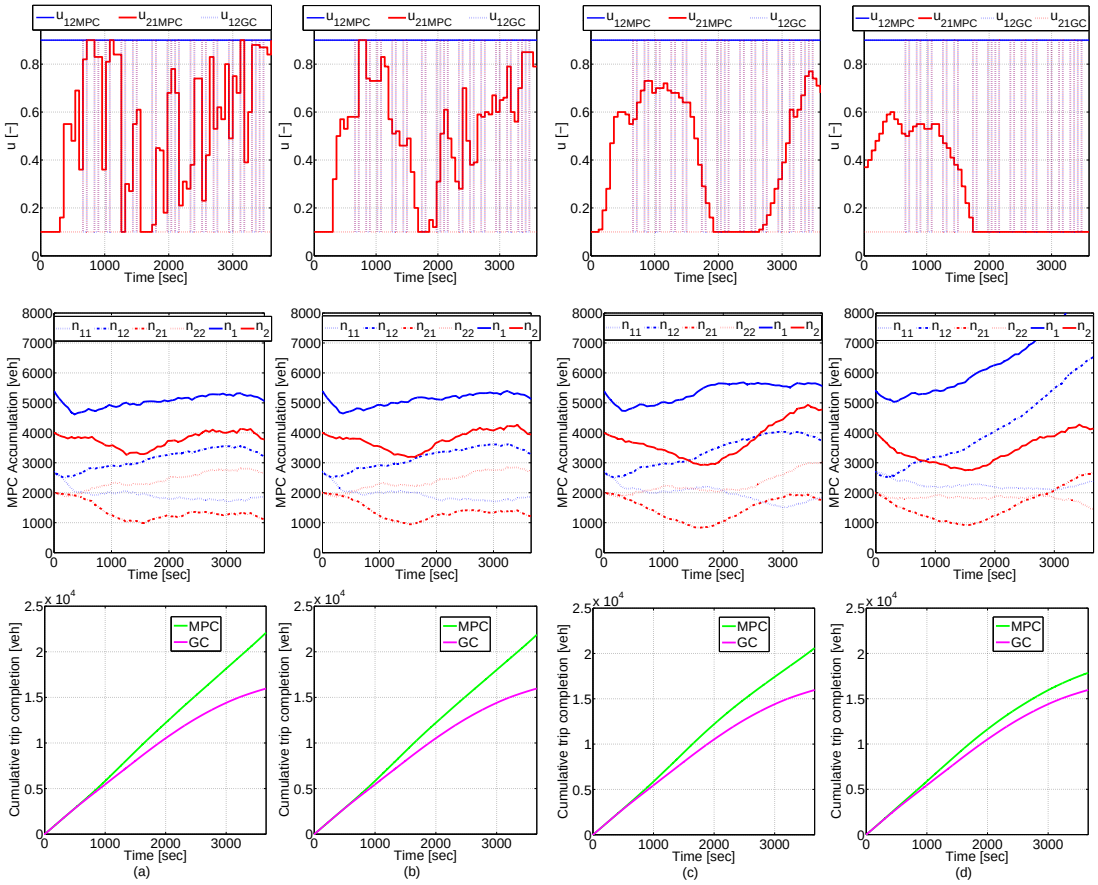


Figure 4.12: Smoothing control by modified objective function (4.31) with: (a) $\beta = 1$, (b) $\beta = 10$, (c) $\beta = 50$, and (d) $\beta = 200$.

4.6 Summary

The optimal perimeter control for two urban regions with MFD representation was formulated and solved by implementing a model predictive control scheme. Comparison results between

the model predictive control and a greedy control show that the model predictive control is superior for all numerical examples presented. These results can be of great importance to practitioners and city managers to unveil simple and robust signal timing planning in such a way that maximize the network capacity and serve the maximum number of people under high demand conditions. The results of this chapter can be utilized to develop efficient hierarchical control strategies for heterogeneously congested cities. A network can be partitioned in homogeneous regions and optimal control methodologies can identify the inter-transfers between regions of a city to maximize the system output, by utilizing the MPC developed in this chapter.

While there are vast contributions in traffic control problems for freeways through ramp metering, the area of control for large urban regions or mixed networks still remains a challenge. Recent findings (e.g. [164], [51]) have shown that MFDs might not be a realistic representation for freeway systems, so in case of mixed arterial-freeway networks, an MFD formulation for the arterial can be combined with a mesoscopic model for the freeway (e.g. a 1st or 2nd order traffic flow model) to describe the dynamics of the system and propose coordinated optimization schemes with ramp metering and perimeter control. Our research provides tools to shed some light toward this direction and Chapter 5 investigates this research problem.

In addition, the effect of density heterogeneity on the MFD is an essential modeling extension, since traffic congestion propagates dynamically over time and space and homogenous conditions might not be valid. To this end, Chapter 6 introduces a functional form to model the effect of heterogeneity on the MFD based on a field data set analysis. Further in Chapter 6, the heterogeneity effect is incorporated into the MFD modeling and perimeter control while two models with different levels of aggregation are used as the optimization model and the plant in the MPC scheme.

A network with multiple homogenous regions (more than two regions) not only requires more state variables to model the dynamics, but also a route choice model is needed to be integrated in the model as vehicles can travel from one region to other with different routes. The route choice strategy can be realized as user equilibrium conditions within the MFD modeling [199]. When the number of homogeneous regions in the network gets larger, the computational complexity and time for solving the MPC problem increase, which might affect the real-time implementation of the proposed methodology. Note that the real-time implementation feasibility is not an issue for the two regions problem presented in this chapter as the computational time was significantly smaller than real time.

5 Cooperative traffic control of a mixed network with two urban regions and a freeway

CURRENTLY most optimization methods for urban transport networks (i) are suited for networks with simplified dynamics that are far from real-sized networks or (ii) apply decentralized control, which is not appropriate for heterogeneously loaded networks or (iii) investigate good-quality solutions through micro-simulation models and scenario analysis, which make the problem intractable in real time. In principle, traffic management decisions for different sub-systems of a transport network (urban, freeway) are controlled by operational rules that are network specific and independent from one traffic authority to another. In this chapter, the macroscopic traffic modeling and control of a large-scale mixed transportation network consisting of a freeway and an urban network is tackled. The urban network is partitioned into two regions, each one with a well-defined Macroscopic Fundamental Diagram (MFD), i.e. a unimodal and low-scatter relationship between region density and outflow. The freeway is regarded as one alternative commuting route which has one on-ramp and one off-ramp within each urban region. The urban and freeway flow dynamics are formulated with the tool of MFD and asymmetric cell transmission model, respectively. Perimeter controllers on the border of the urban regions operating to manipulate the perimeter interflow between the two regions, and controllers at the on-ramps for ramp metering are considered to control the flow distribution in the mixed network. The optimal traffic control problem is solved by the Model Predictive Control (MPC) approach to minimize total delay in the entire network. Several control policies with different levels of urban-freeway control coordination are introduced and tested to scrutinize the characteristics of the proposed controllers. Numerical results demonstrate how different levels of coordination improve the performance once compared with independent control for freeway and urban network. The approach presented in this chapter can be extended to implement efficient real-world control strategies for large-scale mixed traffic networks.

5.1 Introduction

Metropolitan transportation networks have a hierarchical structure which essentially consists of freeways and urban roads providing the interrelated infrastructure for mobility and accessi-

bility. The freeway and the urban network are inherently coupled, but they have dissimilar traffic flow dynamics which challenge the traffic control of mixed networks of two interconnected (urban and freeway) traffic control entities. Although integrating the two entities through an effective mixed control policy could enhance the network performances during heavy congestion conditions, lack of coordination among the urban and freeway network jurisdictions and/or limited means of traffic monitoring and data communication might impede such mixed traffic network ideal goal. To overcome such deficiency, cooperative decentralized or if possible centralized control schemes can be developed as potential solutions, which oblige us to inquire into the traffic dynamics and control of the freeway and urban network to model the mixed traffic network.

Currently most optimization methods for urban transport networks (i) are suited for "toy" networks with simplified dynamics that are far from real-sized networks or (ii) apply decentralized control, which is not appropriate for heterogeneously loaded networks with short links and spillbacks or (iii) investigate good-quality solutions through detailed micro-simulation models and scenario analysis, which due to computational complexity make the problem intractable in real time. In principle, traffic management decisions for different sub-systems of a transport network (urban, freeway) are controlled by operational rules that are network specific and independent from one traffic authority to another. In some cases, the operational decisions of two sub-systems turn out to be competitive. For example, a ramp metering strategy to retain high flows of the freeway sub-system can create long queues in the access ramps that propagate and block the center of the city. In this paper, several control structures with different levels of coordination between the freeway and the urban network control entities are introduced and elaborated for traffic control of the mixed urban-freeway network. Our objective is to investigate how restrictions in coordination among different controllers (e.g. lack of communication or data) can affect the mobility levels of a city. Nevertheless, optimizing in real time all controllers of a city (traffic lights, changeable message signs, on-ramps, etc.) in a coordinated way is an infeasible solution due to the computational burden of a very complex model, needed to represent traffic dynamics. Our objective is to integrate realistic aggregated models of urban and freeway traffic with efficient control approaches that will allow for coordinated traffic management.

Recently, a large effort for the development of Integrated Corridor Management (ICM) has been promoted by Federal Highway Administration, with many case studies around US metropolitan areas. Most of the implementations and case studies mainly perform scenario analysis and considers alternative routes under extreme events, e.g. accidents, while it is expected that a more formal optimization approach could lead to a better system performance.

In freeways, ramp metering is the most commonly used controller to manipulate the flow entering the freeway from its urban surrounding roads. Local and coordinated control strategies were proposed and implemented for ramp metering. In local control strategies, the control law for an on-ramp is determined according to the traffic conditions downstream and upstream of the on-ramp (e.g. ALINEA controller in [134]). In coordinated strategies, the control law for

multiple on-ramps are determined based on the traffic conditions in multiple areas including several on-ramps and sections in the freeway. The coordinated ramp metering is in fact a multi-regulator controller as all ramp meterings attempt to regulate the freeway traffic conditions near the desired densities. Overviews of local and coordinated ramp metering controllers are presented in [137, 133, 57]. The ramp metering approach (even in the coordinated case) might not efficiently operate in case of downstream bottleneck restrictions, e.g. a high demand off-ramp queue spillbacks in the freeway which blocks mainline lanes. Also, in case a freeway ends inside a congested city center, ramp metering might not be able to increase the outflow. In these cases, the freeway and urban network should be controlled in an integrated manner.

For urban networks, the Macroscopic Fundamental Diagram (MFD) aims at simplifying the urban traffic micro-modeling, where the collective traffic flow behaviors of subnetworks capture the main characteristics of traffic dynamics, such as the evolution of space-mean flows and densities in different regions of the network. The MFD provides a unimodal, low-scatter relationship between network vehicle density (veh/km) and network space-mean flow or outflow (veh/hr) for different network regions, if congestion is roughly homogeneous in the region. Alternatively, the MFD links *accumulation*, defined as the number of vehicles in the region, and *trip completion flow*, defined as the output flow of the region. Urban region flow or trip completion flow increases with accumulation up to a critical point, while additional vehicles in the network cause strong reductions in the flow. The physical model of MFD was initially proposed by [60] and observed with dynamic features in congested urban network in Yokohama by [55], and investigated using empirical or simulated data by [16, 89, 120, 32, 46, 202] and others. Earlier works had looked for MFD patterns in data from lightly congested real-world networks or in data from simulations with artificial routing rules and static demands (e.g. [116, 131] and others), but did not demonstrate that an invariant MFD with dynamic features can arise. Control strategies utilizing the concept of the MFD have been introduced for single-region cities in [28] and later a linear control approach applied for a micro-simulation environment by [97]. These strategies provide some useful insights towards system coordination, but might not operate in an efficient manner and might be far from optimal if congestion is heterogeneously distributed or if many trips have destinations outside the area of analysis, which is the case in many congested cities. Moreover, route guidance strategies with the utilization of MFD have been studied in [102] for grid networks.

In case of link density heterogeneity in an urban network, a possible solution to have a well-defined MFD is to partition the heterogenous network into a number of homogeneous smaller regions with small variance of link densities, see [90]. Chapter 4 introduced an elegant perimeter control approach to improve traffic conditions in an urban network which has been partitioned into two regions with well-defined MFDs (for stability analysis of the perimeter control see [66]). These results encourage us to utilize the MFD and the perimeter control approach for the mixed urban and freeway network.

All of the above approaches provide a first proof of concept that coordinated real-time control strategies with parsimonious models can create a new generation of smarter cities and improve

their mobility. But, still congestion governance in large-scale systems is currently fragmented and uncoordinated with respect to optimising the goals of travel efficiency and equity for multiple entities. Understanding these interactions for complex and congested cities is a big challenge, which will allow revisiting, redesigning, and integrating smarter traffic management approaches to generate cities more livable and sustainable. In this chapter, we follow the approach of "system of systems" (SoS), see e.g. [166], that aims to model and control a system consists of several independent heterogeneous systems. The mixed network is aligned with an SoS as it consists of an urban network and a freeway systems. First, the perimeter controllers are integrated with the ramp metering controllers to respectively manipulate the flow transfer between urban regions and the inflows from the urban network to the freeway.

Recent studies with empirical data, [51, 164], have shown that a freeway system might not be well-described with an MFD because of strong hysteresis phenomena. Thus, the traffic dynamics of the freeway are modeled in a more detailed approach according to the Asymmetric Cell Transmission Model (ACTM) in [61], which is a simplified version of the cell transmission model in [30]. Afterwards, the urban and freeway traffic dynamics are extended and integrated to formulate the traffic control problem of the mixed network. A Model Predictive Control (MPC) approach is proposed to minimize the total delay in the whole network. Our optimization findings highlight the importance of coordination of the freeway and urban control entities of a city when congestion is present.

In this chapter, there are two main contributions related to traffic modeling and control. In the modeling part, we present a new model that integrates the traffic dynamics of urban network and a freeway which is not a trivial task or straightforward. The integration of MFD model with the ACTM arises several unique challenges related to the mixed network as queue spillbacks from the on-ramps to the urban network, the allocation of off-ramp flows, and the route choice behavior as drivers can choose between the urban network and the freeway to travel. Note that utilizing the new macroscopic modeling eases the integration task, which is novel. In addition, this chapter contributes to the control area by addressing the traffic control problem for the mixed freeway and urban network following the control concept of SoS. This work is considered as one step forward and a rigorous contribution towards achieving SoS for transportation infrastructure and networks. In this chapter, it is also shown through a numerical example the importance of cooperation between different traffic control jurisdictions.

This chapter tackles an SoS with two interconnected urban regions and a freeway crossing with a couple of ramps in each region. More complex systems can be addressed with similar modeling and optimization principles if combined with efficient algorithms for partitioning the urban part in homogeneous regions with low scatter MFDs. Traffic control policies that minimize delays of the mixed network can affect the route choice of vehicles, which is the flow distribution between the urban network and the freeway, since for example vehicles can travel from the periphery to the city center by traveling either through the freeway or the urban network. This effect should not be ignored as restricting access to one region, might

transfer demand to another region of a city and change the distribution of congestion (e.g. very aggressive ramp metering might force people to utilize the urban network and vice versa). The route choice models can be classified into two types according to the behavioral supposition of drivers [187]: the User Equilibrium (UE) when users choose the best route with minimum cost and no user has an incentive to change his route; and the System Optimum (SO) when users cooperate with one another to minimize the total cost for the entire network. In uncongested networks, the UE route choice flow tends to the SO [143], though to reach the SO in congested networks, it is essential to manage the users' traveling behaviors, such as departure time, mode or route choice. In the mixed traffic network model, we first propose a dynamic simple route choice, which considers that travelers make choices with respect to the sequence of regions (and not a detailed sequence of links) that they will travel to reach their destination. Further we integrate the proposed MPC controller with an SO route choice model to explore the effect of route choice on the overall performance of the mixed network.

The outline of the chapter is as follows. In Sections 5.2 and 5.3, the mixed urban and freeway network modeling and the control problem formulation are introduced, respectively. Afterwards, various control policies for traffic control of the mixed network are elaborated in Section 5.4 and results of the control policies are presented in Section 5.5. The comparison between the dynamic simple route choice and SO route choice models is investigated in Section 5.6. Finally in the last section, conclusions are drawn and future work is summarized.

5.2 Traffic modeling of a mixed urban and freeway network

Consider a mixed urban and freeway network as schematically shown in Fig. 5.1, where it is assumed that the urban network is partitioned into two homogeneous urban regions, denoted by 1 and 2, having their own MFDs. In addition, there is a freeway, denoted by 3, as an alternative commuting route that passes through both urban regions having one on-ramp and one off-ramp within each region. The freeway also carries demand which is generated and finishes outside the two urban regions. Thus, a 3×3 time-dependent origin destination matrix is associated with the network demand.

To keep the traffic dynamic formulations elegant, the following assumptions are made regarding the trip routes in the mixed network: (A1) the freeway can be used at most once during the trip (exit and re-enter is not allowed), (A2) there is at most one urban region transfer during the trip, e.g. traveling from 1 to 2 then to 1 is not allowed. Under these assumptions, an origin-destination trip might have at most two trip routes since the traveler can choose between using the urban network or the freeway, or using both the urban network and the freeway (if this option exists), e.g. there are two trip routes from 1 to 1: $1 \rightarrow 1$ traveling from 1 to 1 using the urban region 1, and $1 \rightarrow 3 \rightarrow 1$ traveling within region 1 to the on-ramp 1, entering and travelling through the freeway, and then exiting from the off-ramp 1 and complete the trip in region 1. All origin-destination trip routes in the mixed network are summarized in Table 5.1. Note that from region 2 to 2 there is only one route through the urban region, as the

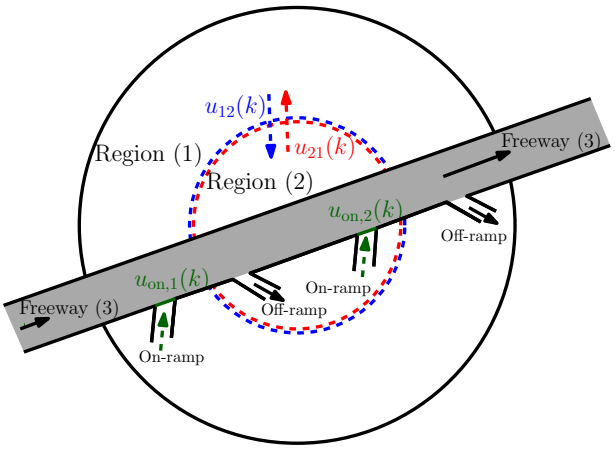


Figure 5.1: A mixed urban and freeway network: two regions 1 and 2 with two perimeter control inputs $u_{12}(k)$ and $u_{21}(k)$, and a freeway with two on-ramps and two off-ramps, and two on-ramp metering control inputs $u_{\text{on},1}(k)$ and $u_{\text{on},2}(k)$.

off-ramp in region 2 is assumed to be prior to its on-ramp, as shown in Fig. 5.1. Alternations of the aforementioned assumptions can be integrated in the methodology, but the formulation might be more tedious. Corresponding to the aforementioned O-D matrix, let $Q_{ij}(k)$ (veh/s) be a priori known demand generated in origin i to destination j at time step k, $i, j = 1,2,3$, $k = 0,1,\ldots,K-1$, which is distributed between two choices for the same origin-destination: $q_{ij}(k)$ (veh/s) denotes a generated demand in origin i with direct destination j at time step k that belongs to the trip route $i \to j$, while $q_{imj}(k)$ (veh/s), $m \neq i, j$, denotes a generated demand in origin i with destination j at time step k that belongs to the trip route $i \to m \to j$. All origin-destination demands are also summarized in Table 5.1. These choices depend on the development and propagation of congestion at different parts of the network. Errors in $Q_{ij}(k)$ are discussed later.

Table 5.1: Trip routes and demands in the mixed network.

O\ D	1	2	3
1	$q_{11}: 1 \to 1$ $q_{131}: 1 \to 3 \to 1$	$q_{12}: 1 \to 2$ $q_{132}: 1 \to 3 \to 2$	$q_{13}: 1 \to 3$ $q_{123}: 1 \to 2 \to 3$
2	$q_{21}: 2 \to 1$ $q_{231}: 2 \to 3 \to 1$	$q_{22}: 2 \to 2$	$q_{23}: 2 \to 3$ $q_{213}: 2 \to 1 \to 3$
3	$q_{31}: 3 \to 1$ $q_{321}: 3 \to 2 \to 1$	$q_{32}: 3 \to 2$ $q_{312}: 3 \to 1 \to 2$	$q_{33}: 3 \to 3$

In the following subsections, the traffic dynamics of the mixed freeway and urban network are elaborated. First, the traffic dynamics of the urban regions and the freeway are respectively introduced and modified according to the ACTM and MFD. A challenge is to integrate the two models (meso and macro), as there are transfer flows at the boundaries of freeway and urban network that depend on traffic conditions in both regions. Details of the dynamic route choice modeling between the freeway and the urban network are also discussed. Later, the entire mixed network traffic control problem will be presented in Section 5.3.

5.2.1 The urban traffic modeling

Corresponding to the aforementioned traffic demands, see Table 5.1, six accumulation states are introduced to model the dynamic equations of the urban network: $n_{ij}(k)$ (veh), $i = 1,2$; $j = 1,2,3$, where $n_{ij}(k)$ is the total number of vehicles in region i with next destination j at time step k. Let us denote $n_i(k)$ (veh) as the accumulation or the total number of vehicles in region i at time step k, i.e. $n_i(k) = \sum_{j=1}^{3} n_{ij}(k)$. The output of MFD, $G_i(n_i(k))$ (veh/s), represents the trip completion flow of region i for $n_i(k)$ which is the sum of transfer flows, i.e. external flows as trips from i with destination j, $i \neq j$, plus the internal flow as trips from i with destination i. The transfer flow from i to j is calculated corresponding to the proportion between accumulations, i.e. the external flow is $M_{ij}(k) = n_{ij}(k)/n_i(k) \cdot G_i(n_i(k))$, $i \neq j$, while the internal flow from i to i is calculated by $M_{ii}(k) = n_{ii}(k)/n_i(k) \cdot G_i(n_i(k))$. These relationships assume that trip length of *regional trips* (internal or external) are similar. For a description of different cases (which will not alter the methodology) the reader can refer to [52].

Simulation and empirical results in [55] show that the shape of MFD can be approximated by a non-symmetric unimodal curve skewed to the right, i.e. critical density that maximizes network flow is smaller than half of the jam density. Thus, we utilize a 3rd-order function of $n_i(k)$, e.g. $G_i(n_i(k)) = a_i \cdot n_i^3 + b_i \cdot n_i^2 + c_i \cdot n_i$, where a_i, b_i, c_i can be estimated from real data. Analytical ways to estimate MFDs as a function of topology and signal setting have been presented in [33, 75, 53]. The laws of mass conservation for the six urban state variables of the mixed network are derived in Section 5.3, see (5.18)–(5.23).

5.2.2 The freeway modeling

Given that an MFD cannot consistently describe the dynamics in a freeway, we consider the traffic dynamics of the freeway in the mixed network based on the ACTM. The mass conservation equations of the on-ramps are adjusted to fit the mixed network problem, as the input demands of the on-ramps are the output of the MFDs. Traditionally, the off-ramp flows are considered independent of the applied control and defined a priori. Given the interactions between the freeway and urban system, the freeway off-ramp flows are the input demand for the urban network, therefore, new equations are derived to split the off-ramp flows to different urban accumulation states. In the sequel, a brief description of the ACTM is presented, followed by mixed network modeling adjustments. The reader can refer to [61, 30] for a full description of ACTM. Note that the CTM (and consequently ACTM) does not reflect the capacity drop phenomenon during congestion. This issue can be addressed by integrating inverse λ fundamental diagrams in the CTM or by employing a higher order freeway model that can capture the capacity drop. Moreover, ACTM simplifies the CTM modeling of merging behavior by assuming independence between the on-ramp and mainline flows during merging, which is not fully consistent with the physics of traffic. While this simplification is applied to derive linear on-ramp control in [61], still with proper parameter calibration of ACTM

(parameter ξ_l) an accurate enough behavior of merging in both the mainline and the on-ramps can be observed, especially under congested conditions.

In the ACTM, the freeway is divided to L cells, where each cell l of the freeway contains at most one on- or one off-ramp. Five cells of the freeway are schematically shown in Fig. 5.2 as cell l has an on-ramp belonging to urban region i, and cell $l+2$ has an off-ramp. The number of vehicles in cell l at time step k, $k = 0, 1, \ldots, K-1$, is denoted by $x_l(k)$ (veh), while $f_l(k)$ (veh) is the number of vehicles moving from cell l to $l+1$ during time step k. The on-ramp is fed from $M_{i3}(k)$ (veh/s), the maximum output that can flow from region i of the urban network with destination to the freeway at time step k and calculated using the MFD. Let $n_{\text{on},i}(k)$ (veh) be the queue length of the on-ramp in region i at time step k, and $n_{\text{on},i,\max}$ (veh) be the maximum queue length of the on-ramp in region i. It is assumed that each cell l has a triangular fundamental diagram with the following parameters: $w_l \in [0,1]$ is the normalized congestion wave speed, $v_l \in [0,1]$ is the normalized free-flow speed, $x_{l,\max}$ (veh/lane) is the jam accumulation, and $\bar{f}_l$ (veh/hr/lane) is the mainline capacity.

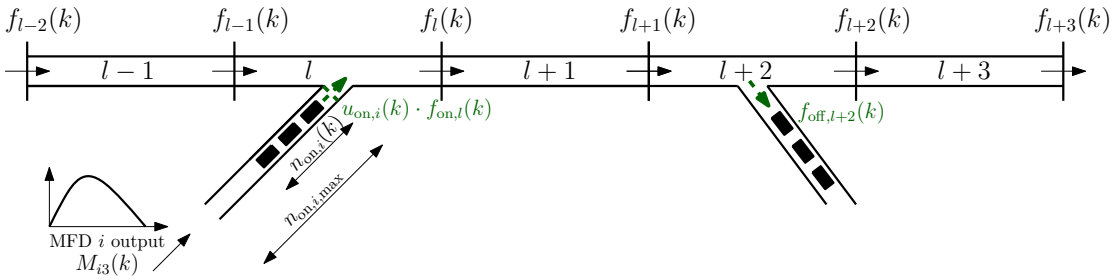


Figure 5.2: Representative freeway cells in the ACTM. Cell l has an on-ramp belonging to region i, while cell $l+2$ has an off-ramp. The on-ramp is fed from a demand generated in urban region i and calculated from the MFD, and flows exit at the off-ramp continuing their trip to reach their final destinations.

The freeway is integrated with the urban network through its on- and off-ramps. In the following, we first determine the on-ramp freeway flow taking into account the MFD output, and afterwards the off-ramp flows are split as inputs to different urban accumulations.

The unmetered on-ramp flow $f_{\text{on},l}(k)$ (veh) is the number of vehicles that can enter cell l from its on-ramp during time step k. It is calculated as follows

$$f_{\text{on},l}(k) = \min\left[n_{\text{on},i}(k) + M_{i3}(k) \cdot T_k,\, \xi_l \cdot (x_{l,\max} - x_l(k)),\, s_{\text{on},i} \cdot T_k\right], \tag{5.1}$$

where i is the region that the on-ramp belongs to, T_k (s) is the time step size, $s_{\text{on},i}$ (veh/s) is the maximum number of vehicles that can enter the freeway from the on-ramp belonging to region i, and ξ_l $(-) \in [0,1]$ is the on-ramp flow allocation parameter, see [61]. The on-ramp metering control inputs, denoted by $u_{\text{on},i}(k)$ $(-)$, $i = 1,2$, are introduced on the entrance of the freeway to control the flow entering from the region i to the freeway, see Fig. 5.1. The queue dynamic for the on-ramp belonging to region i with ramp metering control input, $u_{\text{on},i}(k)$,

taking into the account the on-ramp maximum queue length, is as follows

$$n_{\text{on},i}(k+1) = \min\big(n_{\text{on},i}(k) + M_{i3}(k)\cdot T_k - u_{\text{on},i}(k)\cdot f_{\text{on},l}(k),\, n_{\text{on},i,\max}\big). \tag{5.2}$$

The mainline flow in the freeway is calculated as follows

$$\begin{aligned} f_l(k) = \min\big[&(1-\beta_l(k))\cdot v_l\cdot(x_l(k)+\gamma\cdot u_{\text{on},i}(k)\cdot f_{\text{on},l}(k)), F_l(k),\\ &w_{l+1}\cdot(\bar{x}_{l+1} - x_{l+1}(k) - \gamma\cdot u_{\text{on},i}(k)\cdot f_{\text{on},l+1}(k))\big], \end{aligned} \tag{5.3}$$

where $\beta_l(k)$ $(-)$ is the split ratio for the off-ramp (if exists) in cell l, γ $(-) \in [0,1]$ is the on-ramp (if exists) flow blending coefficient, and $F_l(k) \triangleq \min\{\bar{f}_l, (1-\beta_l(k))/\beta_l(k)\cdot\bar{f}_{\text{off},l}\}$, where $\bar{f}_{\text{off},l}$ (veh) is the off-ramp capacity. The exit flow of the off-ramp in cell l, $f_{\text{off},l}(k)$ (veh), is determined as follows

$$f_{\text{off},l}(k) = \frac{\beta_l(k)}{1-\beta_l(k)}\cdot f_l(k). \tag{5.4}$$

Finally, the mainline mass conservation is

$$x_l(k+1) = x_l(k) + f_{l-1}(k) + u_{\text{on},i}(k)\cdot f_{\text{on},l}(k) - f_l(k) - f_{\text{off},l}(k), \tag{5.5}$$

for $l = 1,2,\ldots,L$, and $k = 0,1,\ldots,K-1$, where $f_{\text{on},l}(k) = 0$ and/or $f_{\text{off},l}(k) = 0$ if cell l does not contain an on-ramp and/or an off-ramp, respectively.

