University of Cyprus

Department of Electrical and Computer Engineering

Traffic Demand Management in the Era of Connected Vehicles

Charalambos Menelaou

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the University of Cyprus

June, 2019

VALIDATION PAGE

Doctoral Candidate: Charalambos Menelaou

Doctoral Dissertation Title: Traffic Demand Management in the Era of Connected Vehicles

The present Doctorate Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the ***Department of*** *Electrical and Computer Engineering, and was approved on June 10, 2019 by the members of the* ***Examination Committee****.*

Examination Committee:

Committee Chair _____
Dr. Georgios Ellinas, Professor

Research Supervisor _____
Dr. Christos Panayiotou, Professor

Research Supervisor _____
Dr. Stelios Timotheou, Assistant Professor

Committee Member _____
Dr. Marios Polycarpou, Professor

Committee Member _____
Dr. Nikolas Geroliminis, Associate Professor

Committee Member _____
Dr. Ioannis Papamichail, Associate Professor

DECLARATION OF DOCTORAL CANDIDATE

The present doctoral dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Charalambos Menelaou

ΠΕΡΙΛΗΨΗ

Η κυκλοφοριακή συμφόρηση έχει σημαντικές οικονομικές και κοινωνικές επιπτώσεις τόσο στις σύγχρονες πόλεις όσο και στους πολίτες που ζουν σε αυτές, με τις κυριότερες συνέπειες να περιλαμβάνουν την σπάταλη καυσίμων, την μείωση της παραγωγικότητας καθώς και τον εκνευρισμό των οδηγών. Η κυκλοφοριακή συμφόρηση οφείλεται κατεξοχήν στο γεγονός ότι η ζήτηση υπερβαίνει την χωρητικότητα σε κάποιες περιοχές του οδικού δικτύου. Πάρα το γεγονός ότι διαφορά συστήματα ελέγχου και διαχείρισης της οδικής κυκλοφορίας έχουν η ήδη προταθεί, το φαινόμενο της κυκλοφοριακής συμφόρησης παραμένει καθώς τα μέχρι τώρα προτεινόμενα συστήματα προσπαθούν μόνο να αναδιανείμουν τις ροές κυκλοφορίας στα κύρια τμήματα του οδικού δικτύου μειώνοντας ή καθυστερώντας την δημιουργία κυκλοφοριακής συμφόρησης χωρίς όμως να επιτυγχάνουν πάντα την πλήρη αποτροπή της.

Η παρούσα διδακτορική διατριβή στοχεύει στην πλήρη αποτροπή της κυκλοφοριακής συμφόρησης και στη μεγιστοποίηση της απόδοσης του οδικού δικτύου συνδυάζοντας τα υπάρχοντα μέτρα ελέγχου (π.χ., καθοδήγηση διαδρομής και έλεγχος της κυκλοφοριακής ροής) με καινοτόμες στρατηγικές διαχείρισης της ζήτησης. Η διαχείριση της ζήτησης επιτυγχάνεται μέσω μιας ενιαίας αρχιτεκτονικής κρατήσεων η οποία καθορίζει την πρόσβαση στο δίκτυο ούτος ώστε ένα όχημα (ή ροή οχημάτων) να ταξιδεύει μόνο διαμέσου οδικών αρτηριών (ή τμημάτων του δικτύου) όπου δεν υπάρχει κυκλοφοριακή συμφόρηση. Το πρόβλημα διερευνάται τόσο σε μικροσκοπικό όσο και σε μακροσκοπικό επίπεδό.

Σε μικροσκοπικό επίπεδο, η προτεινόμενη αρχιτεκτονική κρατήσεων παρέχει οδηγίες σε κάθε όχημα σχετικά με τη διαδρομή που θα ακολουθήσει και τον χρόνο αναχώρησης του, με αποτέλεσμα να βελτιστοποιεί μία ή περισσότερες μετρικές απόδοσης (π.χ. η διαδρομή που επιτυγχάνει είτε το συντομότερο χρόνο άφιξης στον προορισμό, είτε την διαδρομή με την πιο μικρή απόκλιση από μια επιθυμητή ώρα άφιξης) διασφαλίζοντας ότι τα οχήματα θα ταξιδέψουν διαμέσου οδικών αρτηριών στις οποίες δεν επικρατεί κυκλο-

φοριακή συμφόρηση. Η προτεινόμενη αρχιτεκτονική ενδέχεται να ενημερώνει τα οχήματα να καθυστερήσουν την αναχώρηση τους έτσι ώστε να δώσουν χρόνο να αποτραπεί η συμφόρηση στις αρτηρίες που θα χρησιμοποιήσουν ή να ακολουθήσουν εναλλακτικές διαδρομές που ελαχιστοποιούν τις προεπιλεγμένες μετρικές απόδοσης. Στην συνέχεια με την χρήση μαθηματικής μοντελοποίησης, γίνεται διατύπωση του προβλήματος το οποίο φαίνεται να είναι ένα δισεπίλυτο πρόβλημα. Παρά τη πολυπλοκότητα του προβλήματος, διάφορες λύσεις έχουν αναπτυχθεί, βασισμένες στον μαθηματικό προγραμματισμό, στο δυναμικό προγραμματισμό, καθώς και στη θεωρία των γράφων οι οποίες προσφέρουν διαφορετικό διακανονισμό μεταξύ υπολογιστικού κόστους και ποιότητας της λύσης. Για την περαιτέρω μείωση της πολυπλοκότητας, προτείνεται ένα σύστημα συνάθροισης στο οποίο το οδικό δίκτυό χωρίζεται σε επιμέρους περιφέρειες όπου ένας γράφος επικάλυψης δημιουργείτε με στόχο να καθοδήγει τα οχήματα με βάση οδηγιών σε περιφερειακό επίπεδο.

Σε μακροσκοπικό επίπεδο, με την χρήση μακροσκοπικών μοντέλων προτείνετε ο συνδυασμός των συστημάτων καθοδήγησης ροών και διαχείρισης ζήτησης στοχεύοντας τον έλεγχο δικτύων ευρείας κλίμακας, τα οποία αποτελούνται από πολλαπλές περιφέρειες. Η καθοδήγηση διαδρομής σε περιφερειακό επίπεδο, χρησιμοποιείται για την εύρεση των βέλτιστων ροών μεταξύ γειτονικών περιφερειών, με στόχο την μεγιστοποίηση των αριθμών των ταξιδιών που ολοκληρώνουν την διαδρομή τους ανά περιφέρεια. Η διαχείριση της ζήτησης χρησιμοποιείται για τον έλεγχο των ροών κυκλοφορίας που προτίθενται να εισέλθουν στο δίκτυο, προτρέποντας μέρος των ροών ζήτησης να περιμένουν στο σημείο προέλευσής τους. Το συγκεκριμένο πρόβλημα διαμορφώνεται ως πρόβλημα μη γραμμικού προγραμματισμού και με την χρήση μη γραμμικού ρυθμιστή προβλεπτικού μοντέλου στοχεύει στην ελαχιστοποίηση του συνολικού χρόνου ταξιδιού (συμπεριλαμβανομένου του χρόνου αναμονής κατά την προέλευση) για όλες τις ροές κίνησης, βελτιστοποιώντας από κοινού τις ροές ζήτησης που επιτρέπεται να εισέλθουν στο δίκτυο, καθώς και την αναλογία των ροών μεταφοράς μεταξύ περιφερειών. Παρά το γεγονός ότι το πρόβλημα είναι μη γραμμικό και μη κυρτό και ως εκ τούτου πολύ δύσκολο να επιλυθεί, αυτή η εργασία αναπτύσσει δύο γραμμικούς ρυθμιστές προβλεπτικού μοντέλου που παρέχουν στενά κατώτερα και ανώτερα όρια σε σχέση με την βέλτιστη λύση. Ο γραμμικός ρυθμιστής προβλεπτικού μοντέλου ανώτερου ορίου υλοποιείται κάτω από τον περιορισμό ότι κάθε περιοχή λειτουργεί πάντοτε σε κατάσταση ελεύθερης ροής η οποία εξαλείφει τους μη γραμμικούς και μη κυρτούς περιορισμούς από τη μοντελοποίηση του προβλήματος. Ο περιορισμός αυτός μπορεί να

αποφέρει άριστα αποτελέσματα όταν η βέλτιστη λύση απαιτεί την λειτουργία της κάθε περιοχής σε κατάσταση ελεύθερης ροής για κάποια χρονική περίοδο. Ο γραμμικός ρυθμιστής προβλεπτικού μοντέλου κατώτερου ορίου επιτυγχάνεται με τη χαλάρωση των μη κυρτών περιορισμών σε πιο χαλαρούς αλλά γραμμικούς περιορισμούς. Επειδή οι ρυθμιστές προβλεπτικού μοντέλου που προκύπτουν είναι γραμμικοί, μπορούν να λυθούν με βέλτιστο τρόπο πολύ γρηγορά για όλα τα σενάρια κυκλοφορίας, καθιστώντας την πρακτική τους εφαρμογή πολύ ελκυστική.

Όλες οι προτεινόμενες μεθοδολογίες και αλγόριθμοι αξιολογούνται μέσω εκτεταμένων ρεαλιστικών προσομοιώσεων λαμβάνοντας υπόψη είτε μακροσκοπικά είτε μικροσκοπικά μαθηματικά πρότυπα κυκλοφοριακής ροής. Τα αποτελέσματα καταδεικνύουν τις σημαντικές βελτιώσεις που μπορούν να επιτευχθούν με την εφαρμογή της ενσωμάτωσης της καθοδήγησης δρομολόγησης με την διαχείριση της ζήτησης όσον αφορά την αύξηση της αποδοτικής λειτουργίας του δικτύου και την μείωση του χρόνου ταξιδιού.

Abstract

Traffic congestion has significant economic and social consequences in modern cities and their citizens, including fuel waste, productivity loss, and driver frustration. Congestion mainly occurs because the traffic demand exceeds the capacity of a certain area of a road transportation network. Although several traffic management and control schemes have been proposed, the phenomenon still exists because current strategies regulate or redistribute traffic flows through different road segments aiming only to reduce or delay the effect of congestion without preventing traffic overload in high demand scenarios.

This Ph.D. thesis aims to develop a framework that completely eliminates congestion while at the same time, it maximizes the efficiency of the road network by combining existing control measures (such as route guidance and traffic flow control) with innovative demand management strategies. Demand management is achieved through a novel reservation architecture that grants access to the network only in case that it is ensured that the requested vehicle (or traffic flow) will travel only through congestion-free road segments (or network regions). The problem is investigated at microscopic and macroscopic levels.

At the microscopic level, the proposed reservation architecture provides instructions to each vehicle regarding the route to follow and the departure time from the origin in order to optimize one or more performance metrics (e.g., earliest destination arrival time, deviation from on-time arrival) without passing through congested road segments. This implies that vehicles may be instructed to delay their departure until some road segments become uncongested or even follow alternative routes that minimize the considered metrics. The problem is formulated in rigorous mathematical terms and shown to be NP-complete in most of the cases. Despite the difficulty of the problem, several solution methods are developed based on mathematical

and dynamic programming, as well as on graph theory, which exhibit a different trade-off between computational cost and optimality. To further reduce complexity, an aggregation scheme is also proposed for multi-region large-scale networks that constructs an overlay graph and derives instructions at the regional level.

At the macroscopic level, the proposed scheme aims to provide both regional route guidance and demand management to control vehicles in a multi-region network considering macroscopic traffic dynamics. Regional route guidance is used to identify the optimal transfer flows between neighboring regions so that the trip completion rate across all regions is maximized. Demand management is utilized to control the traffic flows entering the network by allowing a portion of the demand flows to wait at their origin. The considered problem is formulated as a non-linear Model Predictive Control (MPC) problem that aims to minimize the total travel time (including the waiting time at the origin) over all flows by jointly optimizing the demand flows allowed to enter in the network, and the ratio of transfer flows between regions. Despite the fact that the problem is highly non-convex and hence very challenging to solve, this thesis develops two linear programming MPC formulations that provide tight lower and upper bounds to the optimal solution. The upper bounding linear MPC formulation is obtained by restricting each region to always operate in the free-flow regime which eliminates the non-linear constraints from the formulation but may yield sub-optimal results if the optimal solution requires operation in the congested regime at some time period. The lower bounding linear MPC formulation is obtained by relaxing the non-convex constraints to looser but linear constraints. Because the resulting MPC formulations are linear programs, they can be solved in a fast and optimal manner under all traffic scenarios, making their practical implementation very attractive.

All proposed methodologies and algorithms are evaluated through extensive realistic simulations considering different microscopic and macroscopic level traffic dynamics. The provided results demonstrate the significant improvements that can be realized by applying the proposed integration of routing guidance with demand management in terms of network efficiency and travel time reduction.

Acknowledgments

This Ph.D. dissertation was performed under the supervision of Dr. Christos Panayiotou and Dr. Stelios Timotheou, Professor and Assistant Professor at the Department of Electrical and Computer Engineering of the University of Cyprus, respectively.

First and foremost, I would like to express my most profound gratefulness and appreciation to my advisors, Prof. Panayiotou and Prof. Timotheou, for their patience, open-mindedness and their invaluable guidance, continuous encouragement and their full support during my Ph.D. studies. I would also give special thanks to Prof. Panayiotou for trusting me and offering me the opportunity to pursue a Ph.D. degree at KIOS CoE. Furthermore, I am thankful to Prof. Timotheou for his kindness, and his insightful comments and ideas that were of high importance for completing this Ph.D. dissertation. Both are not only great advisors and scientists but also excellent friends. It was a great honor and privilege for me to work and cooperate with both of them.

Special thanks to Prof. Panayiotis Kolios also for his unconditional support, assistance, friendship, and the sharing of research experience during all these years. The interaction with him, the helpful discussions, and his comments were valuable for completing my Ph.D. dissertation. I would also like to thanks all my co-authors and especially Prof. Marios Polycarpou for the very effective collaboration and the sharing of research experience during all these years.

I would like to express my sincere thanks to my committee members: Prof. Georgios Ellinas, Prof. Nikolas Geroliminis, and Prof. Ioannis Papamichail for their time to review and handle this Ph.D. dissertation.

It was a great pleasure to be a researcher at KIOS CoE as I have enjoyed the pleasant and motivating environment during my Ph.D. studies. Many thanks to all

friends that I have made in KIOS and also to all my childhood friends with whom I have great memories and delightful times. I hope that they will remain a part of my life.

Furthermore, I want to express the most important and the sincerest gratitude to my parents (Menelaos and Panayiota) for their unconditional love and their continuous support throughout my whole life. Thank you for being there any time I need you. Most importantly, I would like my heartfelt gratitude to my brother and sister (Andreas and Ioanna) and also to my beloved girlfriend (Maria) for their love, support, and understanding.

Finally, I would like to dedicate my Ph.D. dissertation in memory of little Menelaos that we had lost him during my Ph.D. studies.

Charalambos Menelaou

Publications

Refereed Archival Journal Publications

1. C. Menelaou, P. Kolios, S. Timotheou, C.G. Panayiotou, and M.M. Polycarpou, "Controlling road congestion via a low-complexity route reservation approach", Transportation Research Part C: Emerging Technologies, vol. 81, pp. 118–136, 2017.

2. C. Menelaou, S. Timotheou, P. Kolios, and C.G. Panayiotou, "Improved Road Usage Through Congestion-Free Route Reservations", Journal of the Transportation Research Board, vol. 2621, pp. 71–80, 2017.

3. C. Menelaou, S. Timotheou, P. Kolios, C.G. Panayiotou, and M.M. Polycarpou, "Minimizing traffic congestion through continuous-time route reservations with travel time predictions", IEEE Transactions on Intelligent Vehicles, vol. 4, no. 1, pp. 141–153, March 2019.

Refereed Archival Conference Proceedings

1. C. Menelaou, P. Kolios, S. Timotheou, and C.G. Panayiotou, "Congestion free vehicle scheduling using a route reservation strategy", IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas de Gran Canaria, Spain, Sep. 15, 2015, pp. 2103-2108.

2. C. Menelaou, P. Kolios, S. Timotheou, and C.G. Panayiotou, "On the complexity of congestion free routing in transportation networks", IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas de Gran Canaria, Spain, Sep. 15, 2015, pp. 2819-2824.

3. C. Menelaou, P. Kolios, S. Timotheou, and C.G. Panayiotou, "A congestion-free vehicle route reservation architecture", in 18th Mediterranean Electrotechnical Conference, Limassol, Cyprus, April 18, 2015, pp. 1-6.

4. C. Menelaou, P. Kolios, S. Timotheou, and C.G. Panayiotou, "Improved road usage through congestion-free route reservations", Transportation Research Board (TRB), 96th Annual Meeting, Washington DC. USA, Jan. 8, 2017

5. C. Menelaou, S. Timotheou, P. Kolios, C.G. Panayiotou, and M.M. Polycarpou, "Optimal Path Selection in a Continuous-Time Route Reservation Architecture", IEEE 20th International Conference on Intelligent Transportation Systems, Yokohama, Japan, Oct. 16, 2017, pp. 1-6.

6. C. Menelaou, P. Kolios, S. Timotheou, and C.G. Panayiotou, "Effective Prediction of Road Segment Occupancy for the Route-Reservation Architecture", 15th IFAC Symposium on Control in Transportation Systems, Savona, Italy, Jun. 6, 2018, pp. 470-475.

7. C. Menelaou, P. Kolios, S. Timotheou, and C.G. Panayiotou, "Effective Multi-region Traffic Control and Demand Management Using an Overlay Route-Reservation Scheme" IEEE 21th International Conference on Intelligent Transportation Systems, Maui Hawaii, USA, Nov. 4, 2018, pp. 1852-1857.

8. C. Menelaou, S. Timotheou, P. Kolios, and C.G. Panayiotou, "Estimating the Critical Density of Road Transportation Networks using Infinitesimal Perturbation Analysis of Hybrid Systems." IEEE 57th Conference on Decision and Control, Miami Beach, FL, USA, Dec. 17, 2018, pp. 1809-1814.

9. C. Menelaou, S. Timotheou, P. Kolios, and C. Panayiotou, "Joint route guidance and demand management for multi-region traffic networks." In 2019 European Control Conference (ECC),Napoli, Italy, June 25 2019, pp. 2183-2188, IEEE.

10. C. Menelaou, S. Timotheou, P. Kolios, C.G. Panayiotou, and M.M. Polycarpou, "Path-based joint demand management and route guidance for multi-region traffic networks." To appear in the Proceedings of 22th International Conference on Intelligent Transportation Systems (IEEE ITSC'2019), Auckland, New Zealand, Oct. 27-30 (2019).

11. C. Menelaou, S. Timotheou, P. Kolios, C.G. Panayiotou, and M.M. Polycarpou, "Scheduling Vehicles for On-Time Arrival using Route-Reservations." To appear in the Proceedings of 22th International Conference on Intelligent Transportation Systems (IEEE ITSC'2019), Auckland, New Zealand, Oct. 27-30 (2019).

Refereed Archival Journal Publications (submitted)

1. C. Menelaou, S. Timotheou, P. Kolios, and C.G. Panayiotou, "Joint route guidance and demand management for real-time control of multi-regional networks", submitted to IEEE Transactions on Intelligent Transportation Systems (June 2019).