The ACTM does not keep track of the origin and the destination of vehicles as off-ramps are usually the final destinations of the trip. However, in the mixed network vehicles exiting from the off-ramps can have various destinations, e.g. finishing their trip in the urban region connected to the off-ramp or continuing their trip to the other urban region. To calculate the off-ramp flow distribution, we assume that the split ratios are similar to the instantaneous corresponding O-D demands, i.e. the off-ramp exit flows denoted by "hat" variables are calculated as follows

$$[\hat{q}_{31}(k), \hat{q}_{312}(k), \hat{q}_{231}(k)] = f_{\text{off},1}(k)\cdot\frac{[q_{31}(k)+q_{131}(k), q_{312}(k), q_{231}(k)]}{q_{31}(k)+q_{131}(k)+q_{312}(k)+q_{231}(k)}, \tag{5.6}$$

$$[\hat{q}_{32}(k), \hat{q}_{321}(k), \hat{q}_{132}(k)] = f_{\text{off},2}(k)\cdot\frac{[q_{32}(k), q_{321}(k), q_{132}(k)]}{q_{32}(k)+q_{321}(k)+q_{132}(k)}. \tag{5.7}$$

The above "hat" demands will be further integrated in the dynamic equations of the mixed network in Section 5.3.1. A more accurate estimation would require that in each cell we keep track of the final destination of each vehicle. Moreover, (5.6) and (5.7) are precise when O-D patterns are not varying rapidly. If O-D patterns change significantly the off-ramp flow distribution can be derived according to the experienced corresponding O-D demands, i.e. the current split ratios at time step k can be associated with the O-D demands of time step k minus mean of O-D experienced travel time. However, both enhancements would make ACTM dynamics more tedious. These can be a future direction of research.

5.2.3 The route choice modeling

Real-time control can affect the route choices of users. To be able to integrate the modeling framework in a real-time control environment, a parsimonious model of route choice has to be developed. Instead of considering a detailed route choice, a path choice is modeled as a sequence of regions during a trip, as expressed in Table 5.1. Nevertheless, given that MFD (or ACTM) modeling assumes that all vehicles traveling at a region (or cell) at time step k have equal speeds, we decide to add stochasticity in the urban region's trip length to make the model more realistic (otherwise all or nothing choices will be made). By assuming that all origins and destinations are homogeneously distributed in each of the urban regions and that the space is continuous, the Trip Length Distributions (TLDs) within each urban region for different trips can be estimated either in an analytical or experimental way. The TLD between x and y is the distribution of the distance between a random point in x and a random point in y. The possible pairs of (x,y) according to the trip routes of Table 5.1 are: (I) (i,i) two random points inside urban region i, (II) $(i,\partial(ij))$ a random point inside urban region i and the nearest point on the border between regions i and j, (III) (i,on_i) a random point inside urban region i and the on-ramp of urban region i, (IV) $(\partial(ij),\text{on}_j)$ the border between urban regions i and j and the on-ramp of region j, and (V) $(\text{off}_i,\partial(ij))$ the off-ramp of urban region i and the nearest point on the border between regions i and j. If GPS data are available for a fraction of the vehicles these TLDs can be estimated more accurately. Let us denote $TTD^{\text{u}}_{x,y}(k)$ the Travel Time Distribution (TTD) in the urban region for pairs (x,y) at step k (the continuous time t is divided to K time steps, as each time step k has the time duration of T_k (s)). Then, the TTDs are estimated as follows

$$TTD^{\text{u}}_{x,y}(k) = \frac{TLD^{\text{u}}_{x,y}}{V_i(k)}, \tag{5.8}$$

where $V_i(k)$ (m/s) is the average speed in urban region i at k, which is calculated according to the MFD of region i as follows

$$V_i(k) = \frac{G_i(n_i(k)) \cdot \bar{l}_i}{n_i(k)}, \tag{5.9}$$

where $G_i(n_i(k))$ is the MFD value of region i at time step k, $n_i(k)$ is the accumulation of region i at time step k, and $\bar{l}_i$ (m) is the average region trip length (which can be estimated from real data, see [55]). Note that trip length within a region l_i, is considered constant, but as a regional route choice is embedded in the model, trip length between origin i and destination j varies with the traffic conditions.

Let $TT^{\text{f}}_{l_{\text{s}},l_{\text{e}}}(k)$ (s) be the instantaneous travel time along the freeway at time step k from cell l_{s} to cell l_{e} (note that both cells take values between 1 and L, and that the mainline freeway cell associated with on- or off-ramp i is denoted as on_i or off_i for simplicity), and $TT^{\text{r}}_{\text{on}_i}(k)$ (or $TT^{\text{r}}_{\text{off}_i}(k)$) (s) be the travel time in on-ramp i (or off-ramp i) at time step k. Then, $TT^{\text{r}}_{\text{on}_i}(k)$ is

approximated by

$$TT^{\mathrm{r}}_{\mathrm{on}_i}(k) = \frac{n_{\mathrm{on},i}(k) \cdot T_k}{u_{\mathrm{on},i}(k) \cdot f_{\mathrm{on},l}(k)}, \tag{5.10}$$

where $n_{\mathrm{on},i}(k)$ (veh) is the queue length of on-ramp i at time step k, $u_{\mathrm{on},i}(k)$ $(-)$ is the ramp metering control value ($0 \le u_{\mathrm{on},i}(k) \le 1$) at k that constraints $f_{\mathrm{on},l}(k)$ (veh), i.e. the unmetered outflow of on-ramp i at k that enters to its connected freeway cell l, see (5.3) and Fig. 5.2. Further, the travel time of off-ramp i at k, $TT^{\mathrm{r}}_{\mathrm{off}_i}(k)$ (s), is assumed to be equal to the free flow travel time (note that if this assumption is relaxed then one needs to keep track of the destination of vehicles in order to model the traffic state of the off-ramps, which is not the case in ACTM. Keeping track of the origin and the destination of vehicles hinders the macroscopic modeling of freeways).

To have a dynamic route choice modeling, one should split the a priori known $Q_{ij}(k)$ O-D demand between possible alternative trip routes. Let $\theta_{ij}(k)$ $(-)$ be the route choice proportion for direct trip between origin i and destination j, $\theta_{ij}(k) \in [0,1]$, where $q_{ij}(k) = \theta_{ij}(k) \cdot Q_{ij}(k)$ and $q_{imj}(k) = (1-\theta_{ij}(k)) \cdot Q_{ij}(k)$ if $q_{imj}(k)$ exists, otherwise $q_{ij}(k) = Q_{ij}(k)$. e.g. $q_{11}(k) = \theta_{11}(k) \cdot Q_{11}(k)$, and $q_{131}(k) = (1-\theta_{11}(k)) \cdot Q_{11}(k)$, while $q_{22}(k) = Q_{22}(k)$. The value of $\theta_{ij}(k)$ is calculated dynamically according to the traffic conditions in the urban regions and the freeway, based on the estimation of the trip route travel times at time step k. Note that there is no need for route choice calculations of $2 \to 2$ and $3 \to 3$ O-D pairs, since they only have a single trip route choice.

Now let us consider an O-D from i to j with trip routes $i \to j$ and $i \to m \to j$. Recall that $TTD^{\mathrm{u}}_{x,y}$ denotes the travel time distributions in the urban regions between x and y, $TT^{\mathrm{f}}_{l_s,l_e}$ denotes the total travel time along the freeway from cell l_s to cell l_e, and $TT^{\mathrm{r}}_{\mathrm{on}_i}$ (or $TT^{\mathrm{r}}_{\mathrm{off}_i}$) denotes the travel time in on-ramp i (or off-ramp i). Note that the travel time variability is only considered in the urban regions. The two trip route Travel Time Distributions TTD_{ij} and TTD_{imj} at time step k are calculated as follows

$$TTD_{ii}(k) = TTD^{\mathrm{u}}_{i,i}(k) \qquad i = 1 \tag{5.11}$$

$$TTD_{ij}(k) = \begin{cases} TTD^{\mathrm{u}}_{i,\partial(ij)}(k) + TTD^{\mathrm{u}}_{\partial(ij),j}(k) & \text{if } ij \in \{12,21\}, \\ TTD^{\mathrm{u}}_{i,\mathrm{on}_i}(k) + TT^{\mathrm{r}}_{\mathrm{on}_i}(k) + TT^{\mathrm{f}}_{\mathrm{on}_i,L}(k) & \text{if } ij \in \{13,23\}, \\ TT^{\mathrm{f}}_{1,\mathrm{off}_j}(k) + TT^{\mathrm{r}}_{\mathrm{off}_j}(k) & \text{if } ij \in \{31,32\}, \end{cases} \tag{5.12}$$

$$TTD_{imj}(k) = \begin{cases} TTD^{\mathrm{u}}_{i,\mathrm{on}_i}(k) + TT^{\mathrm{r}}_{\mathrm{on}_i}(k) + TT^{\mathrm{f}}_{\mathrm{on}_i,\mathrm{off}_j}(k) + TT^{\mathrm{r}}_{\mathrm{off}_j}(k) & \\ \qquad\qquad \text{if } imj \in \{131,132,231\}, & \\ TTD^{\mathrm{u}}_{i,\partial(im)}(k) + TTD^{\mathrm{u}}_{\partial(im),\mathrm{on}_m}(k) + TT^{\mathrm{r}}_{\mathrm{on}_m}(k) + TT^{\mathrm{f}}_{\mathrm{on}_m,L}(k) & \\ \qquad\qquad \text{if } imj \in \{123,213\}, & \\ TT^{\mathrm{f}}_{1,\mathrm{off}_m}(k) + TT^{\mathrm{r}}_{\mathrm{off}_m}(k) + TTD^{\mathrm{u}}_{\mathrm{off}_m,\partial(mj)}(k) + TTD^{\mathrm{u}}_{\partial(mj),j}(k) & \\ \qquad\qquad \text{if } imj \in \{321,312\}. & \end{cases} \tag{5.13}$$

Ultimately, given the distributions of travel time of both alternative trip routes, one can calculate $\theta_{ij}(k)$ as follows:

$$\theta_{ij}(k) = \Pr(TTD_{ij}(k) - TTD_{imj}(k) < 0). \tag{5.14}$$

Equation (5.14) determines θ_{ij} as the proportion of vehicles of O-D from i to j, whose travel time of trip route $i \to j$ is less than trip route $i \to m \to j$. Equation (5.14) requires first, to reflect the $TTD_{imj}(k)$ about vertical axis with zero travel time because of the negative sign, then convolve it with $TTD_{ij}(k)$. This procedure reads

$$\big(TTD_{ij}(k) - TTD_{imj}(k)\big)(\tau) = \int_{-\infty}^{\infty} (TTD_{ij}(k))(\omega) \cdot (TTD_{imj}(k))(-\tau + \omega) \mathrm{d}\omega \tag{5.15}$$

$$\theta_{ij}(k) = \int_{-\infty}^{0} \big(TTD_{ij}(k) - TTD_{imj}(k)\big)(\tau) \mathrm{d}\tau. \tag{5.16}$$

Finally, the probability of $(TTD_{ij}(k) - TTD_{imj}(k)) < 0$ is the θ_{ij}. The above equations assume independence between the values of the two distributions, which can be relaxed if real data is available for TLD estimation. An example of such procedure is shown in Fig. 5.3(b) for O-D from region 1 to 1 at $k = 0$. This dynamic simple route choice model assumes that drivers (users) have the real-time information to select their trip route, though this real-time information is based on the current condition of the network, see (5.8)–(5.16). It is worth to mention that more detailed dynamic traffic assignment strategies can also be proposed in a way to keep the control problem tractable from an optimization point of view. Following up (5.16) the trip length between origin i and destination j (given that there are two possible routes for every pair) can be estimated as $\bar{l}_{ij} = \theta_{ij}(k) \cdot E\{TLD_{ij}(k)\} + (1 - \theta_{ij}(k)) \cdot E\{TLD_{imj}(k)\}$.

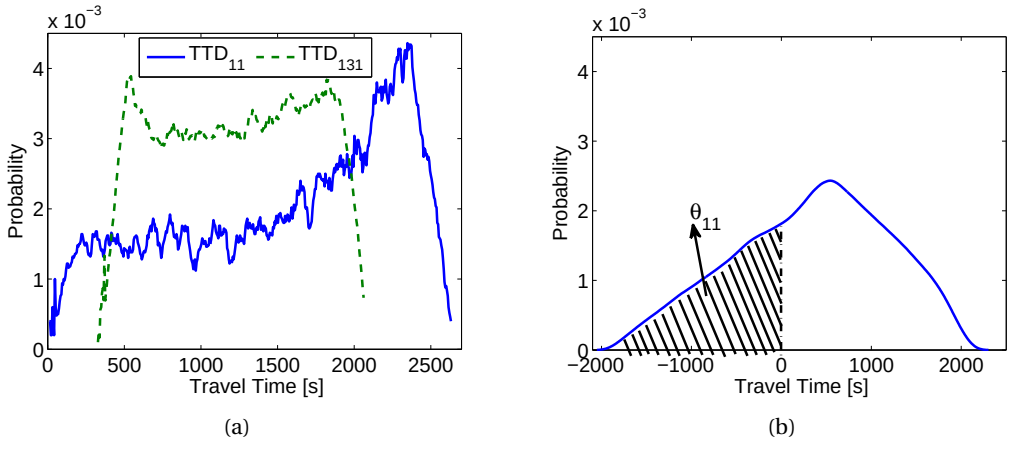


Figure 5.3: An example of route choice procedure for origin 1 to destination 1: (a) the TTD of both trip routes at $k = 0$, (b) the probability distribution function of $(TTD_{11}(0) - TTD_{131}(0))$. The hatched area is $\theta_{11}(0) = 0.35$.

5.3 The mixed network control problem

5.3.1 Problem formulation

Given the traffic dynamics of the previous section, we can now formulate the optimization problem. In the mixed network control problem, there are two types of controllers to minimize the network total delay: the perimeter controllers for the urban regions and the on-ramp meterings for the freeway. The perimeter control inputs denoted by $u_{12}(k)$ and $u_{21}(k)$ $(-)$ are on the border between the two urban regions as shown in Fig. 5.1, to control the urban external transfer flows. Since the perimeter controllers exist only on the border of the two regions, the internal flows cannot be controlled or restricted, while the external transfer flows are controlled such that, only a proportion of inter-region demand flows can pass the perimeter at time step k. i.e. $u_{12}(k)$ and $u_{21}(k)$, where $0 \leq u_{12}(k), u_{21}(k) \leq 1$, are the proportion of the transfer flows that goes from region 1 to 2 and region 2 to 1 at time step k, respectively. Likewise, the on-ramp metering control inputs, $u_{\text{on},i}(k)$, $i = 1,2$, are on the entrance of the freeway to control the flow entering the freeway, such that a proportion of demand input from the urban region can enter into the freeway, i.e. $0 \leq u_{\text{on},i}(k) \leq 1$, $i = 1,2$.

In the traffic dynamics of the mixed network, there are several types of state variables enumerated as: six state variables describing the dynamics of the urban regions, two state variables describing the queue dynamics of the freeway on-ramps, and L state variables describing accumulation of the freeway cells. Therefore, the mixed traffic network control problem is formulated as follows:

$$J = \min_{\substack{u_{12}(k),\, u_{21}(k),\\ u_{\text{on},1}(k),\, u_{\text{on},2}(k);\\ \text{for } k=0,\ldots,K-1}} T_k \cdot \Big[\sum_{k=0}^{K-1} \big(n_1(k) + n_2(k)\big) + \Big[\sum_{k=0}^{K-1} \sum_{l=1}^{L} x_l(k) + \sum_{k=0}^{K-1} \sum_{i=1}^{2} n_{\text{on},i}(k) \Big] \Big] \tag{5.17}$$

subject to

$$\begin{aligned} n_{11}(k+1) = n_{11}(k) + T_k \cdot \Big[& \frac{\hat{q}_{321}(k) + q_{21}(k)}{\hat{q}_{321}(k) + q_{213}(k) + q_{21}(k)} \cdot u_{21}(k) \cdot M_{21}(k) + q_{11}(k) \\ & + \hat{q}_{231}(k) + \hat{q}_{31}(k) - M_{11}(k) \Big] \end{aligned} \tag{5.18}$$

$$n_{12}(k+1) = n_{12}(k) + T_k \cdot \Big[q_{12}(k) + q_{123}(k) + \hat{q}_{312}(k) - u_{12}(k) \cdot M_{12}(k) \Big] \tag{5.19}$$

$$\begin{aligned} n_{13}(k+1) = n_{13}(k) + T_k \cdot \Big[& \frac{q_{213}(k)}{\hat{q}_{321}(k) + q_{213}(k) + q_{21}(k)} \cdot u_{21}(k) \cdot M_{21}(k) + q_{13}(k) \\ & + q_{131}(k) + q_{132}(k) - \min(M_{13}(k), C_{\text{on},1}(k)) \Big] \end{aligned} \tag{5.20}$$

$$n_{21}(k+1) = n_{21}(k) + T_k \cdot \Big[q_{21}(k) + q_{213}(k) + \hat{q}_{321}(k) - u_{21}(k) \cdot M_{21}(k) \Big] \tag{5.21}$$

$$\begin{aligned} n_{22}(k+1) = n_{22}(k) + T_k \cdot \Big[& \frac{q_{12}(k) + \hat{q}_{312}(k)}{q_{12}(k) + q_{123}(k) + \hat{q}_{312}(k)} \cdot u_{12}(k) \cdot M_{12}(k) + q_{22}(k) \\ & + \hat{q}_{132}(k) + \hat{q}_{32}(k) - M_{22}(k) \Big] \end{aligned} \tag{5.22}$$

$$\begin{aligned} n_{23}(k+1) = n_{23}(k) + T_k \cdot \Big[& \frac{q_{123}(k)}{q_{12}(k) + q_{123}(k) + \hat{q}_{321}(k)} \cdot u_{12}(k) \cdot M_{12}(k) + q_{23}(k) \\ & + q_{231}(k) - \min(M_{23}(k), C_{\text{on},2}(k)) \Big] \end{aligned} \tag{5.23}$$

$$0 \le \sum_{j=1}^{3} n_{ij}(k) \le n_{i,\text{jam}} \qquad i = 1,2 \tag{5.24}$$

$$u_{\min} \le u_{ij}(k) \le u_{\max} \qquad i = 1,2;\ j = 3-i \tag{5.25}$$

$$u_{\min} \le u_{\text{on},i}(k) \le u_{\max} \qquad i = 1,2 \tag{5.26}$$

$$n_{ij}(0) = n_{ij,0} \qquad i = 1,2;\ j = 1,2,3 \tag{5.27}$$

$$x_l(0) = x_{l,0} \qquad l = 1,2,\ldots,L \tag{5.28}$$

$$n_{\text{on},i}(0) = n_{\text{on},i,0} \qquad i = 1,2 \tag{5.29}$$

and (5.1) – (5.16)

for $k = 0,1,2,\ldots K-1$, where $n_{ij,0}$, $n_{\text{on},i,0}$, and $x_{l,0}$ are the initial accumulations of the urban states, on-ramps, and freeway cells at $k = 0$; $n_{i,\text{jam}}$ (veh) is the jam accumulation of regions i; $u_{\min}$ and $u_{\max}$ are the lower and upper bounds for perimeter and ramp metering controllers; and $C_{\text{on},i}(k)$ (veh/s) is the available flow storage capacity in the on-ramp queue, i.e. $C_{\text{on},i}(k) = (n_{\text{on},i,\max} - n_{\text{on},i}(k))/T_k$.

The term $\min(M_{i3}(k), C_{\text{on},i}(k))$, $i = 1,2$, in (5.20) and (5.23) represents the flow of vehicles that leaves the urban region to enter the freeway, which is the minimum of outgoing flow and the available flow capacity in on-ramp i. Moreover, the term $u_{ij}(k) \cdot M_{ij}(k)$, $i = 1,2; j = 3-i$, expresses the outflow from region i to region j, which comprises of two parts; (i) the part that is from origin i to destination j, and (ii) the other part that is from origin i to destination 3 through region j. Thus to distinguish between these two parts, the urban inter-flow, $u_{ij}(k) \cdot M_{ij}(k)$, are divided accordingly to the ratio of demand flow. e.g. $(q_{12}(k) + \hat{q}_{312}(k))/(q_{12}(k) + q_{123}(k) + \hat{q}_{312}(k)) \cdot u_{12}(k) \cdot M_{12}(k)$ and $q_{123}(k)/(q_{12}(k) + q_{123}(k) + \hat{q}_{321}(k)) \cdot u_{12}(k) \cdot M_{12}(k)$ respectively represent the outflow from origin 1 to destination 2 and outflow from origin 1 to destination 3 through region 2. Note that, equations (5.18)–(5.23) are the conservation of mass laws for urban accumulations, $n_{ij}(k)$, while (5.24) is the lower and upper bound constraints on region i accumulation. Recall that $M_{ij}(k) = n_{ij}(k)/n_i(k) \cdot G_i(n_i(k))$, $i = 1,2; j = 1,2,3$, the equations (5.8)–(5.16) derive the $q_{ij}(k)$ and $q_{imj}(k)$ based on the proposed dynamic route choice modeling, and (5.1)–(5.7) describe the freeway dynamics. If one compares (5.18)–(5.23) for MFD dynamics with (5.1)–(5.3) for ACTM dynamics it might observe some inconsistency in the mass conservation, as the MFD equations do not directly consider downstream restrictions (while ACTM does), but apply an exit function. One could add one

more term, the boundary capacity, which is a function of the receiving region accumulation. Nevertheless, this constraint can be ignored during the optimization process. The physical reasoning besides this assumption is that (i) boundary capacity decreases for accumulations much larger than the critical accumulation (see [54]), and (ii) the control inputs will not allow the system to get close to gridlock.

5.3.2 Solution approach – an MPC controller

The optimal control problem for the mixed network is solved by the MPC approach. The MPC obtains optimal solutions with feedback control for dynamic systems. It can tackle errors between the model and the plant (reality) by utilizing a feedback monitored-information. In our problem, this is a crucial issue due to the scatter in the MFD, mainly in the congested regime, as errors are expected between the MFD model and the plant. In addition, noise in the traffic demand is expected between predicted and actual demand.

The MPC has been used for optimization in different traffic control problems, e.g. ramp metering of freeway networks in [11], variable speed limits and route guidance for freeway networks in [103], [74], [73], signal control for large-scale urban networks in [45], [3], [111], and mixed urban and freeway networks in [178]. A historical survey for industrial applications (other than traffic control) of MPC can be found in [145], while theoretical issues of MPC can be found in [44], [18], [119].

For the sake of brevity, the full description of the solution method is not presented here, however, the reader can refer to Chapter 4 and [49] for further information. In the following, we present the outlines of the mixed network MPC through the block diagrams presented in Fig. 5.4.

On the top of Fig. 5.4, the “mixed urban-freeway plant” block presents the dynamic evolution of the mixed network in reality. This block can be implemented in several ways: (I) real field implementations, e.g. one can apply the proposed scheme on Yokohama network that experiences a well-defined MFD (as was found in [55]) or [16], (II) simulation-based plant, e.g. using micro-simulations of the San-Francisco business district center presented in [54, 89] or [97], and (III) model-based plant, e.g. [49, 28]. In this chapter, we follow the later to present the reality of mixed networks as the model (5.8)–(5.29) is utilized but with integrating noise in demand and errors in MFDs, similar to equations (23)–(27) in [49], such that the dynamics evolutions of the plant and the model are not the same. Demand noise is considered fully random, while errors in the MFD have some heteroscedastic behavior, i.e. the error is expected to be larger when accumulation grows.

One of the main features of the MPC framework is the “rolling time horizon” (or receding horizon), see right of Fig. 5.4. The optimization technique is applied to finite time duration, called prediction horizon, which is much smaller than the total control process time. The control process time is covered by overlapping several time horizons, where at each time step

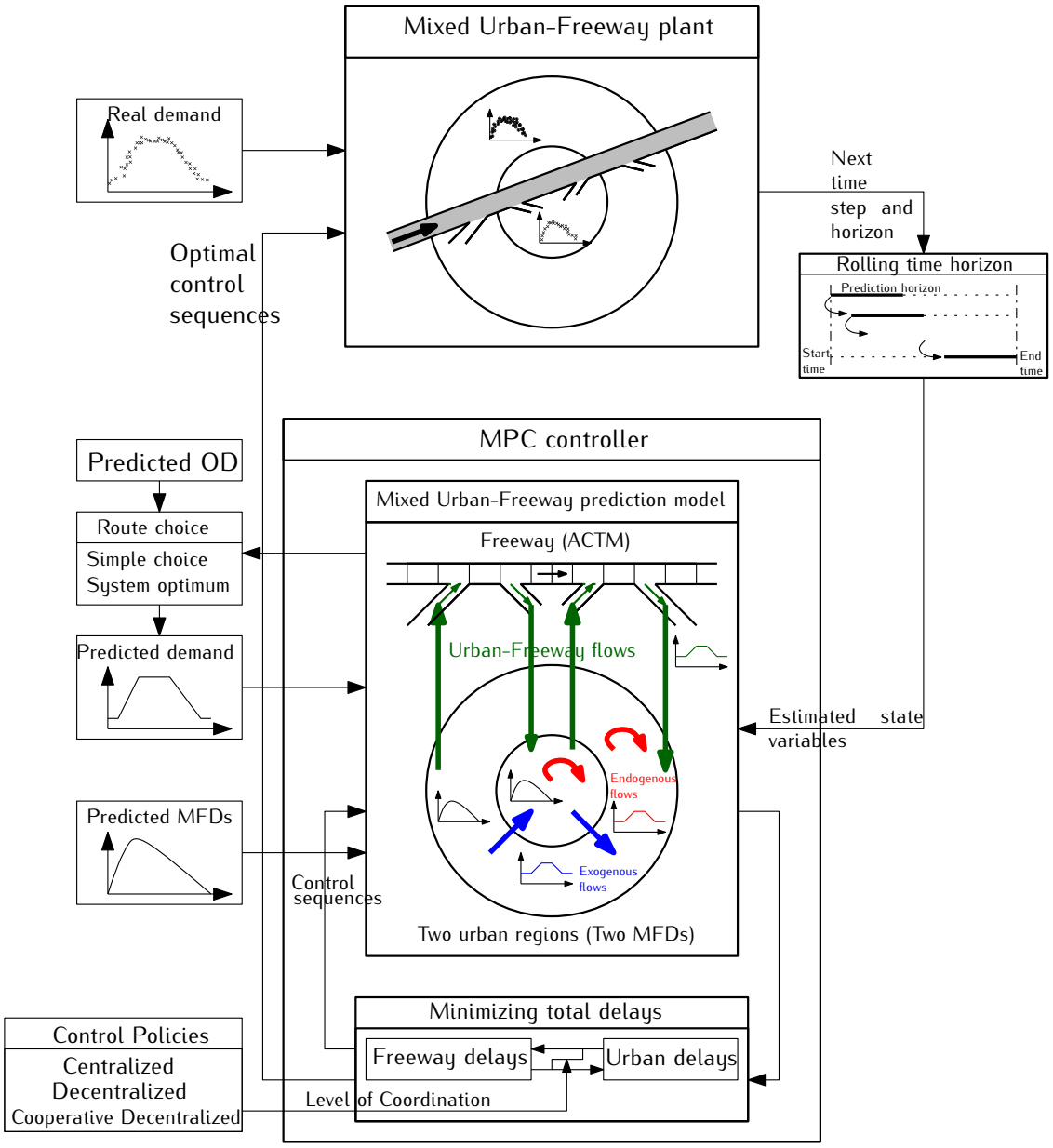


Figure 5.4: Model predictive control scheme for mixed urban-freeway networks.

the open-loop problem of the mixed urban-freeway model is optimized. Then, only the first control sequence is applied to the plant and the same procedure is carried out again till the final time of the problem.

The "MPC controller" (bottom of the figure) contains: (I) the mixed urban-freeway model (5.8)–(5.29) which is used to predict the dynamics of the mixed network, and (II) the optimization technique that minimizes the total delays of the network according to (5.17). The mixed urban-freeway model integrates the ACTM of the freeway and the two-region MFDs model of the urban network through the traffic flows (green arrows in the figure) according to (5.2), (5.6), and (5.7).

The MPC controller obtains the optimal control sequences for the current horizon by solving an optimization problem using the direct sequential method, also referred to single-shooting or control vector parametrization (CVP) in the literature, see [49]. Note that several "control policie", see left-bottom of the figure, which will be introduced in Section 5.4, determine the level of coordination between the two control entities, and the control structure, e.g. centralized or decentralized.

The predicted O-D is given in advance, however, route choice demands are determined corresponding to the chosen strategy of route choice, i.e. a simple route choice (5.8)–(5.16) or system optimum as described in Section 5.6.

5.3.3 Test study example

In this subsection, we present a case study example to investigate the characteristics of the proposed MPC controller. This MPC controller is fully centralized and the comparison with other control policies such as decentralized and cooperative decentralized will be discussed later in Section 5.5. Note that the simple route choice model is integrated in the mixed network dynamics, while an SO route choice model will be integrated later in Section 5.6. In this example, the time varying demand is simulating one hour of morning peak followed by half an hour of low demand. Furthermore, the O-D demands, which are assumed to have trapezoidal shapes, are perturbed by a normal random component (see perturbed demand lines in Fig. 5.5(a)), in order to investigate the robustness of the MPC controller to the accuracy of demand prediction. The MPC controller tries to find the optimal control sequences assuming that the demands are without error (trapezoidal shapes) in the optimization model, while the demands are applied to the plant, i.e. the mixed traffic network in reality, with errors (perturbed lines). Without loss of generality, we assume both regions have the same MFD consistent with the MFD observed in Yokohama, see Fig. 2 in [55]. While the MPC assumes the MFD as aforementioned 3rd-order function (see the two coincide lines in Fig. 5.5(b)), the MFDs of the MPC plant are perturbed by a random component (see red and blue points in Fig. 5.5(b)), in order to both capture the scatter behavior of MFD and test the robustness of the proposed MPC controller to the modeling error of MFD.

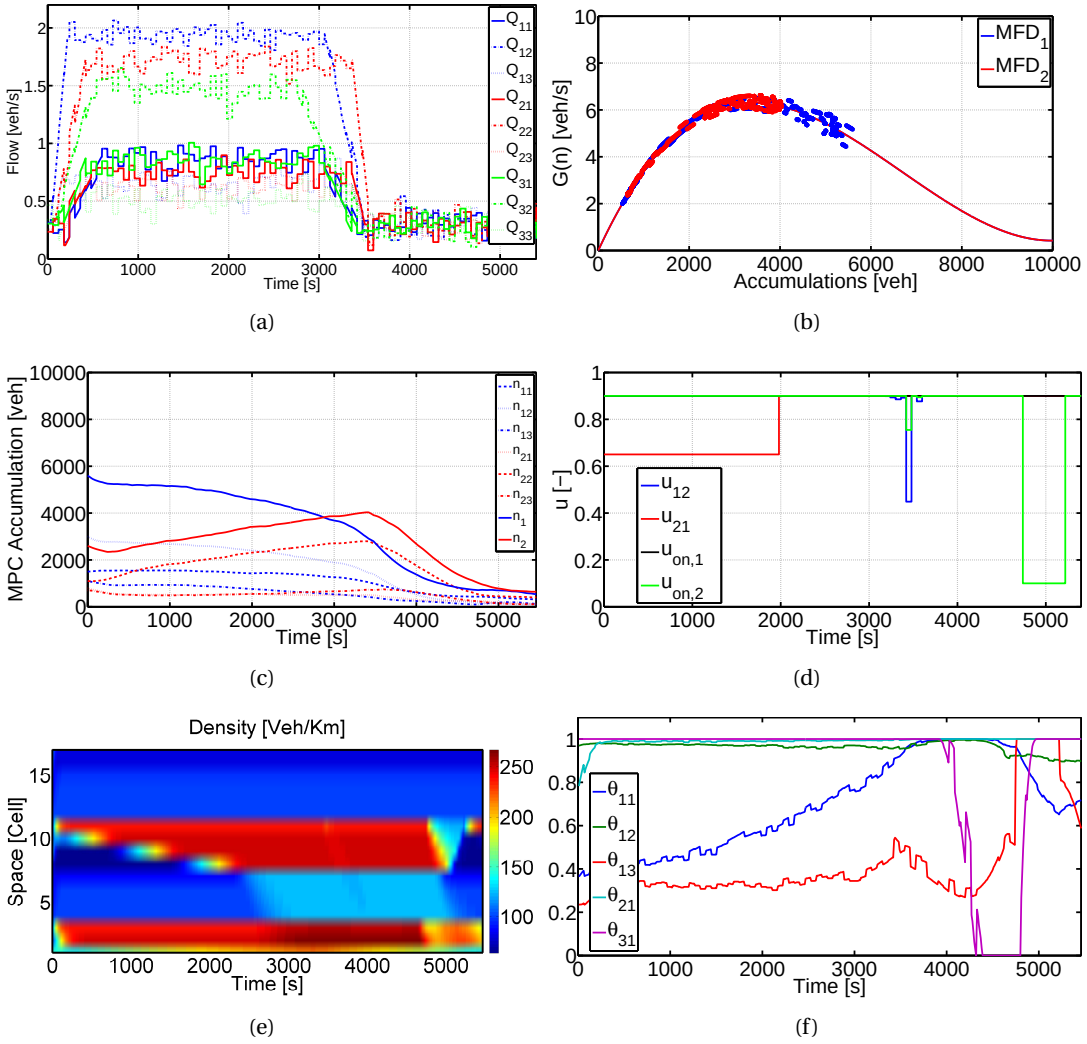


Figure 5.5: The results of case study example for fully coordinated: (a) O-D demands, (b) urban region MFDs, (c) urban accumulations, (d) MPC control sequences, (e) freeway density contour, and (f) evolution of route choice parameters.

The selected MPC controller parameters are: the prediction horizon $N_p = 20$, the control horizon $N_c = 2$, the control lower bound $u_{min} = 0.1$, and the upper bound $u_{max} = 0.9$. Furthermore, the freeway consists of 17 cells each has length of 0.5 (km) except the first cell which is long enough to accommodate all the vehicles in the entrance queue of the network. The parameters of the triangular fundamental diagram of the freeway cells are: the jam accumulation $x_{l,max} = 125$ (veh/km/lane), the mainline capacity $\bar{f}_l = 2200$ (veh/hr/lane), and the free flow speed equals to 88.5 (km/hr). Other setups of the simulation are: the split ratios of off-ramps, $\beta_1 = 0.15$ and $\beta_2 = 0.25$ (−), the maximum queue length of on-ramps, $n_{on,i,max} = 300$ (veh), and the on-ramps capacity flow, $s_{on,i} = 6000$ (veh/hr). The values of the last two parameters might be considered high for a single on-ramp. Indeed, they represent a hyper-ramp per urban region (i.e. all ramps are grouped in one, by considering similar characteristics and

queue lengths). Coordinated ramp metering strategies, e.g. [138, 57], can be utilized to equally distribute the queues among the ramps. The reasoning behind this is that if multiple ramps are considered within each region, then a ramp choice model should be added to the developed formulation. More complex city structures with more urban regions and ramps is a future research priority.

In this numerical example, regions 1 and 2 are initially congested and uncongested, respectively, i.e. the initial accumulations are $n_1(0) = 5600$ (veh) and $n_2(0) = 2600$ (veh), and region 2 as the central business district attracts most of the trips. The evolution of accumulations $n_{ij}(k)$ over 1.5 hours of simulation are presented in Fig. 5.5(c). Note that at the beginning of the control process, the MPC controller decreases the total accumulation in region 1, $n_1(k)$, by choosing $u_{21}(k) = 0.65$ to restrict the flow from region 2 to region 1. Afterwards, the MPC controller tries to keep the both accumulations uncongested by changing u_{21} to $u_{\max}$ at $t = 2000$ (s) (see Fig. 5.5(d)) to let more vehicles enter region 1.

The effect of on-ramp controllers on the condition of freeway is more comprehensible with the help of Fig. 5.5(e), where the density contour of freeway is illustrated. The main points of interest are cells with on-ramps (3 and 11), where actuating $u_{\max}$ for the MPC on-ramp controllers, $u_{\text{on},1}(k)$ and $u_{\text{on},2}(k)$, let more vehicles enter the freeway to avoid on-ramp spillbacks to the urban regions. Consequently, there are fewer vehicles in the urban regions and more vehicles queuing in the freeway, which seems sensible for mixed traffic network controller to keep the vehicles in the freeway instead of urban network during the rush hour. At $t = 4800$ (s), $u_{\text{on},2}(k)$ switches to $u_{\min}$ for 400 (s) leading to a sudden decrement in the density of cell 11 (on-ramp) and its former cells. Note that the critical density of the cells is 99.4 (veh/km) and cells 7 and 15 have off-ramps.

In Fig. 5.5(f), the time series of route choice parameters θ_{11}, θ_{12}, θ_{13}, θ_{21}, and θ_{31} are depicted. The enduring low value of urban travel time between regions 1 and 2 leads θ_{12} and θ_{21} to have almost constant values near to the maximum value. In contrast, θ_{11}, θ_{13}, and θ_{31} follow different trends. At the beginning of simulation as region 1 is congested, its speed is very low, thus about 65% of vehicles of the O-D demand from 1 to 1 prefer to choose the alternative trip route through the freeway ($1 \rightarrow 3 \rightarrow 1$). As the time passes, region 1 accumulation decreases and hence, the speed increases and more vehicles are in favor of choosing trip route $1 \rightarrow 1$ instead of $1 \rightarrow 3 \rightarrow 1$. The same logic is apparent for the case of θ_{13} and θ_{31}. The θ_{23} and θ_{32} are not depicted since they are constantly equal to 1 for all the examples, which is clear as e.g. traveling through trip route $2 \rightarrow 3$ is always faster than trip route $2 \rightarrow 1 \rightarrow 3$. Overall, the outcome of the MPC controller demonstrates that the traffic can be efficiently managed and accommodated in the mixed traffic network with the centralized MPC control scheme.

5.4 Control policies for the mixed network

In this section, several control policies are introduced and elaborated for the decentralized traffic control of the mixed urban-freeway network. The policies have different control struc-

tures and different levels of coordination between the freeway and the urban network control entities. Our objective is to investigate how coordination among different controllers can improve the mobility patterns of a city. In the following, seven control policies (CPs) are introduced from fully decentralized with no communication between urban and freeway entities to fully coordinated, with a comprehensive description for each policy: (CP1) ALINEA control for freeway and $u_{\max}$ for urban network (ALINEA + urban $u_{\max}$), (CP2) ALINEA control with queue constraint for freeway and $u_{\max}$ for urban network (ALINEA Q + urban $u_{\max}$), (CP3) ALINEA control for freeway and MPC for urban network (ALINEA + urban MPC), (CP4) ALINEA control with queue constraint for freeway and MPC for urban network (ALINEA Q + urban MPC), (CP5) Decentralized MPC (D-MPC), (CP6) Cooperative Decentralized MPC (CD-MPC), and (CP7) Centralized mixed network MPC (MPC).

In CP1, the ALINEA control law [134] is applied for the two on-ramps of the freeway, while the perimeter control inputs are set to their upper bounds in order to minimize the restriction for vehicle flows, i.e. $u_{12}(k) = u_{21}(k) = u_{\max}$. The ALINEA policy for the on-ramp metering controllers determines the number of vehicles entering the freeway, $f_{\text{on},l}(k+1)$, corresponding to the difference between the current accumulation of cell l and its desired accumulation, i.e. $x_l(k) - x_{\text{ref}}$, and a preset regulation parameter κ (−),

$$f_{\text{on},l}(k+1) = f_{\text{on},l}(k) + \kappa \cdot (x_l(k) - x_{\text{ref}}). \tag{5.30}$$

CP2 is similar to CP1, except that there are queue constraints on the on-ramps because of the limitation of queue space at the on-ramps. This policy can be considered the simplest step towards coordination as it tries to avoid spillbacks in the urban regions, but without analyzing the direct effect in system delays and outflow. Therefore, the ALINEA control law is modified to tackle the queue constraint as follows [11]

$$f_{\text{on},l}(k+1) = \begin{cases} f_{\text{on},l}(k) + \kappa \cdot (x_l(k) - x_{\text{ref}}) & \text{if } n_{\text{on},i}(k) \leq \eta \cdot n_{\text{on},i,\max}, \\ s_{\text{on},i} \cdot T_k & \text{otherwise,} \end{cases} \tag{5.31}$$

where $\eta \in [0,1]$. In this manner, the ALINEA controller with queue constraint tries to keep the length of on-ramp queue below a predefined ratio η of the maximum queue length $n_{\text{on},i,\max}$, which can effectively hinder the freeway spillbacks to the urban regions.

In CP3, the ramp metering controllers follow the ALINEA control law in (5.30), while the optimal perimeter control values are obtained from solving the mixed control problem using the MPC approach. This means that the MPC solver optimizes the total delay function in (5.17) *only* for $u_{12}(k)$, $u_{21}(k)$ while values of $u_{\text{on},1}(k)$ and $u_{\text{on},2}(k)$ for $k = 0, \ldots, K-1$ are determined separately according to (5.30).

CP4 is similar to CP3, except that the ramp metering controllers follow the ALINEA with queue constraints (control law in (5.31)). In this control policy, the optimal perimeter control is obtained from solving the mixed control problem using the MPC approach, as in CP3 but

values for $u_{\text{on},1}(k)$, $u_{\text{on},2}(k)$ are determined separately according to (5.31).

CP5 is the Decentralized MPC (D-MPC) policy that considers the freeway and urban network controllers as two separate entities without coordination between them, i.e. the D-MPC policy applies two parallel and separate MPC problems for the freeway and the urban network. The D-MPC splits the optimal control problem of the mixed network to two smaller-sized optimization problems: (i) freeway total delay minimization by manipulating *only* the ramp metering controllers, and (ii) urban total delay minimization by manipulating *only* the perimeter controllers. In this manner, the objective function (5.17) is divided to two terms; the urban total delay term, i.e. $T_k \cdot \left[\sum_{k=0}^{K-1}[n_1(k)+n_2(k)]\right]$, which is optimized solely for the perimeter controllers, $u_{12}(k)$ and $u_{21}(k)$, and the freeway total delay term, i.e. $T_k \cdot [\sum_{k=0}^{K-1}\sum_{l=1}^{L} x_l(k) + \sum_{k=0}^{K-1}\sum_{i=1}^{2} n_{\text{on},i}(k)]$, which is minimized solely for the on-ramp metering control inputs $u_{\text{on},1}(k)$ and $u_{\text{on},2}(k)$. Under this policy the two different objectives might be conflicting.

In order to improve the performance of the D-MPC policy, we propose a modification to establish a cooperation between the D-MPC control entities in CP6. The modification is such that each of the two freeway and urban controllers tries to minimize the whole network total delay as formulated in (5.17). Meaning that, the urban controller minimizes (5.17) for the current urban optimization horizon *only* by manipulating the perimeter controllers, $u_{12}(k)$ and $u_{21}(k)$, as the ramp metering controllers, $u_{\text{on},1}(k)$ and $u_{\text{on},2}(k)$, are assumed to be known and constant from the previous step of freeway optimization. Likewise, the freeway controller minimizes (5.17) for the current freeway optimization horizon *only* by manipulating the on-ramp meterings, $u_{\text{on},1}(k)$ and $u_{\text{on},2}(k)$, as the perimeter controllers, $u_{12}(k)$ and $u_{21}(k)$, are assumed to be known and constant from the previous step of urban optimization. This policy is named Cooperative Decentralize MPC (CD-MPC) because the two control entities are not centralized and fully coordinated, yet they communicate with partial information in a way to achieve the mixed network optimal performance. The CD-MPC policy is crucial in case of limited traffic monitoring or data communication.

CP7 is a fully centralized MPC controller for the whole mixed network consistent with the formulated problem (5.1)–(5.29).

5.5 Comparison of control policies

We test all the previous control policies for further investigation on the same example setup, as described in Section 5.3.3. The total delay of all policies for the two urban regions, the freeway, and the on-ramps are summarized in Table 5.2. The primary control policy to inspect is the integration of ALINEA freeway ramp metering and $u_{\max}$ for the urban perimeter controllers. The corresponding evolution of urban accumulations over time is depicted in Fig. 5.6(a), revealing that region 1 encounters the gridlock condition. The ALINEA control sequences are illustrated in Fig. 5.6(b) demonstrating a jumpy non-smooth trend as a result of density alternation of corresponding on-ramp cells, i.e. cells 3 and 11. Figure 5.6(c) shows the density

contour of freeway cells. The cell 3 density is over the desired value from $t = 0$ to $t = 3800$ (s), which forces the ALINEA controller $u_{\text{on},1}$ be equal to $u_{\min}$. As the time passes on, the cell 3 density decreases because of the reduction in the demands, which results in $u_{\text{on},1}$ increasing jumps to regulate the density of cell 3 at the desired value. The same reasoning is valid for on-ramp 2 while the density of cell 11 is always around its preset desired value, which makes the $u_{\text{on},2}$ going up and down to finely regulate the cell 11 density. It is apparent that the freeway is under-utilized because of local control scope of ALINEA, that prevents the vehicle to use the freeway and keeps them in the urban regions, which increases the chance of urban gridlock occurrence.

Table 5.2: Total delay [veh $\cdot$ s $\cdot 10^5$]

	ALINEA + Urban $u_{\max}$	ALINEA Q + Urban $u_{\max}$	ALINEA + Urban MPC	ALINEA Q + Urban MPC	D-MPC	CD-MPC	MPC
Region 1	524	438	504	262	537	254	196
Region 2	121	97.2	162	160	279	133	143
Freeway	129	172	129	189	124	217	218
On-Ramp 1	13.2	12.8	12.6	14.8	16.0	6.55	7.84
On-Ramp 2	16.0	16.0	15.8	16.2	16.3	16.1	15.7
Network	803	736	823	642	972	626	581

The on-ramp queues for both ALINEA and ALINEA with queue constraint freeway controllers are depicted in Fig. 5.6(d). It is apparent that using ALINEA ramp metering law, both the on-ramp queues reach the maximum possible queue length, however, integrating queue constraint within ALINEA forces the queues to be less than the maximum value and be equal to the predefined queue length threshold. For on-ramp 1 this phenomenon is obvious, but for on-ramp 2, since its input demand is so high that even with $u_{\text{on},2} = u_{\max}$ this constraint cannot be met during the rush hour. After $t = 4300$ (s) with reduction in the demand, the on-ramp 2 queue decreases to the predefined queue length and below. The corresponding urban accumulations and control sequences of ALINEA freeway controller with queue constraint and $u_{\max}$ urban perimeter controller are shown in Fig. 5.7(a) and 5.7(b), respectively. The freeway density contour, Fig. 5.7(c), demonstrates that the activation of queue constraint permits more vehicles to enter the freeway, which increases the freeway accumulation and consequently, less chance of urban gridlock and less total delay for the whole network.

The two aforementioned control policies with the urban perimeter fixed control lead the urban network to gridlock. Thus, we propose to replace the fixed controllers with the MPC controllers such that the new control policy consists of ALINEA controller for freeway and MPC for urban perimeter control (CP3). The corresponding evolution of urban accumulations is depicted in Fig. 5.8(a) and once compared with Fig. 5.6(a) reveals that the time that region 1 reaches gridlock is slightly increased, which is not beneficial. The control sequences are illustrated in Fig. 5.8(b) showing the same jumpy non-smooth trend similar to combination of ALINEA with $u_{\max}$ urban perimeter control. The same trend of on-ramp queues and freeway density is also seen which indicates that the integration of freeway ALINEA and urban MPC is not

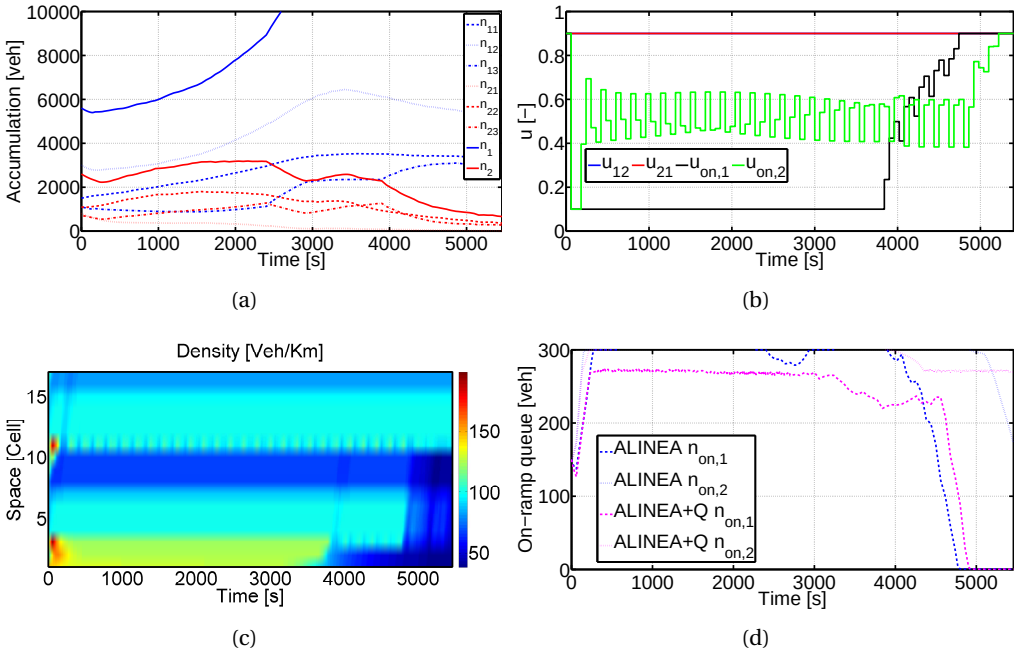


Figure 5.6: The results of case study example for CP1 (ALINEA ramp metering and $u_{\max}$ urban perimeter control): (a) urban accumulations, (b) control sequences, (c) freeway density contour, and (d) comparison of on-ramp queues for CP1 and CP2 (ALINEA with queue constraint ramp metering).

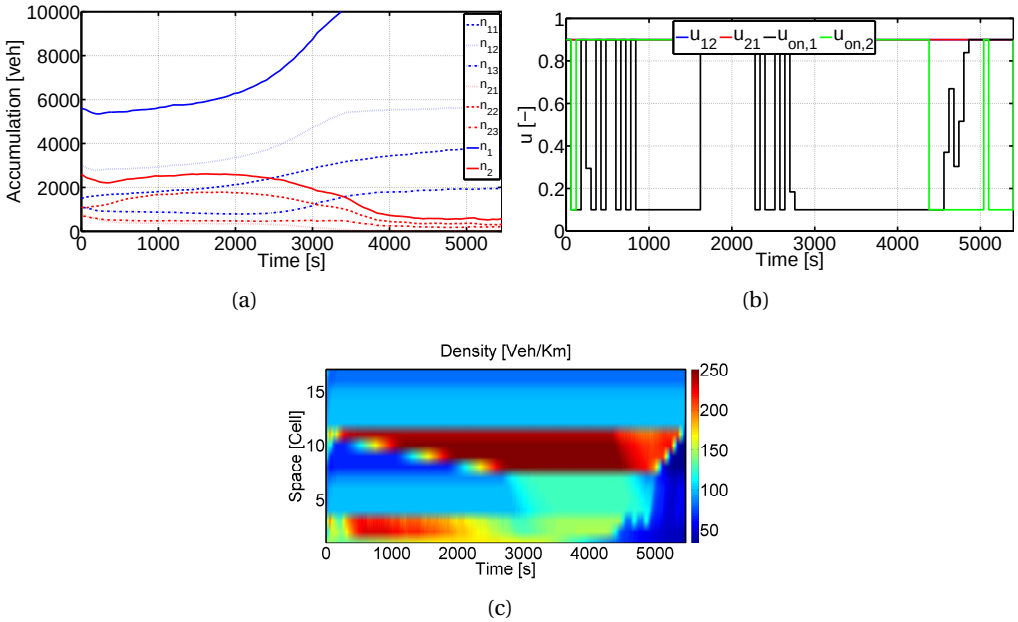
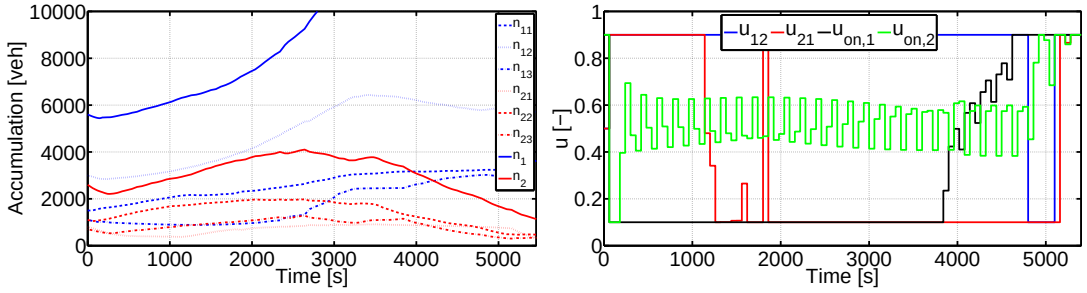


Figure 5.7: The results of case study example for CP2 (ALINEA with queue constraint ramp metering and $u_{\max}$ urban perimeter control): (a) urban accumulations, (b) control sequences, and (c) freeway density contour.

fruitful for the mixed network traffic control and higher coordination level is needed between the two entities.



(a) (b)

Figure 5.8: The results of case study example for CP3 (ALINEA ramp metering and MPC urban perimeter control): (a) urban accumulations and (b) control sequences.

In contrast, the performance of integration of urban MPC approach and ALINEA freeway ramp metering with queue constraint (CP4) demonstrates an advantageous control policy. The urban accumulations are depicted in Fig. 5.9(a) in which both regions end at the uncongested regime implying the effectiveness of the traffic control policy. To investigate more, control sequences are shown in Fig. 5.9(d) and once compared to Fig. 5.7(b), the MPC urban perimeter controllers are different while ALINEA controllers are to some extent the same, still manifests lack of cooperation between urban and freeway controllers. The u_{21} starts with the value of 0.5 for one control step then switches to $u_{\min}$ during which the accumulation of region 1 decreases and the accumulation of region 2 increases. Note the accumulations between $t = 0$ and $t = 600$ (s) in Fig. 5.9(a) and 5.7(a). This not very significant alternation in the urban perimeter control values makes region 1 accumulation decreases from the edge of gridlock to a more stable regime, which yields to stability of both regions during the peak hour. In addition, by inspecting the freeway density contour, Fig. 5.9(g), we observe more vehicles traveling along the freeway which is the reason that the urban regions could operate at the uncongested regime.

To improve the performance of the above control policies, the D-MPC controller which consists of one urban and one freeway control entities is introduced and tested on the same case study example. Each of the urban and freeway controllers tries to minimize the total delay *only* in their subnetwork of interest, hence, the freeway MPC controller forces both the ramp metering controllers, i.e. $u_{\text{on},1}$ and $u_{\text{on},2}$, to be equal to $u_{\min}$ as shown in Fig. 5.9(e). This prevents the vehicles to use the freeway for commuting, causes the freeway to be underutilized (Fig. 5.9(h)) and consequently increases the accumulation in the urban regions rapidly which leads to urban gridlock (Fig. 5.9(b)). To avoid the competitiveness between the D-MPC control entities, we add an additional level of coordination that adjusts the objective function of MPC entities to modify them into CD-MPC configuration. The results in Fig. 5.9(c), 5.9(f), and 5.9(i) show that the CD-MPC can control the traffic demand very efficiently compared to D-MPC. Table 5.2

summarizes the numerical results of all control policies averaged over 10 runs demonstrating that the best policy is the centralized MPC. Yet for cases with larger number of state variables, e.g. cases with more urban regions and/or more freeway cells, the MPC optimization module is not tractable. In this case or when the centralized MPC controller is not applicable because of technical limitations, the proposed CD-MPC controller, which consists of two smaller sized control entities seems promising with insignificant performance loss (less than 7%).

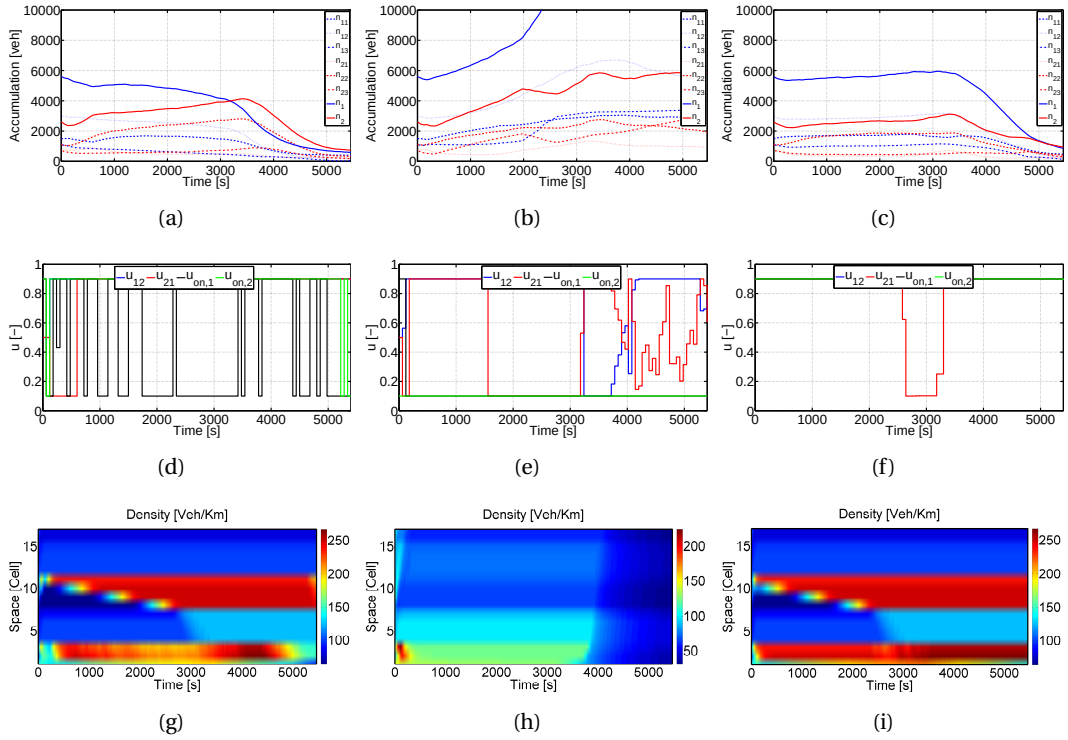


Figure 5.9: The results of case study example for (i) CP4 (ALINEA with queue constraint ramp metering and MPC urban perimeter control): (a) urban accumulations, (d) control sequences, (g) freeway density contour; (ii) CP5 (decentralized MPC control (D-MPC)): (b) urban accumulations, (e) control sequences, (h) freeway density contour; and (iii) CP6 (cooperative decentralized MPC control (CD-MPC)): (c) urban accumulations, (f) control sequences, (i) freeway density contour.

Note that the performances of CPs (percentage of improvement) are fully dependent on the level of congestion, e.g. if urban regions are uncongested the marginal improvement of CP6 (CD-MPC) or CP7 is not significant. This statement was verified as we have run the same example with 10% less demand of O-Ds $1 \rightarrow 2$ and $2 \rightarrow 2$. The results show that with less demand still the CP1, CP3, and CP5 cannot handle the traffic load and lead the urban regions to high congestion. The improvements of CP6 (CD-MPC) and CP7 (MPC) over the CP2 (ALINEA Q + urban $u_{\max}$) are about 3% and 4%, respectively, meaning that the simple cooperative CP2 is good enough to control this scenario. Though in the base scenario, the corresponding improvements are 15% and 21%, respectively.

5.6 System optimum route choice in the mixed network

The introduced simple route choice modeling in Section 5.2.3 is based on the user equilibrium (UE) assumption, indicating that the drivers (users) choose the route with the smallest travel time (cost) among their route choice set. We model the estimation of route travel time based on the current condition of the mixed network, while alternative estimation models (e.g. integrating prediction to capture future evolution of the network) have also great potential for future investigation. Independently of route travel time estimation modeling, the UE assumption still holds as a sound description of non-cooperative route choice behavior of drivers. It is evident that from the stand point of the system, i.e. mixed transport network, the simple route choice cannot provide the optimal solution. Nevertheless, the proposed MPC framework provides the opportunity to integrate the system optimum (SO) route choice in the mixed network traffic control problem. Though, the real-world implementation of SO traffic control policy is not fully operational, because of inherent unpredictability of human behavior [85], theoretically, we assume that it is possible to fully control (guide) the vehicles to a specific route to commute their origin-destination trip. Even if this is not fully implementable with traffic management schemes (variable message signs, pricing, etc.), it provides a comparison of what can be the most efficient way to control a transport network with a given demand and infrastructure.

In the following, we implement an SO route choice model only for the CP7–Centralized mixed network MPC (and not for all control policies presented in Section 5.4), since it provides the upper bound of total improvement. We consider the same example setup to test the implementation of SO route choice model. In this model, the route choice proportions, i.e. θ_{ij} for various origin destinations, are regarded as the output of the MPC optimization module, contrary to the simple route choice model where route choice proportions are the input to the optimization module. Thus, adjustments of mixed traffic dynamic equations are confined to omission of (5.8)–(5.16) and instead addition of all the θ_{ij} as optimization variables in the objective function (5.17). The results of the mixed network control with SO route choice model, averaged over 10 runs and presented in Table 5.3, indicating 12% improvement over the simple route choice model, in terms of decreasing the network total delay. Fig. 5.10(a), 5.10(b), and 5.10(c) depict the urban accumulations, the control sequences, and the freeway density contour of the augmented MPC controller with SO route choice model, respectively.