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and objectives

Road transportation networks are one of the critical infrastructures that significantly contribute to a country's economic growth as they support the movement of people and goods when and where they are needed. For instance, in EU road transport accounts for up to 5% of its gross domestic product (GDP) [1]. Despite the fact that road transportation is a significant contributor to growth, it also has significant adverse effects to cities and society, with traffic congestion being the primary one.

Traffic congestion has become a critical threat in modern city landscapes resulting in multiple adverse effects such as environmental pollution, non-predictable travel times, and unwanted delays [2]. Time wasted in congestion entails many socio-economic problems, while congestion annually costs up to 1% of the EU's GDP (around 100 billion€ each year) [1]. Besides, an increase in congestion leads to higher travel times, with drivers experiencing about 26 hours of travel delays due to the traffic congestion on average per year [3].

Congestion occurs as traffic demand surpasses the infrastructure's available capacity, a phenomenon mostly observed during rush-hours and is characterized by increased vehicular queuing, lower speeds, and hence longer journey times. Currently, traffic congestion countermeasures fall into three categories, as shown in Fig. 1.1:

1. **Modification and expansion of the road infrastructure**, e.g., building new roads and improving current road junctions (e.g., grade separation) [4].

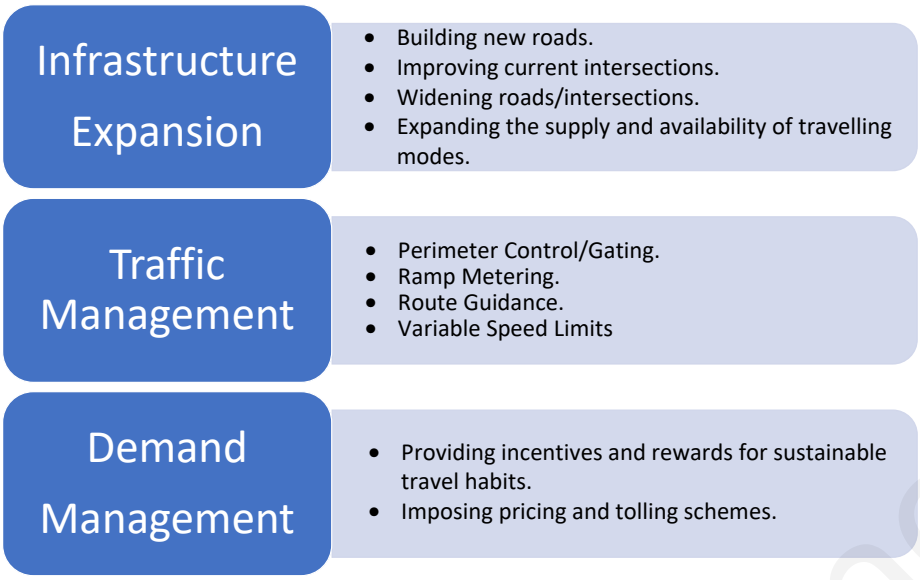


Figure 1.1: The three major categories of traffic congestion countermeasures.

2. **Traffic management** which aims to minimize the travel times of vehicles and also to improve the overall network operation. Recent advances in information and communication technologies (ICT) facilitate the management and control of all vehicular movements within a region of interest, e.g., using ramp metering, gating, perimeter control, and route guidance [5].

3. **Demand management** which intends to alleviate traffic congestion by applying various restriction policies such as economic instruments (road and congestion pricing), regulatory measures and physical restraints (e.g., road closures and parking restrictions) [6]. Switching to alternative modes of transport (buses, trains, etc.) is also considered as demand management.

Nonetheless, traffic congestion does not necessarily occur due to lack of the overall network capacity, and thus, expanding the road infrastructure will be inefficient. Besides, the investment cost and operating expenses of such solutions are quite high, limiting their applicability with recent studies focusing on online demand and traffic management techniques. On the other hand, controlling and managing large-scale transportation networks is a challenging task that becomes even harder to tackle as an increase in demand[1] for mobility results in higher levels of congestion [7].

[1]In this thesis the word "demand" always refers to the number of vehicles that want to access the network.

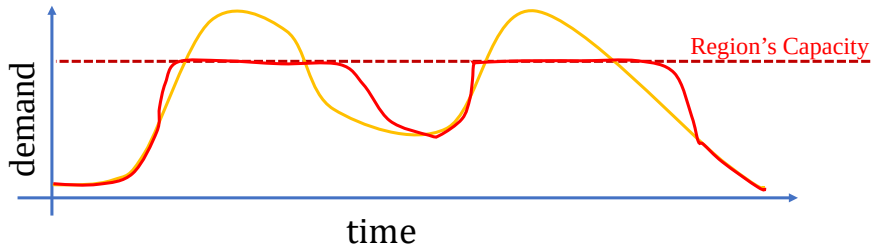


Figure 1.2: A Typical traffic demand pattern observed within an urban area during morning and evening peaks (orange line). Demand management schemes aim to redistribute traffic demand in space and time (red line).

The recent advances in ICT can substantially reduce the congestion problem by employing a plethora of intelligent traffic management strategies [5]. For instance, the existing route guidance strategies can provide real-time trip information to drivers or can offer advice on alternative congestion-free routes in an effort to reduce traffic imbalances across the road network while minimizing travel time [8,9]. Interestingly, offering real-time state information to drivers has been shown to create additional side effects to network utilization, since all rational drivers would opt to follow less congested road segments instead of following the shortest distance paths; this leads to high demand in some road segments, while the rest of the network remains under-utilized [10].

Despite these significant advancements, traffic congestion still persists because the existing route guidance methods seek to improve the **User Optimum** (referred also as the Wardrop Equilibrium [11]). Selfish routing often results in transferring congestion to other road links, which results in network state oscillations, with congestion levels shifting back and forth to different parts of the network [12]. On the contrary, routing solutions that aim to improve the **Social Optimum** can substantially reduce travel times by slightly decreasing traffic demand [10]. Besides social optimum requires some vehicles to be willing to sacrifice some of their travel time or to take longer routes if these actions benefit the network operation. In this way, the number of vehicles sacrificing some of their time for a late departure (i.e., demand management) can be significantly smaller than the number of vehicles that are actually benefiting [13]. Hence, an alternative approach to the problem is to manage driver actions (i.e., routes and departure times) to prevent the occurrence

of congestion so that the increasing demand could be served without building more infrastructure [14].

An example of a traffic demand pattern that can be observed within an urban area is depicted with the orange line in Fig. 1.2. The two peaks represent the morning and the evening peak during which traffic demand surpasses the maximum amount of the traffic capable of being handled by the infrastructure (capacity), in which traffic congestion is unavoidable. Under these circumstances, a demand management scheme intents to influence drivers' travel routine by employing a combination of operational strategies to reduce or redistribute traffic demand in space and time (as depicted with the red line in Fig. 1.2). This problem can be resolved either by promoting alternative travel modes (e.g., public transportation or vehicle sharing) or by suggesting alternative routes to follow or even by providing travel time information to drivers influencing their decisions (in respect to the travel mode choice) [3, 15]. These key facts emphasize that congestion can be potentially alleviated only through innovative policies that can integrate an intelligent demand management scheme with vehicle routing (i.e., traffic management) to support the changing demand while steering traffic away from hot-spot regions [3]. For instance, one approach is to affect drivers' routing decisions (traffic management) and at the same time to control vehicle departure times (demand management). In this thesis demand management refers to the regulation of traffic inflow inside a network, (e.g., by controlling the departure time and route to follow either on a per vehicle basis (microscopic level) or even on a traffic flow basis (macroscopic level)) aiming to eliminate traffic congestion and to minimize travel time.

Interestingly, congestion within an urban area can be reduced by curbing the number of vehicles concurrently using the road infrastructure [16]. Several approaches have been proposed to alleviate the problem. Most of them offer fully connected systems that utilize very detailed link-level information. Hence, their real-life implementation requires various mobility characteristics to be taken into consideration, such as the critical density and the maximum capacity of different links which may be difficult to acquire due to sensor sparsity, the large-scale nature of the network, and the inherent uncertainty that exists due to unpredictable human behavior factors that limit their real-life application [17]. Under these circumstances,

the accurate knowledge of all road segment characteristics is challenging, if not impossible to achieve.

To anticipate the complexity of the detailed link-level information, recent literature proposed the Macroscopic Fundamental Diagram (MFD), which serves as the primary mathematical tool to develop aggregated traffic models for control [18]. The MFD framework offers low complexity modeling of the large-scale urban characteristics capturing the macroscopic relationship between the three main mobility parameters, i.e., speed, flow, and density [18]. The MFD is composed of two distinct regimes, separated at the critical density point: 1) the free-flow regime where traffic flows at its maximum speed (i.e., the free-flow speed) and 2) the congested regime where traffic experiences a speed reduction as congestion emerges. According to the MFD, within the congested regime, an increase in the region's density results in lower vehicle speeds with a higher possibility of gridlock. On the other hand, the probability of gridlock diminishes within the free-flow regime where both driver and network dynamics are well-approximated [16]. The existence of the MFD is shown in [19] (using real data), demonstrating that it can be used to estimate the outflow rate (trip completion rate) of different city regions [18]. Nonetheless, these macroscopic relations are also present when autonomous vehicles are assumed [20].

Guided by the MFD analysis, in this thesis, it is assumed that a critical density exists for every road segment in accordance to the corresponding MFD of the region such that, vehicles can travel with high speeds when operating in the congestion-free regime. On the other hand, if the critical density is overreached (congested regime) then vehicle flows, and speeds become unpredictable. Thus, the key objective of the proposed demand management methodologies is to prevent the vehicle density in a region from exceeding its critical density value. For this Ph.D. thesis, we assume that the critical density of each road segment is known (e.g., through the MFD analysis) however, even if these are not known they can be computed through extensive simulations or other tools like perturbation analysis [21].

Under the above framework, simple control mechanisms can be employed to maintain traffic flow below the critical density of each road segment. As such, in this thesis, we proposed a novel Route Reservation Architecture which ensures

that each vehicle is scheduled to only traverse along *congestion-free*[2] road segments. When a vehicle is about to begin its journey, it sends a route reservation request to a centralized Road Side Unit (RSU) to inform it about its **origin and destination**. The RSU is responsible for all route reservations, by keeping track of the number of vehicles that have reserved different time frames along specific road segments. Road segments become temporarily unavailable whenever reservations reach their critical density value, and they are reconsidered only when the allocated time slots have elapsed. Hence, these route reservations provide estimates of the future state of each road segment, and thus when reserved density of a segment exceeds its critical density during particular time intervals, the RSU flags it as *non-admissible*. Whenever the RSU receives a route request, it computes the Earliest Destination Arrival Time (EDAT) route for the vehicle such that any road segment that reaches its critical density is avoided taking into consideration the fact that it may be better for a vehicle to wait at its origin until certain road segments become *admissible*. Once the RSU determines the best path, it updates its reservation table with the assumptions that: (a) vehicles follow the recommended path and (b) vehicles travel with free flow speed (or the speed at capacity).

Another motivation of this thesis is to use the above reservation principles to distribute the flattening of demand over a larger period (i.e., some vehicles should enter the network earlier or later) such that the peak demand will not exceed the network's capacity. As a result, congestion will be avoided, and vehicles will arrive at their destination on-time, without excessive delays in the road network. In that case, the reservation architecture is similar; however, the optimization problem and objective are different with vehicles sending to the RSU their origin-destination pair and their desired arrival time at the destination. Then, the RSU determines the time that the vehicle should depart from its origin and the path to follow, such that it will arrive at the destination on or before the desired arrival time. In this approach, the objective of the RSU is to minimize the difference between the departure and the desired arrival times such that congested road segments are avoided and travelers are not significantly inconvenienced (e.g., they do not arrive too early at their desti-

[2]Congestion-free road segments, also referred as *admissible road segments*, are those that their critical density value is not reached.

nation). Furthermore, through the reservation architecture, the RSU has a reasonable estimate of the future states of the network; thus, it will route vehicles only through non-congested road segments. Note that this approach and the route reservation architecture operate considering microscopic traffic dynamics (microscopic level).

Furthermore, this Ph.D. thesis proposes a multi-regional level Model Predictive Control (MPC) scheme that integrates route guidance with a demand management method (macroscopic level). Given the origin and destination pairs of the vehicular flows that request to navigate within the considered road network, the proposed scheme tries to find the path that minimizes the destination arrival time of all vehicles. The proposed MPC scheme does not only suggest a path to follow but also manages the entering rate of the external inflow rates, resulting in a congestion-free operation since a portion of the inflows is restricted at their origins (demand management). In this way, route guidance finds the optimal transfer flows across neighboring regions, while demand management regulates the external inflow rates in the same manner as proposed in the route-reservation architecture. The resulting formulation assumes that all routes within a region have a constant length independent of their origin-destination pair. However, this assumption is often violated in practice; thus, in this thesis, we also present a reformulation of the problem that explicitly defines the paths followed for each origin-destination pair. The novelty of this thesis lies both in modeling and solving the resulting problem under these control measures.

An important assumption of the proposed methodologies is that each region's critical density is constant and known in advance. However, in realistic scenarios, the critical density can change over time for a variety of reasons including, changes in demand or *OD* pairs, due to roadworks or accidents and due to environmental factors, such as weather conditions [22]. In this way, the above assumption is relaxed on the intention to estimate its value in an online fashion by employing Infinitesimal Perturbation Analysis (IPA) which can be utilized to capture the dynamic changes in the critical density value.

1.2 Benefits of the proposed demand management methodologies

The introduced methodologies have the following benefits.

- Vehicles are routed through non-congested paths which is a benefit for the individual vehicle in the sense of the experienced travel time.

- By not allowing vehicles to go through segments that are above their capacity, it "protects" other vehicles that have already reserved those segments and it guarantees that they will not experience congestion either; thus the approach has also a social benefit. This is in contrast to other time-depended routing approaches which may allow a vehicle to enter a road but change the road segment cost dynamically (e.g., the time to traverse the road segment). In this case, if the vehicle finds it beneficial, to traverse a slightly congested road in terms of arrival time at the destination, it also adversely affects the delays of all other vehicles that are also scheduled to traverse the same road segment.

- By suggesting vehicles for a delayed departure, means that they are kept away from the road network minimizing their travel time and the cost associated with the lost productivity or the environmental impact. Furthermore, sustaining network operation under free-flow conditions facilitates accurate travel time estimation (i.e., assuming free-flow speed conditions) which in turn enables the estimation of the destination arrival time.

- The late departures ensure that no congested conditions will appear and hence the adverse effects of congestion such as unnecessary fuel consumption, time losses, and health issues can be potentially eliminated.

- By sustaining a region's density below its critical value, the related macroscopic level problems (i.e., the MPC framework which is a Non-Convex Non-Linear program) can be relaxed into a linear formulation that leads to fast and optimal solutions, which enables real-life implementation.

1.3 Thesis contributions

The primary purpose of this thesis is to enhance an urban network operation through effective joint demand management and route guidance methodologies in which vehicles can either be delayed at their origins or routed through longer but non-congested routes in order to minimize their travel time. This results in congestion elimination with beneficial effects for the entire transportation system. Hence, this thesis has made several significant contributions to the field of demand management and control of urban transportation networks, which are briefly summarized below:

- A novel route-reservation architecture is introduced for congestion-free routing, in the context of Intelligent Transportation Systems (ITS). Given the obtained reservations, the mathematical formulation of the related routing problem (referred to as the Earliest Destination Arrival Time (EDAT) problem) is derived in both the continuous and discrete time domains. In the process, we examine the potential of the proposed architecture as an approach to alleviate road congestion and to minimize the time for vehicles to reach their destination. A rigorous complexity analysis of the EDAT problem is derived in Chapter 3 that demonstrates the NP-completeness of the problem. To solve the EDAT problem optimally, a Mixed Integer Linear Programming (MILP) approach is developed that allows delayed departures and considers admissible road segments in the continuous time domain. Although this approach yields to the optimal solution, it is computationally expensive and hence, three alternative heuristic algorithms are proposed to solve the EDAT problem. The proposed algorithms offer different trade-offs between the solution quality and computational complexity with the iterative Dijkstra's-based Route-Reservation Algorithm (RRA) being the primary heuristic used for solving the EDAT problem. Furthermore, to resolve any fairness issue that may arise due to the first-come-first-served execution of the EDAT problem, a load balancing scheme is developed that leads to better network performance in high congestion scenarios. This is achieved by balancing the traffic across different road segments to improve the homogeneity of the network. A Time-Varying Multiple Linear Regression method (TVMLR) is proposed in Chapter 4 to enhance the accuracy

of route reservations through better travel time predictions.

- Furthermore, an extension of the route reservation architecture is presented in Chapter 5 in which an aggregated and scalable route-reservation architecture is proposed that employs an overlay graph to summarise the per link route reservations into regional level metrics. In addition, a load balancing scheme that operates over the overlay graph is proposed in Chapter 5; aiming to enhance the performance of the aggregated route reservation scheme.

- Another extension of the route reservation architecture is presented in Chapter 6 where a novel reservation-based architecture is developed to compute vehicles routes and departure times such that drivers reach their destination at the desired arrival time while guaranteeing a network-wide congestion-free operation.

- The mathematical formulation of the non-linear non-convex joint route guidance and demand management multi-regional MPC scheme. A solution to the problem is provided by an approximate MILP formulation that is derived also in Chapter 7. Towards the global optimality of the proposed non-linear MPC scheme, a novel Linear Programming (LP) MPC formulation is derived that yields tight lower bounds to the optimal solution. A second LP formulation is presented in Chapter 7 aiming to provide an upper bound solution which is also a feasible but sub-optimal solution to the original non-linear MPC scheme, that is achieved by restricting the density of each region within the non-congested regime. Also, a path-based joint route guidance and demand management scheme is introduced in Chapter 8; providing similar LP relaxations as those made for the multi-regional MPC framework.

- An Infinitesimal Perturbation Analysis (IPA) is performed in Chapter 9 to identify the sensitivity estimates of the performance measure of the instantaneous throughput (i.e., number of vehicles that exit a region) of a transportation region with respect to the critical density. The proposed IPA estimator is applied over the actual system to estimate its critical density value.

1.4 Thesis outline

This thesis consists of ten chapters out of which Chapters 3 - 9 describe the technical contributions. The remainder of this thesis is structured as follows.

Chapter 2 discusses the relevant literature work and elaborates on the contributions of this thesis compared to the state-of-the-art.

Chapter 3 introduces the route-reservation architecture for achieving congestion-free routing in the context of Intelligent Transportation Systems. In this chapter, the EDAT problem is mathematically formulated in both continuous and discrete time domains, while according to a rigorous complexity analysis it is shown that the EDAT problem is NP-complete in most cases. Furthermore, in this chapter an optimization problem referred to as the Traffic Load Balancing Problem is also formulated. Several solutions for both formulations with complementary objective functions are proposed. Detailed simulation results across a particular region of the San Francisco area, demonstrate the great benefits that can be realized by applying the proposed solutions.