Table 5.3: Total delay for MPC with system optimum route choice [veh $\cdot$ s $\cdot 10^5$]

Region 1	Region 2	Freeway	On-Ramp 1	On-Ramp 2	Network
167	118	209	10.5	7.60	512

The θ_{11}, θ_{12}, θ_{13}, θ_{21}, and θ_{31} corresponding to both SO and the simple route choice models are illustrated in the Fig. 5.10(d), 5.10(e), 5.10(f), 5.10(g), and 5.10(h), respectively. It is evident that most of the time the SO route choice model forces the vehicles to use their first trip route choice, which is the direct one without any region to region or region to freeway transfer.

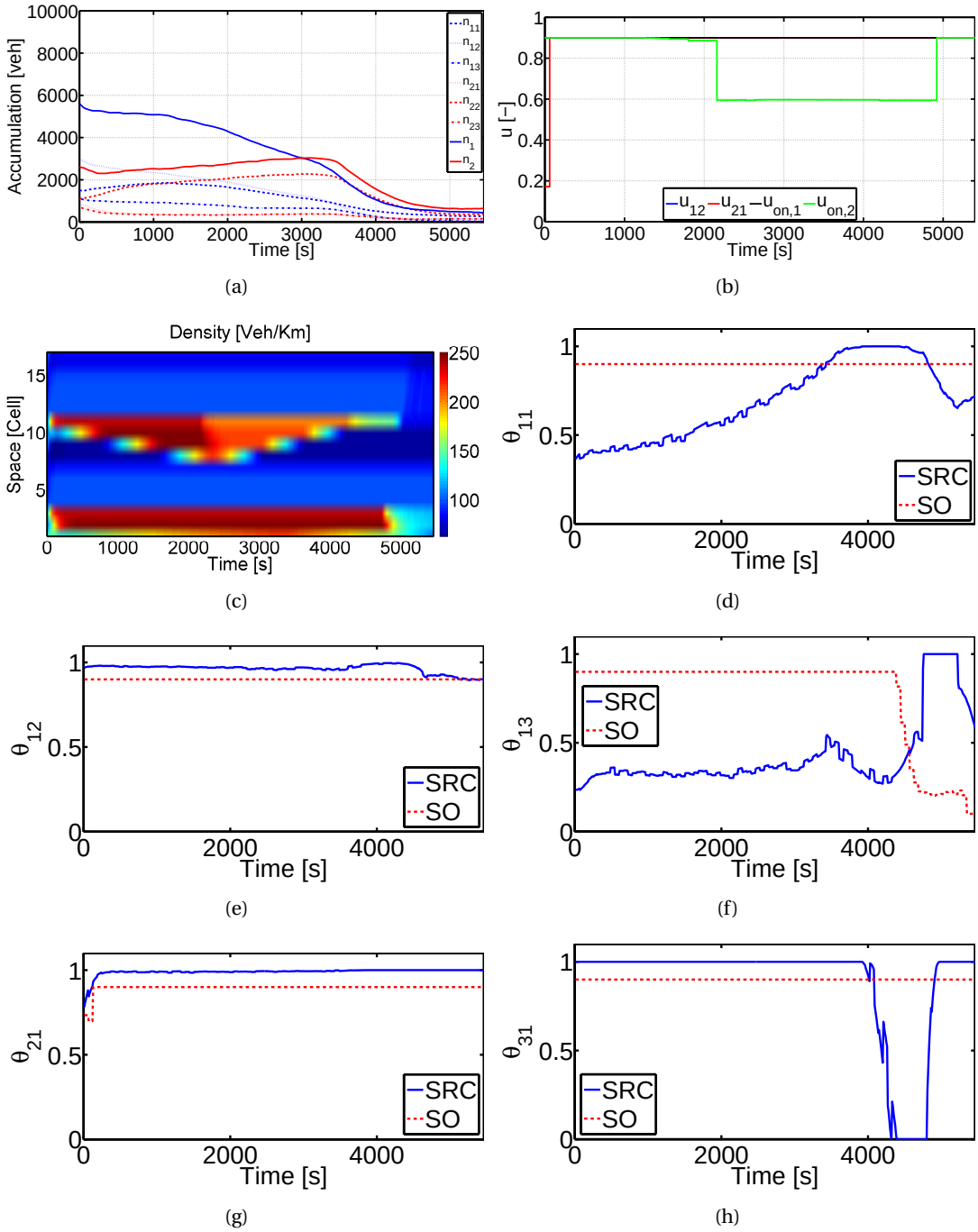


Figure 5.10: The results for the augmented MPC controller with system optimum route choice model: (a) urban accumulations, (b) control sequences, (c) freeway density contour, and the route choice results corresponding to the system optimum (SO) and the simple route choice model (SRC) for: (d) θ_{11}, (e) θ_{12}, (f) θ_{13}, (g) θ_{21}, (h) θ_{31}.

5.7 Summary

The control problem of a large-scale mixed traffic network, consisting of two urban regions with MFD representation, and one alternative freeway route modeled with the asymmetric cell transmission model is formulated. For traffic control purposes, we consider two controllers on the perimeter of regions to manipulate the urban inter-transfer flows; in addition, two on-ramp controllers to control the traffic flow from urban regions to the freeway. The optimal traffic control problem is solved by an MPC scheme. Several control policies with different controller structures and levels of urban-freeway coordination are introduced and scrutinized. The results demonstrate the advantage of centralized control over combination of a simple freeway ramp metering with urban MPC controller, and also the great importance of cooperation in decentralized MPC approach in cases with lack of full data communication and coordination between urban and freeway control entities, i.e. when the centralized MPC is not tractable. In addition, the system optimum route choice is integrated within the centralized MPC, which leads to increase the network performance by 12%.

These results can be beneficial for municipal administrators to develop efficient hierarchical control strategies for metropolitan mixed traffic networks. This work contributes one step forward towards achieving SoS approach for transportation infrastructure and networks. The outcome of this optimization does not provide the exact phase settings for traffic signals in the boundary between the two regions. Nevertheless, recent work for single [97] and multiple regions [4] provide the necessary tools to dynamically change the signal settings to meet the controllers' inputs u_{12} and u_{21} from the aggregated optimization. In case that local queues are developed in the proximity of the controllers (ramps and boundaries between urban regions), analysis of [53] can identify signal parameters in the individual regions of a city in such a way to move traffic smoothly at the desired flows, without concentrating a large number of vehicles at the boundaries of the regions. Traffic control problem of networks with more complex structure and dynamic traffic assignment is ongoing research. Ongoing work also involves the development of efficient control strategies for networks with a larger number of urban regions, on- and off-ramps. In this case, the size of regions might change over time due to congestion propagation, which will require a dynamic partitioning of the city in different parts. This is a challenging problem, both from an optimization (higher computational effort) and modeling (more complex dynamics).

6 Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control

REAL traffic data and simulation analyses reveal that for some urban networks a well-defined Macroscopic Fundamental Diagram (MFD) exists, which provides a unimodal and low-scatter relationship between the network vehicle density and outflow. Recent studies demonstrate that link density heterogeneity plays a significant role in the shape and scatter level of MFD and can cause hysteresis loops that influence the system performance. Evidently, a more homogeneous network in terms of link density can result in higher network outflow, which implies a network performance improvement. In this chapter, we introduce two aggregated models, region- and subregion-based MFDs, to study the dynamics of heterogeneity and how they can affect the accuracy scatter and hysteresis of a multi-subregion MFD model. We also introduce a hierarchical perimeter flow control problem by integrating the MFD heterogeneous modeling. The perimeter controllers operate on the border between urban regions, and manipulate the percentages of flows that transfer between the regions such that the network delay is minimized and the distribution of congestion is more homogeneous. The first level of the hierarchical control problem can be solved by a model predictive control approach, where the prediction model is the aggregated parsimonious region-based MFD and the plant (reality) is formulated by the subregion-based MFDs, which is a more detailed model. At the lower level, a feedback controller of the hierarchical structure, tries to maximize the outflow of critical regions, by increasing their homogeneity. With inputs that can be observed with existing monitoring techniques and without the need for detailed traffic state information, the proposed framework succeeds to increase network flows and decrease the hysteresis loop of the MFD. Comparison with existing perimeter controllers without considering the more advanced MFD modeling of heterogeneity highlights the importance of such approach for traffic modeling and control.

6.1 Introduction

Efficient traffic control and management of large-scale transportation networks still remain a challenge both for traffic researchers and practitioners. Unlike microscopic approaches that usually utilize disaggregate traffic flow models, as behavior of each vehicle is modeled in

detail, e.g. car following and lane changing models, in this chapter, we follow the macroscopic (network level) approach utilizing the macroscopic fundamental diagram (MFD). The MFD aims at simplifying the micro-modeling task of the urban network, where the collective traffic flow dynamics of subnetworks capture the main characteristics of traffic congestion, such as the evolution of space-mean flows and densities in different regions of a city. Nevertheless, it is not a universal law for all network structures and demands and if it is applied in a non-careful way it can hide critical patterns of congestion and result in inefficient control methodologies.

The MFD provides a unimodal, low-scatter relation between network vehicle density (veh/km) and network space-mean flow or outflow (trip completion rate) (veh/hr) for different network regions, if congestion is roughly homogeneous in the region. Recently, the macroscopic (network) traffic modeling has intensively attracted the traffic flow community. The physical model of MFD was initially proposed by [60] and observed with dynamic features in congested urban network in Yokohama by [55], and investigated using empirical or simulated data by [16, 89, 120, 203] and others. Earlier works had looked for MFD patterns in data from lightly congested real-world networks or in data from simulations with artificial routing rules and static demands (e.g. [116, 131] and others), but didn't demonstrate that an invariant MFD with dynamic features can arise. The observability of the MFD with different sensing techniques have been studied by [108, 132].

Studies [120, 58, 46, 118, 101] have shown that networks with heterogeneous distribution of density exhibit network flows smaller than those that approximately meet homogeneity conditions (low spatial variance of link density), especially for high network densities. Networks with small variance of link densities have *a well-defined* MFD, i.e. low scatter of flows for the same densities. Heterogeneously congested networks might exhibit points below the upper envelope of an MFD or strong hysteresis loops (see for example freeway networks in [51, 165]). Recently, in agreement with previous publications in heterogeneity, [118] proposed and calibrated with simulated data an MFD where the effect of heterogeneity decreases the MFD output with a functional relationship. Following these findings, the concept of an MFD can be applied for heterogeneously loaded cities with multiple pockets of congestion, if these cities can be partitioned in a small number of homogeneous clusters. Recent work [90] created clustering algorithms for heterogeneous transportation networks with an objective to obtain small variance of link densities within a cluster. Understanding and modeling the dynamics of heterogeneity is a crucial and challenging question that can shed some light on how to develop smarter hierarchical traffic control schemes for large-scale urban networks. This chapter moves towards this direction.

The MFD can be utilized to introduce elegant real-time control strategies to improve mobility and decrease delays in large urban networks, that local ones are unable to succeed, see pioneer works in [28, 66, 49]. Perimeter control strategies, i.e. manipulating the transfer flows at the perimeter border of the urban region, utilizing the concept of the MFD have been introduced for single-region cities in [28, 97], and for multi-region cities in [66, 49, 5]. Moreover, route guidance strategies with the utilization of MFD have been studied in [102] for grid networks

without traffic lights. [47, 109] also studied simple routing strategies for two-bin or two-route network abstractions.

In [69] different control strategies with different levels of coordination have been introduced for metropolitan transportation networks that have a hierarchical structure that consists of freeways and urban roads. Previous works [33, 53, 203] have shown that traffic-responsive signal control strategies and different signal settings can change the shape of the MFD and the critical accumulations. While Chapters 4 and 5 do not explicitly model the effect of link heterogeneity, in this chapter we aim at studying the effect of heterogeneity by introducing a new model that considers MFD of both regions and smaller subregions. This chapter also models a simple route choice process between paths through subregions, and considers the effect of subregion flow receiving capacity.

The control problems in previous chapters, i.e. Chapters 4 and 5, have been solved by the model predictive control (MPC) approach. It was shown that this control approach can handle different levels of error in traffic demand and noise in MFDs shape. Nevertheless, the optimization model and the plant in the MPC framework were inherently similar, except the errors in demand and the MFD distinguish between the two. A stronger level of dissimilarity between the model and the plant can provide more convincing evidence for the applicability of such approaches in more realistic and complex networks.

The objectives of this chapter are two-fold, in modeling and control aspects. First, we would like to further investigate the relation between the heterogeneity and the MFD. With respect to modeling, we investigate the dynamics of heterogeneity and how it can affect accuracy and scatter of a multi-region MFD model, which consists of variables that can be obtained with existing sensor technology. While there is some work regarding how heterogeneity influences the shape of the MFD, there is no theoretical work to investigate how an asymmetric demand pattern can affect the distribution of congestion over time and space and its dynamic behavior. Existing MFD dynamic models as expressed in various publications are hysteresis-free and as a result the developed control frameworks based on such models cannot be trusted when hysteresis appears. With respect to control, our objective is to integrate the dynamics of heterogeneity in the optimization framework and design perimeter control strategies that can decrease congestion heterogeneity and increase system performance. As our analysis shows, this is a crucial step for efficient control strategies, where strong hysteresis loops appear. Also we will demonstrate that considering an MFD without hysteresis and ignoring the dynamics of heterogeneity can result in situations where perimeter control is not beneficial for the system. Such an advanced model also allows to develop a two-level hierarchical control framework that decreases system delays and hysteresis loops, which are directly related with capacity loss at the network level.

The remainder of the chapter is organized as follows: in Section 6.2 we describe the dynamics of the region-based model which integrates the effect of heterogeneity in the MFD. A more detailed model of sub-regions, which can describe the dynamics of heterogeneity, is developed.

In Section 6.3, a closed form expression of MFD as a function of mean and variance of link accumulation is obtained based on real data, while in Section 6.4 a hierarchical control framework based on the MPC tries to optimize system performance according to the developed models. Decreasing the level of heterogeneity with control has a positive effect in the system delays and the avoidance of large hysteresis loops. Results of a case study and future work conclude the chapter.

6.2 Modeling the dynamics of spatial density heterogeneity in urban regions with perimeter control

In this chapter, we introduce two aggregated models with an objective to integrate the dynamics of heterogeneity in a network: (i) a *region-based model* considers networks partitioned into a small number of regions that might be split by perimeter controllers, and (ii) a *subregion-based model*, where each region of the above model is partitioned into subregions, see Fig. 6.1. Existing region-based dynamic models for single or multi-region networks (e.g. [28, 49, 97]) consider an MFD without hysteresis. Hysteresis creates multivalueness in the network flow for the same value of network accumulations. Given that these values are strongly influenced by the distribution of congestion, such a hysteresis cannot be an external input to the model (e.g. as a functional form), but it has to be integrated within the model and be influenced by the dynamics of heterogeneity. This is a challenging methodological step that requires the interconnection of a region-based and a subregion-based model. At the region-based model the heterogeneity dynamics are integrated in the regional MFDs in two directions: (i) variant regional trip lengths and (ii) an MFD depending on regional accumulation and the heterogeneity in the spatial distribution of congestion. To integrate the dynamics of heterogeneity in the region-based model, a subregion-based model is needed that (a) describes the evolution of subregion accumulations, (b) integrates the heterogeneity dynamics in the subregion MFDs, (c) integrates a route choice model, and (d) models the effect of receiving (or boundary) capacity of the destination subregion.

The constraint of receiving (or boundary) capacity has not been considered during the optimization process in previous control oriented publications (see [97, 49, 4]). The physical reasoning behind this assumption is that (i) boundary capacity decreases for accumulations much larger than the critical accumulation (see [54]), and (ii) the control inputs will not allow the system to get close to gridlock. Nevertheless, perimeter controllers are acting only in the boundaries between regions and not all subregions are integrated in a perimeter control logic. Thus, such a constraint cannot be fully ignored now. Figure 6.1(a) and (b) depict a schematic urban network with (part of) internal and transfer flows for region I and subregions i, j, r in the (i) region- and (ii) subregion-based models, respectively. All the related variables are introduced later in details.

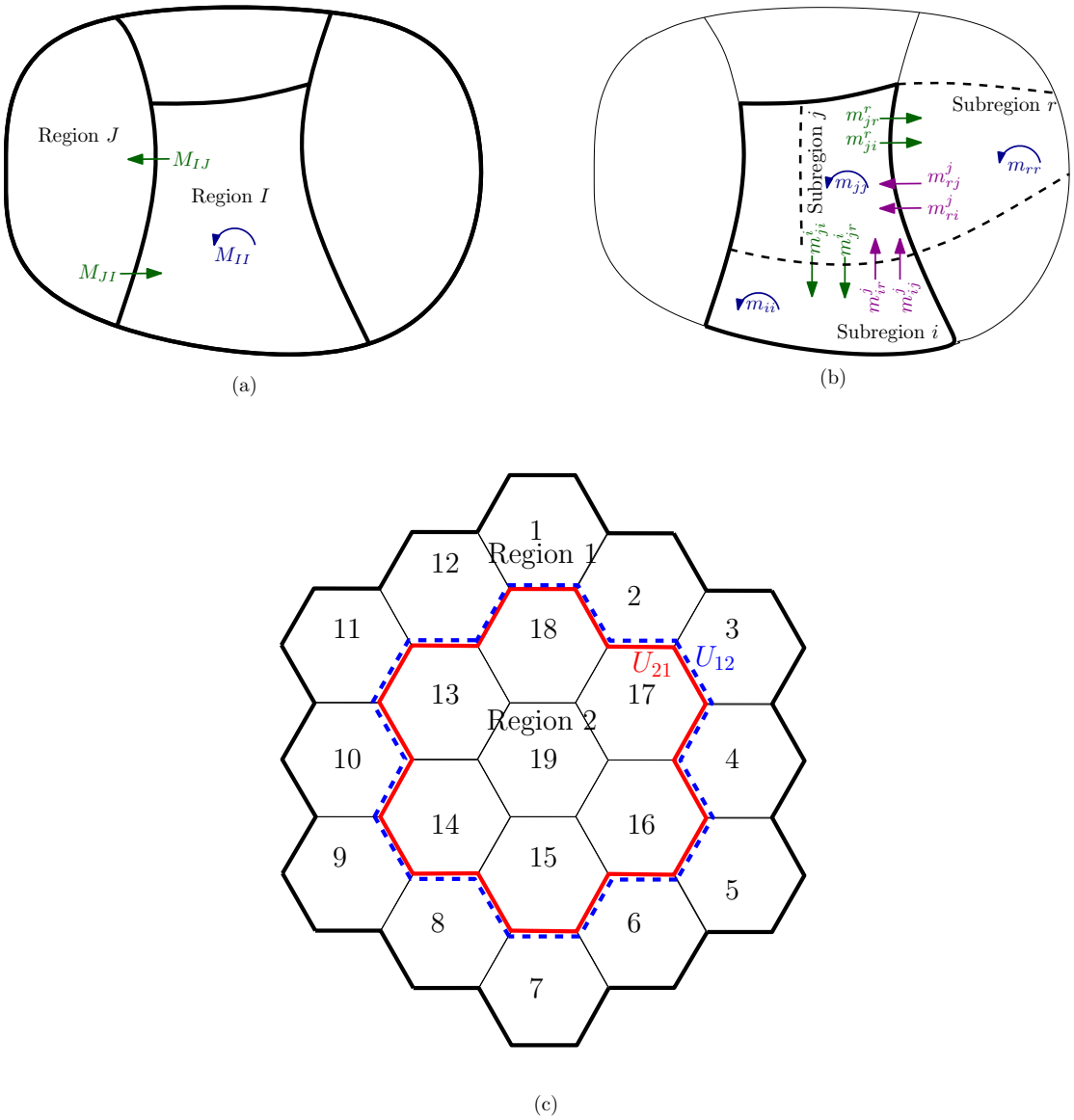


Figure 6.1: A schematic urban network with (part of) internal and transfer flows for region I and subregions i, j, r in the (a) region- and (b) subregion-based models, respectively. (c) A case study network consists of two regions - Region 1 (the periphery) and Region 2 (city center) partitioned respectively into 12 and 7 subregions.

6.2.1 Region-based model

Let us assume that an urban network is partitioned into R regions, $\mathscr{R} = \{1, 2, \cdots, R\}$. Let $Q_{IJ}(t)$ (veh/s) be the traffic demand flow generated in region I with destination to region J, $N_{IJ}(t)$ (veh) be the accumulation in region I with region destination J; $I, J \in \mathscr{R}$, and $N_I(t)$ (veh) be the total accumulation in region I.

The total production $P_I\big(N_I(t), \sigma(N_I(t))\big)$ (veh $\cdot$ distance travelled per unit time) in region I is

a function of the regional accumulation $N_I(t)$ and its variance across all links in the region, $\sigma(N_I(t))$, as has been reported in [120, 118, 58, 101]. The trip completion flow for region I is the sum of transfer flows, i.e. trips from I with direct destination J, $J \in \mathscr{H}_I$, where $\mathscr{H}_I$ is the set of regions that are directly reachable from (adjacent to) region I, plus the internal flow, i.e. trips from I with direct destination I. The *transfer flow* from I with destination to J is denoted by $M_{IJ}(t)$ (veh/s), while $M_{II}(t)$ denotes the *internal flow* from I with destination to I. They are calculated corresponding to the ratio between accumulations as follows

$$M_{II}(t) = \frac{N_{II}(t)}{N_I} \cdot \frac{P_I\big(N_I(t), \sigma(N_I(t))\big)}{L_{II}(t)}, \tag{6.1a}$$

$$M_{IJ}(t) = \frac{N_{IJ}(t)}{N_I} \cdot \frac{P_I\big(N_I(t), \sigma(N_I(t))\big)}{L_{IJ}(t)}, \tag{6.1b}$$

where $P_I(\cdot)$ (veh/s $\cdot$ m) is the MFD production for region I at $N_I(t)$ with heterogeneity variance $\sigma(N_I(t))$, $L_{II}(t)$ (m) is the average trip length (space mean) for trips in region I, and $L_{IJ}(t)$ (m) is the average trip length for trips from region I to J. Note that the variable $\sigma(N_I(t))$ captures the link density spatial *heterogeneity* for an urban region (see later equation (6.12), where function $P_I\big(N_I(t), \sigma(N_I(t))\big)$ is described). Note that for flows $Q_{IJ}(t)$ where I and J are not adjacent, a sequence of regions should be known to develop the transfer flows between I and J.

One of the main objectives of this chapter is to integrate the developed modeling in a control framework and investigate strategies that will decrease heterogeneity and network flow hysteresis loops. To this end, a semi-analytical approximative model of $\sigma(N_I(t))$ is required. To model $\sigma(N_I(t))$, we need first to develop an analytical model for subregion link density heterogeneity then aggregate the subregional heterogeneity into the regional one. Thus, in Section 6.3 we investigate the heterogeneity dynamics for a subregion based on a field dataset and introduce a method to aggregate and scale up the subregional heterogeneities to present link heterogeneity in regions.

Perimeter controllers, $U_{IJ}(t)$ and $U_{JI}(t)$ (-) , and $0 \leq U_{IJ}(t), U_{JI}(t) \leq 1$, might exist between each two regions I and J, $J \in \mathscr{H}_I$, that can constrain the transfer flows from I to J and from J to I, respectively. The mass conservation equations of an R-region MFDs system are as follows:

$$\frac{\mathrm{d}N_{II}(t)}{\mathrm{d}t} = Q_{II}(t) - M_{II}(t) + \sum_{J \in \mathscr{H}_I} U_{JI}(t) \cdot M_{JI}(t), \tag{6.2}$$

$$\frac{\mathrm{d}N_{IJ}(t)}{\mathrm{d}t} = Q_{IJ}(t) - \sum_{J \in \mathscr{H}_I} U_{IJ}(t) \cdot M_{IJ}(t), \tag{6.3}$$

for $I = 1, 2, \ldots, R$ and $\forall J \in \mathscr{H}_I$. Note that $N_I(t) = N_{II}(t) + \sum_{J \in \mathscr{H}_I} N_{IJ}(t)$. These equations are a generalized (R regions instead of two) equations presented in [49] with integrated heterogeneity. Note that route choice modeling is not integrated in the *region-based* dynamic equations and this model is not aware that travellers make route choice decisions when conditions change. This is done on purpose, since traveller behavior might be difficult to be

predicted in real-time. It is also assumed that drivers are not allowed to cross a boundary more than once, e.g. a trip from region I to I by crossing region J is not considered. This will change the dynamic equations (6.2) and (6.3) and more complicated accumulation states have to be developed, which is beyond the scope of this chapter (e.g. number of vehicles in region I with destination I that will cross to region J and return in region I). Nevertheless, the hierarchical control framework that is developed later is not influenced by such constraint.

6.2.2 Subregion-based model

The subregion-based model is a more detailed model since the urban region is considered as a collection of several smaller urban areas, called *subregions*, which still contain a significant number of links to be described by a low-scatter MFD. Each subregion accumulation evolves differently in time which allows to integrate the heterogeneity of the spatial distribution of congestion in the the urban region (see terms related to MFD heterogeneity in (6.1a) and (6.1b)). This modeling approach will give us the opportunity to investigate more rigorously several assumptions in the MFD literature that have been empirically observed, e.g. trip length in a region is about constant, if and how route choice, perimeter control, and O-D affect the heterogeneity and the distribution of congestion. These are challenging research questions that have been raised by many researchers and it is not clear yet under what network conditions an MFD provides a decent representation of network performance. The purpose of the following formulation is to express in a consistent manner the variables in (6.1a)-(6.3), which represent region-based MFDs, accumulations and trip lengths as a function of a more detailed model at the subregional level. While microscopic simulation might be an alternative instead of the described model, this thesis chooses a more methodological path, which allows to create further insights of the dynamics of heterogeneity and the hierarchical control.

Let us consider region $I \in \mathscr{R}$ which is heterogeneous in space link density and consists of subregions, as schematically shown in Fig. 6.1(b). We use capital letters for variables related to regions and lower case letters for variables related to subregions. We denote $\mathscr{SR}$ as the set of all subregions in the urban network, while $\mathscr{SR}_I$ is the set of subregions that belongs to region I. Let $q_{ij}(t)$ (veh/s) be the demand from subregion i to subregion j, $n_{ij}(t)$ (veh) be the accumulation in subregion i with final subregion destination j, $\{i, j\} \in \mathscr{SR}$, and $n_i(t)$ (veh) be the total accumulation in subregion i, i.e. $n_i(t) = \sum_{j \in \mathscr{SR}} n_{ij}(t)$. The MFD production for subregion i, denoted by $p_i(t)$ (veh/s $\cdot$ m), is the total distance traveled for subregion i by all vehicles $n_i(t)$, which is equal to the sum of the transfer and internal flows multiplied by the average trip length in subregion i, $l_i(t)$ (m).

Let $m^h_{ij}(t)$ (veh/s) be the *transfer flow* from subregion i with *final* subregion destination j, $i \neq j$, through the *immediate* next subregion $h \in \mathscr{H}_i$, where $\mathscr{H}_i$ is the set of subregions that are directly reachable from subregion i. The transfer flow is related to the ratio between subregion accumulations and corresponding trip length, i.e. $m^h_{ij}(t) = \theta^h_{ij}(t) \cdot n_{ij}(t)/n_i(t) \cdot p_i(n_i(t))/l_i(t)$, where $\theta^h_{ij}(t)$ (-) is the flow percentage of the total transfer flows from subregion i to destination

j that passes immediately through subregion h, thus $\sum_{h\in\mathscr{H}_i}\theta_{ij}^h(t)=1$. Note that a simple route choice model is integrated in the *subregion-based* model, where $\theta_{ij}^h(t)$ are calculated by a logit model according to the travel times from i to j through the k (current best) shortest paths (sequence of subregions), which are calculated using Dijkstra's algorithm. The travel time for each path is calculated by summing travel times through subregions, where each subregion travel time is calculated as the fraction between the distance travelled inside the subregion (through its center) and its average speed $v_i(t)$ (m/s) calculated from the subregion MFD at the beginning of the trip, i.e. $v_i(t)=p_i(n_i(t))/n_i(t)$. Trip length within subregion i is assumed to be independent of origin, destination, and route choice, which is consistent with the field data in [55] and the assumptions made for the region-based models of previous Chapters 4 and 5.

The *internal flow* from subregion i with destination to subregion i, denoted by $m_{ii}(t)$ (veh/s), is calculated by $m_{ii}(t)=n_{ii}(t)/n_i(t)\cdot p_i(n_i(t))/l_i(t)$. For instance, Fig. 6.1(b) illustrates part of the transfer and internal flows for a network considering only three subregions i, j, and r. The transfer flows between subregions j and r are m_{ji}^r, m_{jr}^r, m_{rj}^j, m_{ri}^j, and m_{ji}^i, m_{ij}^j, m_{ir}^j, m_{jr}^i are the transfer flows between subregions j and i, while the internal flows for subregions i, j, r are m_{ii}, m_{jj}, m_{rr}, respectively. Note that the region-based model implicitly assumes that the internal regional trips never leave the region and also external trips cross the boundary between the regions only once. The route choice of subregion-based model meets these assumptions.

The subregion-based model also integrates the effect of flow receiving capacity of the destination subregion. In other words, flow transferring into a subregion might be restricted since accumulation at subregion destination is such high that there is not enough space to fully receive the incoming transfer flows. Receiving capacity is not integrated in the region-based model as the controllers at the boundary are expected to avoid these situations. Equation (6.4) expresses the transfer flow as the minimum of two terms, (i) the sending flow upstream of the boundary (from region i) which depends on the accumulations of region i and the (ii) receiving flow which depends on the accumulation of region h. Such an approach has been integrated in mass conservation equations for 1st and 2nd order models of traffic flow, e.g. the Cell Transmission Model [30]. The difference is that the 2nd term is an analogy of the remaining storage capacity of the receiving region. Therefore, we introduce a receiving capacity term into the transfer flow dynamic equations as follow

$$\hat{m}_{ij}^h(t)=\min\left(m_{ij}^h(t),\frac{m_{ij}^h(t)}{\sum_k m_{ik}^h(t)}\cdot r_{ih}(n_h(t))\right), \tag{6.4}$$

where $k\in\mathscr{SR}, k\neq i$, and $r_{ih}(\cdot)$ (veh/s) is the receiving flow capacity of subregion h, $h\in\mathscr{H}_i$, from subregion i. We consider that the receiving capacity is a piecewise function of $n_h(t)$ with

two pieces, a constant value and a decreasing function, as follows

$$r_{ih}(n_h(t)) = \begin{cases} r_{ih}^{\max} & \text{if } 0 \le n_h(t) < \alpha \cdot n_h^{\text{jam}}, \\ -\frac{r_{ih}^{\max}}{(1-\alpha)\cdot n_h^{\text{jam}}} \cdot n_h(t) + \frac{r_{ih}^{\max}}{1-\alpha} & \text{if } \alpha \cdot n_h^{\text{jam}} \le n_h(t) \le n_h^{\text{jam}}, \end{cases} \tag{6.5}$$

where $r_{ih}^{\max}$ (veh/s) is the maximum value of the receiving capacity and boundary capacity, n_h^{jam} (veh) is the jammed accumulation of subregion h, and $0 < \alpha < 1$ is a parameter that defines the critical accumulation when the receiving capacity starts to decrease and can be estimated if real data from sensors are readily available.

The transfer flows might be controlled by subregion perimeter controllers on the border between subregions, e.g. $0 \le u_{ih}(t)\,(-) \le 1$ denotes the perimeter control input between subregions i and h. The mass conservation equations for the subregions are as follows

$$\frac{\mathrm{d}n_{ii}(t)}{\mathrm{d}t} = q_{ii}(t) - m_{ii}(t) + \sum_{h \in \mathscr{H}_i} u_{hi}(t) \cdot \hat{m}_{hi}^{i}(t)\,, \tag{6.6}$$

$$\frac{\mathrm{d}n_{ij}(t)}{\mathrm{d}t} = q_{ij}(t) - \sum_{h \in \mathscr{H}_i} u_{ih}(t) \cdot \hat{m}_{ij}^{h}(t) + \sum_{h \in \mathscr{H}_i;\, h \neq j} u_{hi}(t) \cdot \hat{m}_{hj}^{i}(t) \quad \forall\, j \in \mathscr{H}_i\,, \tag{6.7}$$

$$\frac{\mathrm{d}n_{ir}(t)}{\mathrm{d}t} = q_{ir}(t) - \sum_{h \in \mathscr{H}_i} u_{ih}(t) \cdot \hat{m}_{ir}^{h}(t) + \sum_{h \in \mathscr{H}_i} u_{hi}(t) \cdot \hat{m}_{hr}^{i}(t) \quad i \neq r;\, \forall\, r \notin \mathscr{H}_i\,. \tag{6.8}$$

Equations (6.6)–(6.8) assume that perimeter controllers exist between each neighbour subregions, however, they still hold if the assumption is relaxed by setting the control inputs to be equal to 1. Note that our intension is not to control inter transfers between any two subregions, but only in the boundaries of the region-based model, see Fig. 6.1(c). In this way we will keep the computational effort small and we will not rely on information which is difficult to be obtained with real data, e.g. n_{ij} for each subregion. Nevertheless, as stated before the more detailed model will shed light on the dynamics of heterogeneity and how it can affect the performance of an MFD region-based model, which consists of variables that can be obtained with existing sensors more accurately.

Finally, the region internal and external average trip lengths described in (6.1a) and (6.1b), L_{II} and L_{IJ}, respectively, are estimated as follows (considering a steady state law as the ratio of travel production over outflow)

$$L_{II}(t) = \frac{\sum_{i \in \mathscr{SR}_I} \sum_{j \in \mathscr{SR}_I} n_{ij}(t)}{\sum_{i \in \mathscr{SR}_I} n_i(t)} \cdot \frac{\sum_{i \in \mathscr{SR}_I} p_i(n_i(t))}{\sum_{i \in \mathscr{SR}_I} m_{ii}(t)}\,, \tag{6.9a}$$

$$L_{IJ}(t) = \frac{\sum_{i \in \mathscr{SR}_I} \sum_{j \in \mathscr{SR}_J} n_{ij}(t)}{\sum_{i \in \mathscr{SR}_I} n_i(t)} \cdot \frac{\sum_{i \in \mathscr{SR}_I} p_i(n_i(t))}{\sum_{i \in \mathscr{SR}_I} \sum_{h \in \mathscr{SR}_J} m_{ij}^{h}(t)}\,. \tag{6.9b}$$

The estimation of L_{II} and L_{IJ} is based on the assumption that the region- and subregion-based models should be consistent and have the same internal and external region outflows in case of perfect information. Thus, (6.9a) and (6.9b) have respectively similar logic to (6.1a) and

(6.1b), while the right hand sides are expressed in terms of detailed variables of subregion-based model. I.e., the internal outflow M_{II} in the region-based model is equivalent to the sum of all m_{ii}, $i \in \mathscr{SR}_I$, and the external outflow M_{IJ} in the region-based model is equivalent to the sum of all $m_{ij}^h(t)$, $i \in \mathscr{SR}_I$ and $h \in \mathscr{SR}_J$.

6.3 A functional form of the effect of heterogeneity on MFDs: Field data analysis

The region-based MFD dynamic model of (6.1a) and (6.1b) requires a functional form of the regional production depending on accumulation of the region and standard deviation of the spatial distribution of link accumulation. While various studies have investigated the effect of heterogeneity in the MFD [51, 120, 101, 118], a functional form is necessary as these MFD dynamics have to be integrated in a control framework. While [118] provides a functional form based on simulated data, in this section we investigate such a relation with real data. More specifically, we re-scrutinize the Yokohama field data, investigated in [55, 58], to obtain further insights into the dynamics of link occupancy heterogeneity and its effect on the MFD. Our objective is to propose an analytical distribution that models the first two statistical moments of individual link occupancy distribution. This approach is motivated by previous publications towards this direction, which are described in more detail in Section 6.1.

The developed models of this work in Section 6.2 can also be implemented for different functional forms and the reader can skip this section without loss of continuity. Nevertheless, this analysis provides useful empirical analysis for heterogeneity. An interesting finding is that the spatial distribution of congestion has similar functional form with other physical systems that experience spatial correlation.

The data consist of the occupancy of 540 links every 5 minutes from early morning to the end of the day. We are interested in mean and standard deviation (STD) of link occupancy, since mean occupancy is an indicator of network congestion level and STD of link occupancies can be regarded as the heterogeneity indicator of the network. Analyses demonstrate that the negative binomial (NB) distribution can provide accurate estimations for mean and STD of occupancies for the Yokohama network data. NB distributions can describe well the spread of different phenomena with spatial correlations, such as infectious diseases [113], tree growth [21], and others. Note that [58] have derived a semi-analytical model of estimating this distribution based on spatial correlations between links, which might be difficult to be integrated in a control scheme. We choose to utilize the NB distribution, due to the ease of numerical calculations. NB is a discrete probability distribution of the number of successes in a sequence of binomial trials with probability of success, p, before a pre-specified number of failures, r, occurs. NB distribution is useful in modeling count data similar to Poisson distribution, however, NB is more general and accurate to capture dispersion with spatial correlations than Poisson distribution because its variance is greater than its mean. The NB probability mass

function is

$$\Pr(X = x) = NB(x, r, p) = \binom{x + r - 1}{x} (1 - p)^r p^x . \tag{6.10}$$

Note that r can be interpreted as the number of congested links (failures), while p can be related to the occupancy that indicates congested state (probability of success or failure, if occupancy is normalized between zero and one).

Fig. 6.2 depicts the field data and the best NB fit (in maximum likelihood sense) representing link occupancy distribution at four different times during a day covering a wide range of traffic conditions from early uncongested to mid-day congested and evening mild-congested. The four cases have different mean, STD, and distribution of occupancies, while the NB estimation accurately models the link occupancy distribution. Moreover, it has been observed in [58] that there is a well-defined relationship between the average network occupancy and the STD of individual detector occupancy for Yokohama data. The NB estimation can reproduce a well-defined MFD similar to the MFD based on the field data even if link FD has significant scatter.

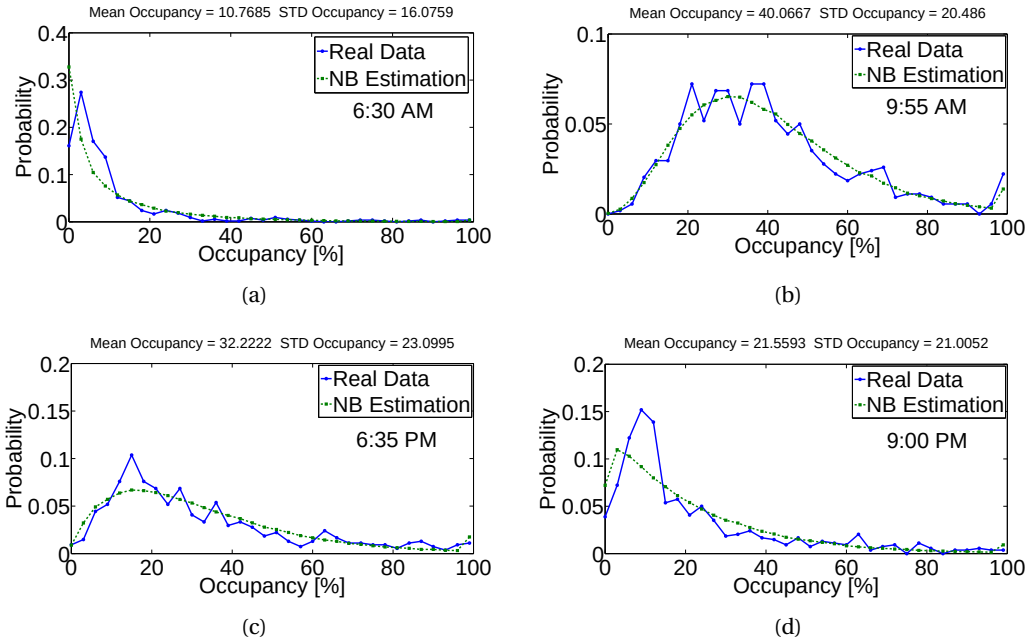


Figure 6.2: Field data and the best NB estimated link occupancy distribution at four different times.

6.3.1 Effect of link occupancy heterogeneity on subregion MFD

Aforementioned observations confirm that the NB distribution can be regarded as a proper estimator of link occupancy distribution in a homogeneous subregion. Nevertheless, given

the low scatter MFD of Yokohama, it is not possible to investigate the effect of heterogeneity for a large range of STD for a given subregion mean occupancy. To succeed this objective, we draw NB distributions with a range of sensible STD, o_{std}, for various average subregion occupancies, i.e. $o_{\text{u}} \in [5\% - 75\%]$, and then estimate the subregion average flow with a low scatter link FD. The outcomes for $o_{\text{u}} = [5, 10, 15, \ldots, 75\%]$ are depicted in Fig. 6.3. To obtain a closed-form expression relating subregion average flow, q_{u}, to mean occupancy, o_{u}, and occupancy STD, o_{std}, we fit an exponential function to the data, i.e.

$$q_{\text{u}}(o_{\text{u}}, o_{\text{std}}) = (d_3 \cdot o_{\text{u}}^3 + d_2 \cdot o_{\text{u}}^2 + d_1 \cdot o_{\text{u}}) \cdot (a \cdot e^{b \cdot o_{\text{std}}} + c), \tag{6.11}$$

where a, b, c, d_1, d_2, and d_3 are estimated parameters. The results reveal the function that is product of a 3-degree polynomial, representative of a low-scatter MFD, and a exponential function, representative of heterogeneity effect, models accurately the MFD. Hence, the same decomposition of MFD function to two terms (i) an upper bound (low-scatter) MFD and (ii) the heterogeneity degradation is used in modeling the effect of link heterogeneity on region MFD, as we describe in next subsection. The results show that in case of light conditions

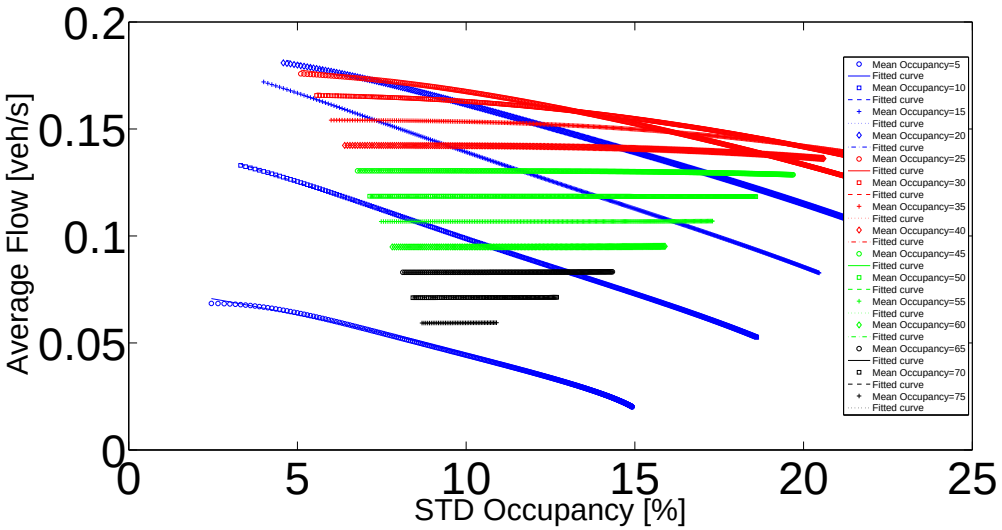


Figure 6.3: Subregion average flow for different mean and STD occupancies.

that the average subregion occupancy is low, increase in STD, i.e. the subregion becomes more heterogeneous, decreases the subregion average flow. As the subregion occupancy increases, the degrading effect of heterogeneity on the subregion average flow becomes less severe. About the subregion occupancy 50%, the best fit becomes almost a line with slope zero revealing that the subregion average flow is independent of the link occupancy heterogeneity.

Note that the validity of theses observations is based on the Yokohama field data which does not comprise very congested situations or particular cases which are unlikely in reality, e.g. a case with low network occupancy and high STD. This limitation prevents us to make a general statement, however with more field data specifically for congested situations a

better understanding of heterogeneity effect on the MFD can be concluded. For example for cases close to gridlock, simulations from [120] showed that the distribution of congestion is bimodal with a fraction of links being around jam occupancy and another fraction close to zero occupancy, which might not be described well by an NB distribution.

6.3.2 Effect of link occupancy heterogeneity on region MFD

To obtain link density heterogeneity for urban region I, we assume that a well-defined relationship between the mean occupancy and the STD of link occupancies for each subregion i exists, $i \in \mathscr{SR}_I$, where $\mathscr{SR}_I$ denotes the set of subregions in region I. Thus, the STD of link occupancies for every subregion i, $i \in \mathscr{SR}_I$, and consequently, the link occupancy distribution, based on the NB distribution assumption, can be estimated given the mean occupancy of subregion i. Afterwards, NB distributions for all $i \in \mathscr{SR}_I$ are summed to capture the link occupancy distribution in region I and the STD of summation of NB distributions is an approximation of $\sigma(N_I(t))$. In a direct analogy with Eq. (6.11), the MFD production of region I considering the link heterogeneity in region I is

$$\begin{aligned} P_I(N_I(t),\sigma(N_I(t))) = &|\mathscr{SR}_I| \cdot \Big(D_3 \cdot \big(\frac{N_I(t)}{|\mathscr{SR}_I|}\big)^3 + D_2 \cdot \big(\frac{N_I(t)}{|\mathscr{SR}_I|}\big)^2 + D_1 \cdot \big(\frac{N_I(t)}{|\mathscr{SR}_I|}\big)\Big) \times \\ &\Big(A \cdot e^{B\cdot(\sigma(N_I(t))-\sigma_{\mathrm{h}})} + (1-A)\Big), \end{aligned} \tag{6.12}$$

where $|\mathscr{SR}_I|$ denotes the number of subregions in region I, σ_{h} is the STD of summation of $|\mathscr{SR}_I|$ NB distributions with mean occupancy $N_I(t)/|\mathscr{SR}_I|$, and A, B, D_1, D_2, and D_3 are the estimated parameters that regulate the extent of link density heterogeneity effect on the region production. Note that (6.12) assumes that the region I production can be regarded as the product of two terms, the exponential term considering the heterogeneity and the production term which assumes homogeneous conditions corresponding to the upper bound (low-scatter) MFD. Figure 6.4 illustrates the exponential term (blue curves) for different region mean occupancies, where red bars depict the range.

It is clear that such a functional form can approximate well the heterogeneity in the spatial distribution of congestion and it can be integrated in the remaining of the study to develop more advanced perimeter control strategies that can treat explicitly this effect.

6.4 Hierarchical control for heterogeneous networks

The previous section has provided, based on real-data analysis, some findings related to the effect of heterogeneity on MFDs, which result in integrating heterogeneity dynamics in urban network modeling. This section aims at utilizing the integration of heterogeneity dynamics for control purposes.

The optimal perimeter control problem formulation and solution have been introduced for homogeneous networks, showing that applying a perimeter control strategy can improve the

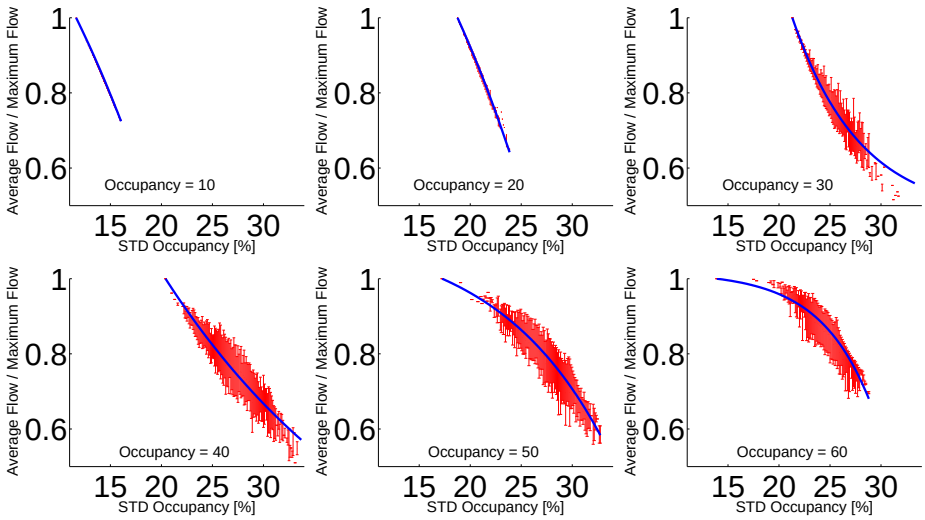


Figure 6.4: The exponential term (blue curves) for different region mean occupancies, where red bars depict the range. Maximum flow is the flow associated with the homogeneous condition, where the heterogeneity exponential term is equal to 1.

network performance of urban regions. Physically speaking a perimeter control strategy that assumes that all sub-regions in the one or the other boundary of the control have equal accumulations, it might move to erroneous interpretations if this is not the case. Applying similar control restrictions (e.g. same amount of transfer flows) in sub-regions with significantly different levels of congestion, might further increase congestion in some of them. Thus, ignoring the effect of heterogeneity in the development of control might lead in non-optimal results as we will show later. In Chapters 4 and 5, an MPC approach solution has been applied to minimize the total network delay, without considering the effect of the regions heterogeneity.

While this might be successful when congestion is uniformly distributed, not all cities have such a property. Clustering algorithms have been proved efficient in decreasing the spatial heterogeneity, but in principle this is a feature of mobility that cannot totally disappear due to complex demand characteristics. Similar to homogeneous networks, in the current study the aim of perimeter control for heterogeneous networks is also to minimize the total network delay. However, giving the negative effect of heterogeneity in the network flow, one can introduce a new control scheme that is crafted for heterogeneous networks. In the following, we introduce *a hierarchical perimeter control framework* for heterogeneous networks, having two levels of control: a high-level controller aims at minimizing total delay with the help of MPC approach, and a low-level (feedback) controller that aims at minimizing the region accumulation heterogeneity.

6.4.1 The high-level (MPC) controller

The aim of optimal perimeter control for heterogeneous networks is to minimize the network delay, defined as the integral of the network accumulation with respect to time, by manipulating the perimeter controllers. We utilize the MPC approach to solve the optimal control problem. Nevertheless, in the previous Chapters 4 and 5 the model and the plant were very similar with the only difference being some unknown stochastic term in demand. Developing an MPC framework with a very different model than plant is challenging and will shed more light in the possible application of such approaches in real life, where many network characteristics are unknown (e.g. route choice and sub-regional O-D tables).

While microscopic simulation might be an alternative choice for the plant, this chapter chooses the direction of two different types of aggregation (sub-regional and regional model), which allows to create further insights of the dynamics of heterogeneity and the methodological framework of traffic flow analysis. Other works (see for example, [97, 4]) have shown in a microsimulation environment that perimeter control strategies can significantly decrease network delays.

Both models, the subregion- and region-based models, are utilized in the MPC framework. The subregion-based model describes the traffic flow dynamics in detail (MPC-plant), while the region-based model is utilized to calculate the optimal control inputs in the optimization loop (MPC-model). Recall that the subregion-based model describes in more detail the mass conservation dynamics based on subregional MFDs that also integrates the constraints on the transfer flows by the receiving capacity, while the region-based model is the MPC model that is suitable for performing tractable optimization. Note that the region-based model considers the effect of link heterogeneity, while this information is provided by the subregion-based model. Some of the variables of the sub-regional model might require significant estimation efforts and high density of sensors, which make the real-time implementation challenging. Nevertheless, the regional model, which is utilized in the optimization framework, is based on information that can be estimated readily with standard monitoring and sensing techniques.

The MPC controller determines the optimal control inputs in a receding horizon manner, meaning that at each time step an objective function is optimized over a prediction horizon of K_p steps and a sequence of optimal control inputs are derived. Then the first sample of the control inputs is applied to the system and the procedure is carried out again with a shifted horizon. The closed-loop optimal control scheme in the MPC framework takes into account not only the errors between the plant and the model, but also the disturbances, e.g. variations in the expected demands, that might affect the system.

The optimal control problem is directly formulated as an MPC problem. Let k_c (−) and T_c (s) be the control time step and the control sample time, respectively. Then, the overall optimization

problem is formulated as follows:

$$\min_{\tilde{U}_{IJ}(k_c),\tilde{U}_{JI}(k_c)} \quad T_c \cdot \sum_{I\in\mathscr{R},J\in\mathscr{H}_I} \sum_{0}^{K_p-1} N_{II}(k_c) + N_{IJ}(k_c) \tag{6.13}$$

subject to

$$N_{II}(k_c+1) = N_{II}(k_c) + T_c \cdot \big(Q_{II}(k_c) - M_{II}(k_c) + \sum_{J\in\mathscr{H}_I} U_{JI}(k_c)\cdot M_{JI}(k_c)\big), \tag{6.14}$$

$$N_{IJ}(k_c+1) = N_{IJ}(k_c) + T_c \cdot \big(Q_{IJ}(k_c) - \sum_{J\in\mathscr{H}_I} U_{IJ}(k_c)\cdot M_{IJ}(k_c)\big), \tag{6.15}$$

$$U_{IJ,\min} \le U_{IJ}(k_c) \le U_{IJ,\max}, \tag{6.16}$$

for $I = 1,2,\ldots,R$ and $\forall J \in \mathscr{H}_I$.

The problem (6.13)–(6.16) is a nonlinear optimization problem and it can be solved using nonlinear optimization algorithms. $U_{IJ,\min}$ and $U_{IJ,\max}$ $(-)$ are respectively the lower and upper bounds for the perimeter control inputs between regions I and J. The optimization variables defined over the prediction horizon K_p are $\tilde{U}_{IJ}(k_c) = [U_{IJ}(k_c),\ldots,U_{IJ}(k_c+K_p-1)]^T$, where $U_{IJ}(k_c+l)$ for $l = 0,\ldots,K_p-1$ are the perimeter control inputs obtained by the MPC framework at every control time step k_c. The following subsection further elaborates the low-level feedback controller, which utilizes the high-level control inputs such that the regions become more homogeneous.

6.4.2 The low-level feedback homogeneity controller (FHC)

The goal of the high-level (MPC) controller of the hierarchical control framework is to minimize the total network delay. However, there is no explicit consideration of regional accumulation heterogeneity in the high-level control. We also aim at minimizing the regional heterogeneity, and to achieve this goal, a low-level feedback homogeneity controller (FHC) is introduced, where the high-level MPC controller defines the set values for the control inputs such that the MPC control inputs, i.e. the regional transfer flow percentage U_{IJ}, are forwarded to the FHC. The FHC determines the subregional perimeter controls, u_{ij}, to control the subregional accumulations and minimize the accumulation heterogeneity. Note that subregions, which are not attached to the boundary between the regions cannot be directly controlled, e.g. subregion 19 in Fig. 6.1(c). While n_{ij} variables might be more difficult to estimate, subregional accumulations without information for the final destination, i.e. n_i are easier.

The high- and low-level controllers are not conflicting, but are complementary. The FHC will try to homogenize the region so that circulating flow increases for trips within the region, independently if the MPC controller increases or decreases the transfer flow. This hierarchical scheme can have significant benefits in real-life applications, where the perimeter controller might create strong local heterogeneities and spillbacks for the intersections in the proximity of the border. More homogeneous networks can improve both the average travel time (because of higher network outflow), but also the travel time reliability, as shown by [117].

The FHC is a state feedback controller that its control law is based on the feeded subregional accumulations n_i. The FHC aims at manipulating the subregional controllers u_{ij} to bring subregional accumulations as close as possible to desired accumulations. Defining appropriate desired accumulations should achieve our control goal to homogenize the whole region. It is clear that defining desired accumulations is not a trivial task, since they are not a priori known and may change over time. However, utilizing information from the high-level controller helps us in this task. Given that the high-level controller at each time step k_c predicts the regional accumulation states for K_p step ahead, these predicted accumulations can be considered, after dividing them by the number of subregions within the region, as the subregional accumulation set points for the FHC. The FHC control law for u_{ij} should consider both subregional accumulations n_i and n_j, because u_{ij} affects both subregional accumulations. Hence, the following control law, which is a multivariable integral discrete controller, is proposed:

$$\begin{aligned} u_{ij}(k_c) = u_{ij}(k_c-1) &+ K_1 \cdot \Big(\frac{N_J(k_c+K_p-1)}{|\mathscr{S}\mathscr{R}_J|} - n_j(k_c) \Big) \\ &- K_2 \cdot \Big(\frac{N_I(k_c+K_p-1)}{|\mathscr{S}\mathscr{R}_I|} - n_i(k_c) \Big), \end{aligned} \tag{6.17}$$

where K_1 and K_2 are positive designed parameters. Note that $N_I(k_c+K_p-1)/|\mathscr{S}\mathscr{R}_I|$ and $N_J(k_c+K_p-1)/|\mathscr{S}\mathscr{R}_J|$ are the controller set points that change over time. Following the regulating problem, the control gains K_1 and K_2 are designed in this chapter assuming that the set points are known constant. When we apply the designed controller, it might be unstable as the set points change over the time, but in this problem since the regional accumulations change smoothly with time (as we see later in the results section) and do not experience strong fluctuations, this allows the controller to be effective in tracking the time-varying set points.

The MPC output, U_{IJ}, can be applied in two different options: (i) where each u_{ij}, $i \in I$ and $j \in J$, is equal (or be very close) to U_{IJ} without considering that each sub-region can be treated differently, and (ii) where the collective effect of u_{ij}, $i \in I$ and $j \in J$, is almost equal to U_{IJ}, i.e.

$$\Big| \frac{\sum_{i\in I, j\in J} u_{ij}}{|u_{IJ}|} - U_{IJ} \Big| < \delta, \tag{6.18}$$

where $|u_{IJ}|$ denotes the number of subregional controllers between region I and J, and δ is a prescribed positive value, e.g. 0.2. The second option empowers the controller to manipulate each u_{ij} differently and individually in order to minimize the regional heterogeneity. We utilize the second option in the hierarchical framework and compare numerical results, in the following section, with and without considering the low-level controller to highlight its importance. It is known that feedback regulators of type (6.17) cannot handle directly constraints in the optimization. The FHC first determines the u_{ij} s based on (6.17), and then truncates the control outputs to satisfy (6.16) and then if necessary add or subtract a value from all u_{ij} s to also satisfy (6.18).

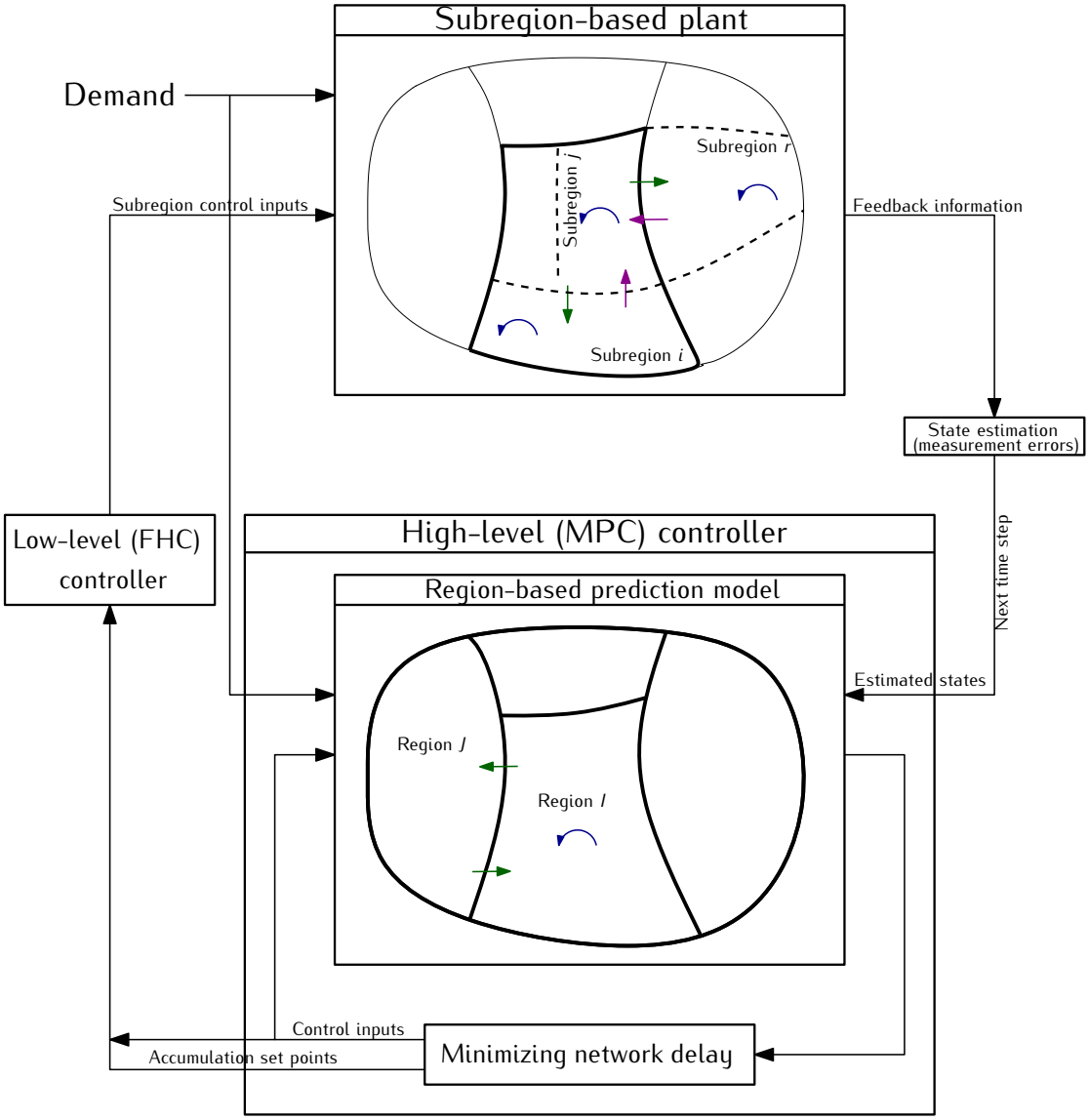


Figure 6.5: The hierarchical perimeter control framework.