Chapters 4, 5 and 6 introduce several extensions and variations of the RRA architecture. More specifically, Chapter 4 enhances the route reservation architecture by considering the modeling uncertainties of road segments travel times through a time-varying regression method, enabling real-time accurate, travel-time predictions which leads to the minimization of reservations errors. Chapter 5 suggests an aggregation of the route reservation architecture scheme where a heterogeneous urban area is partitioned into multiple homogeneous regions. The proposed approach creates an overlay graph which is used by the reservation architecture to control the traffic flow within each region. Chapter 6 addresses the problem of scheduling vehicle departures from their origin in order to arrive at their destination on times taking into consideration the route reservation architecture. The mathematical formulation of the proposed problem is presented, and an efficient algorithmic solution is derived. Microscopic simulation results demonstrate the substantial improvements obtained by applying the proposed approach in realistic scenarios.

Chapter 7 introduces a MPC framework that combines the multi-regional route guidance with a novel demand management method. A formulation of the related non-linear non-convex MPC problem is presented while a Mixed Integer Linear

Programming solution and a Linear Programming solution that approximates the original non-linear problem are developed. By allowing each region to operate in the free-flow regime (in a similar manner with the route reservation architecture) this thesis proposes a second linear formulation that offers a feasible upper bound solution to the original non-linear MPC problem. In a similar way, Chapter 8 introduces a path-based formulation of the above multi-regional MPC problem which provides similar MPC formulations as the multi-regional approach but in a different modeling framework. Extensive simulation results demonstrate that the linear MPC approaches (multi-regional and path-based level) execute in real-time and yield near-optimal results even under heavy traffic scenarios.

In Chapter 9, a stochastic Fluid Modeling framework is adopted to estimate the critical density value of a homogeneous region, where the route-reservation scheme is employed to control the traffic within the related region. The estimation of the critical density value, is based on an Infinitesimal Perturbation Analysis scheme, which can be employed in an online fashion to capture the dynamic changes in the critical density value.

Finally, Chapter 10 summarizes the main contributions of this thesis and concludes with an outline of directions for future work.

Chapter 2

Literature Review

2.1 Gating, Perimeter Control and Ramp Metering

Currently, the *gating* and *perimeter control* methods constitute the state-of-the-art solutions for addressing the traffic congestion problem [23, 24]. Gating control aims to regulate the amount of traffic that resides inside a homogeneous [25] region. This is done by allowing external traffic to enter if the critical density of the region's MFD has not been reached [26, 27], and this can be achieved, for example, by using street closures or by controlling the traffic lights phases. Fig. 2.1 provides an illustrative example where traffic lights installed at the boundaries of a protected region are used to control the external flows entering the region. To avoid the extensive amount of data that is required for the characterization of each regions' MFD, recent works attempt to use a reduced MFD constructed using real-time measurements [28].

Similarly, the two-level perimeter-and-boundary control is applied in multi-region networks to regulate the traffic exchange between regions and the outside world [23], [24]. At the first level, an urban area is clustered into inter-connected homogeneous regions that maintain modeling accuracy at the macroscopic level. At the second-level, similar to gating, vehicles are allowed to enter in the region only if the critical density has not been reached [29]. By discriminating between different areas of the network based on their homogeneity, more accurate decisions can be made. Furthermore, decision making using feedback control at the macroscopic level is simpler to implement and computationally efficient, since it does not require extensive traffic information (e.g., the per-link densities, speeds, and flows) [24], [29].

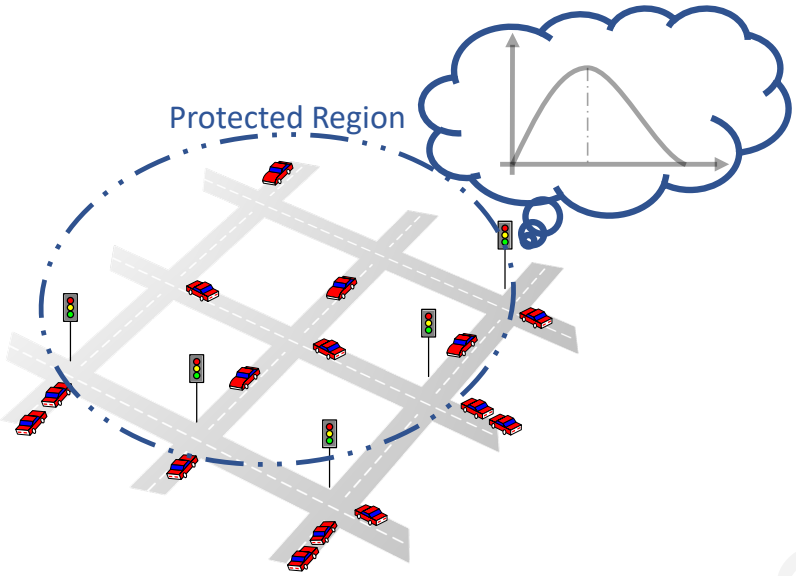


Figure 2.1: Example of gating control schemes.

Other recent efforts that formulate hierarchical perimeter control problems for multi-region urban areas also use the MFD dynamics, making easier the employment of efficient Model Predictive Control (MPC) frameworks [30,31]. These methods yield better performance compared to feedback control because they are more robust to traffic demand and modeling uncertainties.

Similarly with perimeter and gating control, ramp metering is also an admission control approach that aims to improve freeway efficiency by regulating the inflow from on-ramps to the freeway mainstream [32,33]. Ramp metering is a traffic responsive strategy that considers real-time measurements to coordinate ramp metering actions to control either a single on-ramp [33] (i.e., uncoordinated approach) or more consecutive on-ramps (i.e., coordinated approach) [34]. The work in [35] presents an efficient real case implementation of a ramp metering scheme that controls six consecutive inbound on-ramps on the Monash Freeway in Melbourne, Australia achieving significant reductions of travel times and congestion levels.

The major drawback of admission techniques is that, if gated links do not have sufficient space for queuing, queues due to gating may obstruct the upstream network destinations. In this way, the benefits of these control policies are reduced [36,37]. To anticipated this issue, the work in [36] proposed a balanced queue strategy that reduced remarkably the length of the observed queues by balancing the flow proportionally to the saturation flow of each gated segment. Additionally,

the work in [36] analyzes the queuing behavior at the gated segments indicating that the queue length may not necessarily increase compared to when no perimeter control is applied. By the same token, the work in [38] proposes a hierarchical control scheme that combines perimeter control (at a higher level) with a lower-level control scheme that is applied at intersections to improve the system's performance when spill-back phenomena occur. Likewise, a ramp metering approach suggested in [35] utilizes a threshold method that can potentially avoid the creation of long queues in on-ramps. Despite that efforts, queues can be generated within a region (called as "artificial inter-regional queues") which can contribute to unwanted delays. Recent work presented in [37] utilizes an on-line adaptive optimization scheme that promises a better congestion distribution as it tries to anticipate the inter-regional queuing problem by considering how generated queues affect the vehicular movements.

The proposed route-reservation architecture does not distinguish endogenous and exogenous flows and applies a more targeted control over all vehicles preventing the formation of long queues and excessive delays. Furthermore, the aforementioned approaches do not have a reliable mechanism for predicting the future state of the network, which is something achieved through the proposed reservation architecture.

2.2 Routing techniques based on travel time prediction

Recent developments in Intelligent Transportation Systems provide a plethora of complementary solutions to minimize travel times by guiding drivers via the shortest-travel-time paths [39]. Initial attempts towards this direction are the Dynamic Traffic Assignment (DTA) and the Route Guidance schemes that have attracted a lot of attention to accomplish real-time dynamic traffic management, aiming to improve either User Equilibrium or System Optimum under time-varying demand conditions [8,9]. Hence, route guidance and Dynamic Traffic Assignment (DTA) constitute the primary routing advisory schemes that seek to reroute traffic flows towards alternative routes aiming to reduce traffic imbalance across the road network. However, the appeal of route guidance methods is strengthened by the recent advancements in in-

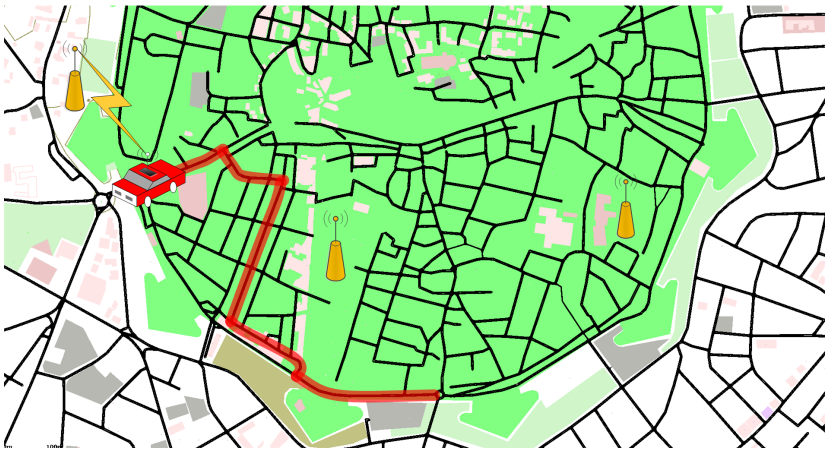


Figure 2.2: Example of Route Guidance schemes. In the depicted figure vehicles are equipped with onboard unit that informed drivers about their shortest time route.

formation and communication capabilities of onboard units which are now capable of providing real-time traffic state information to drivers and recommend alternative routes to follow. An example of such a scheme is depicted in Fig. 2.2.

Other routing efforts investigated in the literature consider how departure time choices can affect the network operation and the possibility to alleviate congestion [40, 41]. The state-of-the-art routing method in this domain is the Decreasing Order of Time (DOT) algorithm [42], which efficiently finds the time-dependent shortest path (using travel-time) within a user-chosen time window. Similarly, the work in [43] that is based on Bellman's principle [44], calculates the shortest paths from all nodes to a given destination in a network with time-dependent generalized link costs.

Even though the above approaches are of particular interest to ITS applications, most of them do not consider the unpredictability of driver behavior that is observed, especially in the congested regime [10]. In their majority, scheduling decisions are made through shortest-travel-time paths according to these state estimates but neglect the adverse effects that may occur when the selected road segments become congested [45]. Hence, these algorithms do not consider the changes in the traffic state when scheduling decisions are made, and thus, there is no guarantee that the traffic state will not experience congestion. Along these lines, the work presented in [46] indicates that route travel times are affected by the segment's traversal time

and the delays observed at the intersections (expressed as travel time penalties). An additional disadvantage of using traffic state estimation in routing is that estimation is required for all road segments, which is usually not the case as state estimation is only available for the main road segments.

However, when there is an inconsistency between the observed travel times and travel time predictions, then the performance achieved by routing methods diminishes [47]. Clearly, the issue mentioned above can impact other state-dependent schemes and thus, a large volume of work has focused on achieving accuracy in travel time predictions; using analytical or statistical methods [48]. Both methods can be employed to predict travel times in a deterministic or stochastic manner [41, 49], always using data obtained through various traffic surveillance sensors (e.g., loop detectors, mobile detectors, radars, and cameras [50]). With respect to the analytical methods, Kalman filter algorithms constitute the primary state estimation method where recent observations that become available are used to update the state variables continuously. Unfortunately, the Kalman filter cannot be utilized in our approach as there is no simple and accurate model that can be applied alongside route reservations [51].

To address the above shortcomings this Ph.D. thesis proposes a prediction-based approach. First, it should be noted here that the proposed solution differs from the current state-of-the-art methods on time-depended routing, as it does not only change each road segment's costs according to the segment's density but also restrict the segments inflow. Besides, the key advantage of the proposed demand management methods is that the network's density is sustained below its critical value, ensuring a congestion-free operation. By doing this, travel times can be minimized while an accurate and straightforward travel time prediction method can be employed to predict travel times.

2.3 Routing methods based on the Macroscopic Fundamental Diagram

Similar with perimeter control approaches, the regional-level route guidance frameworks partitioned the network into smaller homogeneous [25] regions within which

vehicles are responsible for following a regional-level path to better spread the traffic load across a larger area of the network [38, 52]. Furthermore, the work in [53] investigates the properties of a dynamic regional-level traffic assignment method with departure time choices. More specifically, a state-depend optimal control problem is formulated aiming to minimize the total travel time under the constraints of a fix traffic volume that should be served within a pre-specified period; demonstrating that the mismatch of improper flow propagations can be avoided by considering the time lag between traffic inflow and system response. Along the same lines, an aggregated and approximate dynamic traffic assignment model is introduced in [54] that incorporates the MFD dynamics to establish regional routing under stochastic user equilibrium conditions.

The majority of regional-level route guidance frameworks are centralized, and they are implemented at the macroscopic level without considering the detailed lower-level traffic dynamics. Thereby, the MFD is used, as it can offer low complexity modeling of large urban networks. In this direction, a route choice strategy was developed in [55] to alleviate congestion in urban areas, by considering the effect of aggregated regional and partially known sub-regional dynamics. Also, the latter work investigates the impact of drivers behavior on the MFD model, demonstrating its superiority compared with route guidance schemes that do not consider the drivers' behavior. Apart from this, the use of advanced variational methods can successfully estimate MFDs resulting from different driver route choices, as indicated in [56]. In line with the aforementioned works, the study in [57] demonstrated that the shape of the MFD and the size of the hysteresis loop could be affected by the redistribution of traffic achieved through online travel information. Hence, despite their high efficiency, macroscopic route guidance schemes have been challenged about their real-case implementation because of their aggregate control decisions. To address this issue the work in [58] implements a hierarchical MFD based route guidance framework that can translate the aggregate regional-level control actuation into lower-level traffic decisions.

Another crucial issue of MFD-base route guidance schemes is that their majority assumes that all paths passing through a region have equal trip lengths [59], an assumption which is not always valid. This issue has been addressed in [54] where

a dynamic assignment method is derived that considers different trip lengths. An extension of the latter work is presented in [59] where trip length distributions are explicitly estimated to calibrate the MFD model.

Regrettably, these solutions are not able to cope well with heavy congestion levels; usually, such approaches aim to control restricted areas (e.g., the city center), so that performance improvements occur only for scenarios with relatively light traffic. This is due to the fact that in high demand, a load balancing method can only delay the emergence of congestion but not prevent it. The latter can only be achieved by sustaining the total number of vehicles in all regions below their critical density [16]. Recent attempts trying only to control the total number of vehicles result in travel time imbalances since traffic is not evenly distributed across the regions [58, 60].

2.4 Model Predictive Control methods for transportation systems

Model predictive control (MPC) approaches are increasingly being employed to control traffic congestion, with the MFD serving as the prediction model. MPC can optimize the current states while its ability to consider future implications through the region's MFD model [61]. Model predictive control (MPC) has been employed as the primary control mechanism for route guidance problems due to its ability to optimize the current control actions by considering future state estimates [61]. The works in [62] and [30] initially used a non-linear MPC framework to control a freeway system and a two-region urban network, respectively while the work in [63] use an MPC framework to coordinated the ramp metering actions in a freeway network. A hybrid MPC scheme is presented in [64] for an urban region, equipped with the time switching plans together with perimeter control where, the non-linear MPC problem is approximated to a MILP, showing the importance of the approximate model regarding the required computation times for real-case implementation. Furthermore, the work in [65] and [66] utilized an Extended Kalman Filter framework to provide real-time traffic state estimates to the MPC, a step that transforms the non-linear problem into a linear-parameter-varying model. Additionally, the work presented in [67] applies a gradient-based optimization approach to sub-optimally

solve the non-linear MPC problem that aims to balance the trade-off between the level of congestion and the reduction of emissions.

In this direction, this thesis proposes an MPC framework to jointly solve the region-level route guidance and demand management problem in order to find the best alternative routing strategies which minimize the cumulative total time of all vehicles, where the total time of each vehicle accounts for both the waiting time outside the network and vehicles travel time. However, demand management tackles congestion by sustaining the region's density below the critical values and in doing so minimize the observed travel times. The formulated non-linear MPC problem is approximated into a Linear Program (LP) formulation. The resulting LP formulation can be solved using standard LP solvers very fast, as the derived solution does not depend on the initialization of the solver, which is the case for non-linear solvers.

2.5 Tolling systems and Demand Management schemes

Congestion pricing (CP) has been a recurrent measure in trying to alleviate congestion [68], where charges are applied to regulate the entering rate of vehicles in a controlled area of interest. Another approach is road pricing; an economic policy that controls road usage while constituting a credible long-term option for maintaining the road infrastructure [6] (an example depicted in Fig.2.3). Other methods force drivers to pay a cost proportional to the road infrastructure they are using in such a way that may help to alleviate traffic congestion while aim to reduce their impact on environmental pollution [69]. Even though pricing schemes have long been appreciated due to their efficient properties (demand management), their acceptability constitute a significant issue [70]. However, congestion pricing has found successful applications in many places around the globe. For instance, the City of London uses automatic vehicle license plate recognition to impose a charge on the driving vehicle within the charging zone between 07:00 and 18:00, Monday to Friday.

The work in [71] proposed a credit-based congestion pricing scheme where road tolls based on the negative externalities associated with driving under congested conditions (a revenue-neutral policy) within which the generated fees are returned

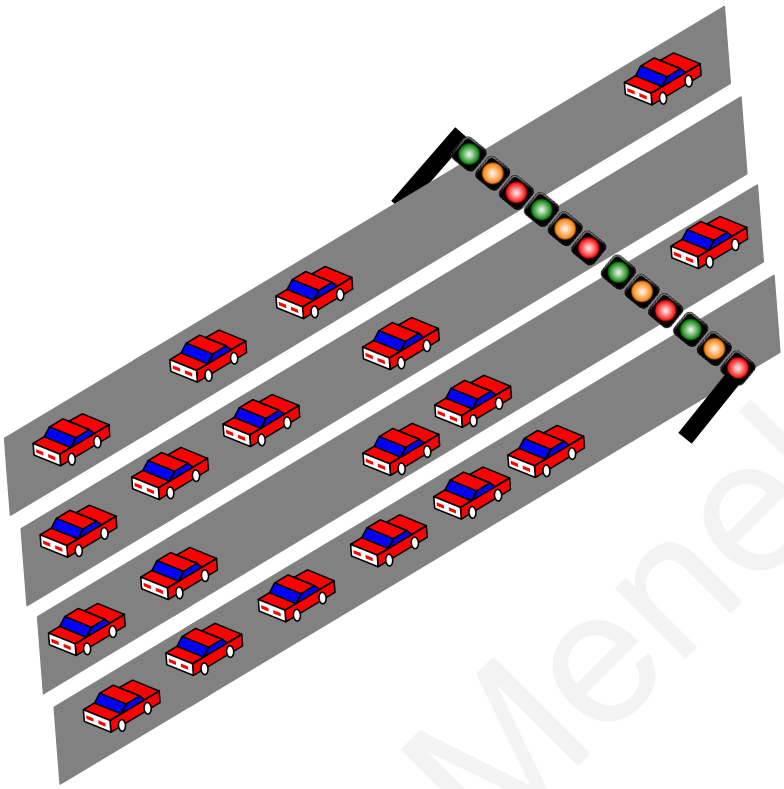


Figure 2.3: Example of tolling/pricing scheme in freeways.

to all licensed drivers uniformly. Under this setting, the frequent long-distance peak-period drivers subsidize average drivers, in effect paying them to stay off congested roads. The work in [72] proposes a new tax rule where not all links of a congested transportation network should be tolled to enhance its practical applicability. For instance, the work in [73] combines traffic assignment with congestion toll, aiming to reduce the size of the tolled area.