6.5 Comparison of control strategies

In this section, we present a case study example to explore the characteristics of the proposed region-based and subregion-based models along with the hierarchical control scheme. Moreover, we investigate the effect of heterogeneity controller, the low-level FHC, on the control strategy performances. Note that the main modeling contribution of this chapter is developing two different models with different scales of aggregation and utilize them in the MPC framework as the prediction model and the plant, in contrast to Chapters 4 and 5 in which the dynamics of model and plant in the MPC frameworks were inherently similar, but the demand prediction errors and the MFD noisy scatter distinguish between them. With

respect to control, we highlight the importance of the hierarchical multi-layer control structure enhanced with advanced traffic modeling. This section emphasizes that if some parts of the modeling and/or the control are not fully implemented, the network traffic states are worse, especially under congested scenarios.

The case study network consists of two regions, designating the periphery and city center of an urban network, each comprises of 12 and 7 subregions, respectively, as schematically shown in Fig. 6.1(c). Without loss of generality, we assume every subregion has the same MFD (production) consistent with the MFD (production) observed in Yokohama, and consequently, the well-defined relationship between mean and STD of subregion link occupancy exists, and the subregional average trip length is constant. Note that the *region* average trip lengths are varying as the model evolves, see equations (6.9a) and (6.9b).

In all numerical runs, every subregion accumulation is initially identical and uncongested that results in initial regional accumulations as $N_1(0) = 29000$ (veh) (71% of N_1^{crit}) and $N_2(0) = 19000$ (veh) (80% of N_2^{crit}), where N_I^{crit} denotes the region I accumulation that maximizes the production MFD, P_I. To model the observation errors, a normal random component is added to the measurements from the plant (subregion-based model), i.e. the errors are introduced in $N_{II}(t)$, $N_{IJ}(t)$, $L_{II}(t)$, and $L_{IJ}(t)$, see Fig. 6.5. Also, a uniform random component is added to the perimeter control sequence to model the flow stochasticity of supply and demand at the boundary of regions, see the small fluctuations in Fig. 6.10(a)–6.10(d).

The exogenous time varying demand, shown in Fig. 6.6(a), simulates one hour of morning peak followed by two and half hours of low demand to fully clear the network, while region 1 generates most of the demand towards region 2 that as the central business district attracts trips. Moreover, we consider medium and high demands, where in the high demand the exogenous demand from region 1 to region 2 is 10% higher than the medium demand shown in Fig. 6.6(a). The selected MPC controller parameters are as follows: the prediction horizon $K_{\text{p}} = 20$, the control lower bound $U_{\min} = 0.1$, and the upper bound $U_{\max} = 0.9$.

We compare five control strategies that are essential to investigate the importance of proper heterogeneity modeling and hierarchical control: (i) no control where there is no restriction on the perimeter transfer flows, (ii) MPC, only the high-level control, ignoring the modeling of heterogeneity and assuming an MFD without any hysteresis (similar to [49]), (iii) MPC+HET, only the high-level control that utilizes the heterogeneity modeling of Section 6.2, (iv) MPC+FHC, that is the hierarchical control ignoring the heterogeneity modeling by assuming an MFD with no hysteresis, and (v) MPC+HET+FHC, that is the hierarchical control structure that applies MPC at the upper level and FHC at the lower level considering the heterogeneity modeling. These extensive tests enables to highlight the importance of heterogeneity modeling and control on the performance of proposed traffic control for heterogeneous urban networks.

Figure 6.6(b)–6.6(f) depict the evolution of subregional accumulations $n_i(t)$ over the simulation duration for the five control strategies in case of medium demand. The gridlock is apparent for no control case as shown in Fig. 6.6(b), while the rest of strategies manage to

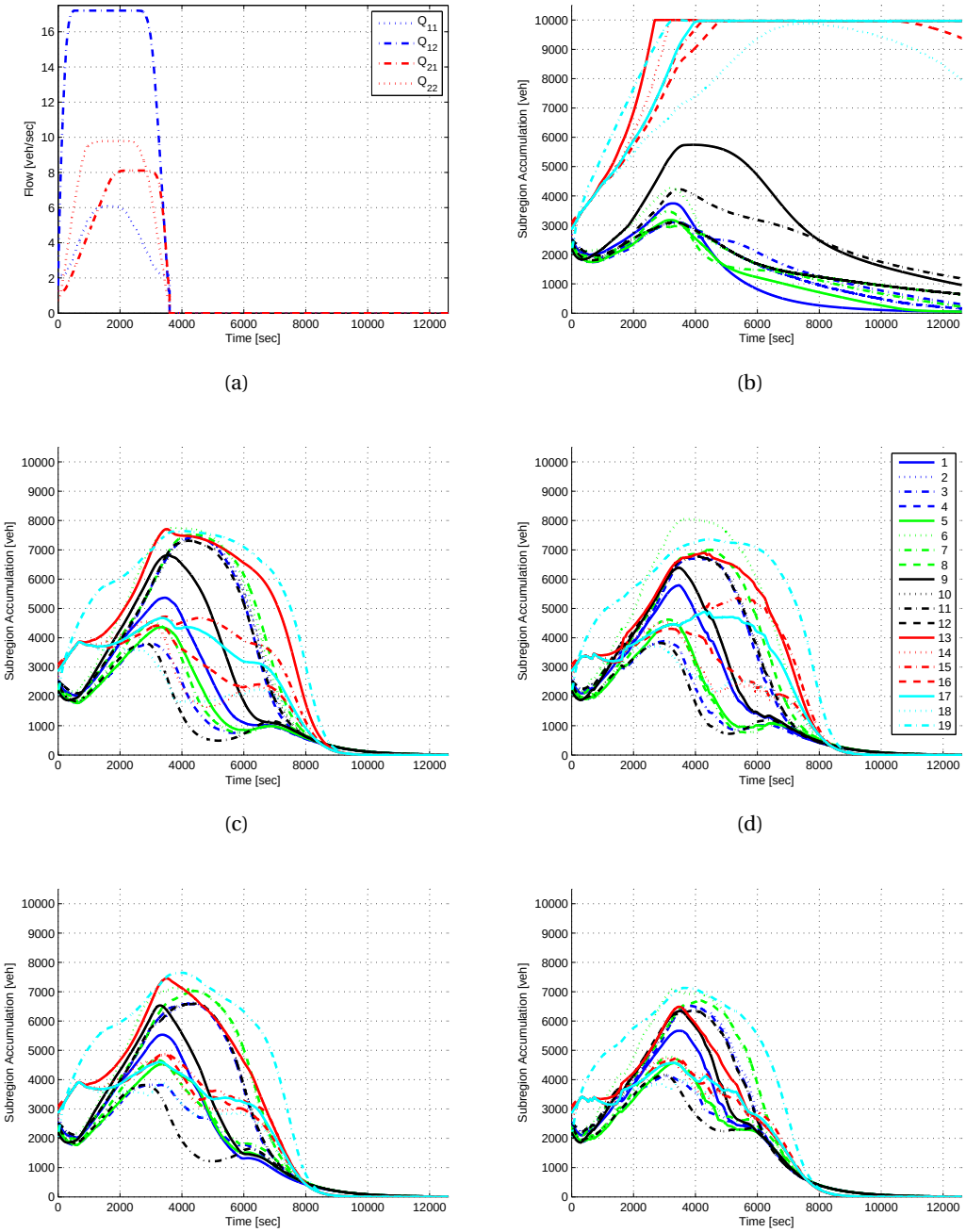


Figure 6.6: The case study example with medium demand: (a) the time-varying demand profile, and subregional accumulations with (b) no control, (c) MPC, (d) MPC+HET, (e) MPC+FHC, and (f) MPC+HET+FHC. Note that (b)–(f) have similar legends.

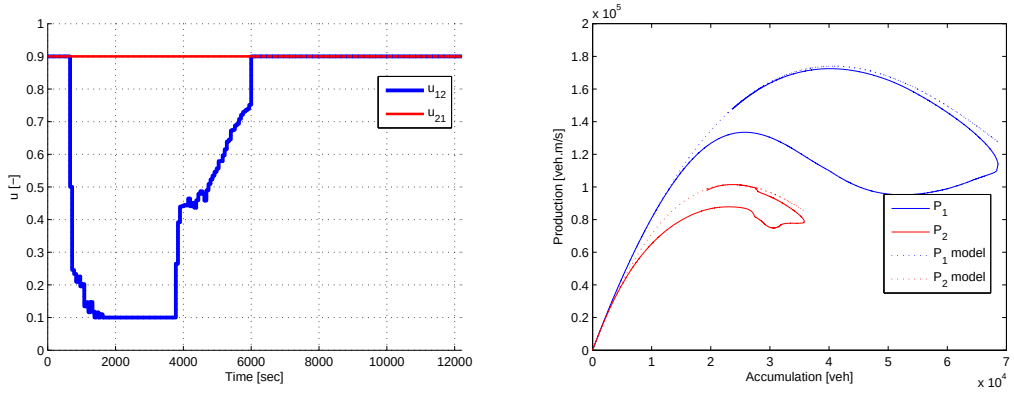
clear the network, see Fig. 6.6(c)–6.6(f). Figure 6.6(b) also demonstrates that the accumulation of each subregion cannot exceed the jam accumulation, i.e. 10000 (veh), which is captured by considering the receiving capacity of subregions in the subregion-based model. A more careful investigation is necessary to compare the different control strategies. Table 6.1 lists the total network delay (averaged over 5 runs) for the control strategies with the two different levels of demand where the values in parenthesis designate the improvement over the MPC strategy without heterogeneity modeling and without lower level control. It is apparent that utilizing the heterogeneity modeling without integrating the FHC controller to decrease the level of heterogeneity, will improve the system delays, but obviously is worse than the two-level hierarchical control. In addition, the hierarchical control without heterogeneity modeling (i.e. MPC+FHC) provides similar performance with MPC+HET. Thus a careful consideration of heterogeneity in the modeling and control frameworks is essential to decrease the total network delay.

Table 6.1: Total network delay ($\text{veh} \cdot \text{sec} \cdot 10^6$).

Demand	No Control	MPC	MPC+HET	MPC+FHC	MPC+HET+FHC
Medium	1069.4	573.8	546.2	541.9	518.0
		(-)	(4.8%)	(5.6%)	(9.7%)
High	1204.4	930.5	881.5	851.0	636.8
		(-)	(5.3%)	(8.5%)	(31.6%)

To understand why the strategies provide different performance, we initially investigate the MPC strategy. The MPC control strategy ignores the heterogeneity effect in the regional MFD, i.e. the exponential term in (6.12) is equal to 1. Figure 6.7 depicts the control sequences and MFD for medium demand with the MPC control strategy, where strong hysteresis loops are apparent and notably the model MFD is identical in the loading and unloading phases. The Table 6.1 results indicate 10% and 30% increase in control performance with the MPC+HET+FHC strategy over the MPC strategy respectively in case of medium and high demand. Moreover, assuming the lower envelop of MFD as the optimization model, i.e. the exponential term in (6.12) is equal to its maximum value, results in worse outcomes. Note that [97, 98] are able to keep accumulation at the critical values but cannot avoid hysteresis loops in the MFD with control. Avoiding hysteresis loops can be very beneficial for the overall network performance as we see in the analysis.

Figures 6.8, 6.9, and 6.10 highlight in detail the importance of the proposed hierarchical control scheme. Note that the following illustrative comparisons are between the MPC+HET control strategy and the two-level hierarchical control strategy (MPC+HET+FHC). MPC+HET is superior to the MPC strategy which has been shown to be superior to standard simple control strategies, such as the "bang bang" control approach [49], thus this is a strict test. Figure 6.8 presents time-series of accumulations, Fig. 6.9 shows the corresponding MFDs, and Fig. 6.10 demonstrates the control actions. This is described in more detail in the following paragraphs.



(a) (b)

Figure 6.7: The case study example for medium demand with the MPC control strategy (without heterogeneity consideration in modeling): (a) control inputs and (b) MFD productions.

The corresponding regional accumulations are illustrated in Fig. 6.8 for region 1 (center) and region 2 (periphery), while the lines above (or below) of the accumulation curve represent plus (or minus) one STD of the accumulation, a heterogeneity index. The accumulation results obtained with the MPC+HET and MPC+HET+FHC control strategies for the medium demand are depicted in Fig. 6.8(a) and 6.8(b), while the results for the high demand are depicted in Fig. 6.8(c) and 6.8(d), respectively. Accumulations look almost identical in the onset of congestion when heterogeneity index is small (remember that initial conditions have small spatial heterogeneity), however the mean accumulation and the accumulation heterogeneity are different in the offset of congestion. For region 2, the duration of congestion period is shorter and this influences also region 1 in the offset of congestion. Note that region 2 attracts more trips than region 1, even if it has smaller size. In case of high demand scenario with MPC+HET strategy, some subregions face gridlock and the network is not fully cleared, note the residual accumulation at the end of simulation. As MPC+HET+FHC can avoid high level of congestion in subregions, the improvement over MPC+HET is 28%. If initial conditions are also more heterogeneous an even better performance is expected for the advanced controller. While accumulations do not provide a full picture on how the hierarchical control framework improves the mobility levels, this is clear once MFDs are described for the regions.

The production MFDs obtained with the MPC+HET and MPC+HET+FHC strategies for the medium demand are shown in Fig. 6.9(a) and 6.9(b), and for the high demand in Fig. 6.9(c) and 6.9(d), respectively. Note that the figures show the MFDs obtained by the subregion model (noted as P_1 and P_2) and the MFDs of the region model (noted as P_1 model and P_2 model) as estimated by (6.12). It is clear that (6.12) can capture well the effect of heterogeneity in the production MFD of the regions. Evidently, FHC improves the performance of the urban network by minimizing the extent of the hysteresis in both region MFDs during the unloading of the network, as shown in Fig. 6.9. Hysteresis can be considered as a strong inefficiency of the system as for the same level of vehicles in the network, the performance is worse and

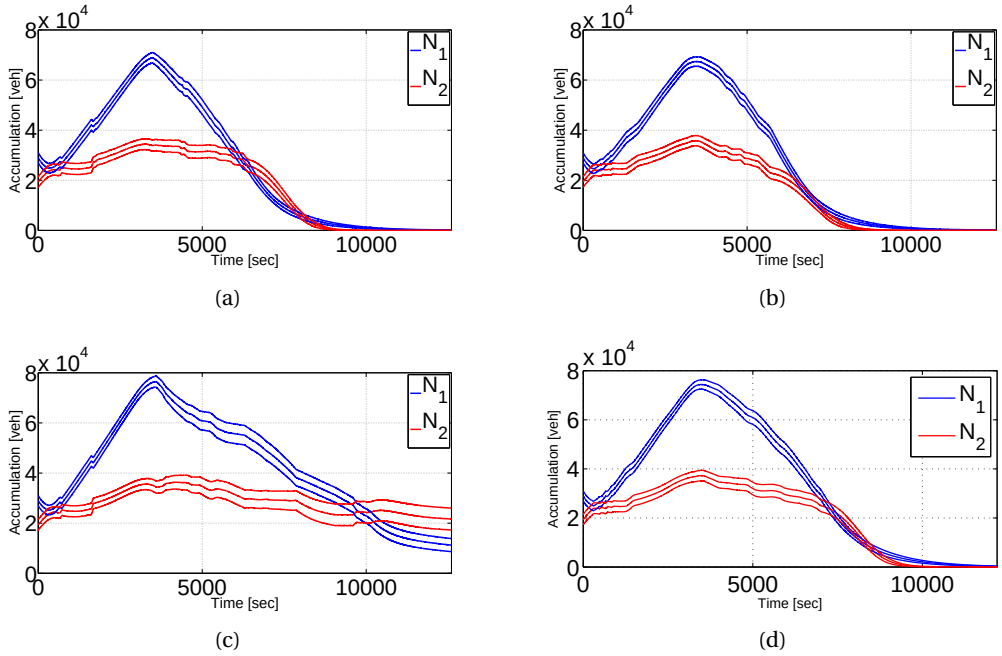


Figure 6.8: Accumulation results obtained with MPC+HET control strategy and MPC+HET+FHC for the medium demand in (a) and (b), and for the high demand in (c) and (d), respectively.

vehicles have to spend in the system longer times. Note that while the controller tries to protect region 2, it also succeeds to improve the production (and also the trip completion) for region 1, by decreasing the level of hysteresis to small values. This is succeeded by distributing accumulations in a more uniform pattern by controlling the subregional inter-transfers. This is an important finding as many previous investigations in the shape of the MFD with real data and simulations observed hysteresis loops in many cases that result in a decrease in the network performance. It is worth to mention that, the production and outflow MFDs are related with the time-varying region average trip length. This variable affects the dynamics of MFDs, notably the hysteresis loop shape, i.e. clockwise vs. counterclockwise. Average trip length has a significant role in our understanding of urban traffic dynamics and reveals substantial information on the human mobility pattern in urban areas. Further research on this direction could shed more light on properties of urban average trip length distribution.

Figure 6.10 depicts the control sequences of MPC+HET and MPC+HET+FHC strategies at the regional level (i.e. U_{12} and U_{21}) for medium (a and b) and high demand (c and d). Each figure contains the estimated values of U_{12} and U_{21} by the optimization model (6.13)–(6.16), while U_{12} applied and U_{21} applied show the ones implemented utilizing the FHC (6.17)–(6.18). Fig. 6.10(a) and 6.10(c) do not contain the FHC control thus the two lines are almost identical (except some small random error during implementation). Nevertheless, in the FHC case (Fig. 6.10(b) and 6.10(d)) differences are substantial as the FHC tries to equalize subregion accumulations and it is allowed to deviate by the factor δ (see (6.18)) from the estimated values. The control sequences show similar trends. For instance, at the very beginning of

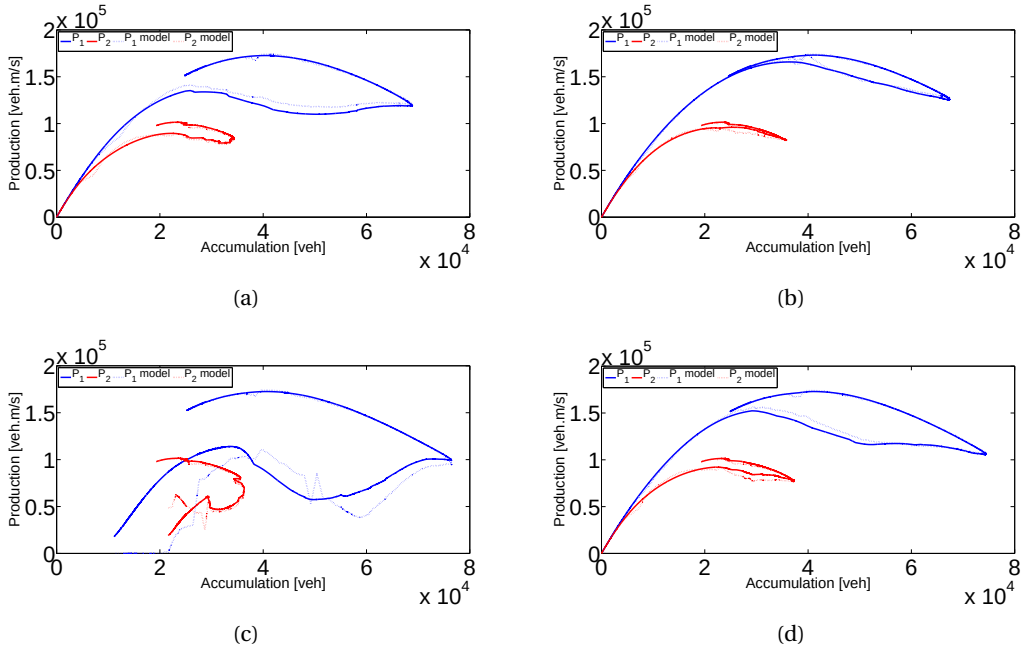


Figure 6.9: Production MFD for MPC+HET and MPC+HET+FHC for the medium demand in (a) and (b), and for the high demand in (c) and (d), respectively.

the control process, the controllers do not restrict inter flow transfers since both regions are uncongested. While afterwards, as region 2 becomes more congested and attracts more trips, the controllers attempt to control region 2 accumulation by changing U_{12} from $U_{\max}$ to $U_{\min}$ in a smooth manner, since without any restriction region 2 will face gridlock, see the subregional accumulation with no control in Fig. 6.6(b). The overall situation remains invariant till the end of morning demand $t = 3600$ (s), then because of decrease in the demand (unloading phase), regions shift towards the uncongested state. Thus, the controllers gradually permit more vehicles to enter to the city center, region 2, by altering U_{12} to $U_{\max}$. Though the similar trend, the MPC+HET+FHC applied control sequence (dashed lines in Fig. 6.10(b) and 6.10(d)) deviate around the MPC control value, see (6.18). The FHC modifies the MPC control sequence to manipulate transfer flows between subregions. This offers a flexible framework to directly control the heterogeneity. The dashed lines are the applied perimeter control inputs that achieve the two-fold objectives of the hierarchical (MPC+HET+FHC) control strategy, to minimize the total network delay by simultaneously maximizing the network outflow and minimizing the regional accumulation heterogeneity. We test the same experiment study with δ equal to infinity, see (6.18), which means no restriction on FHC to follow the MPC control values. This results in less heterogeneity in regions however the total network delay is worse than the case with $\delta = 0.2$. The outcome demonstrates that selecting the goal of the controller solely to make regions more uncongested in a network and ignore other aspects (e.g. overall network state, future prediction of traffic state, etc.) is not beneficial.

While results in Fig. 6.8–6.10 focus on the regional characteristics of the system, to further

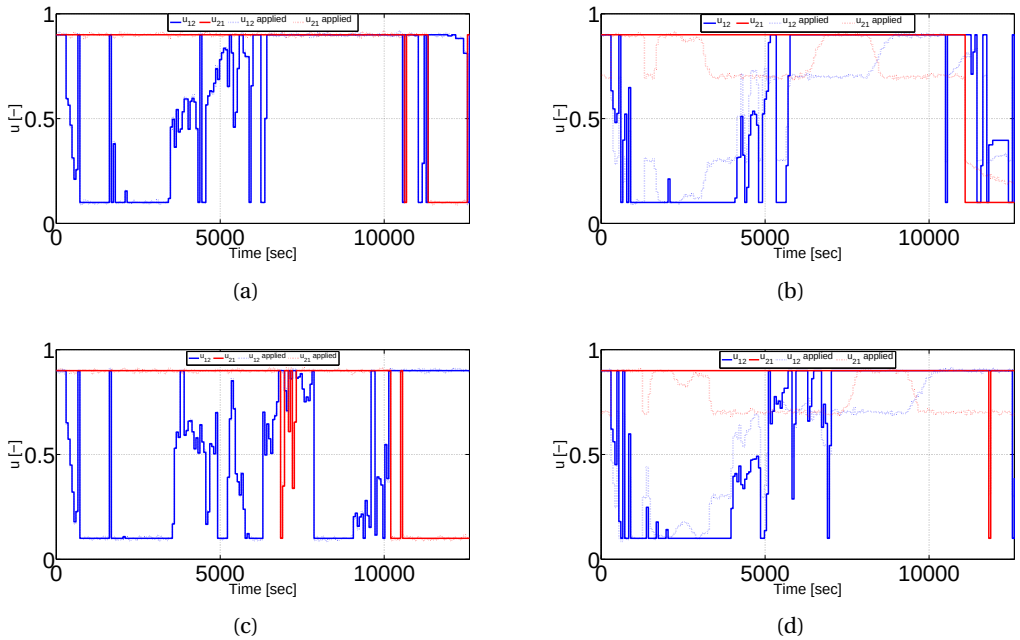


Figure 6.10: Control input results obtained with MPC+HET and MPC+HET+FHC for the medium demand in (a) and (b), and for the high demand in (c) and (d), respectively.

investigate the effect of FHC, the accumulation of subregions in region 2 for high demand case with MPC+HET+FHC (left column) and MPC+HET (right column) are shown in Fig. 6.11. Apparently, MPC+HET+FHC brings the accumulation of all subregions to zeros by the end of the experiment whereas without FHC, subregion 13 goes to gridlock. To investigate the dynamics of subregion 13, Fig. 6.11(c) and 6.11(d) depict subregion 10, 11, 12 (the neighbours of subregion 13 in region 1) accumulations and Fig. 6.11(e) and 6.11(f) depict control inputs between subregion 13 and subregion 10, 11, 12, respectively for MPC+HET with and without FHC. As expected, control inputs between region 1 and region 2, i.e. $u_{10,13}, u_{11,13}, u_{12,13}$, are identical in case of MPC without FHC (similarly for control inputs between region 2 and region 1). However, FHC provides a traffic-responsive perimeter control strategy that is based on subregion traffic states to distribute the traffic congestion more efficiently among subregions. Consequently, FHC homogenizes region 2 accumulation and stabilizes subregion 13 accumulation by manipulating control inputs (see Fig. 6.11(e)), such that the dispersion of subregion accumulations over time is smaller compared to Fig. 6.11(b). It is evident that subregion 19 accumulation is less consistent with other subregions as it is uncontrollable with perimeter controllers. Note that in Fig. 6.11(e) and 6.11(f) the noise in control input is eliminated for illustration purposes.

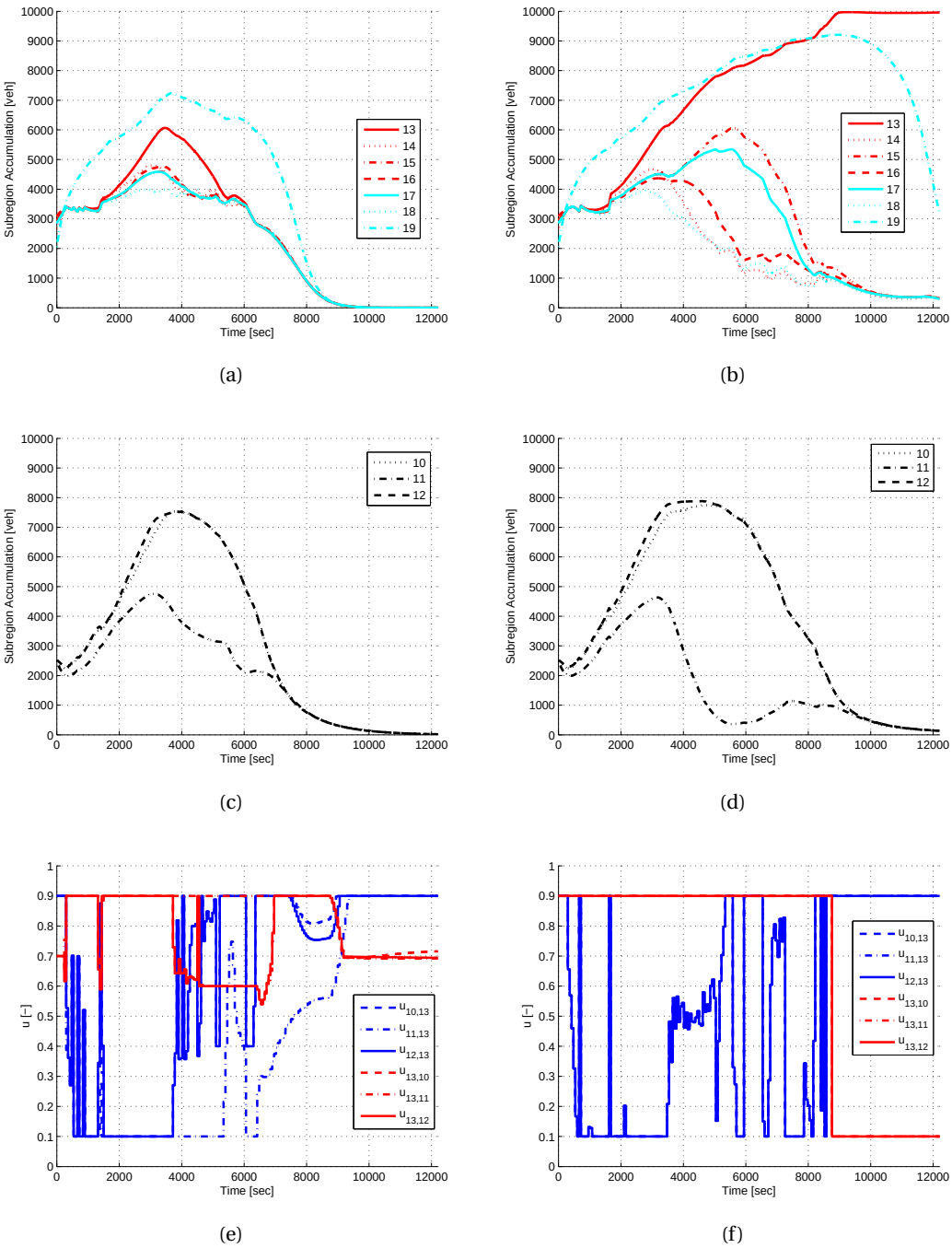


Figure 6.11: The case study example with high demand: (a) subregion accumulations in region 2 with FHC, (b) subregion accumulations in region 2 without FHC, (c) subregion 10, 11, 12 accumulations with FHC, (d) subregion 10, 11, 12 accumulations without FHC, (e) control inputs between subregion 13 and 10, 11, 12 with FHC, and (f) control inputs between subregion 13 and 10, 11, 12 without FHC.