As traffic demand increases in road networks, demand management strategies are employed to meet the increased traffic demand with the best cost-effective and efficient manner [69]. Initial attempts on travel demand management schemes tried to motivate the drivers of single-occupant vehicle to use alternative modes of transport. Over time, demand management literature shifted to infrastructure based approaches where for example availability restrictions were imposted on the road network or distance-based pricing schemes are investigated [3]. A real-case application of a demand management scheme is the High-Occupancy Vehicle (HOV) Lanes in which some lanes of the road infrastructure are restricted to be used by vehicles with more than two passengers including buses and carpools. These restrictions

may be imposed on a full-time basis or only during the peak hours always preserving travel time reductions for their users as compared to the general purpose lanes [74]. Unfortunately, the number of people who are willing to use HOV lanes is limited as the benefits of HOV lanes are significant only in the case that HOV lanes are well-utilized while at the same time all the general purposes lanes are over saturated [74].

The literature also reports incentive reward programs to encourage travelers to try alternative transportation choices such as ride sharing or alternative means (i.e., public transports, cycles, etc.) [3]. Similarly, a large number of recent studies are investigating the concept of shared vehicles where users can access them at any time applying charges based on their travel time, or distance traveled [75]. Nevertheless, shared vehicles are established as a more flexible option for travelers that primarily rely on public transport avoiding the necessity to bear the costs of vehicle ownership while diminishing pollutant emissions and the need for parking areas [76].

Demand management schemes are not limited to influencing traveling modes but also to promote alternative routes to follow and even to suggest late trip departure times (distributing demand in space and time) [3]. Besides, the congestion levels during peak hours can successfully be determined based on traveler departure times, and thus departure-time demand strategies aim to regulate drivers departure time choices [77,78]. Similarly, route guidance schemes can be combined with departure time strategies pursuing alternate routes when their usual route is expected to be congested [3,13].

Despite the substantial efficiency of toiling and demand management methods, they are only part of a larger set of approaches to curb congestion [69]. Interestingly, the work in [79] presents a survey on how real-time travel information can alter a traveler's initial decision on the choice of mode, travel route and departure time in the cities of Pittsburgh and Philadelphia (USA). The Survey indicates that 68% of travelers in Pittsburgh and 86% of users in Philadelphia changed their original travel route, while 47% of users in Pittsburgh and 66% of users in Philadelphia changed their initial departure time. On the other hand, the effect on mode choice was less noticeable, with 6% of travelers in Pittsburgh and 2% in Philadelphia changed their mode of transport [79]. Considering these insights, this thesis also investigates

how route reservations and route guidance schemes can jointly be looked at with demand management and traffic management in a way that the network operation is sustained at its maximum (social/system optimum) while at the same time travel times are minimized.

2.6 Infrastructure reservation based approaches

Time-slot reservations are not new in the literature. The proposed route-reservation method has been conceived based on time-slot reservation models used to solve the ground holding problem for Air Traffic Management and Control Systems (ATM/ATC) since airport utilization increases while runway capacity remains constant [80]. Specifically, to increase their runway capacity, airport ground control allocates specific time slots for each airplane that requests take-off or landing. Time slots are shared among arrival and departure flights, and planes are instructed to follow their schedules without any delays or deviations. In doing so, the airport efficiency is significantly improved, as shown in [81] and [82].

This concept is also introduced in road transportation networks, with the initial work in that direction appearing in [83] within which trip reservations are proposed to relieve the holiday congestion problem on a rural motorway similar to the train seat reservations. More specifically, [83] studied how the trip reservations can be affected from adjustable departure times by quantitatively evaluating reservations with a stated-preference survey. Along the same lines, the work in [84] conceptually proposed a highway booking system that operates alongside other driver information systems. In [84] for each time period each road segment has an available capacity where the number of available seats in each vehicle is considered within the problem's capacity constraints. Both works mentioned above [83, 84] indicate that the trip reservations are promising as they can significantly improve the transportation systems efficiency.

Trip booking methods also are investigated in [85] where an infrastructure manager uses a slot allocation algorithm to manage the demand (departure time allocation) for pre-specified routes. Similarly, a congestion pricing alternative is proposed in [86] where a reservation system is developed to control the capacity of vehicles

that enter a protected region; by managing the vehicles' flows passing through a cordon.

The idea of trip reservations on highways is also investigated in recent literature, as shown in [87], [88] confirming that the efficiency of a reservation system can overpass the existing traffic management methods. Extensions of the last two works are presented in [89] and [90] where reservations are used only by users that are willing to pay in order to access a high priority lane providing a better quality of service and travel time guarantees. Also, in the latter work, an auction-based reservation is proposed to reduce the inefficiencies due to user heterogeneity.

A major contribution of this thesis is that it explicitly uses a reservation system and formulates and solves an optimization problem that allows vehicles to arrive at their destination avoiding road segments that are expected to be at their capacity.

2.7 On Time Arrival approaches

Several approaches have already been proposed to address the on-time arrival problem by determining each vehicle's departure time and the associated route aiming to either maximize the vehicle's on-time arrival probability or to minimize the expected traversal time [91,92]. The majority of the literature considers link-level dynamics, assuming that each link's travel time distribution is known [93]. However, as mentioned above, this approach is not easy to be implemented, especially during congested conditions since travel time distributions can hardly be predicted [16].

There is also a great interest in practical aspects of stochastic routing that aim at finding the least expected travel time paths or the most reliable paths, where the travel-time on each road segment is a random variable with an associated probability distribution [91,92]. The objective of the most reliable path problem is to reduce the risk of arriving late rather than to minimize the expected travel time [94]. For instance, some travelers tend to sacrifice travel-time to take a more reliable route when hard deadlines are considered. The primary issue with these approaches is that routing, and scheduling decisions determine the traversal path without considering the adverse effects of congestion and the changes in traffic state due to unreliable estimates of the travel times [45].

The stochastic on-time arrival problem is formulated as a stochastic dynamic programming problem [95] and solved by determining the optimal path at each node based on the travel-time realized on that node [93]. Nonetheless, these approaches are computationally expensive, making them non-practical for real applications since all the detailed link-level dynamics should be taken into consideration for all routing decisions. Such guarantees can be only provided through the use of the reservation architecture as proposed by this thesis as the unique solution to the problem is to prevent congestion altogether by restricting the number of vehicles in the network below its critical density [16]. This thesis derives an extension of the proposed route reservation architecture in which the departure times of each vehicle is controlled (i.e., apply demand management) in an effort to sustain travel times around those achieved assuming free-flow speed conditions while ensuring the on-time arrival at the destination [96, 97].

2.8 Infinitesimal Perturbation Analysis methods

Stochastic fluid models (SFMs) have been developed and used for control and optimization of dynamic networks even though modelling accuracy might sometimes be less than ideal. The SFM modeling enables the abstraction of the system to a fluid queue and derives gradient estimators for the performance measures of interest (e.g., queue throughput and packet delay) with respect to an assigned control parameter (e.g., buffer maximum content). Then, Infinitesimal Perturbation Analysis (IPA) is employed in order to compute the gradient of a performance metric which in turn can be used to optimize the selected control parameter (as depicted in Fig. 2.4). In the works presented in [98, 99] SFM is proven as an efficient technique to identify the optimal buffer size of a single queue system (i.e., single node SFM). The derived IPA gradient estimators are proven to be unbiased and non-parametric and are able to estimate the optimum value in an online fashion [100]. In addition to network optimization solutions, the IPA framework is also utilized for performance-regulation purposes as introduced in [101].

Recent works in transportation networks employing the SFM framework and IPA analysis, including [102, 103] trying to solve the traffic-light control problem for a

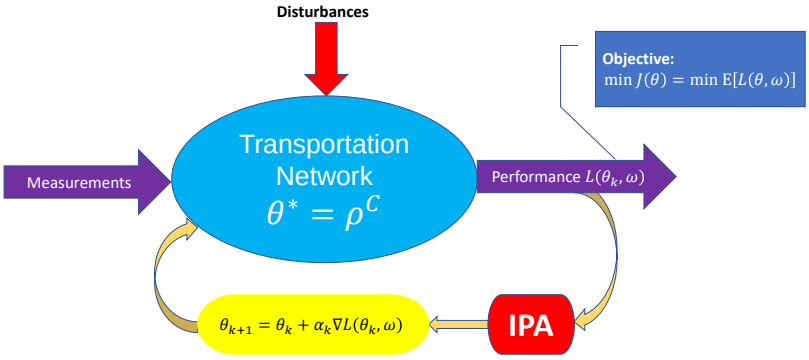


Figure 2.4: Execution Procedure of Infinitesimal Perturbation Analysis scheme.

single intersection. A recent work presented in [104] extended the approaches above to solve the related problem considering multiple intersections. In these works, the on-line gradient estimators are used to iteratively adjust the optimum light cycle length over a traffic congestion metric with respect to the controllable variables that in turn define the green and red cycle phases. The work in [105] tries to control the red/green phases over a signalizing intersection to regulate congestion under a given reference level (queue length). The major advantage of these approaches is that vehicles flow rates are measured only when specific events occur with the gradient estimators obtained only by counting the traffic light switching plans.

Findings in [106] [25] suggest that various parameters such as the spatial distribution of congestion in the network can potentially affect the MFD's shape and its scatter. Having said this, the controlled strategies that rely on pre-defined critical density value may be inconsistent [22]. On this direction, the work in [22] proposes a Kalman filter estimation scheme that utilizes real-time measurements of circulating flow and accumulation of vehicles to produce accurate estimates of the critical density value showing that the developed estimation algorithm coupled with the proposed adaptive perimeter flow control strategy may be valuable whenever the MFD is not well-defined [106]. Inspired by works done by [98, 99], in this thesis a region of the road network is modeled as a hybrid system using the SFM framework and then an Infinitesimal Perturbation Analysis (IPA) is employed in order to compute the gradient of a performance metric which in turn can be used to optimize the estimated region's critical density. The aim is to estimate that value in an online fashion.

Chapter 3

Route Reservation methods

3.1 Introduction

This chapter introduces a novel route-reservation architecture that utilizes the obtained reservations in order to determine the best possible path subject to avoiding road segments that are expected to be at their capacity (microscopic level). Road segments become temporarily unavailable whenever reservations reach their critical density and are reconsidered only when the allocated time has elapsed. In the proposed architecture, a centralized Road Side Unit (RSU) is considered that assumes responsibility of all route reservations. Therefore, when a vehicle is about to begin its journey it sends out a reservation request to the RSU indicating its origin-destination pair. Once the RSU receives a request, it needs to solve the routing problem with the objective of determining the path that will allow the vehicle to reach its destination at the earliest possible time while avoiding unavailable road segments. The RSU could also delay the vehicle's departure time if that action minimizes its destination arrival time. However, by allowing waiting only at the origin means that the delayed vehicles do not occupy space in the transportation network; in this way, the proposed architecture removes waiting in congested situations. Therefore, given the past requests, the RSU has an estimate of the number of vehicles that are expected to be in each road segment, during any interval from the current time into the future. Based on these reservations, the RSU knows which road segments are expected to be below their critical capacity and thus unavailable.

On those premises, two route reservation problems are proposed, which con-

stitute the major contribution of this chapter. The first problem seeks to navigate vehicles through non-congested road segments while each vehicle's destination arrival time is minimized, that has been shown to be an NP-complete problem. The second problem seeks to navigate vehicles through congestion-free road segments while minimizing the load variance of the overall traffic (Traffic Load Balancing (TLB)). Waiting at the origin is considered for both of the above problems. Furthermore, in this chapter, various solution approaches for the Earliest Destination Arrival Time (EDAT) and TLB problems are conducted in both continuous and discrete time domains with simulation results demonstrating the superior performance compared with the state of the art algorithms.

The rest of this chapter is organized as follows. Section 3.2 introduces the route reservation architecture and Sections 3.3 and 3.4 mathematically formulate the EDAT problem providing also a detailed complexity analysis of it. Section 3.5 proposes various solutions approaches of the EDAT problem with simulation results demonstrating the benefits of each proposed solution. The TLB problem is mathematically formulated in Section 3.6 where a solution of it is presented in Section 3.7. Extensive simulation results that are also included in Section 3.7 demonstrate the benefits of the proposed solution considering micro-simulations. Finally Section 3.8 concludes this chapter.

3.2 Route Reservation Architecture

The proposed architecture is used to support efficient route reservations while preventing congestion by ensuring that the traffic of each road segment is sustained up to its critical density. The proposed architecture is depicted in Fig. 3.1, showing an RSU that acts as a central controller responsible of navigating vehicles, monitoring the utilization of each road segment and for reserving routes for arbitrary origin-destination pairs. To do this, each road segment is associated with a time series starting from the current time into the future. In this way, the RSU keeps an estimate of the number of vehicles that are expected to traverse each road segment. Hence, to ensure a congestion-free operation, the density of each road segment should be maintained below the critical density and, in this thesis, this is enforced by limiting

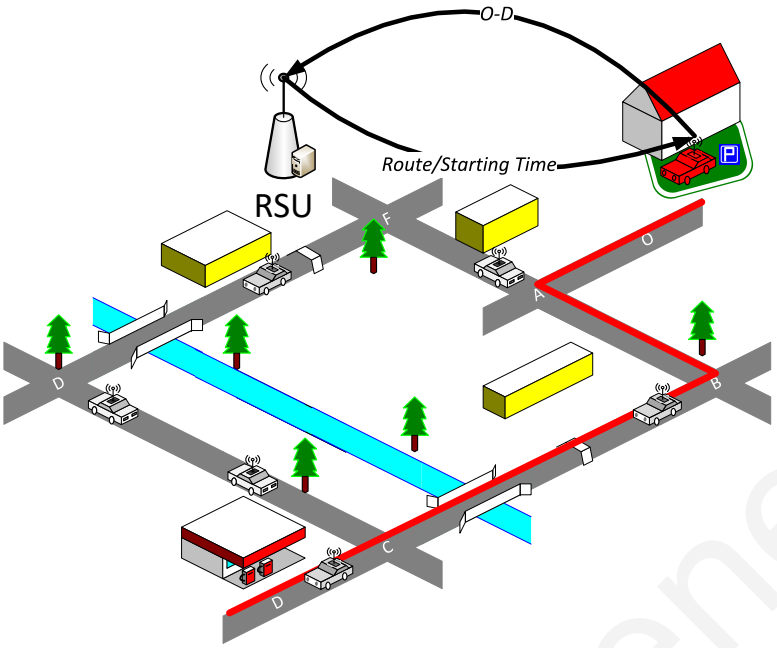


Figure 3.1: The Route-Reservation Architecture.

the reservation availability of each road segment.

As a vehicle plans to start its journey (or even earlier if "pre-bookings" will be allowed), it sends a request to the RSU in order to obtain a path from its current location (i.e., its origin) to the required destination. Given the current reservation state, the RSU responds to the vehicle request, giving the starting time of the journey and the route that the vehicle should follow (e.g., the red line as indicated in Fig. 3.1). Thereafter, the vehicle is responsible for traveling along the allocated route within the time constraints imposed without any deviations. At the same time, the RSU updates the reservation state of each road segment at the time frame that the vehicle is expected to traverse it; assuming that the vehicle will be traveling at a constant speed (i.e., free-flow speed). Assuming that the region's MFD [106] is available and considering that each segment's reservations will no surpass its critical density, then one can use either the free-flow speed or the speed at capacity to also account for some possible delays. If the MFD is not available, then the speed to be used can be obtained from historical data or predicted values. At this point, it is worth pointing out that it is unrealistic to expect that all vehicles will actually travel at the same constant speed, thus in practice, it is expected that there will be significant deviations. Despite these deviations, our simulation results indicate that the whole

approach still works well and is robust with respect to such inaccuracies.

In this way, the RSU determines the best possible path for the vehicle such that it will arrive at its destination at the earliest possible time while avoiding road segments that are expected to surpass their critical density at the time when the vehicle is expected to traverse them. Furthermore, the RSU may impose a waiting period only at the origin if the destination arrival time is minimized by doing so. Note that, reservation decisions are made by the route-reservation algorithms running at the RSU and routes are communicated to requesting vehicles which are in turn responsible to traverse them. In order to compute its response, the RSU formulates and solves the routing problem as indicated in the subsequent sections.

3.3 Mathematical formulations

A homogeneous region is expressed as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V}$, $N_V = |\mathcal{V}|$, representing the road junctions and edges $\mathcal{E}$, $N_E = |\mathcal{E}|$, representing the road segments. Each road segment $(i, j) \in \mathcal{E}$, $\{v_i, v_j\} \in \mathcal{V}$ is described by parameters λ_{ij}, denoting the number of lanes and l_{ij} representing the segment's length.

All traffic dynamics of each region are defined according to a well define MFD [106] with parameters ρ^C, ρ^J, u_c and u_f, representing the critical density corresponding to the maximum flow, jam density, speed at capacity and free-flow speed, respectively. The traffic dynamics of each road segment $(i, j) \in \mathcal{E}$, $\{v_i, v_j\} \in \mathcal{V}$ are described according to parameters ρ^J_{ij}, ρ^C_{ij} and $\rho_{ij}(t)$ indicating the jam density, the critical density, and the instantaneous density at time t, respectively. Note that, the critical density denotes the maximum density that a road segment can accommodate in order to operate at u^f, i.e., $\rho_{ij}(t) \leq \rho^C_{ij}$ and thus, to avoid the derivation of fundamental diagrams (FDs) for each road segment, in this thesis we approximate ρ^C_{ij} with the quantity $(\rho^C/\rho^J)\rho^J_{ij}$, which is derived using MFD parameters and the geometry of the road. In the free-flow regime i.e., $\rho_{ij}(t) \leq \rho^C_{ij}$ the speed ranges from the free-flow speed to the speed-at-capacity; and hence congestion-free road segments are those for which $\rho_{ij}(t) \leq \rho^C_{ij}$ with vehicles can be assumed to travel with speed-at-capacity u_c. The speed at capacity assumption is used instead of the free-flow speed so that travel time estimates account for the possible delays due to driver imperfection and

the delays observed across non-priority road junctions. Congestion-free routing can be achieved if vehicles traverse the network only through road segments that are expected to be below their critical density (i.e., admissible road segments). For this reason, the RSU utilizes the admissibility states of each road segment and needs to formulate and solve an optimization problem to determine the shortest path for a vehicle such that the *admissibility condition* for each link is always satisfied. For the formulations that follow the variables t_0 and d_{v_i} are required which denote the routing request time and the vehicle arrival time at road junction v_i, respectively.

3.3.1 Discrete time formulation

The proposed reservation architecture requires the monitoring of the cumulative number of the expected vehicle arrivals and departures at road segment $(i, j) \in \mathcal{E}$ up to time t, $\alpha_{i,j}(t)$ and $\beta_{i,j}(t)$, respectively. In addition, it requires to monitor the accumulate number of vehicle reservation of road segments (i, j) (i.e., $n_{ij}(t) = \alpha_{i,j}(t) - \beta_{i,j}(t)$) for time unit t. Hence, based on route reservations, the proposed reservation scheme keeps track of the expected accumulated number of vehicles within each road segment over time. Along the same lines the expected instantaneous density of a road segment $(i, j) \in \mathcal{E}$ at time-slot t is expressed by the variable $\hat{\rho}_{ij}(t)$ and mathematically defined as:

$$\hat{\rho}_{ij}(t) = n_{ij}(t)/(\lambda_{ij} l_{ij}). \tag{3.1}$$

According to the reservation architecture a road segment $(i, j) \in \mathcal{E}$ is denoted as admissible if a vehicle entering road junction v_i at time unit t can traverse segment (i, j) without making the expected accumulated density larger than the critical density during any time-slot for which the vehicle will travel on the segment.

In discrete time formulation, the time is quantized into time-slots of duration T so that the number of time-slots required to traverse road segment $(i, j) \in \mathcal{E}$ is $\bar{\tau}_{ij} = \left\lfloor l_{ij}/u_f/T \right\rceil$, where $\lfloor z \rceil$, is the nearest integer to z.

We denote the *admissibility state* of a road segment $(i, j) \in \mathcal{E}$ at time-slot t with the variable $x_{ij}(t)$, and let a road segment be considered as *admissible* (i.e., $x_{ij}(t) = 1$) if a vehicle starting from road junction i at time-slot t can traverse road segment (i, j) without making the accumulated reserved density larger than the critical density at any point within the traversal time, and $x_{ij}(t) = 0$ otherwise. Mathematically $x_{ij}(t)$ in

discrete time can be defined as:

$$x_{ij}(t) = \begin{cases} 1, \text{ if } \hat{\rho}_{ij}(t) \leq \rho_{ij}^{C}, \ \forall \ \tau = t, \ldots, t + \bar{\tau}_{ij} \\ 0, \text{ otherwise} \end{cases} \tag{3.2}$$

Considering the above notation, the cost of traversing a road segment $c_{ij}(t)$ at each discrete time-slot t can mathematically be expressed as follows:

$$c_{ij}(t) = \begin{cases} \bar{\tau}_{ij}, \text{ if } x_{ij}(t) = 1, i \neq O \\ \bar{\tau}_{ij} + w, \text{ if } x_{ij}(t) = 0, i = O \\ \infty, \text{ if } x_{ij}(t) = 0, i \neq O \end{cases} \tag{3.3}$$

where, w denotes the number of time-slots that a vehicle may wait at the origin such that the path found to traverse from origin (i.e., O) to destination (i.e., D) is admissible.

3.3.2 Continuous time formulation

Similar to the discrete time formulation, let the variable $n_{ij}(t)$ defines the accumulated number of reservations at time t while the traversal time for each road segment $(i, j) \in \mathcal{E}$ be equal to $\bar{\tau}_{ij} = l_{ij}/u_f$. Thus, the state variable $n_{ij}(t^l, t^u)$ denotes the accumulated number of reservations within (i, j) during time interval (t^l, t^u) where t^l and t^u denote the lower and upper time bounds *i.e.*, $t^l \leq t^u$. Accordingly, a road segment (i, j) is considered as *admissible* within time interval (t^u, t^l) if the expected instantaneous density i.e.,

$$\hat{\rho}_{ij}(t^u, t^l) = n_{ij}(t^u, t^l)/(\lambda_{ij} l_{ij}), \tag{3.4}$$

during the time interval (t^u, t^l) is not larger than the segment's critical density as,

$$\hat{\rho}_{ij}(t^u, t^l) \leq \rho_{ij}^{C}. \tag{3.5}$$

Hence, the RSU constructs the admissible sets $\mathcal{S}_{ij}(t_c) = \{(t^l_{ij1}, t^u_{ij1}), \ldots, (t^l_{ijK_{ij}(t_c)}, t^u_{ijK_{ij}(t_c)})\}$, of each road segment $(i, j) \in \mathcal{E}$, which define the admissible time intervals (t^l_{ijk}, t^u_{ijk}), $k \in \mathcal{K}_{ij}(t_c) = \{1, \ldots, K_{ij}(t_c)\}$, where $K_{ij}(t_c)$ denotes the number of admissible time intervals of segment (i, j) at time t_c. Note also that $t^l_{ijk} < t^u_{ijk} < t^l_{ijk+1}$ where t^l_{ijk} and t^u_{ijk} denote the lower and upper bounds of the k-th admissible time interval of link $(i, j) \in \mathcal{V}$, respectively. These time intervals are determined by the RSU, given the

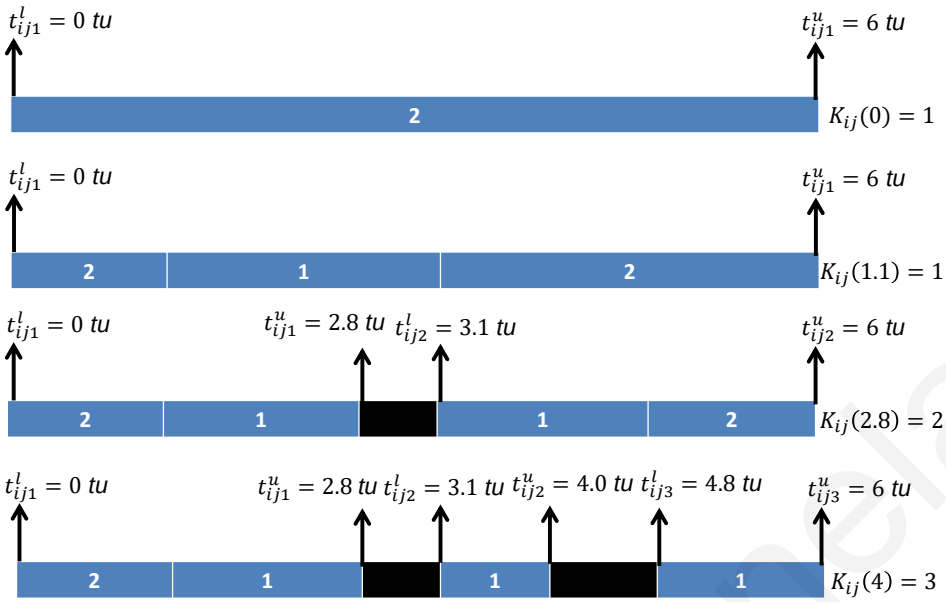


Figure 3.2: Example depicting the evolution of the admissible set of a particular road segment with transit time 2 tu, following three vehicle requests at 1.1 tu, 2.8 tu and 4 tu; the black regions denote non-admissibility.

past reservations and the path of a vehicle, under the assumption that the vehicle will travel at a constant speed u^f[1].

Fig. 3.2 depicts a toy example of how the admissible time intervals evolve in time for link (i, j). It is assumed that the link can be traversed in 2 time units (tu), while its critical density is equal to 2 vehicles for its entire length. Also, the considered time horizon is assumed to be 6 tu. In the figure, the number in the shaded area represents the remaining density of the link for each particular time interval. Observing, the figure at $t_0 = 0$ tu there are no vehicle requests for a paths, hence the link has a single admissible interval such that $t^l_{ij1} = 0$ tu and $t^u_{ij1} = 6$ tu. Next, the first vehicle is expected to arrive at $t_0 = 1.1$ tu and hence the RSU makes a reservation from $t = 1.1$ tu until $t = 3.1$ tu. After the first request, that link is admissible for the same time duration (whole interval) as the link's critical density is 2 vehicles. The second vehicle is expected to arrive at $t = 2.8$ tu and reserves the link from $t = 2.8$ tu until $t = 4.8$ tu. Both requests share the link during the interval 2.8 tu to 3.1 tu so

[1] As indicated in the simulation results, the approach achieves significant reduction in congestion even if the vehicle deviates from the assumed speed u^f.

that the link becomes non-admissible for that period. This yields to two admissible time intervals for the remaining time (i.e. from $t^l_{ij1} = 0$ tu to $t^u_{ij1} = 2.8$ tu and from $t^l_{ij2} = 3.1$ tu to $t^u_{ij2} = 6$ tu). In the same way, a request by a third vehicle allows a reservation within the interval $t = 4$ tu to $t = 6$ tu, yielding to three separate time intervals.

The cost of traversing the road segment (i, j), $c_{ij}(t)$ in this continuous time approach is:

$$c_{ij}(t) = \begin{cases} \bar{\tau}_{ij}, \text{ if } d_{v_i} + t \in \mathcal{S}_{ij}(t_0), \forall\ 0 \leq t \leq \bar{\tau}_{ij},\ i \neq O \\ \bar{\tau}_{ij} + w, \text{ if } t_0 + w + t \in \mathcal{S}_{ij}(t_0), \forall\ 0 \leq t \leq \bar{\tau}_{ij},\ i = O \\ \infty, \text{ otherwise,} \end{cases} \tag{3.6}$$

where, w denotes the time interval for which the vehicle will have to wait at its origin before starting its trip ($i = O$).

3.3.3 Congestion free routing under admissibility states

Congestion-free routing is ensured whenever vehicles traverse the network without violating the capacity constrains (i.e., vehicles traverse only admissible road segments). The admissibility of road segments is an essential aspect of the route reservation scheme as it allows vehicles to always traverse through non-congested road segments at free-flow speed conditions, leading to a congestion-free routing. Nonetheless, it introduces the additional challenge of dealing with non-admissible road segments. To that end, in both continuous and discrete time formulations, vehicles are allowed to traverse road segments only during times where the admissibility is feasible since the RSU can only make any reservations along those times to enable congestion-free routing.

Two alternative options arise in a case that the shortest path of vehicles includes a non-admissible road segment. The first prompts vehicles to wait at their origin until all road segments in the shortest path become admissible. The second chooses an alternative path where all links are admissible. Combining both option may yield to better a solution (e.g., wait for a short period of time at O and then follow an alternative route). This chapter investigates two problems to address route-reservation under admissible or non-admissible road segments and where waiting is allowed at the origin. The first problem, called the Earliest Destination Arrival Time (EDAT)

problem, aims to find the path arriving at the destination at the earliest possible time. The second problem, called Traffic Load Balancing (TLB), aims at finding a path that provides a good trade-off between early destination arrival and traffic load balancing which can create additional robustness to the proposed architecture. The following Sections present a detail mathematical formulation of both the EDAT and TLB problems, also providing their algorithmic solution.

3.4 The Earliest Destination Arrival Time (EDAT) problem

This section formulates and solves an optimization problem for determining the path that will allow a vehicle to arrive at its destination at the earliest possible arrival time while avoiding non-admissible links. This is referred to as the Earliest Destination Arrival Time (EDAT) problem. The EDAT problem seeks to route vehicles only through admissible road segments, and thus to reach their destination at the earliest time while a detail complexity analysis of the problem indicates that the EDAT shows to be NP-complete. Note that the EDAT problem is formulated and solved in both discrete and continuous time domains.

3.4.1 Discrete time

Given an origin-destination ($O-D$) pair, the time-stamp t_0 at which the routing request is made, and the admissibility states $x_{ij}(t)$, $(i,j) \in \mathcal{E}$, for each time-slot $\forall t \geq t_0$, then the EDAT problem requests the earliest-arrival-time-at-destination (from O to D). Let p_h denote the h-th path from origin (O) to destination (D) denoted as $p_h = (v_0^h, v_1^h),\ (v_1^h, v_2^h),\ (v_2^h, v_3^h), \cdots (v_{L_h-1}^h, v_{L_h}^h)$, where $v_j^h \in \mathcal{V}$ is the j-th visited vertex in the h-th path, with $v_0^h = O$, $v_{L_h}^h = D$, and L_h is the length of the path p_h in terms of the number of hops. Additionally, let w and $d_{v_j}^h$ denote the waiting time at the origin and the earliest arrival time at junction v_j (assuming the vehicle was delayed by w at the origin), respectively. Then, the earliest arrival time to each vertex of the path

can be expressed as:

$$
\begin{aligned}
d^h_{v^h_0} &= t_0 \quad w \geq 0 \\
d^h_{v^h_1} &= d^h_{v^h_0} + c_{v^h_0, v^h_1}(d^h_{v^h_0}) \\
&\vdots \\
d^h_{v^h_{L_h}} &= d^h_{v^h_{L_h-1}} + c_{v^h_{L_h-1}, v^h_{L_h}}(d^h_{v^h_{L_h}})
\end{aligned} \tag{3.7}
$$

Hence, the EDAT problem can be expressed in compact form as:

$$
(\Pi_d)\ d^*_D = \min_{w \geq 0,\ p_h} d^h_D \tag{3.8}
$$

$$
\text{s.t. Constraints (3.1) } - \text{ (3.3) and (3.7) are satisfied.}
$$

A detailed complexity analysis of the problem follows while two solutions are presented in Section 3.5.

Earliest Destination Arrival Time Problem complexity analysis

This section provides a rigorous complexity analysis of the resulting EDAT problem in discrete time (i.e., Problem (Π_d)) that aims to provide vehicles routes under the proposed reservation protocol. This complexity analysis indicates that for some instances, the Problem (Π_d) reduces to an NP-complete problem.

At a first glance, the formulated EDAT problem looks similar to the well investigated time-dependent route planning problem [107]. Nevertheless, the EDAT problem differs from the time-dependent route planning problem since EDAT introduces road segments with infinite cost (non-admissible road segments) and also allows for waiting intervals that may be observed at the originating junction.

For notation simplicity let the discrete time EDAT problem (i.e., Problem (Π_d)) also be denoted as (Π). The complexity analysis of (Π) requires to examine the complexity of two variations of the particular problem, that we denote as the (Π_{AW}) and (Π_{NW}) problems. (Π_{AW}) has a similar objective function as (Π) but it allows vehicles to wait at all road junctions until they become available. Clearly the solution to this problem is not implementable since physically there is not space for vehicles to park and wait until a road section becomes available, however, the solution to this

problem can serve as a lower bound to the solution of (Π) while (as we will show) it can be solved in polynomial time. The other related problem is the (Π_{NW}) that does not allow for vehicle waiting neither at the origin nor at any other junction.

Starting from the (Π_{AW}) problem, the cost $c_{ij}(t)$ of traversing road segment (i, j) can mathematically be expressed as:

$$c_{ij}(t) = \begin{cases} \bar{\tau}_{ij}, \text{ if } x_{ij}(t) = 1 \\ \bar{\tau}_{ij} + w_{ij}, \text{ if } x_{ij}(t) = 0 \end{cases} \tag{3.9}$$

where, w_{ij} denotes the number of time-slots that a vehicle may wait at i such that the path, found to traverse from O to destination D, is admissible. Thus, (Π_{AW}) can mathematically be expressed as:

$$(\Pi_{AW})\ d^*_{D_{AW}} = \min_{w_{ij} \geq 0,\ p_h} d^h_D \tag{3.10}$$

$$\text{s.t. Constraints (3.1) – (3.2), (3.7) and (3.9) are satisfied.}$$

The cost $c_{ij}(t)$ of traversing a road segment for problem (Π_{NW}) can mathematically be stated as follows:

$$c_{ij}(t) = \begin{cases} \bar{\tau}_{ij}, \text{ if } x_{ij}(t) = 1 \\ \infty, \text{ if } x_{ij}(t) = 0 \end{cases} \tag{3.11}$$

Hence, (Π_{NW}) can be mathematically stated as follows:

$$(\Pi_{NW})\ \ d^*_{D_{NW}} = \min_{p_h} d^h_D \tag{3.12}$$

$$\text{s.t. Constraints (3.1) – (3.2), (3.7) and (3.11) are satisfied.}$$

The two aforementioned variants (Π_{AW}) and (Π_{NW}) are used to prove that the (Π) problem can be categorized as an NP-complete problem. The NP-completeness of (Π) is derived using the *restriction method* [108]. The restriction method requires to prove that problem (Π) can be reduced to a special case of a known NP-complete problem. Hence, the examined proof reduces the (Π) problem to the *Number Partitioning* Problem (Π′) (described in [108]) which is defined as follows.

Number partitioning problem: Let the set $\mathcal{A}$ consist of n integer numbers $\mathcal{A} = \{a_1, a_2, ..., a_n\}$, $a_j \in \mathbb{Z}^+$ and let an integer number $b \in \mathbb{Z}^+$. (Π′) requires to identify the

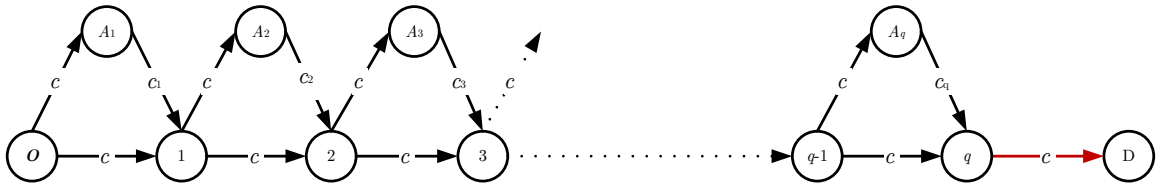


Figure 3.3: Special case of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (with edge (q, D) attain to non admissible state).

subset $\mathcal{A}'$ where $\mathcal{A}' \subseteq \mathcal{A}$, such that the sum of the numbers in $\mathcal{A}'$ is equal to a given number b. Equivalently, this problem can be expressed using variables $y_j = \{0, 1\}$ that indicate whether a_j is in $\mathcal{A}'$ ($y_j = 1$) or not ($y_j = 0$), as follows:

$$\sum_{j=1}^{n} a_j y_j = b, \text{ where } y_j = \{0, 1\} \tag{3.13}$$

Note that problem (Π') is an NP-complete problem [108].

Lemma 1. (Π_{NW}) *is an NP-complete problem in the case where at least one road segment attains a non-admissible state.*

Proof. To prove Lemma 1 we need to show that (Π_{NW}), can be reduced to a special case of (Π'). For this purpose a special case of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed as shown in Fig. 3.3. As illustrated in Fig. 3.3, the traversal cost of each road segment is defined by c and c_j values which are predefined integer constants, i.e., $c \neq c_j$ and $c_j \neq c_k$. Considering the structure of the graph, the cost to traverse the edge from node j to node $j + 1$, (i.e., $\hat{c}_{j,j+1}$) can mathematically be stated as:

$$\hat{c}_{j,j+1} = \begin{cases} c + c_{j+1}, & \text{if path passes from } A_{j+1} \\ c, & \text{otherwise} \end{cases} \tag{3.14}$$

Let (q, D) (indicated with red color) be the single edge on $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that attains a non-admissible state as follows:

$$x_{qD}(t) = \begin{cases} 1, \text{ for } t = cq + b \\ 0, \text{ otherwise} \end{cases} \tag{3.15}$$

while all other edges always attain an admissible state. According to constraints (3.11), (3.14) and (3.15) the traversal cost $c_{qD}(t)$ of (q, D) can be expressed as follows:

$$c_{qD}(t) = \begin{cases} \tau_{qD}, \text{ for } t = cq + b \\ \infty, \text{ otherwise} \end{cases} \tag{3.16}$$

where, $\tau_{qD} = c$ as indicated in Fig. 3.3.