6.6 Summary

This chapter has presented two urban traffic models based on the MFD at different levels of spatial aggregations to model the dynamics of density heterogeneity. A heterogeneous urban region can be partitioned into homogeneous subregions as the detailed model aims at modeling the accumulation dynamics of subregions, while the dynamics of urban regions are modeled in an aggregated manner.

We utilize the subregion- and region-based model as the plant and the optimization model in the MPC framework to formulate the optimal perimeter control for urban regions. We integrate variable perimeter control inputs for each subregion in the region boundary to actively control the density heterogeneity. The results in this chapter can be utilized to develop efficient hierarchical control strategies for heterogeneously congested cities. A challenging modeling direction is how to describe the aggregated modeling dynamics of regions, i.e. (6.2) and (6.3), when routes pass through the same regions more than once (e.g. a trip sequence of subregions in Fig. 6.1).

7 Conclusions and future research

THIS dissertation thoroughly studied traffic state estimation methods based on probe vehicle data in Part I (Chapters 2 and 3) and investigated a hierarchical control strategy for heterogenous large-scale urban networks in Part II (Chapters 4, 5, and 6). This final chapter briefly summarizes the findings and main contributions of each part, elaborates the potential field applications, and outlines the directions for the future research. The detailed conclusions and contributions are also provided at the end of each chapter.

7.1 Part I: Traffic state estimation

7.1.1 Queue profile estimation with probe data

Chapter 2 main contributions are as follows:

- A queue shockwave profile estimation method is proposed.
- The estimation method utilizes probe vehicle data within an integrated framework of LWR traffic theory and data mining techniques.
- The method relaxes the assumption of uniform arrival flow and explicit knowledge of signal settings.
- The method is based on an optimization formulation that incorporates the physics of traffic, longitudinal kinematics, and traffic state interdependencies between adjacent links.
- The optimization part of the proposed method is formulated as a least squares problem that is inherently robust to noisy measurements.
- The proposed method is applicable to oversaturated conditions containing a probabilistic procedure of spillover occurrence inference.

The proposed traffic estimation method provides comprehensive dynamics of the urban traffic flow that can be utilized for delay analysis (mean and distribution) and estimation of queue

length, arrival flow distribution, and signal setting. The proposed method can be also used for vehicle trajectory reconstruction that is not limited only to probe vehicles but particularly can be applied to all vehicles. Estimation of detailed trajectory of vehicles further provides (rough) inputs to emission or fuel consumption estimation models. Nevertheless, further research is needed to investigate and improve the accuracy of these models. Estimation of vehicles trajectories also enables to tackle the travel time decomposition problem [77], which is about how to decompose the travel time between two successive probe data with high sampling interval (e.g. 3 min) to individual link travel time. This is another future direction of research.

A future research priority is to extend the queue profile estimation model to consider multilane roads, where in case of lanes with different channelization characteristic (e.g. left turn), the queue development and dissipation varies among different lanes. The challenge is that while the input data of the proposed method includes location and velocity of probe vehicles, the state-of-practice technology is immature to provide accurate enough lane specific measurements. Further research can attempt to infer the lane features of probe vehicles based on their path characteristics [147] or possible data fusion with loop detector data that are able to provide lane information. In addition, for queue shockwave estimation in multimodal networks, vehicle classifications based on loop detector data [189] or probe data is an intermediate step.

Even though the method does not require arrival traffic flow distribution, outflow measurements from upstream input links can provide significant information for queue estimation in case of limited probe data. Data fusion of sparse probe data with loop detector data (e.g. see [175, 179]) can also facilitate the developed method in case of small penetration rates (e.g. less than 5%) and improve the accuracy of the model. Integration of the proposed method in real time traffic control is a tall order, but a feasible solution given the new era of data for vehicle positioning and crowdsourcing. Existing traffic control strategies that require queue size information (e.g. see [181, 2]) can utilize this method.

7.1.2 Route travel time distribution estimation

The main contributions of Chapter 3 are summarized as:

- A travel time distribution (TTD) estimation method is proposed.
- The estimation approach is to address spatial and temporal traffic state correlations in signalized arterials.
- The method utilizes probe vehicle link travel times to infer the arterial route TTD.
- A Markov chain procedure is incorporated into the method to consider traffic progression and correlation in consecutive links.
- A graphical tool is introduced to represent the joint distributions of successive link travel times that in combination with a heuristic grid clustering algorithm provides state definitions, and initial and transition probabilities of the Markov chain procedure.

- The proposed method demonstrates a sound performance capturing the fundamental characteristics of travel time variability even with sparse probe data.

The TTD estimation method provides a thorough travel time reliability performance measure of traffic conditions on signalized arterials. The estimated TTD can be integrated in traffic state monitoring tools for both traffic system managers to evaluate the transport network and provide travelers with routing information. The method is straightforward in implementation and has demonstrated promising performance and accuracy through numerous experiments.

The next crucial step for future studies is to investigate under what traffic conditions and route structures the proposed method significantly outperforms convolution estimation, i.e. when correlation plays a main role in TTD estimation. For instance, in case of very long links because of platoon dispersion and different driver characteristics, correlation might be negligible. More empirical experiments with real data are needed to examine the influence of different traffic regimes. Moreover, the method would be helpful to inspect how the signal timing, offset, site topology, etc., affect the travel time variability. Another major future research is, as GPS data become available in real sites, to determine travel time reliability in urban networks in case of limited data coverage, i.e. investigating adequate area coverage and integration of historical and real-time data.

The proposed method can be integrated in a real-time implementation of TTD estimation. In this case, by utilizing historical data and partitioning the data set for different times of days and level of congestion, one can apply the method for each temporal partition and estimate a priori TTD offline. In the online part of the real-time implementation, as more data become available, a learning procedure should be developed (e.g. based on Bayesian update), which will update the parameters of the model (states and initial and transition probabilities) to improve the estimation accuracy. Moreover, studying the effect of TTD information on drivers' behavior in the route and departure time choice frameworks will shed light to the implementation of TTD estimation for route guidance purposes.

7.2 Part II: Network-level traffic modeling and control

7.2.1 Model predictive Perimeter control for two urban regions

Chapter 4 main contributions are listed as follows:

- A novel large-scale control strategy for urban networks is proposed, where the control strategy identifies the inter-transfer flows between regions of the network to maximize the network output.
- The control problem is tackled by the model predictive control (MPC) scheme.
- The open-loop optimal control problem is solved using the direct sequential (or control vector parameterization (CVP)) method that converts the optimal control problem into a

finite-dimensional nonlinear problem through discretization of the control variables with piecewise constant controls, while the ODEs are embedded in the nonlinear problem.

- The modeling of the optimization model and plant (reality) are based on the macroscopic fundamental diagram (MFD).
- Two methods based on augmenting a constraint into the optimization problem and a penalty term into the objective function are investigated to smooth the control inputs.

The perimeter control strategy is the cornerstone of the hierarchical urban traffic control scheme to tackle traffic congestion in heterogeneously congested cities. The perimeter control strategy provides the upper-level control considering the network-level traffic states to change the spatial distribution of congestion in such a way that the network outflow increases. Thus, it is of great importance to practitioners and city managers to unveil simple and robust signal timing plans to maximize the network capacity and serve the maximum number of people under congested conditions.

Improving the real-time implementation feasibility for networks with multiple regions is a research priority. In addition, further analysis is needed to identify individual signal timing plans in each region of the city to move traffic smoothly at the desired flow rates, without concentrating a large number of vehicles at the boundaries of the regions and avoiding local spillbacks in the proximity of the perimeter [96]. This is a challenging task that requires knowledge on how the network flow for a region of a city changes as a function of topology, control, and level of congestion.

By restricting access to congested cities, one can significantly improve system output, highlighting the importance of a reliable estimator of subnetwork/route capacity. Towards this direction, the work [53] investigates the effect that have in the MFD, different degrees of variability in link lengths and signal characteristics for different city topologies and signal structures. Nevertheless, the effect of perimeter control in the heterogeneity of density in each region (and the boundaries) cannot be investigated with the developed macroscopic plant. An application of these strategies in the field or in a micro-simulation environment can provide more insights towards this direction and identify the necessary local control schemes to succeed smooth boundary conditions. This is a research priority.

7.2.2 Cooperative large-scale traffic control of a network with two urban regions and a freeway

The main contributions of Chapter 5 are as following:

- The control problem of a large-scale mixed traffic network, consisting of two urban regions and one alternative freeway route is formulated. The perimeter controller on the boundary of regions manipulates the inter-transfer flows among regions, while two on-ramp controllers control the traffic flow from urban regions to the freeway.

- The urban region traffic modeling is based on the MFD and the freeway traffic dynamics are based on the asymmetric cell transmission model. The two different models are integrated to form the mixed traffic network dynamics.
- The optimal traffic control problem is solved by the MPC scheme.
- Several control policies with different controller structures and levels of urban-freeway coordination are introduced and investigated.
- The cooperative decentralized MPC approach seems promising as an alternative of the centralized MPC in case of lack of full data communication and coordination between urban and freeway control entities.
- The system optimum route choice is incorporated within the centralized MPC framework that leads to increase the network performance compared to the user equilibrium condition.

This study contributes one step forward toward the system of systems approach in transportation modeling and control framework. The developed decentralized and centralized controllers in this chapter are fitted in the hierarchical traffic control structure to tackle large-scale mixed urban-freeway networks. The findings offre coordinated decentralized control strategies for city traffic managers as an alternative resolution when the centralized approach is intractable because of lack of coordination between urban and freeway jurisdictions.

A future research direction is to investigate traffic control problem of networks with more complex structure and dynamic traffic assignment procedure. Further, the continuation of this research involves the development of control strategies for networks with a larger number of urban regions, on- and off-ramps. In this case, the size of regions might change over time due to congestion propagation that requires a dynamic partitioning of the city. Also a more detailed route choice (specifically ramp choice) procedure is required. This is a challenging problem, both from an optimization (higher computational needs) and modeling (more complex dynamics) points of view.

In this study, the controller outcomes do not provide the exact phase plan settings for traffic signals on the boundary of the regions. Nevertheless, recent works for single [97] and multiple regions [4] address the tools to dynamically modify the signal settings to meet the controllers' input actions. In case, the controller produces local queues in the proximity of the control locations (ramps and boundaries between urban regions), analysis of [53] can identify signal parameters in each region of the city to manage traffic smoothly at the desired flows, without concentrating a large number of vehicles at the boundaries of the regions. However, thorough investigations in integration of local queue equalization strategies within the proposed control method is essential.

In addition, a field implementation and/or detailed micro-simulation of proposed control policies are also important to identify the effect of the control and the heterogeneity of spatial distribution of congestion on the shape of the MFD and the efficiency of the control method.

Another research direction is related to the traffic monitoring with the objective to process real-time data from multiple sensors to estimate the necessary state variables (accumulations, travel times, etc.) involved in the control problem.

7.2.3 Aggregated traffic modeling and hierarchical control in heterogenous urban networks

The list below summarizes the main contributions of Chapter 6.

- Two urban traffic models are introduced based on the MFD at different levels of spatial aggregations (regional and subregional) to model the dynamics of density heterogeneity and average trip length.
- The subregional receiving boundary capacity is augmented in the vehicle conservation model.
- A functional form is developed to model the effect of heterogeneity on subregion and region MFDs based on real data analysis.
- A hierarchical control framework based on the perimeter control strategy is proposed for congestion management in heterogeneously congested cities.
- The upper-level of the hierarchical control structure is solved by the MPC scheme, while the subregion- and region-based models are respectively the plant (reality) and the optimization model in the MPC framework.
- The lower-level of the hierarchical control system consists of a feedback controller that manipulates the perimeter control inputs for each subregion in the region boundary to actively control the spatial distribution of traffic density in regions.
- The significance of heterogeneity modeling and the feedback homogeneity controller in control applications is demonstrated.
- It is demonstrated that efficient control strategies could decrease the magnitude of hysteresis loop in the MFD.

The introduced control strategy is a complementary segment of the hierarchical control framework that tackles vehicle density heterogeneity in congested cities. The results also stressed the importance of density heterogeneity considerations for filed implementations.

A challenging modeling extension is to relax the assumption in the aggregated region modeling that O-D routes cannot pass through the boundary between the regions more than once, i.e. to allow drivers to cross the regions of the city without any limitation [200]. Another research direction is to augment the “current best” route choice strategy with the experienced travel time estimation. In addition, considering user equilibrium conditions in MFD modeling

and the interaction between the adaptive perimeter control and the equilibrium state are important issues to be addressed in future works. These problems could be overcome in a day-to-day assignment framework, where drivers adapt to unexpected traffic conditions (created by the new control strategy) by taking different route choice decisions over days.

Moreover, a route guidance advisory control system can be integrated within the proposed control strategy to further improve congestion homogeneity and network performance. This involves the effect of regional route choice in the MFD modeling and further explores implications of drivers' route choice behaviour in performance of the control strategy.

All the efforts related to control and MFD modeling for multi-region networks assume a stationary partitioning boundary in time and space. With dynamic congestion propagation in different regions of a city (see e.g. [91]), a dynamic partitioning associated with a dynamic boundary adjustment framework should be studied to improve the controllability of the proposed control method. A control strategy with dynamic boundary adjustment is expected to further handle congestion heterogeneity in cities. This is a research priority. Furthermore, a field test would provide more insights about the applicability and implications of the proposed hierarchical traffic control strategy.

Bibliography

[1] K. Aboudolas, M. Papageorgiou, and E. Kosmatopoulos. Store-and-forward based methods for the signal control problem in large-scale congested urban road networks. *Transportation Research Part C*, 17:163–174, 2009.

[2] K Aboudolas, M Papageorgiou, and E Kosmatopoulos. Store-and-forward based methods for the signal control problem in large-scale congested urban road networks. *Transportation Research Part C: Emerging Technologies*, 17(2):163–174, 2009.

[3] K. Aboudolas, M. Papageorgiou, A. Kouvelas, and E. Kosmatopoulos. A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks. *Transportation Research Part C*, 18:680–694, 2010.

[4] Konstantinos Aboudolas and Nikolas Geroliminis. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. *Transportation Research Part B: Methodological*, 55(0):265 – 281, 2013.

[5] Konstantinos Aboudolas and Nikolas Geroliminis. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. *Transportation Research Part B: Methodological*, 55:265–281, 2013.

[6] Rahmi Akçelik. A queue model for hcm 2000. Technical report, ARRB Transport Research Ltd, Vermont South, Australia, 1998.

[7] Richard E Allsop. Delay at a fixed time traffic signal-i: Theoretical analysis. *Transportation Science*, 6(3):260–285, 1972.

[8] Constantinos Antoniou and Haris N Koutsopoulos. Estimation of traffic dynamics models with machine-learning methods. *Transportation Research Record: Journal of the Transportation Research Board*, 1965(1):103–111, 2006.

[9] Constantinos Antoniou, Haris N. Koutsopoulos, and George Yannis. Dynamic data-driven local traffic state estimation and prediction. *Transportation Research Part C: Emerging Technologies*, 34(0):89 – 107, 2013.

[10] Xuegang (Jeff) Ban, Peng Hao, and Zhanbo Sun. Real time queue length estimation for signalized intersections using travel times from mobile sensors. *Transportation Research Part C: Emerging Technologies*, 19(6):1133 – 1156, 2011.

[11] T. Bellemans, B. De Schutter, and B. De Moor. Model predictive control for ramp metering of motorway traffic: A case study. *Control Engineering Practice*, 14:757–767, 2006.

[12] Robert L Bertini and Sutti Tantiyanugulchai. Transit buses as traffic probes: Use of geolocation data for empirical evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, 1870(1):35–45, 2004.

[13] John T. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming (Second Edition)*. Society for Industrial and Applied Mathematics, 2010.

[14] Ashish Bhaskar, Edward Chung, and AndrÃ©-Gilles Dumont. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. *Computer-Aided Civil and Infrastructure Engineering*, 26(6):433–450, 2011.

[15] Burak Boyaci, Konstantinos G. Zografos, and Nikolas Geroliminis. An optimization framework for the development of efficient one-way car-sharing systems. *European Journal of Operational Research*, (0):–, 2014.

[16] C. Buisson and C. Ladier. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams. *Transportation Research Record*, 2124:127–136, 2009.

[17] E. F. Camacho and C. Bordons. *Model Predictive Control in the Process Industry*. Springer-Verlag, Berlin, Germany, 1995.

[18] E. F. Camacho and C. Bordons. *Model Predictive Control*. Springer-Verlag, Berlin, Germany, 1999.

[19] Yang Cheng, Xiao Qin, Jing Jin, and Bin Ran. An exploratory shockwave approach to estimating queue length using probe trajectories. *Journal of Intelligent Transportation Systems*, 16(1):12–23, 2012.

[20] Eleni Christofa, Juan Argote, and Alexander Skabardonis. Arterial queue spillback detection and signal control based on connected vehicle technology. *Transportation Research Record: Journal of the Transportation Research Board*, 2356(1):61–70, 2013.

[21] James S Clark. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, 152(2):204–224, 1998.

[22] G. Comert and M. Cetin. Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data. *Intelligent Transportation Systems, IEEE Transactions on*, 12(2):563–573, June 2011.

[23] European Commission. Roadmap to a single european transport area - towards a competitive and resource efficient transport system. In *White Paper, Brussels*. 2011.

[24] Goncalo Homem De Almeida Correia, Diana Ramos Jorge, and David Marques Antunes. The added value of accounting for users flexibility and information on the potential of a station-based one-way car-sharing system: An application in lisbon, portugal. *Journal of Intelligent Transportation Systems*, 18(3):299–308, 2014.

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[26] M Cremer and J Ludwig. A fast simulation model for traffic flow on the basis of boolean operations. *Mathematics and Computers in Simulation*, 28(4):297–303, 1986.

[27] Graham Currie and Herman Lai. Intermittent and dynamic transit lanes: Melbourne, australia, experience. *Transportation Research Record: Journal of the Transportation Research Board*, 2072:49–56, 2008.

[28] C. F. Daganzo. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B*, 41(1):49–62, 2007.

[29] Carlos F Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, 28(4):269–287, 1994.

[30] Carlos F. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B*, 28(4):269–287, 1994.

[31] Carlos F Daganzo. A behavioral theory of multi-lane traffic flow. part i: Long homogeneous freeway sections. *Transportation Research Part B: Methodological*, 36(2):131–158, 2002.

[32] Carlos F. Daganzo, Vikash V. Gayah, and Eric J. Gonzales. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. *Transportation Research Part B*, 45(1):278–288, 2011.

[33] Carlos F. Daganzo and Nikolas Geroliminis. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transportation Research Part B*, 42(9):771–781, 2008.

[34] JN Darroch, Gordon Frank Newell, and RWJ Morris. Queues for a vehicle-actuated traffic light. *Operations Research*, 12(6):882–895, 1964.

[35] C. Diakaki, M. Papageorgiou, and K. Aboudolas. A multivariable regulator approach to traffic-responsive network-wide signal control. *Control Engineering Practice*, 10(13):183–195, 2002.

[36] Francois Dion, Hesham Rakha, and Youn-Soo Kang. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transportation Research Part B: Methodological*, 38(2):99 – 122, 2004.

[37] Jing Dong and Hani S Mahmassani. Flow breakdown and travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 2124(1):203–212, 2009.

[38] Lili Du, Srinivas Peeta, and Yong Hoon Kim. An adaptive information fusion model to predict the short-term link travel time distribution in dynamic traffic networks. *Transportation Research Part B: Methodological*, 46(1):235 – 252, 2012.

[39] Michael Eichler and Carlos F. Daganzo. Bus lanes with intermittent priority: Strategy formulae and an evaluation. *Transportation Research Part B: Methodological*, 40(9):731 – 744, 2006.

[40] Alan L Erera, Tim W Lawson, and Carlos F Daganzo. Simple, generalized method for analysis of traffic queue upstream of a bottleneck. *Transportation Research Record: Journal of the Transportation Research Board*, 1646(1):132–140, 1998.

[41] Yiheng Feng, Gary A Davis, and John Hourdos. Arterial travel time characterization and real-time traffic condition identification using gps-equipped probe vehicles. In *90th Transportation Research Board Annual Meeting*, pages 11–2018, 2011.

[42] MARTIN A. FERMAN, DENNIS E. BLUMENFELD, and XIAOWEN DAI. An analytical evaluation of a real-time traffic information system using probe vehicles. *Journal of Intelligent Transportation Systems*, 9(1):23–34, 2005.

[43] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

[44] C. E. Garcia, D. M. Prett, and M. Morari. Model predictive control: Theory and practice – a survey. *Automatica*, 25(3):335–348, May 1989.

[45] N. H. Gartner, F. J. Pooran, and C. M. Andrews. Optimized policies for adaptive control strategy in real–time traffic adaptive control systems, implementation and field testing. *Transportation Research Record*, 1811:148–156, 2002.

[46] V. V. Gayah and C. F. Daganzo. Clockwise hysteresis loops in the macroscopic fundamental diagram: An effect of network instability. *Transportation Research Part B*, 45(4):643–655, 2011.

[47] Vikash V Gayah and Carlos F Daganzo. Effects of turning maneuvers and route choice on a simple network. *Transportation Research Record: Journal of the Transportation Research Board*, 2249(1):15–19, 2011.

[48] Denos C Gazis, Robert Herman, and Richard W Rothery. Nonlinear follow-the-leader models of traffic flow. *Operations Research*, 9(4):545–567, 1961.

[49] N. Geroliminis, J. Haddad, and M. Ramezani. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. *Intelligent Transportation Systems, IEEE Transactions on*, 14(1):348–359, March 2013.

[50] N. Geroliminis and A Skabardonis. Real time vehicle reidentification and performance measures on signalized arterials. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pages 188–193, 2006.

[51] N. Geroliminis and Jie Sun. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. *Transportation Research Part A: Policy and Practice*, 45(9):966–979, November 2011.

[52] Nikolas Geroliminis. Dynamics of peak hour and effect of parking for congested cities. In *Transportation Research Board Annual Meeting*, number 09-1685, Washington, D.C., 2009.

[53] Nikolas Geroliminis and Burak Boyacı. The effect of variability of urban systems characteristics in the network capacity. *Transportation Research Part B*, 46(10):1607–1623, December 2012.

[54] Nikolas Geroliminis and Carlos F. Daganzo. Macroscopic modeling of traffic in cities. In *86th Annual Meeting of the Transportation Research Board*, Washington, DC, 2007. Paper No. 07-0413.

[55] Nikolas Geroliminis and Carlos F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. *Transportation Research Part B*, 42(9):759–770, 2008.

[56] Nikolas Geroliminis and A. Skabardonis. Identification and analysis of queue spillovers in city street networks. *IEEE Transactions on Intelligent Transportation Systems*, 12:1107–1115, 2011.

[57] Nikolas Geroliminis, Anupam Srivastava, and Panos Michalopoulos. A dynamic-zone-based coordinated ramp-metering algorithm with queue constraints for minnesota's freeways. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1576–1586, December 2011.

[58] Nikolas Geroliminis and Jie Sun. Properties of a well-defined macroscopic fundamental diagram for urban traffic. *Transportation Research Part B: Methodological*, 45(3):605 – 617, 2011.

[59] S. Ghosh-Dastidar and H. Adeli. Neural network-wavelet microsimulation model for delay and queue length estimation at freeway work zones. *Journal of Transportation Engineering*, 132(4):331–341, 2006.

[60] J. W. Godfrey. The mechanism of a road network. *Traffic Engineering and Control*, 11(7):323–327, 1969.

[61] G. Gomes and R. Horowitz. Optimal freeway ramp metering using the asymmetric cell transmission model. *Transportation Research Part C*, 14:244–262, 2006.

[62] Bruce Douglas Greenshields, JT Thompson, HC Dickinson, and RS Swinton. The photographic method of studying traffic behavior. In *Highway Research Board Proceedings*, volume 13, 1934.

[63] N. Groot, B. De Schutter, and H. Hellendoorn. Integrated model predictive traffic and emission control using a piecewise-affine approach. *Intelligent Transportation Systems, IEEE Transactions on*, 14(2):587–598, June 2013.

[64] N. Groot, B. De Schutter, and H. Hellendoorn. Toward system-optimal routing in traffic networks: A reverse stackelberg game approach. *Intelligent Transportation Systems, IEEE Transactions on*, PP(99):1–12, 2014.

[65] Feng Guo, Hesham Rakha, and Sangjun Park. Multistate model for travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 2188(1):46–54, 2010.

[66] Jack Haddad and Nikolas Geroliminis. On the stability of traffic perimeter control in two-region urban cities. *Transportation Research Part B*, 46(1):1159–1176, 2012.

[67] Jack Haddad, Mohsen Ramezani, and Nikolas Geroliminis. Model predictive perimeter-control for two-region urban cities. In *Transportation Research Board Annual Meeting*, number 12-1445, Washington, D.C., 2012.

[68] Jack Haddad, Mohsen Ramezani, and Nikolas Geroliminis. Model predictive perimeter control for urban areas with macroscopic fundamental diagrams. In *American Control Conference (ACC)*, pages 5757–5762, Montrèal, Canada, June 2012.

[69] Jack Haddad, Mohsen Ramezani, and Nikolas Geroliminis. Cooperative traffic control of a mixed network with two urban regions and a freeway. *Transportation Research Part B: Methodological*, 54(0):17 – 36, 2013.

[70] Peng Hao, Xuegang Ban, K.P. Bennett, Qiang Ji, and Zhanbo Sun. Signal timing estimation using sample intersection travel times. *Intelligent Transportation Systems, IEEE Transactions on*, 13(2):792–804, June 2012.

[71] HCM. *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington DC, 2000.

[72] Qing He, K. Larry Head, and Jun Ding. Pamscod: Platoon-based arterial multi-modal signal control with online data. *Transportation Research Part C: Emerging Technologies*, 20(1):164 – 184, 2012. Special issue on Optimization in Public Transport+ISTT2011 Special issue on Optimization in Public Transport+International Symposium on Transportation and Traffic Theory (ISTTT), Berkeley, California, July 18â€“20, 2011.

[73] A. Hegyi, B. De Schutter, and H. Hellendoorn. Model predictive control for optimal coordination of ramp metering and variable speed limits. *Transportation Research Part C*, 13(3):185–209, June 2005.

[74] A. Hegyi, Bart De Schutter, and J. Hellendoorn. Optimal coordination of variable speed limits to suppress shock waves. *IEEE Transactions on Intelligent Transportation Systems*, 6(1):102–112, March 2005.

[75] D. Helbing. Derivation of a fundamental diagram for urban traffic flow. *The European Physical Journal B*, 70(2):229–241, 2009.

[76] Dirk Helbing. Traffic and related self-driven many-particle systems. *Reviews of modern physics*, 73(4):1067, 2001.

[77] Bruce Hellinga, Pedram Izadpanah, Hiroyuki Takada, and Liping Fu. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C: Emerging Technologies*, 16(6):768 – 782, 2008.

[78] Robert Herman and Ilya Prigogine. A two-fluid approach to town traffic. *Science*, 204(4389):148–151, 1979.

[79] Juan C. Herrera, Daniel B. Work, Ryan Herring, Xuegang (Jeff) Ban, Quinn Jacobson, and Alexandre M. Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568 – 583, 2010.

[80] R. Herring, A Hofleitner, P. Abbeel, and A Bayen. Estimating arterial traffic conditions using sparse probe data. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 929–936, Sept 2010.

[81] A Hofleitner, R. Herring, P. Abbeel, and A Bayen. Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. *Intelligent Transportation Systems, IEEE Transactions on*, 13(4):1679–1693, Dec 2012.

[82] Jun-Wei Hsieh, Shih-Hao Yu, Yung-Sheng Chen, and Wen-Fong Hu. Automatic traffic surveillance system for vehicle tracking and classification. *Intelligent Transportation Systems, IEEE Transactions on*, 7(2):175–187, June 2006.

[83] http://mobility.tamu.edu/ums/.

[84] Pedram Izadpanah, Bruce Hellinga, and Liping Fu. Automatic traffic shockwave identification using vehicles' trajectories. In *Proceedings of the 88th Annual Meeting of the Transportation Research Board (CD-ROM)*, 2009.

[85] Olaf Jahn, Rolf H.Möhring, Andreas S.Schulz, and Nicolás E.Stier Moses. System-optimal routing of traffic flows with user constraints in networks with congestion. *Operations Research*, 53(4):600–616, 2005.

[86] Vedran Jelaca, Jorge Oswaldo Nino Castaneda, Aleksandra Pizurica, and Wilfried Philips. Image projection clues for improved real-time vehicle tracking in tunnels. In *IS&T/SPIE Electronic Imaging*, pages 83010C–83010C. International Society for Optics and Photonics, 2012.

[87] Erik Jenelius. The value of travel time variability with trip chains, flexible scheduling and correlated travel times. *Transportation Research Part B: Methodological*, 46(6):762 – 780, 2012.

[88] Erik Jenelius and Haris N. Koutsopoulos. Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, 53(0):64 – 81, 2013.

[89] Y. Ji, W. Daamen, S. Hoogendoorn, S. Hoogendoorn-Lanser, and X. Qian. Macroscopic fundamental diagram: Investigating its shape using simulation data. *Transportation Research Record*, 2161:42–48, 2010.

[90] Yuxuan Ji and Nikolas Geroliminis. On the spatial partitioning of urban transportation networks. *Transportation Research Part B*, 46(10):1639–1656, December 2012.

[91] Yuxuan Ji, Jun Luo, and Nikolas Geroliminis. Empirical observations of congestion propagation and dynamic partitioning with probe data for large-scale systems. *Transportation Research Record (in press)*, (14-0816), 2014.

[92] X. Jiang and H. Adeli. Dynamic wavelet neural network model for traffic flow forecasting. *Journal of Transportation Engineering*, 131(10):771–779, 2005.

[93] Ioannis Kaparias, Michael G.H. Bell, and Heidrun Belzner. A new measure of travel time reliability for in-vehicle navigation systems. *Journal of Intelligent Transportation Systems*, 12(4):202–211, 2008.

[94] Boris S Kerner, Sergey L Klenov, and Dietrich E Wolf. Cellular automata approach to three-phase traffic theory. *Journal of Physics A: Mathematical and General*, 35(47):9971, 2002.

[95] Arne Kesting, Martin Treiber, and Dirk Helbing. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4585–4605, 2010.

[96] M. Keyvan-Ekbatani, M. Yildirimoglu, N. Geroliminis, and M. Papageorgiou. Traffic signal perimeter control with multiple boundaries for large urban networks. In *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on*, pages 1004–1009, Oct 2013.