According to the above setup, the only possible solution consists of a path from O to D, (i.e., p) where, the arrival time at node q is exactly equal to $cq + b$. Therefore, the arrival time at junction q must be:

$$d_q = cq + b \tag{3.17}$$

Let, p contain the subpath p' from vertices O to q. There are in total 2^q possible combinations that can constitute subpath p' and the total travel time of each combinations (i.e., c_{Oq}) can be defined as follows:

$$c_{Oq} = \sum_{j=1}^{q} \hat{c}_{j-1,j} y_j = \sum_{j=1}^{q} c(1 - y_j) + \sum_{j=1}^{q} (c + c_j) y_j = cq + \sum_{j=1}^{q} c_j y_j \text{ where,} \tag{3.18}$$

$$y_j = \begin{cases} 1, \text{ if path passes from } A_j \\ 0, \text{ otherwise} \end{cases} \tag{3.19}$$

Considering Eq. (3.14), cq time-slots can be provided from all of the 2^q paths while the remaining b time-slots must be identified by the summation of $\sum_{j=1}^{q} c_j y_j$. Therefore, the solution returned according to the selected y_j values, provides a solution to the number partitioning problem since a subset of values (that sum up exactly to b) is required to be selected from the range of c_j, and this completes the proof. □

The second variant assumes that waiting intervals are allowed at all road junctions. This assumption is not feasible along real transportation networks due to lack of adequate buffering space where vehicles will wait. Nonetheless, this case can be considered as a lower bound solution and is part of the subsequent proof of theorem 3.4.1 used to prove that the EDAT can be reduced to (Π') as a special case.

Lemma 2. *The problem (Π_{AW}), i.e., finding the earliest arrival time while waiting at every junction is allowed, can be solved in polynomial time.*

Proof. In the case when a vehicle can wait at all intersections, the problem becomes significantly easier and can be solved to optimality using a simple modification of the Dijkstra's shortest path algorithm [109] which can converge in polynomial time. Specifically for an arbitrary graph, at every step of the algorithm, given the time of the earliest vehicle arrival at any node p (through the previous steps of the

algorithm), if the next link (p, q) is unavailable until u time units later, its cost $c_{p,q}$ is simply undated to $c_{p,q} + u$, while if the vehicle's earliest arrival at p is during a time period when the link is available, then its cost is simply $c_{p,q}$ which corresponds to the time needed to traverse the link. A detailed correctness proof can be shown using the Dijkstra's proof of correctness based on the contradiction method [44]. □

The third case completes the complexity analysis of the formulated EDAT(Π) problem as a combination of the two previous cases of problems (Π_{AW}) and (Π_{NW}).

Theorem 3.4.1. *The problem (Π), i.e., vehicles are only allowed to wait at the origin O, is an NP-complete problem when more than one road-segments become non-admissible during certain time-slots.*

The proof of theorem 3.4.1 is divided into two special cases. This distinction is required in order to find a special case in which (Π) can be stated as an NP-complete problem. The first case illustrates the situation where (Π) can always be solved in polynomial time and the second case covers the scenario where the problem (Π) can be reduced to the Number Partitioning Problem.

Special Cases 1:
Considering Theorem 3.4.1 and assuming that only one road segment (which should be a part of the path) has to attain a non-admissible reservation state, then (Π) can be solved in polynomial time.

Proof. Consider the same example of lemma 1 (shown in Fig. 3.3) illustrating edge (q, D) that attains a non-admissible reservation state. The solution requires a vehicle to depart from node q exactly at time $cq + b$ as indicated by equality constraint (3.17). When only one road segment attains a non-admissible state, the problem can be adequately expressed through Lemma 2, since (3.17) can be transformed to an inequality constraint. As shown in Lemma 2, a solution can easily be found with a feasible path from vertex O to q where $d_q \leqslant cq + b$ according to constraint. If the solution results in arriving at q on an earlier time then the vehicle can wait for the remaining time-slots to the originating junction to satisfy constraint (3.17). □

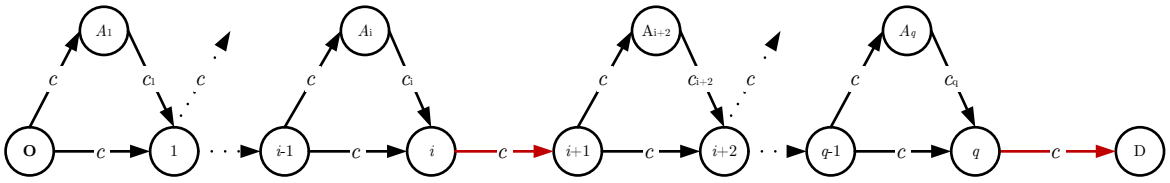


Figure 3.4: Special case of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (with edges $(i, i+1)$ and (q, D) attain to non admissible state).

Special Cases 2:

Considering Theorem 3.4.1 and assuming that more than one road segments attain a non-admissible state during certain time-slots (which should be a part of the path) then (Π) results to an NP-complete problem.

Proof. Consider the special case in $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as shown in Fig. 3.4 where the cost to traverse the link from node i to node $i+1$, is based according to Eq. (3.14) (i.e, $\hat{c}_{i,i+1}$). Fig. 3.4 indicates that, in total, two road segments attain a non-admissible state (i.e., edges $(i, i+1)$ and (q, D)) as follows:

$$x_{i,i+1}(t) = \begin{cases} 1, \text{for } t = ci + b_1 \\ 0, \text{ otherwise} \end{cases} \tag{3.20}$$

$$x_{qD}(t) = \begin{cases} 1, \text{for } t = cq + b_1 + b_2 \\ 0, \text{ otherwise} \end{cases} \tag{3.21}$$

while all other links always attain an admissible state. According to constraints (3.3), (3.20), (3.21), the traversal cost of both links can be expressed as follows:

$$c_{ii+1}(t) = \begin{cases} \tau_{ii+1}, \text{ for } t = ci + b_1 \\ \infty, \text{ otherwise} \end{cases} \tag{3.22}$$

$$c_{qD}(t) = \begin{cases} \tau_{qD}, \text{ for } t = cq + b_1 + b_2 \\ \infty, \text{ otherwise} \end{cases} \tag{3.23}$$

where, $\tau_{i,i+1} = c$ and $\tau_{qd} = c$ as indicated in Fig. 3.4.

According to Fig. 3.4, (Π) has a feasible solution only if an admissible $O - D$ path exists while the departure times at node i and q should be exactly at time-slots $ci + b_1$

and $cq + b_1 + b_2$, respectively. Following the analysis in the proof of Lemma 1, let p contain sub-paths p' and p'' where, p' is the sub-path from node O to i and p'' is the sub-path from node $i+1$ to q. Similar to Lemma 1, the total travel time costs of both sub-paths $c_{p'}(t)$ and $c_{p''}(t)$ can be expressed as follows:

$$\begin{aligned} c_{Oi} &= \sum_{j=1}^{i} c(1-y_j) + \sum_{j=1}^{i}(c+c_j)y_j = ci + \sum_{j=1}^{i} c_j y_j \\ c_{i+1,q} &= \sum_{j=i+1}^{q} c(1-y_j) + \sum_{j=i+1}^{q}(c+c_j)y_j = c(q-i) + \sum_{j=i+1}^{q} c_j y_j \end{aligned} \tag{3.24}$$

As indicated in Lemma 2, the first time constraint can be easily satisfied since there are 2^i possible paths from node O to i with $d_i \leqslant ci + b_1$ since waiting can take place at the originating junction in such a way as to achieve $d_i = ci + b_1$. Nonetheless, the second time constraint (i.e., $d_q = cq + b_1 + b_2$) is addressed by Lemma 1. Thus, a solution of sub-path p'' can be reduce to a problem addressed by Lemma 1.

Same as before, considering Eq. (3.14), the amount of $cq + b_1$ time-slots can be provided from all of the 2^q paths, while the remaining b_2 time-slots must be identified by the summation of Eq. (3.24) (*e.g.*, $\sum_{j=i+1}^{q} c_j y_j$). Therefore, to select y_j values the number partitioning problem needs to be solved; completing the NP-completeness proof. □

3.4.2 Continuous time

Similar to the discrete time formulation, given an origin-destination ($O-D$) pair let p_h denote the h-th path from source O to destination D denoted as $p_h = (v_0^h, v_1^h)$, (v_1^h, v_2^h), (v_2^h, v_3^h), $\cdots$, $(v_{L_h-1}^h, v_{L_h}^h)$, where $v_j^h \in \mathcal{V}$ is the j-th visited node in the h-th path, with $v_0^h = O$ and $v_{L_k}^h = D$. Then, the arrival time at each road junction v_j of the path can be expressed in continuous time as:

$$\begin{aligned} d_{v_0^h}^h &= t_0, \quad w \geq 0 \\ d_{v_1^h}^h &= d_{v_0^h}^h + c_{v_0^h, v_1^h}(d_{v_0^h}^h) \\ &\vdots \\ d_{v_{L_h}^h}^h &= d_{v_{L_h-1}^h}^h + c_{v_{L_h-1}^h, v_{L_h}^h}(d_{v_{L_h}^h}^h) \end{aligned} \tag{3.25}$$

Given the vehicle routing request time t_0 and the admissible sets $\mathcal{S}_{ij}(t_0), \forall\ (i,j) \in \mathcal{E}$, then the EDAT problem can be expressed in compact form as:

$$(\Pi_c)\ d_D^* = \min_{w \geq 0,\ p_h} d_D^h \tag{3.26}$$

$$\text{s.t. Constraints (3.4) } - \text{ (3.6) and (3.25) are satisfied.}$$

The complexity analysis of Section 3.4 can also hold for Problem (Π_c) as (Π_d) is a special case of (Π_c). Nonetheless, an optimal solution to the continuous time EDAT problem is mathematically derived and solved using a Mixed Integer Linear Program (MILP) which can only solve small instances of the problem due to its NP-completeness. In this manner, the following sections also provide various heuristic solutions (for both continuous and discrete time formulations) that can converge in a reasonable time.

3.5 Proposed solutions for EDAT

3.5.1 Discrete time solutions for EDAT

Many alternative algorithms are proposed for solving the formulated EDAT problem (either in the discrete or continuous time). In the subsequent solutions, we assume that as new journey requests are issued by soon-to-be-departing vehicles, decisions should be made on which route to take and where should vehicles wait in order to arrive at their destination on the earliest possible time. When decisions are made, vehicles are responsible for following the assigned route within the scheduled time constraints from the origin to the destination.

Time expansion approach

In this section we utilize the "time expansion" approach to demonstrate some of the complexities associated with solving the (Π_d) problem optimally. Fig. 3.5 shows a simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where road segments (A, B) and (B, F) attain two non-admissible time-slots (from time intervals 1s-10s and 1s-3s, respectively). Notably, time-dependent networks can easily be transformed to static networks using time-expansion as discussed in [110] and allow the problem to be solved in the space

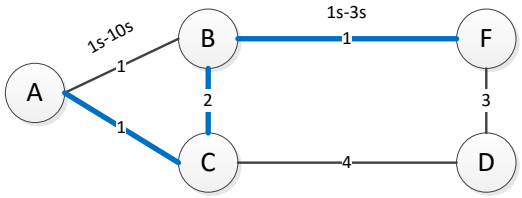


Figure 3.5: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

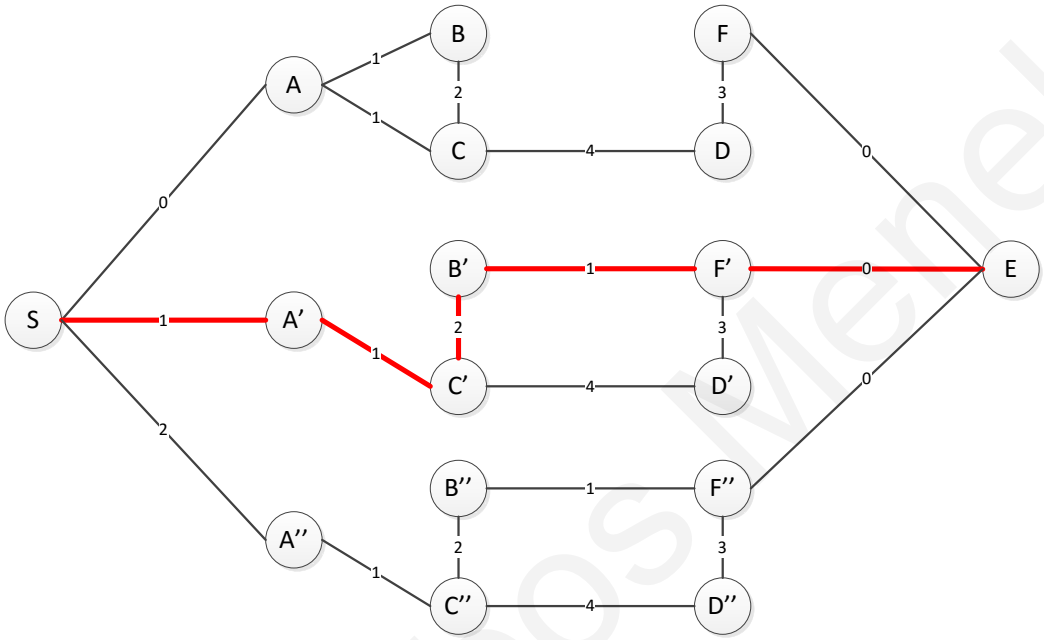


Figure 3.6: Time expanded $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

dimension only. In this way, the problem is solved in two stages. In the first stage, the graph is expanded to future time-steps considering incremental waiting intervals at the originating junction. Thereafter, a static shortest path algorithm (e.g. Dijkstra [109]) is used to provide a solution. Fig. 3.6 illustrates the time-expanded graph for the network provided by Fig. 3.5. Fig. 3.5 illustrates the optimal solution with a blue line and total cost of 4*s* while Fig. 3.6 shows the shortest path solution over the time-expanded graph indicated with a red line and total cost of 5*s*. As Fig. 3.6 indicates, the shortest path algorithm (e.g., Dijkstra) miss the optimal solution since the non-admissible time slots are not considered to the time-expanded graph. Note that the earliest arrival time at each junction does not ensure the optimal choice based on the label setting property discussed in [44]. The possibility of selecting a junction a little bit later may reduce the destination arrival time since a currently

non-admissible segment may become admissible in future time-slot. Therefore, all possible arrival times must be examined at each intermediate junction in order to ensure that an optimal solution is reached.

Route Reservation Algorithm (RRA)

A heuristic solution to the discrete time EDAT problem is derived through the Route Reservation Algorithm (RRA) which also allows an initial wait at the origin. The RRA algorithm employs the Dijkstra's algorithm which is commonly used on static (non-constrained) networks. The proof of correctness of Dijkstra's algorithm indicates two basic properties. The first one is that Dijkstra's algorithm is a label-setting algorithm since on each iteration a label (i.e., d_{v_j}) becomes the actual shortest path from the origin to junction v_j and the algorithm terminates when all nodes are permanently labeled. Labeled nodes are those which an optimal path is found and all the permanently labeled nodes are stored in a predecessor array [44]. The second property is a result of the first property known as the relaxation technique[2], where in each iteration the cost of all non-labeled nodes is $d_{v_i} = \min(d_{v_i}, d_{v_j} + c_{ij}(d_{v_j}))$. Therefore, using the label-setting property and the relaxation technique, Dijkstra's algorithm calculates the earliest-arrival-time from origin to each other road junction v_i. RRA adopts the above properties and returns a feasible solution to the EDAT problem accounting also for possible waiting that can take place at the origin.

The RRA algorithm executes in two stages (the inner and outer loop). The inner loop returns the earliest-destination-arrival-time path, from O to D, by allowing vehicles to wait at any intermediate junction until the road segment's state changes from non-admissible to admissible (i.e., it solves the (Π_{AW}) problem). As shown by Lemma 2, if waiting intervals are allowed to all intermediate road junctions (nodes) a polynomial time optimal solution can be found. This relaxed solution, which is not practically implementable, is a lower bound solution to the discrete time EDAT problem.

Subsequently, the outer loop, checks if the solution computed by the inner loop involves waiting intervals at any intermediate junction. If the resulting shortest

[2]The term "relaxation" is used in a way such that an upper bound solution is found by amending the shortest path as explained in [44].

Algorithm 1 Inner loop of the discrete time RRA (IL-RRA).

Input: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $x_{ij}(t)$, $O - D$, t_0
Initialization
$P[v_i] \leftarrow NULL \ \forall \, v_i \in \mathcal{V}$ ▷ Sets the predecessor matrix
$Q \leftarrow v_i \ \forall \, v_i \in \mathcal{V}$ ▷ Sets all non-labeled junctions
$d_{v_i} = \infty \ \forall \, v_i \in \mathcal{V}$ ▷ Sets the arrival time at i
$P[O] \leftarrow 0$ ▷ Origin Predecessor
$d_O \leftarrow t_0$ ▷ Arrival time at Origin
$\epsilon = 10^{-6}$ ▷ Waiting Coefficient
$w_{min} = \infty$
End of Initialization
while $Q \neq \emptyset$ **do**
    $\forall \, v_i \in Q$ **Extract** v_i with $min(d_{v_i})$
    Set v_i as **labeled** ▷ Since $d_{v_i} = d^*_{v_i}$
    for $\forall \, (i, j) \in \mathcal{E}$ **do**
        if $x_{ij}(d^*_{v_i}) == 1$ **then**
            $w_{ij}(d^*_{v_i}) = 0$
        else
            Calculate $w_{ij}(d^*_{v_i})$ ▷ Required waiting-slots
            $c_{ij}(d^*_{v_i}) = \tau_{ij} + w_{ij}(d^*_{v_i}) + \epsilon$ ▷ Update $c_{ij}(d^*_{v_i})$
        end if
        if $d_{v_j} > d^*_{v_i} + c_{ij}(d^*_{v_i})$ **then**
            $d_{v_j} = d^*_{v_i} + c_{ij}(d^*_{v_i})$; $P[v_j] = v_i$ ▷ Update d_{v_j}
            if $w_{ij}(d^*_{v_i}) < w_{min}$ **then**
                $w_{min} = w_{ij}(d^*_{v_i})$ ▷ Update w_{min}
            end if
        end if
    end for
end while
return (w_{min} and path)

path from the relaxed problem (inner loop) does not require any waiting at any intermediate node, then the algorithm ends. The obtained solution is considered as the shortest path that the vehicle should follow after waiting the accumulated waiting interval at the origin. On the other hand, if the solution of the relaxed problem involves waiting at one or more intermediate nodes, the outer stage transfers the minimum waiting interval among all nodes to the origin and updates the vehicle's start time (i.e., $t_0 = t_0 + w_{min}$), where $w_{min} = \min(w_{ij}(d^*_{v_i}))$, $w_{i,j} > 0$ is the minimum waiting at an intermediate node in the obtained relaxed solution. Given the updated waiting time at the origin t_0, the relaxed problem is solved again. This procedure repeats until a path is found that does not include any links that are at their capacity (given the estimated reservations) nor any waiting at any intermediate node.

The execution procedure of the inner loop is illustrated in Algorithm 1. Algorithm 1 is similar to the Dijkstra's algorithm but road segment costs are calculated dynamically, since edges cannot be traversed if they are non-admissible. In that case, vehicles are forced to wait at a starting junction (v_i) of the road segment (i, j), until their admissibility state changes, thus their cost is updated to also include that waiting time.

The initialization of Algorithm 1 is identical to Dijkstra's algorithm with the predecessor matrix initiated as empty (line 3), all junctions initiated as non-labeled (line 4), all variables initiated to have an infinite cost (line 5) and the arrival time at the destination set to t_0 (line 7). Thereafter, the inner loop is executed for all non-labeled junctions and the one with the earliest arrival time is set as labeled (line 11 and 13). Evidently, the first junctions that the algorithm sets in the route is the originating node since all others have infinite cost while in subsequent iterations a new labeled junction is set to be the one that has the earliest possible arrival time ($d^*_{v_i}$) according to the label-setting property. With every new set junction, a dynamic calculation of the traversal cost from the new labeled junction to its neighbors is performed (lines 15 to 28). This dynamic calculation is performed in those cases where segment (i, j) is non-admissible at $d^*_{v_i}$ (line 15). The minimum number of time-slots that may be required $w_{ij}(d^*_{v_i})$ can be calculated based on both the reservation status of the concerned segment (i, j) and the arrival time at junction v_i (line 18). In every other case, when the segment attains admissible states, no waiting is necessary (line 16).

Therefore, in every iteration the edge cost traversal function $c_{ij}(d^*_{v_i})$ is calculated using only the constant travel time cost (free-flow conditions) and the waiting time duration (i.e., $c_{ij}(d^*_i) = \tau_{ij} + w_{ij}(d^*_{v_i}) + \varepsilon$) (line 19). After all costs have been calculated, a relaxations is performed (lines 20 to 23). If the traversal cost is lower than the arrival time d_{v_j} then the arrival time at junction v_j is relaxed to d_{v_j} (i.e., $d_{v_j} = c_{ij}(d^*_{v_i})$) and junction v_i is characterized by the predecessor of v_j. By doing so, RRA updates the earliest arrival time d_{v_i} to each non-labeled neighboring junction and stores the minimum waiting interval among all junctions (w_{min}) (lines 22 and 23). The above procedure repeats until all road junctions are characterized as labeled. Finally, the inner loop returns to the outer loop the w_{min} and the identified path.

The outer loop determines if any waiting has been included in the path computed by the inner loop (i.e. $w_{min} \neq 0$). The execution of the outer loop is illustrated in Algorithm 2 where, as a first step the total delay that may be observed at the origin (i.e., w_{total}) is initiated to zero (line 2) and afterwards the inner loop is executed (line 3). Thereafter, w_{total} is updated according to the returned w_{min} (line 5). Whenever waiting is identified, the procedure repeats until no waiting is necessary within the computed path (lines 5 to 9). Waiting is added to the origin (i.e., the entry point to the region) and the start time is updated (i.e., $t_0 = t_0 + w_{total}$) (line 7) before the inner loop re-executes with the new starting time (line 8). With each inner loop execution, the waiting intervals that are required are summed to $w_{total} = w_{total} + w_{min}$ (line 9) and repeats until no waiting is necessary.

Observations There are cases where two or more feasible solutions for the discrete time EDAT problem may exists with equal cost. In those cases if one of the two does not require any waiting while the other does, then the algorithm chooses the path with no intermediate node waiting and discards the other one since the algorithm terminates by the first iteration. In the case where both alternative paths experience some waiting at intermediate nodes, then the inner loop should re-iterate at least one more time to identify if the waiting interval can be allocated only at the originating junction. To overcome the selection problem between the alternative solutions, a constant $\varepsilon = 10^{-6}$ is added in case where waiting is required at each particular road segment. Thus, the coefficient ε is added to $c_{ij}(t)$ (i.e., $c_{ij}(d^*_i) = \tau_{ij} + w_{ij}(t) + \epsilon$, with

Algorithm 2 Iterative Dijkstra Algorithm Discrete Time.

1: **Input:** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $O - D$, t_0, w_{min}, P
2: $w_{total} = 0$
3: Execute IL-RRA($\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $x_{ij}(t)$, $O - D$, t_0)
4: $w_{total} = w_{total} + w_{min}$
5: **while** $w_{min} \neq 0$ **do**
6: $\quad w_{min} = 0$
7: $\quad t_0 = t_0 + w_{total}$
8: $\quad$ Execute IL-RRA($\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $x_{ij}(t)$, $O - D$, t_0)
9: $\quad w_{total} = w_{total} + w_{min}$
10: **end while**
11: **return** (Path and w_{total})

$t = d^*_{v_i}$) (line 18) whenever waiting at a junction is required. This additional cost ensures that when equivalent paths exist, the algorithm will choose the one with the least waiting.

As emphasized above, RRA is a heuristic solution that can be executed efficiently in real time to provide either the optimal or a near optimal path. For example the RRA algorithm will miss the optimal solution for the example illustrated in Fig.3.5. As already mentioned, the optimal solution is indicated with a blue line and has a total cost of 4s. The inner loop of RRA will first return as a solution the path consisted from road segment (A, B), (B, F) with a waiting delay of 2s at junction B. The returned solution is equivalent to the optimal one, however, the outer loop of RRA requires to clarify if that waiting can be transferred to the origin. Hence, the RRA inner loop re-executes with the new starting time and returns the path consisted from road segment (A, C), (C, B), (B, F) as the final solution with total cost 6s.

The complexity of RRA is $O(ME^2V)$, where $M < \infty$ denotes the number of iterations that the outer loop of RRA needs before converging to a solution. At this point it is worth pointing out that the RRA algorithm will always terminate in a finite number of iteration. Note that in any scenario, there is a finite number of vehicles which means that there is a finite number of reservations which also means that the intervals for which any link is non-admissible are also finite. Let T_{max}

denote the maximum time when a non-admissible interval for any link ends. For any execution of the RRA T_{max} is fixed, while the vehicle initial waiting time t_0 is monotonically increasing at discrete steps which are associated with the end of some non-admissible interval. Clearly, a non-congested path will always be found when $t_0 > T_{max}$, thus $M < \infty$.

The extensive simulation results that follow, demonstrate the superiority of the RRA algorithm compared to the uncontrolled scenario as it can achieve substantial improvements in terms of road utilization and travel times.

3.5.2 Performance evaluation

Setup

The road network under consideration is an 1.8 km^2 unsignalized homogeneous region of downtown San Francisco as illustrated in Fig. 3.7. The spatial compactness and homogeneity of this area was initially investigated in [29] and [111], while a similar region is used in [24]. The selected region consists of 99 road junctions and 208 single-lane road segments with lengths varying from 100 m to 400 m.

To simulate mobility along this region, SUMO micro-simulator [112] is employed using Krauss' car-following model [113]. Standard car-following parameters were used, including: vehicle length of 5 m, maximum speed 15 m/s, acceleration 2.5 m/s^2, deceleration 4.5m/s^2, driver imperfection 5%, driver reaction time 0.5 s, and minimum gap distance 2.5 m. The simulation time-step in SUMO was set to 0.1 s, while the time step of the algorithm is set equal to $T = 1$ s.

Finally, vehicles follow strictly their reservation routes, but not their reservation times. The reason is that in the conducted simulations, the route reservation scheme makes all reservations and routing decisions based on computed travel times without consideration of the actual network state at the time a vehicle request arrives. It is important to note that in SUMO environment, even under free-flow conditions the vehicle speeds vary significantly due to various random events that occur due to acceleration and deceleration of vehicles, and queuing delays at intersections due to the passing of higher priority vehicles; hence, travel times may vary from those computed by the proposed algorithms due to various sources of uncertainty such as

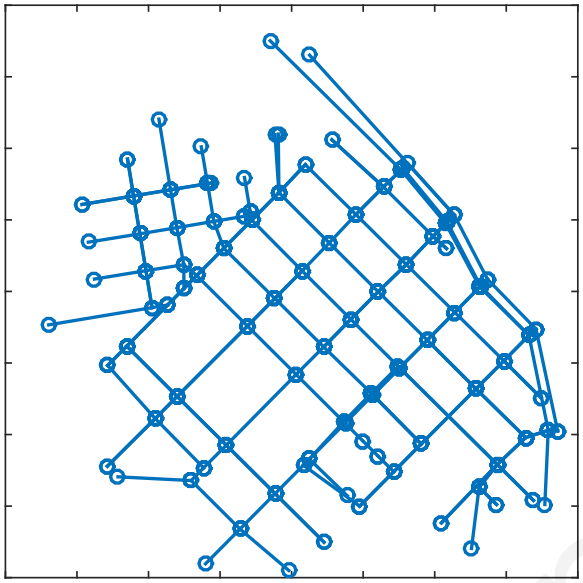


Figure 3.7: San Francisco road network under consideration.

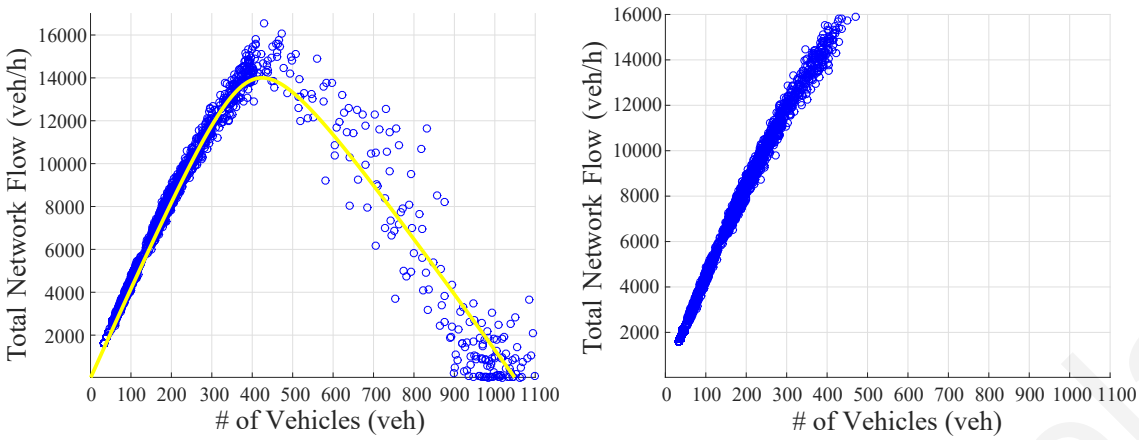
driver imperfection and junction priorities.

MFD analysis

As a first step, the region's MFD is analyzed in order to identify the parameters to be used by the RRA, including u_c and ρ_{ij}^C. To do so, a 6 hours scenario was simulated within which for the first 4 hours the input flow was set to 2000 veh/h and incrementally increased by 2000 veh/h for the next three hours while only exogenous[3] flows entering the network. Thereafter, the input flow was set to 4000 veh/h and 2000 veh/h for the last two hours in order to discharge the network. For the results presented hereafter, 10 Monte Carlo simulations were conducted within which the $O-D$ pairs and inter-arrival times were randomly generated.

Fig. 3.8 (a) depicts the Macroscopic Fundamental Diagram of the uncontrolled scenario (US) (i.e., where vehicles select their path strictly based on shortest path) which illustrates the total flow as a function of the total density of the network. In the figure, each point corresponds to 1 min measurements. The calibrated model shown by the solid yellow line is derived through the automated calibration method proposed by [114] for the single-regime *Van Aerde* model [115]. As detailed in [114], an initial set of free-flow-speed (u_f), speed-at-capacity (u_c), capacity and jam density

[3]The flows that generated and destined outside the considered network as they are entering and exiting from the sides.



(a) (b)

Figure 3.8: Region MFD of: (a) US; (b) RRA.

(ρ^J) values are used together with an iterative procedure to update u_f, u_c and ρ^J to compute the best fit values of the varying parameters which minimize the sum of squared orthogonal errors. From the figure, the following model parameters are obtained: $u_f = 47$ km/h, $u_c = 40.5$ km/h and $\rho^J = 1050$ veh.

Hence, the RRA algorithm was set to use $\rho_{ij}^C = 40$ veh/km/lane (i.e., around 40% of the regions total density) and travel time calculations are estimated using $u_c = 40.5$ km/h. We emphasize that even though for the purposes of computing the reservations for each vehicle, the constant u_c was used, the actual speed of each vehicle is determined by the simulator based on the assumed model. Fig. 3.8 (b) depicts the resulting MFD when the RRA algorithm is employed demonstrating the absence of the congested regime. This is achieved by restricting the number of vehicles allowed to simultaneously traverse the network.

To demonstrate the performance of RRA, the average volume of total network flow, the average volume of total network density and average volume of network speeds, obtained from each Monte Carlo realization of the aforementioned network scenario, are depicted in Figs. 3.9, 3.10, and 3.11. For comparison, the performance of US is also superimposed in these figures. Specifically, Fig. 3.9 illustrate the average volume of the total network flow for both US and RRA, as a function of the simulation over the Monte Carlo simulations. Similarly, Fig. 3.10 illustrates the average total network density and Fig. 3.11 the average of the mean network speeds over the

Monte Carlo simulations, for both US and RRA. Comparing these three figures, it is evident that using RRA the density decreases (near 330 veh for RRA compared to more than 500 veh for US as shown in Fig. 3.11) but the traveling speeds remain high and thus the flow is similar to that of US. Additionally, as Fig. 3.11 illustrates RRA always maintains traffic below critical capacity ρ^C (near 350 veh) even when demand is high (i.e., simulation time 200-240 min). At the same time, RRA maintains vehicles speeds near the speed-at-capacity at all times, as shown in Fig. 3.10.

To demonstrate the improvements obtained by RRA, Figs. 3.12 and 3.13 depict the percentage of per road segment density in relation with ρ^C and the per road segment speed as a function of the simulation time, respectively, for the case of US. Similarly, Fig. 3.14 and Fig. 3.15 illustrate identical results for the case of RRA. As Figs. 3.12, 3.13, 3.14 and 3.15 indicate, at low flow-demands the performance of both US and RRA is similar while at high flow-demands RRA outperforms US by avoiding congestion. Clearly, this is due to the fact that at low demands there are no significant restriction in the admissibility of particular road segments and so both approaches yield similar results; on the other hand, as demand increases, there is limited admissibility on road segments and RRA ensures that vehicles wait at their origins until an admissible path can be identified. As shown in Fig. 3.12, without a control mechanism, a subset of the road segments exceed their critical density and some of them get fully loaded especially in high densities (indicated with the magenta color in the figure). For these road segments, speed drops to near zero (as indicated with blue color in Fig. 3.13). On the contrary, with RRA road segment densities are maintained below the critical capacity (as shown in Fig. 3.14), allowing vehicles to maintain their speed near the free-flow speed. Hence, despite the increase in demand, RRA can greatly improve the overall network utilization.

Results

The proposed route-reservation architecture, that uses the RRA algorithm, is compared against US and with the state-of-the-art Decreasing Order of Time (DOT) algorithm [42]. DOT finds the time-dependent shortest travel time path according to a user-chosen time window for departure. As such, in this chapter the waiting time at the origin for both RRA and DOT is not considered in the total travel-time

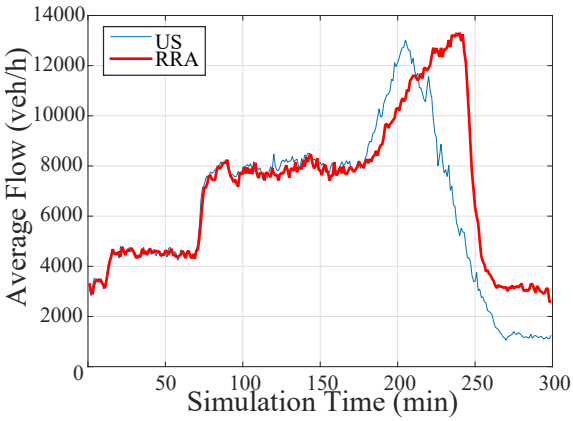


Figure 3.9: Average network flow over time for US and RRA.

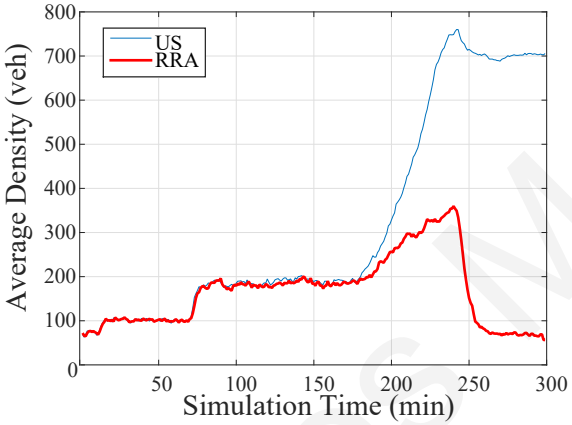


Figure 3.10: Average network density over time for US and RRA.

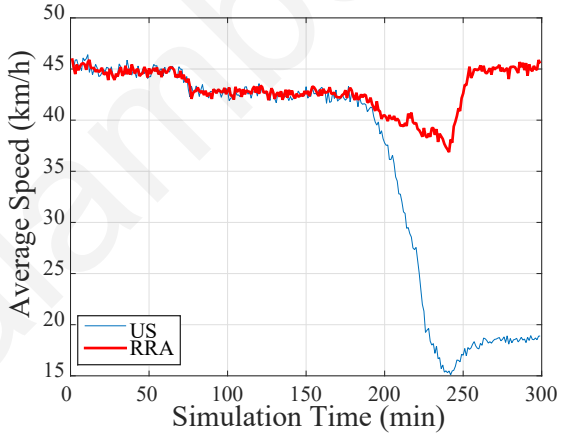


Figure 3.11: Average network speed over time for US and RRA.

for a fair comparison. For the same reason, the travel time estimates for the DOT algorithm were done according to the route reservation requests and using identical $O - D$ pairs. Finally, the maximum allowed waiting interval for DOT was set up to 1 min (i.e., half the average trip length for the considered network).

It should be noted here that, in the proposed solution, new route reservations are

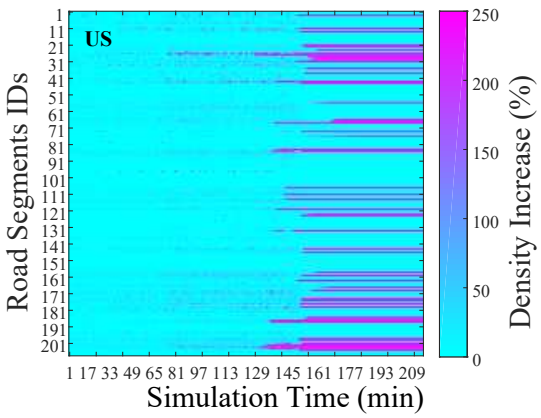


Figure 3.12: Evolution of traffic density for each road segment over time for US.

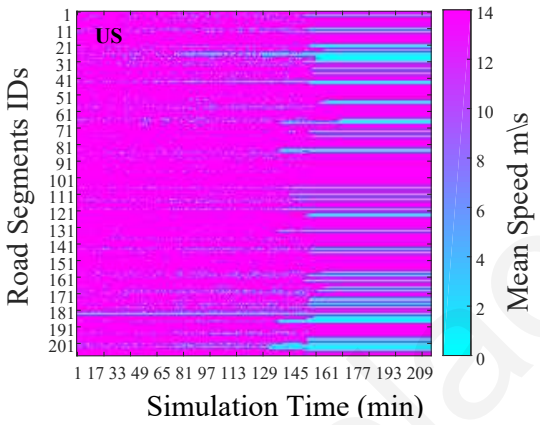


Figure 3.13: Evolution of speed for each road segment over time for US.