[97] Mehdi Keyvan-Ekbatani, Anastasios Kouvelas, Ioannis Papamichail, and Markos Papageorgiou. Exploiting the fundamental diagram of urban networks for feedback-based gating. *Transportation Research Part B: Methodological*, 46(10):1393 – 1403, 2012.

[98] Mehdi Keyvan-Ekbatani, Markos Papageorgiou, and Ioannis Papamichail. Urban congestion gating control based on reduced operational network fundamental diagrams. *Transportation Research Part C: Emerging Technologies*, 33(0):74 – 87, 2013.

[99] Frank Killmann and Elart von Collani. A note on the convolution of the uniform and related distributions and their use in quality control. *Economic Quality Control*, 16(1):17–41, 2001.

[100] V. L. Knoop, S.P. Hoogendoorn, and J. W. C. Van Lint. Routing strategies based on the macroscopic fundamental diagram. In *Transportation Research Board Annual Meeting*, Washington, D.C., 2012.

[101] Victor L Knoop, Serge Hoogendoorn, and JWC Van Lint. Impact of traffic dynamics on macroscopic fundamental diagram. In *Transportation Research Board 92nd Annual Meeting*, 2013.

[102] VL Knoop, SP Hoogendoorn, and J WC Van Lint. Routing strategies based on macroscopic fundamental diagram. *Transportation Research Record: Journal of the Transportation Research Board*, 2315(1):1–10, 2012.

[103] A. Kotsialos, M. Papageorgiou, M. Mangeas, and H. Haj-Salem. Coordinated and integrated control of motorway networks via non-linear optimal control. *Transportation Research Part C*, 10:65—84, 2002.

[104] Jaimyoung Kwon, Tiffany Barkley, Rob Hranac, Karl Petty, and Nick Compin. Decomposition of travel time reliability into various sources. *Transportation Research Record: Journal of the Transportation Research Board*, 2229(1):28–33, 2011.

[105] Karric Kwong, Robert Kavaler, Ram Rajagopal, and Pravin Varaiya. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies*, 17(6):586 – 606, 2009.

[106] Jean-Patrick Lebacque. The godunov scheme and what it means for first order traffic flow models. In *Internaional symposium on transportation and traffic theory*, pages 647–677, 1996.

[107] Jean-Patrick Lebacque, Salim Mammar, and Habib Haj Salem. Generic second order traffic flow modelling. In *Transportation and Traffic Theory 2007. Papers Selected for Presentation at ISTTT17*, 2007.

[108] Ludovic Leclercq, Nicolas Chiabaut, and Béatrice Trinquier. Macroscopic fundamental diagrams: A cross-comparison of estimation methods. *Transportation Research Part B: Methodological*, 62(0):1 – 12, 2014.

[109] Ludovic Leclercq and Nikolas Geroliminis. Estimating mfds in simple networks with route choice. *Transportation Research Part B: Methodological*, 57:468–484, 2013.

[110] M. J. Lighthill and G. B. Whitham. On kinematic waves. i: flood movement in long rivers. ii: a theory of traffic flow on long crowded roads. *Proceedings of Royal Society*, A229:281–345, 1955.

[111] Shu Lin, Bart De Schutter, Yugeng Xi, and Hans Hellendoorn. Fast model predictive control for urban road networks via milp. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):846–856, September 2011.

[112] Henry X. Liu and Wenteng Ma. A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. *Transportation Research Part C: Emerging Technologies*, 17(1):11 – 26, 2009.

[113] James O Lloyd-Smith, Sebastian J Schreiber, P Ekkehard Kopp, and WM Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.

[114] Yingyan Lou, Yafeng Yin, and Jorge A. Laval. Optimal dynamic pricing strategies for high-occupancy/toll lanes. *Transportation Research Part C: Emerging Technologies*, 19(1):64 – 74, 2011.

[115] J. M. Maciejowski. *Predictive Control with Constraints*. Prentice-Hall, Harlow, UK, 2002.

[116] H. Mahmassani, J.C. Williams, and R. Herman. Performance of urban traffic networks. In N.H. Gartner and N.H.M. Wilson, editors, *Proceedings of the 10th International Symposium on Transportation and Traffic Theory*, Amsterdam, The Netherlands, 1987. Elsevier.

[117] Hani S Mahmassani, Tian Hou, and Meead Saberi. Connecting networkwide travel time reliability and the network fundamental diagram of traffic flow. *Transportation Research Record: Journal of the Transportation Research Board*, 2391(1):80–91, 2013.

[118] Hani S. Mahmassani, Meead Saberi, and Ali Zockaie K. Urban network gridlock: Theory, characteristics, and dynamics. *Procedia - Social and Behavioral Sciences*, 80(0):79 – 98, 2013. 20th International Symposium on Transportation and Traffic Theory (ISTTT 2013).

[119] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36:789–814, 2000.

[120] Amin Mazloumian, Nikolas Geroliminis, and Dirk Helbing. The spatial variability of vehicle densities as determinant of urban network capacity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4627–4647, 2010.

[121] Donald R McNeil. A solution to the fixed-cycle traffic light problem for compound poisson arrivals. *Journal of Applied Probability*, 5(3):624–635, 1968.

[122] Monica Menendez and Carlos F. Daganzo. Effects of {HOV} lanes on freeway bottlenecks. *Transportation Research Part B: Methodological*, 41(8):809 – 822, 2007.

[123] Panos G. Michalopoulos, Gregory Stephanopoulos, and George Stephanopoulos. An application of shock wave theory to traffic signal control. *Transportation Research Part B: Methodological*, 15(1):35 – 51, 1981.

[124] Pitu Mirchandani and Larry Head. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies*, 9(6):415 – 432, 2001.

[125] B.T. Morris and M.M. Trivedi. Learning, modeling, and classification of vehicle track patterns from live video. *Intelligent Transportation Systems, IEEE Transactions on*, 9(3):425–437, Sept 2008.

[126] Kai Nagel and Michael Schreckenberg. A cellular automaton model for freeway traffic. *Journal de physique I*, 2(12):2221–2229, 1992.

[127] Gordon Frank Newell. Approximation methods for queues with application to the fixed-cycle traffic light. *Siam Review*, 7(2):223–240, 1965.

[128] Gordon Frank Newell et al. Queues for a fixed-cycle traffic light. *The Annals of Mathematical Statistics*, 31(3):589–597, 1960.

[129] NGSIM. *Next Generation Simulation. http://ngsim.fhwa.dot.gov/, 2006.*

[130] Robert B Noland and John W Polak. Travel time variability: a review of theoretical and empirical issues. *Transport Reviews*, 22(1):39–54, 2002.

[131] Piotr Olszewski, Henry S. L. Fan, and Yan-Weng Tan. Area-wide traffic speed-flow model for the singapore cbd. *Transportation Research Part A*, 29A(4):273–281, 1995.

[132] Javier Ortigosa, Monica Menendez, and Hector Tapia. Study on the number and location of measurement points for an mfd perimeter control scheme: a case study of zurich. *EURO Journal on Transportation and Logistics*, pages 1–22, 2013.

[133] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 91(12):2043–2067, December 2003.

[134] M. Papageorgiou, H. Haj-Salem, and J. Blosseville. ALINEA a local feedback control law for on-ramp metering. *Transportation Research Record*, (1320):58–64, 1991.

[135] Markos Papageorgiou, Christina Diakaki, Vaya Dinopoulou, Apostolos Kotsialos, and Yibing Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 91(12):2043–2067, 2003.

[136] Markos Papageorgiou, Habib Hadj-Salem, and Jean-Marc Blosseville. Alinea: A local feedback control law for on-ramp metering. *Transportation Research Record*, (1320), 1991.

[137] Markos Papageorgiou and Apostolos Kotsialos. Freeway ramp metering: An overview. *IEEE Transactions on Intelligent Transportation Systems*, 3(4):271–281, December 2002.

[138] I Papamichail and M. Papageorgiou. Balancing of queues or waiting times on metered dual-branch on-ramps. *Intelligent Transportation Systems, IEEE Transactions on*, 12(2):438–452, June 2011.

[139] Ioannis Papamichail, Apostolos Kotsialos, Ioannis Margonis, and Markos Papageorgiou. Coordinated ramp metering for freeway networks – a model-predictive hierarchical control approach. *Transportation Research Part C*, 18:311–331, 2010.

[140] Sangjun Park, H. Rakha, and Feng Guo. Multi-state travel time reliability model: Impact of incidents on travel time reliability. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, pages 2106–2111, Oct 2011.

[141] H. J. Payne. Models of freeway traffic and control. volume 1, pages 51–61. Simulation Council Proc., 1971.

[142] Louis A Pipes. An operational analysis of traffic dynamics. *Journal of applied physics*, 24(3):274–281, 1953.

[143] Joseph N. Prashker and Shlomo Bekhor. Some observations on stochastic user equilibrium and system optimum of traffic assignment. *Transportation Research Part B*, 34:277–291, 2000.

[144] Ilya Prigogine and Frank C Andrews. A boltzmann-like approach for traffic flow. *Operations Research*, 8(6):789–797, 1960.

[145] S. Joe Qin and T. A. Badgwell. A survey of industrial model predictive control technology. *Control Engineering Practice*, 11:733–764, 2003.

[146] Mohammed A. Quddus, Washington Y. Ochieng, and Robert B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312 – 328, 2007.

[147] Mahmood Rahmani and Haris N Koutsopoulos. Path inference from sparse floating car data for urban networks. *Transportation Research Part C: Emerging Technologies*, 30:41–54, 2013.

[148] H. Rakha, I El-Shawarby, M. Arafeh, and F. Dion. Estimating path travel-time reliability. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pages 236–241, Sept 2006.

[149] M. Ramezani and N. Geroliminis. Markov chain procedure for arterial route travel time distribution estimation. In *2nd international conference on Models and Technologies for Intelligent Transportation Systems*, 2011.

[150] M Ramezani and N Geroliminis. Estimation of arterial route travel time distribution with markov chains. In *Transportation Research Board 91st Annual Meeting*, 2012.

[151] M. Ramezani and N. Geroliminis. On the estimation of arterial route travel time distribution with markov chains. *Transportation Research Part B*, 46(10):1576–1590, 2012.

[152] M. Ramezani and N. Geroliminis. Exploiting probe data to estimate the queue profile in urban networks. In *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on*, pages 1817–1822, Oct 2013.

[153] M Ramezani and N Geroliminis. Queue profile estimation with probe vehicle data for urban arterials. In *Transportation Research Board 93rd Annual Meeting*, 2014.

[154] M. Ramezani, J. Haddad, and N. Geroliminis. Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control. *under review in Transportation Research Part B*.

[155] Mohsen Ramezani and Nikolas Geroliminis. Arterial route travel time distribution estimation with a markov chain procedure. In *11th Swiss Transportation Research Conference*, 2011.

[156] Mohsen Ramezani and Nikolas Geroliminis. Queue profile estimation in signalized arterials by utilizing probe vehicle data. In *Swiss Transport Research Conference*, 2013.

[157] Mohsen Ramezani and Nikolas Geroliminis. Queue profile estimation in congested urban networks with probe data. *Computer-Aided Civil and Infrastructure Engineering*, 2014.

[158] Mohsen Ramezani, Jack Haddad, and Nikolas Geroliminis. Macroscopic traffic control of a mixed urban and freeway network. In *Swiss Transport Research Conference*, 2012.

[159] Mohsen Ramezani, Jack Haddad, and Nikolas Geroliminis. Macroscopic traffic control of a mixed urban and freeway network. In *13th IFAC Symposium on Control in Transportation Systems*, 2012.

[160] Mohsen Ramezani, Jack Haddad, and Nikolas Geroliminis. Cooperative traffic management policies for mixed freeway and urban networks. In *Transportation Research Board 92nd Annual Meeting*, number 13-1094, 2013.

[161] Mohsen Ramezani, Jack Haddad, and Nikolas Geroliminis. Integrating heterogeneity dynamics in aggregated urban network modeling and control. In *Swiss Transport Research Conference*, 2014.

[162] Mohsen Ramezani, Jack Haddad, and Nikolas Geroliminis. Integrating the dynamics of heterogeneity in aggregated network modeling and control. In *Transportation Research Board 93rd Annual Meeting*, number 14-0710, 2014.

[163] P. J. Richards. Shock waves on the highway. *Operations Research*, 4:42–51, 1956.

[164] Meead Saberi and Hani S. Mahmassani. Exploring properties of network-wide flow-density relations in a freeway network. In *Transportation Research Board Annual Meeting*, Washington, D.C., 2012.

[165] Meead Saberi and Hani S Mahmassani. Hysteresis and capacity drop phenomena in freeway networks: Empirical characterization and interpretation. *Transportation Research Record: Journal of the Transportation Research Board*, 2391(1):44–55, 2013.

[166] T Samad and AM Annaswamy. The impact of control technology. *IEEE Control Systems Society*, 2011.

[167] Donald C. Shoup. Cruising for parking. *Transport Policy*, 13(6):479 – 486, 2006. Parking.

[168] A. Skabardonis and N. Geroliminis. Prediction of arrival profiles and queue lengths along signalized arterials by using a markov decision process. *ransportation Research Record: Journal of the Transportation Research Board*, 1934:116–124, 2005.

[169] Alexander Skabardonis and Nikolas Geroliminis. Real-time monitoring and control on signalized arterials. *Journal of Intelligent Transportation Systems*, 12(2):64–74, 2008.

[170] Alexander Skabardonis, Nikolas Geroliminis, and Eleni Christofa. Prediction of vehicle activity for emissions estimation under oversaturated conditions along signalized arterials. *Journal of Intelligent Transportation Systems*, 17(3):191–199, 2013.

[171] Zhanbo Sun and Xuegang (Jeff) Ban. Vehicle trajectory reconstruction for signalized intersections using mobile traffic sensors. *Transportation Research Part C: Emerging Technologies*, 36(0):268 – 283, 2013.

[172] K. L. Teo, C. J. Goh, and K. H. Wong. *A Unified Computational Approach to Optimal Control Problems*. Longman Scientific & Technical, New York, 1991.

[173] Dušan Teodorović and Panta Lučić. Intelligent parking systems. *European Journal of Operational Research*, 175(3):1666 – 1681, 2006.

[174] Alejandro Tirachini, David A. Hensher, and John M. Rose. Multimodal pricing and optimal design of urban public transport: The interplay between traffic congestion and bus crowding. *Transportation Research Part B: Methodological*, 61(0):33 – 54, 2014.

[175] Martin Treiber, Arne Kesting, and R Eddie Wilson. Reconstructing the traffic state by fusion of heterogeneous data. *Computer-Aided Civil and Infrastructure Engineering*, 26(6):408–419, 2011.

[176] Theodore Tsekeris and Alexander Skabardonis. On-line performance measurement models for urban arterial networks. In *Transportation Research Board Annual Meeting*, volume 935, page 24, 2004.

[177] Nobuhiro Uno, Fumitaka Kurauchi, Hiroshi Tamura, and Yasunori Iida. Using bus probe data for analysis of travel time variability. *Journal of Intelligent Transportation Systems*, 13(1):2–15, 2009.

[178] M. van den Berg, A. Hegyi, B. De Schutter, and J. Hellendoorn. Integrated traffic control for mixed urban and freeway networks: A model predictive control approach. *European Journal of Transport and Infrastructure Research*, 7(3):223–250, September 2007.

[179] JWC Van Lint and Serge P Hoogendoorn. A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways. *Computer-Aided Civil and Infrastructure Engineering*, 25(8):596–612, 2010.

[180] Femke van Wageningen-Kessels, Hans van Lint, Kees Vuik, and Serge Hoogendoorn. Genealogy of traffic flow models. *EURO Journal on Transportation and Logistics*, pages 1–29, 2014.

[181] Pravin Varaiya. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 36:177–195, 2013.

[182] Georgios Vigos, Markos Papageorgiou, and Yibing Wang. Real-time estimation of vehicle-count within signalized links. *Transportation Research Part C: Emerging Technologies*, 16(1):18 – 35, 2008.

[183] Francesco Viti and Henk J. van Zuylen. Probabilistic models for queues at fixed control signals. *Transportation Research Part B: Methodological*, 44(1):120 – 135, 2010.

[184] Eleni I. Vlahogianni and Matthew G. Karlaftis. Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties. *Computer-Aided Civil and Infrastructure Engineering*, 28(6):420–433, 2013.

[185] Fei-Yue Wang. Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications. *IEEE Transactions on Intelligent Transportation Systems*, 11:630–638, 2010.

[186] Yibing Wang and Markos Papageorgiou. Real-time freeway traffic state estimation based on extended kalman filter: a general approach. *Transportation Research Part B: Methodological*, 39(2):141 – 167, 2005.

[187] J. G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II*, 1(2):352–362, 1952.

[188] F.V. Webster. Traffic signal settings. *Road Research Technical Paper*, (39), 1958. Great Britain Road Research Laboratory, London.

[189] Heng Wei, Hao Liu, Qingyi Ai, Zhixia Li, Hui Xiong, and Benjamin Coifman. Empirical innovation of computational dual-loop models for identifying vehicle classifications against varied traffic conditions. *Computer-Aided Civil and Infrastructure Engineering*, 28(8):621–634, 2013.

[190] R Eddie Wilson and Jonathan A Ward. Car-following models: fifty years of linear stability analysis–a mathematical perspective. *Transportation Planning and Technology*, 34(1):3–18, 2011.

[191] TH Witte and AM Wilson. Accuracy of non-differential gps for the determination of speed over ground. *Journal of biomechanics*, 37(12):1891–1898, 2004.

[192] GCK Wong and SC Wong. A multi-class traffic flow model–an extension of lwr model with heterogeneous drivers. *Transportation Research Part A: Policy and Practice*, 36(9):827–841, 2002.

[193] Xinkai Wu and Henry X. Liu. A shockwave profile model for traffic flow on congested urban arterials. *Transportation Research Part B: Methodological*, 45(10):1768 – 1786, 2011.

[194] Xinkai Wu, Henry X. Liu, and Nikolas Geroliminis. An empirical analysis on the arterial fundamental diagram. *Transportation Research Part B: Methodological*, 45(1):255 – 266, 2011.

[195] Xinkai Wu, Henry X. Liu, and Douglas Gettman. Identification of oversaturated intersections using high-resolution traffic signal data. *Transportation Research Part C: Emerging Technologies*, 18(4):626 – 638, 2010.

[196] Hao Xu, Hongchao Liu, Chin-Woo Tan, and Yuanlu Bao. Development and application of an enhanced kalman filter and global positioning system error-correction approach for improved map-matching. *Journal of Intelligent Transportation Systems*, 14(1):27–36, 2010.

[197] Hai Yang and Michael G.H Bell. Traffic restraint, road pricing and network equilibrium. *Transportation Research Part B: Methodological*, 31(4):303 – 314, 1997.

[198] Jiyoun Yeon, Lily Elefteriadou, and Siriphong Lawphongpanich. Travel time estimation on a freeway using discrete time markov chains. *Transportation Research Part B: Methodological*, 42(4):325 – 338, 2008.

[199] Mehmet Yildirimoglu and Nikolas Geroliminis. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. *Transportation Research Part B, in press*.

[200] Mehmet Yildirimoglu, Mohsen Ramezani, and Nikolas Geroliminis. Equilibrium analysis and route guidance in large-scale networks with mfd dynamics. In *submitted to 21st International Symposium on Transportation and Traffic Theory*.

[201] Shihong Yue, Miaomiao Wei, Jeen-Shing Wang, and Huaxiang Wang. A general grid-clustering approach. *Pattern Recognition Letters*, 29(9):1372 – 1384, 2008.

[202] Lele Zhang, Timothy M Garoni, and Jan de Gier. A comparative study of macroscopic fundamental diagrams of arterial road networks governed by adaptive traffic signal systems. *Transportation Research Part B: Methodological*, 49(0):1 – 23, 2013.

[203] Lele Zhang, Timothy M Garoni, and Jan de Gier. A comparative study of macroscopic fundamental diagrams of arterial road networks governed by adaptive traffic signal systems. *Transportation Research Part B: Methodological*, 49(0):1 – 23, 2013.

[204] Fangfang Zheng and Henk Van Zuylen. Uncertainty and predictability of urban link travel time. *Transportation Research Record: Journal of the Transportation Research Board*, 2192(1):136–146, 2010.

[205] Nan Zheng, Rashid A. Waraich, Kay W. Axhausen, and Nikolas Geroliminis. A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. *Transportation Research Part A: Policy and Practice*, 46(8):1291 – 1303, 2012.

CURRICULUM VITAE
Mohsen Ramezani

Education

- **EPFL**, Lausanne, Switzerland — 2010 - October 2014
  Ph.D., Civil and Environmental Engineering
  Thesis: Traffic Modeling, Estimation and Control for Large-Scale Congested Urban Networks
  Advisor: Prof. Nikolas Geroliminis
- **University of Tehran**, Tehran, Iran — 2008 - 2010
  M.Sc., Electrical Engineering, Control Systems
  Thesis: Traffic Flow Control Using Multi Agent Reinforcement Learning
  Advisor: Prof. Ashkan Rahimi-Kain
- **University of Tehran**, Tehran, Iran — 2004 - 2008
  B.Sc., Electrical Engineering, Control Systems
  Thesis: Bio-inspired Algorithms for Vehicle Routing Problem
  Advisor: Prof. Masoud Asadpour

Research Experience

- **Research Assistant** — 2010 - 2014
  LUTS, EPFL, Lausanne, Switzerland
- **Visiting Scholar** — Jan. 2014 - Mar. 2014
  Prof. Pravin Varaiya
  University of California, Berkeley, CA, USA
- **Master Thesis Co-supervisor** (LUTS, EPFL) — 2012 - 2014
  - *Nicolas de Lamberterie, "Real-time control of queue spillbacks in signalized arterials" (in collaboration with Prof. Alexander Skabardonis, UC Berkeley)*
  - *Ralph Burgener, "Public transport priority schemes in arterials"*
- **Semester Project Supervisor** (LUTS, EPFL) — 2011 - 2014
  - *Paul Anderson, "Ramp metering and variable speed limit for coordinated control"*
  - *Anne Koymans, "Travel time distributions for congested transportation networks"*
  - *Sohrab Sahaleh, "Analysis for emissions estimation along signalized arterials"*
- **Research Assistant** — 2006 - 2010
  Control and Intelligent Processing Center of Excellence
  University of Tehran, Tehran, Iran

Teaching experience

Fundamental of traffic operations and control (M.Sc.) *lecturer: Prof. N. Geroliminis*	Teaching Assistant	EPFL	2011-2014
Analysis I (B.Sc.) *lecturer: Dr. B. Buffoni*	Tutor	EPFL	2013

Honors and Awards

- ***1st*** among more than 15000 EE students participating in nationwide university entrance exam (leading to M.Sc. scholarship, University of Tehran) 2008
- 2nd best paper award, IEEE student paper contest, University of Tehran 2009
- Gold prize, faculty of engineering alumni association, University of Tehran 2009
- 18th in national EE olympiad among all Iranian EE students 2008
- Exceptional talent student scholarship, University of Tehran 2005

Peer-Reviewed Journal Publications

[J1] **M. Ramezani** and N. Geroliminis, "Queue profile estimation in congested urban networks with probe data," *Computer-Aided Civil and Infrastructure Engineering*, DOI: 10.1111/mice.12095, 2014.

[J2] J. Haddad, **M. Ramezani**, and N. Geroliminis, "Cooperative traffic control of a mixed network with two urban regions and a freeway," *Transportation Research Part B*, vol. 54, pp. 17-36, 2013.

[J3] N. Geroliminis, J. Haddad, and **M. Ramezani**, "Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 348-359, 2013.

[J4] **M. Ramezani** and N. Geroliminis, "On the estimation of arterial route travel time distribution with Markov chains," *Transportation Research Part B*, vol. 46, no. 10, pp. 1576-1590, 2012.

Articles under review:

[J5] **M. Ramezani**, J. Haddad, and N. Geroliminis, "Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control," *under 2nd round of review in Transportation Research Part B.*

[J6] M. Yildirimoglu, **M. Ramezani**, and N. Geroliminis, "Integrating aggregated route choice in network modeling and control with MFDs," *under 2nd round of review in 21st International Symposium on Transportation and Traffic Theory (ISTTT).*

[J7] **M. Ramezani**, N. de Lamberterie, A. Skabardonis, and N. Geroliminis, "A clustering approach for real-time control of queue spillbacks on signalized arterials," *under review in Transportmetrica Part B.*

Refereed Conference Publications

[C1] **M. Ramezani**, R. Burgener, and N. Geroliminis, "Optimization of traffic signals for transit priority in arterials with dedicated bus lanes and stochastic arrivals: a system-oriented approach," *to be presented in 94th Annual Meeting of the Transportation Research Board*, Washington D.C., January 2015.

[C2] M. Yildirimoglu, **M. Ramezani**, and N. Geroliminis, "A rolling horizon approach for route guidance in large-scale networks," *to be presented in 94th Annual Meeting of the Transportation Research Board*, Washington D.C., January 2015.

[C3] R. Burgener, **M. Ramezani**, and N. Geroliminis, "Traffic signal control for preferential treatment of public transport," *3rd Symposium of the European Association for Research in Transportation*, Leeds, UK, September 2014.

[C4] **M. Ramezani**, J. Haddad, and N. Geroliminis, "Integrating the dynamics of heterogeneity in aggregated network modeling and control," *93rd Annual Meeting of the Transportation Research Board*, Washington D.C., January 2014.

[C5] **M. Ramezani** and N. Geroliminis, "Queue profile estimation with probe vehicle data for urban arterials," *93rd Annual Meeting of the Transportation Research Board*, Washington D.C., January 2014.

[C6] N. de Lamberterie, **M. Ramezani**, A. Skabardonis, and N. Geroliminis, "A clustering approach for real-time control of queue spillbacks on signalized arterials," *93rd*

Annual Meeting of the Transportation Research Board, Washington D.C., January 2014.

[C7] **M. Ramezani** and N. Geroliminis, "Exploiting probe data to estimate the queue profile in urban networks," *16th International IEEE Conference on Intelligent Transportation Systems*, The Hague, The Netherlands, October 2013.

[C8] N. de Lamberterie, **M. Ramezani**, A. Skabardonis, and N. Geroliminis, "Real-Time management of queue spillovers along an arterial," *2nd Symposium of the European Association for Research in Transportation*, Stockholm, Sweden, September 2013.

[C9] **M. Ramezani**, J. Haddad, and N. Geroliminis, "Cooperative Traffic Management Policies for Mixed Freeway and Urban Networks," *92nd Annual Meeting of the Transportation Research Board*, Washington D.C., January 2013.

[C10] **M. Ramezani**, J. Haddad, and N. Geroliminis, "Macroscopic Traffic Control of a Mixed Urban and Freeway Network," *13th IFAC Symposium on Control in Transportation Systems*, Sofia, Bulgaria, September 2012.

[C11] J. Haddad, **M. Ramezani**, and N. Geroliminis, "Model predictive perimeter control for urban areas with macroscopic fundamental diagrams," *proceedings of IEEE 2012 American Control Conference*, Montreal, Canada, pp. 5757-5762.

[C12] **M. Ramezani** and N. Geroliminis, "Estimation of arterial route travel time distribution with Markov chains," *91st Annual Meeting of the Transportation Research Board*, Washington D.C., January 2012.

[C13] J. Haddad, **M. Ramezani**, and N. Geroliminis, "Model predictive perimeter-control for two-region urban cities," *91st Annual Meeting of the Transportation Research Board*, Washington D.C., January 2012.

[C14] **M. Ramezani** and N. Geroliminis, "Markov chain procedure for arterial route travel time distribution estimation," *2nd international conference on Models and Technologies for Intelligent Transportation Systems*, Leuven, Belgium, June 2011.

[C15] **M. Ramezani**, H. Hajimirsadeghi, and C. Lucas, "Discrete Invasive Weed Optimization Algorithm: Application to Cooperative Multiple Task Assignment of UAVs," *48th International IEEE Conference on Decision and Control*, Shanghai, China, December, 2009.

Presentations

[P1] **M. Ramezani**, J. Haddad, and N. Geroliminis,"Integrating heterogeneity dynamics in aggregated urban network modeling and control," 14th Swiss Transportation Research Conference, Ascona, Switzerland, May 2014.

[P2] **M. Ramezani** and N. Geroliminis,"Queue profile estimation in signalized arterials by utilizing probe vehicle data," 13th Swiss Transportation Research Conference, Ascona, Switzerland, April 2013.

[P3] **M. Ramezani**, J. Haddad, and N. Geroliminis, "Macroscopic Traffic Control of a Mixed Urban and Freeway Network," 12th Swiss Transportation Research Conference, Ascona, Switzerland, May 2012.

[P4] **M. Ramezani** and N. Geroliminis, "Arterial route travel time distribution estimation with a Markov chain procedure," 11th Swiss Transportation Research Conference, Ascona, Switzerland, May 2011.

Training Courses

- Advanced modeling and simulation of transportation networks, Italy 2014
  Instructors: Prof. Moshe Ben-Akiva, Prof. Ennio Cascetta
- 10th Short course on dynamic traffic flow modeling and control, Greece 2012
  Instructor: Prof. Markos Papageorgiou
- Nearctis summer school - Traffic modelling for traffic management and cooperative systems, The Netherlands 2011

Research Projects

- METAFERW - Modeling and controlling traffic congestion and propagation in large-scale urban multimodal networks 2014-
  Sponsor: European Research Council, Starting Grants
- Development of perimeter traffic signal control for the city of Sydney 2013-2014
  Sponsor: Roads and Maritime Services, New South Wales, Australia (collaboration with UNSW(Australia))
- Real-time hierarchical control and monitoring of urban traffic systems 2010-2014
  Sponsor: Swiss National Science Foundation
- Coordinated control of ramps with variable speed limits for Swiss motorways
  Sponsor: Swiss Federal Roads Office 2012-2014

Academic Services

Reviewer:

- IEEE Transactions on Intelligent Transportation Systems
- IEEE Transactions on Vehicular Technology
- EURO Journal on Transportation and Logistics
- Swarm and Evolutionary Computation
- Journal of Zhejiang University SCIENCE A
- IEEE Conference on Intelligent Transport Systems
- Transportation Research Part B
- Transportation Research Part C
- Operations Research
- Journal of Advanced Transportation
- Transportation Research Board
- American Control Conference

Organizer:

- Member of local organizing committee of LATSIS Symposium 2012 - 1st European symposium on quantitative methods in transportation systems, Sep. 2012, EPFL, Switzerland