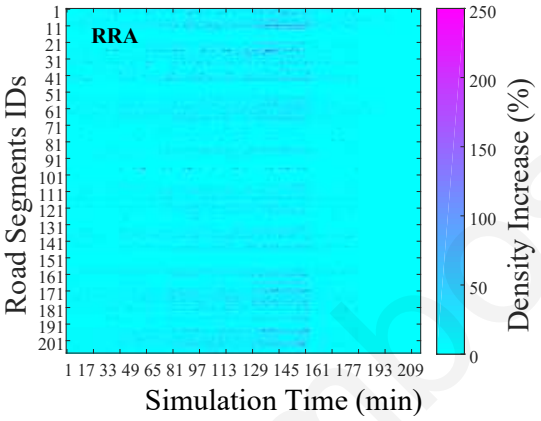


Figure 3.14: Evolution of traffic density for each road segment over time for RRA.

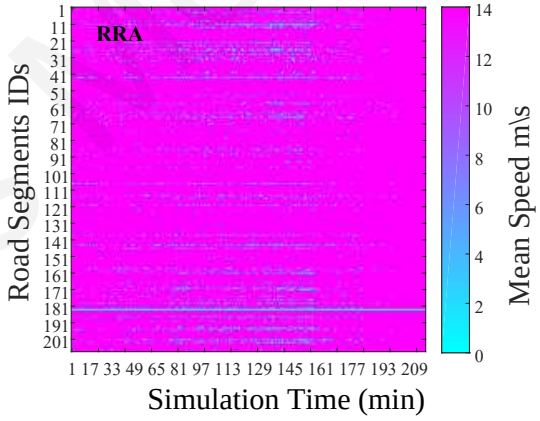


Figure 3.15: Evolution of speed for each road segment over time for RRA.

computed solely based on information from previous reservations made and not the actual network state. Since a number of different factors can affect vehicle journeys (including waiting at intersections and other vehicle interactions) the actual traversal of the reserved road segments can occur at time periods not anticipated. Hence, all vehicles follow their pre-computed reservation routes while actual travel times may vary due to various sources of uncertainty. These uncertainty errors are thoroughly examined in the sequel.

As before, 10 Monte Carlo simulations were executed with random $O - D$ pairs and with flow rates varying between $1000 - 8000$ veh/h over a period of 2 hours.

Figs. 3.16, and 3.17 show the average vehicle travel time, and the average total

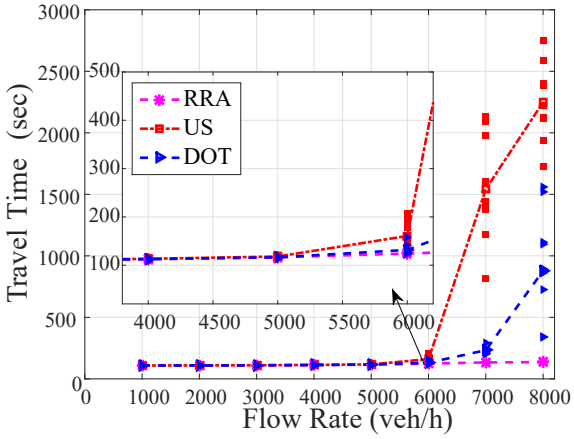


Figure 3.16: Average vehicle travel time.

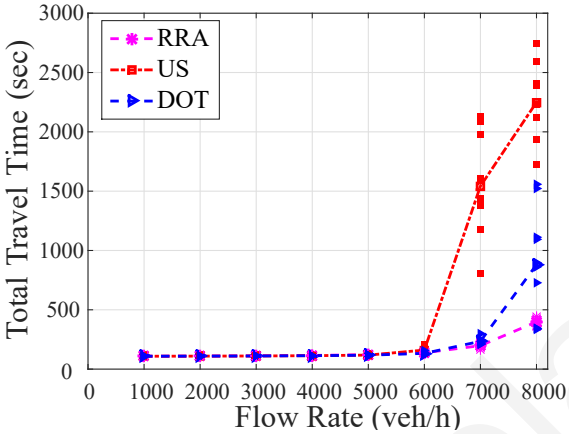


Figure 3.17: Average total travel time.

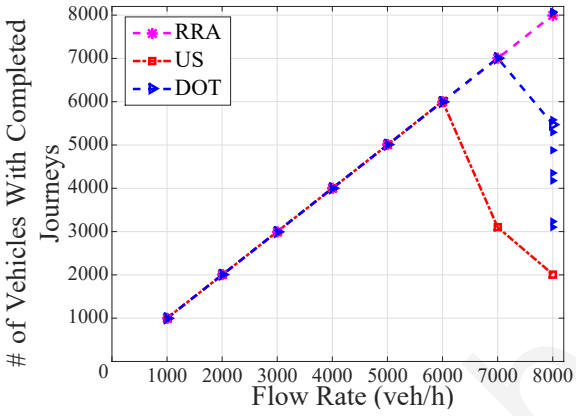


Figure 3.18: Number of vehicles with completed journeys.

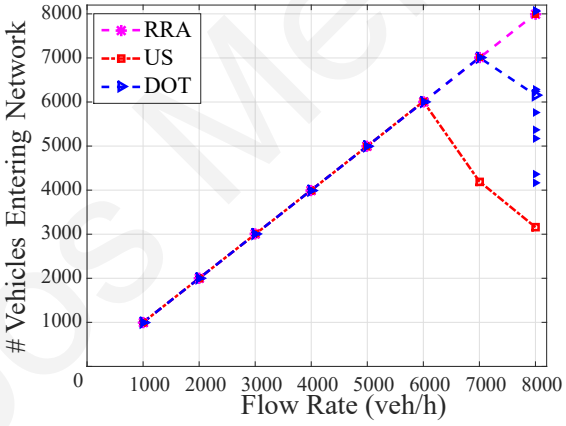


Figure 3.19: Number of loaded vehicles.

time including both the waiting at the origin and the travel time experienced across the network, as a function of the different flow rates. Additionally, 3.18, and 3.19 illustrate the average number of vehicles that completed their journeys and the number of vehicles entering the network within the simulation time, as a function of the different flow rates. The scattered plots in Figs. 3.16, and 3.17 depict the mean travel and total time of each realization, while the dashed lines represent the mean travel time and the mean total time for all realizations, respectively.

Similarly, the dashed lines in Figs. 3.18 and 3.19 illustrate the average number of vehicles that have finished their journey within the simulation time and the average number of vehicles entering the network, respectively. The scattered plots represent the realizations obtained by each simulation run. Figs. 3.16, 3.17,

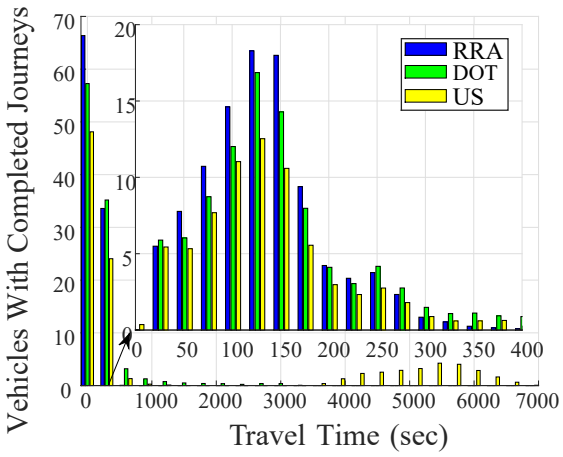


Figure 3.20: Travel time distribution of 7000 *veh/h*.

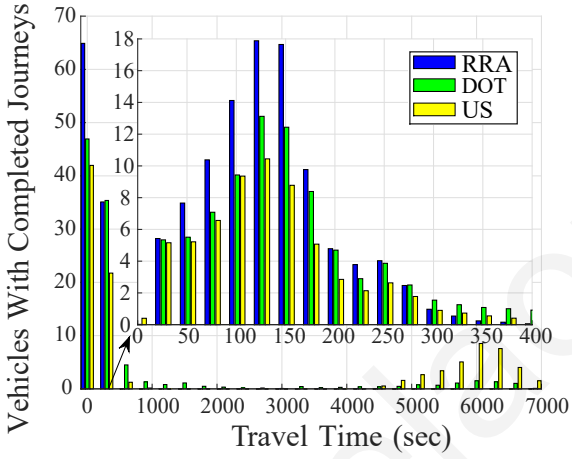


Figure 3.21: Travel time distribution of 8000 *veh/h*.

3.18, and 3.19 illustrate the overall network behavior considering different flow rates. As indicated in Figs. 3.16, 3.18, and 3.19, for low flow rates (ranging from 1000 veh/h to 6000m veh/h), there is minimal congestion and thus both algorithms have similar behavior to US. At higher flow rates, congestion emerges and RRA is shown to greatly outperform DOT since the travel time remains short for RRA and all vehicles arrive at their destination within the investigated simulation time. Fig. 3.17 indicates that both RRA and DOT algorithms in low flow rates observe minor waiting intervals at the origin. On the other hand, as congestion increases, the mean origin waiting time for RRA increases while the mean origin waiting time for DOT remains almost constant (around 5 s increase). Nonetheless, Fig. 3.17 indicates that RRA outperforms both US and DOT; with waiting times observed at the origin being only a small percentage to the waiting times caused due to congestion.

Figs. 3.20 and 3.21 illustrate the travel time distribution for all vehicles that manage to reach their destination during the simulation time for flow rates of 7000 veh/h and 8000 veh/h. As illustrated, RRA greatly improved travel time compared to DOT. As shown in Fig. 3.21, the mean travel time for RRA is 135.9 s, for DOT is 695 s and for US is 2163.5 s. The standard deviation for RRA is 64.8 s, for DOT is 1536.8 s and for US is 2774.1 s demonstrating that as congestion of the road segments increases, RRA is more stable and accurate than DOT. Further, RRA is more resilient to the increase in flow rate since travel times do not significantly deviate.

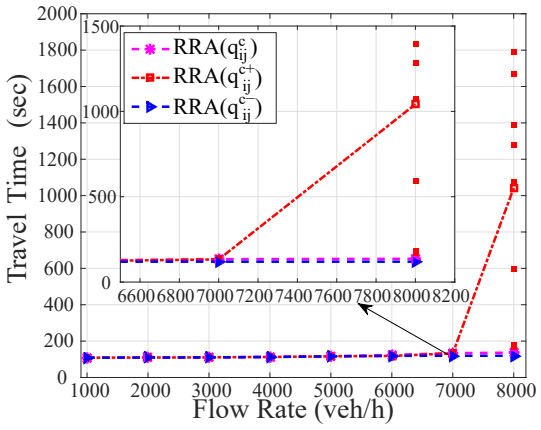


Figure 3.22: Average travel time for RRA with varying critical capacity values.

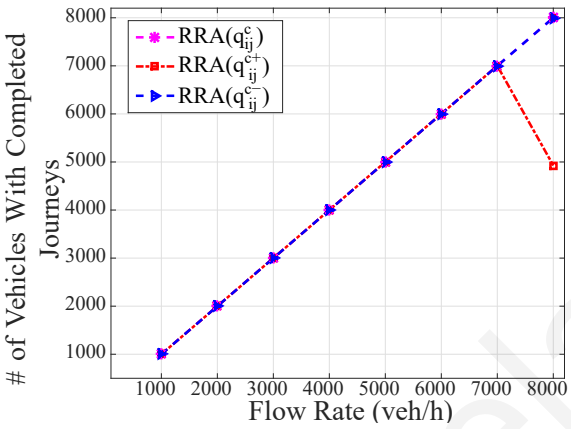


Figure 3.23: Number of vehicles with completed journeys using RRA with varying critical capacity values.

The RRA performance for different values of ρ_{ij}^{C} is also examined. Figs. 3.22 and 3.23 show the average vehicle travel time and the average number of vehicles that completed their journeys for the cases where $\rho_{ij}^{C} = 0.4\rho_{ij}^{J}$, $\rho_{ij}^{C-} = 0.3\rho_{ij}^{J}$, $\rho_{ij}^{C+} = 0.5\rho_{ij}^{J}$ where ρ_{ij}^{C-} and ρ_{ij}^{C+} deviate by -25% and 25%, respectively from the selected critical capacity value (*i.e.*, ρ_{ij}^{C}). Both figures indicate that a 25% increase over the ρ_{ij}^{C} result to a drop in algorithm performance since travel times increase and a lower number of vehicles manages to complete their journeys. Interestingly, using lower capacities the observed algorithm performance is similar to that of ρ_{ij}^{C} since no congestion occurs and travel times are similar since segment densities do not exceed their critical values.

Nevertheless, a lower ρ_{ij}^{C} value increases the waiting time at the origin. This is illustrated in Fig. 3.24 which shows the waiting-time that vehicles need to wait before departing for their journeys in the form of a box-plot[4]. This behavior is expected since a decrease of the allowed capacity reduces the number of vehicles that simultaneously traverse the network. Additionally, as illustrated in Fig. 3.24 a higher ρ_{ij}^{C} value decrease the waiting time at the origin affecting the algorithm performance

[4]The bottom and top of each box indicate the first and third quartiles (25% and 75%) of a ranked data set, while the horizontal line inside the box indicates the median value (second quartile). The horizontal lines outside the box indicate the lowest/highest datum still within 1.5 inter-quartile range of the lower/upper quartile.

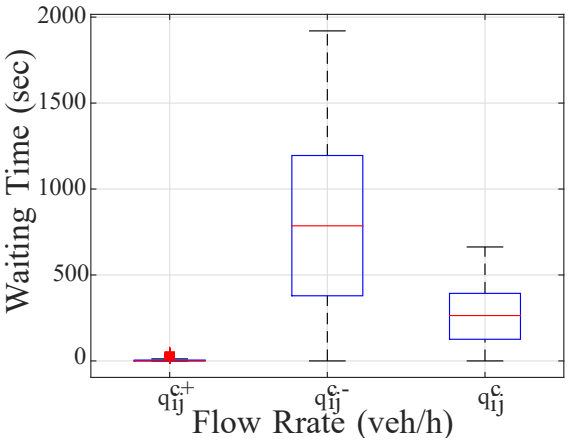


Figure 3.24: Waiting intervals for ρ_{ij}^{c+}, ρ_{ij}^{c-} and ρ_{ij}^{c} (8000veh/h).

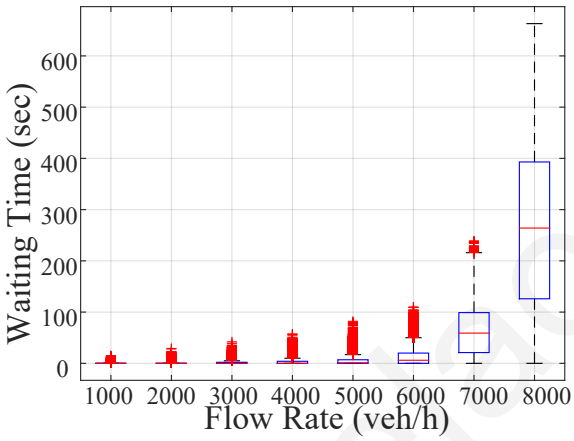


Figure 3.25: RRA origin waiting times.

as congestion occurs. Therefore, the late depart can affected the network behavior as congestion is avoided.

Notably, as demand increases, a higher number of vehicles request to traverse the network. Since the allowed density is restricted below the critical value, vehicles prefer to wait at their origin until an admissible path is feasible. Fig. 3.25 demonstrates that as flow rates increase, waiting time increases exponentially. However, this is expected since in high-demand scenarios, significant waiting needs to be incurred to maintain high network flows. Even so, the average waiting is within acceptable levels (5 min) and therefore, a small departure delay could prove sufficient for the overall network operation.

Moreover, Fig. 3.26 illustrates the mean distance traveled by all vehicles as a function of different flow rates in relation to the shortest distance path (computed using Dijkstra's algorithm). In fact, RRA paths appear to maintain constant travel times (close to the shortest distance path) irrespective of the flow demand, as illustrated in Fig. 3.26. Looking at the findings of both Figs. 3.26 and 3.25 whenever there are non-admissible road segments, the RRA algorithm tends to postpone departures and enable vehicles to traverse through shortest distance paths instead of taking longer routes. This is also verified in Fig. 3.27, which illustrates the percentage of vehicles that travel through paths other than the shortest distance path. The figure assumes a flow rate of 8000 veh/h. As shown, the majority of vehicles (around 75%) were guided through the shortest paths. All these findings indicate that whenever there

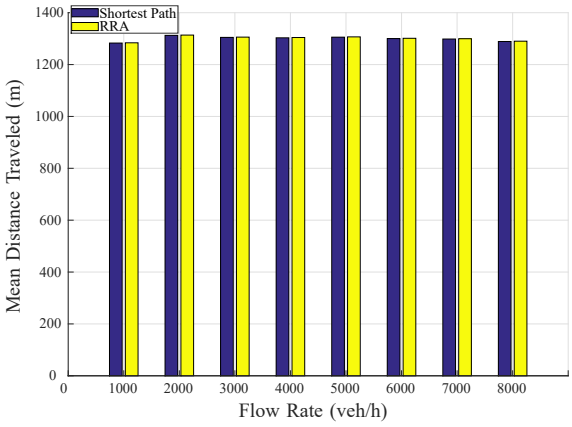


Figure 3.26: Mean distance traveled comparison for RRA and shortest path.

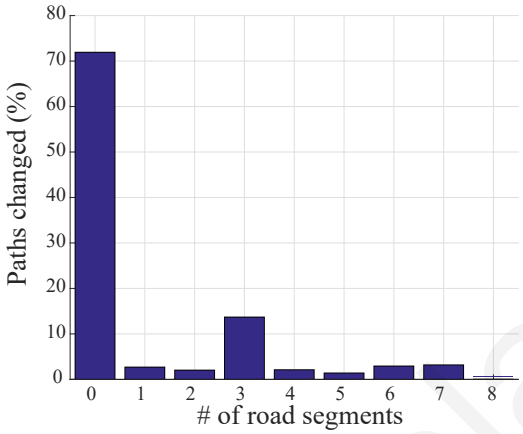


Figure 3.27: Road segments that changed.

are non-admissible road segments, the RRA algorithm prefers to instruct vehicles to wait at their origin before departing, instead of scheduling vehicles through longer distance paths. This indicates that through RRA longer paths are avoided so that both travel time and cost are minimized.

3.5.3 Continuous time solutions for EDAT

For comparison purposes this section adopts the continuous time formulation presented in Section 3.3.2 to develop a Mixed Integer Linear Program (MILP) formulation to solve the continuous time EDAT problem optimally. The developed MILP formulation is used to investigate the quality of the solution achieved form the RRA algorithm that heuristically solves the EDAT problem. For a fair comparison, a continuous-time RRA algorithm (i.e., RRAC) is also being derived with simulation results indicating that the RRAC algorithm provides fast and close-to-optimal results. Comparison of the continuous-time RRAC with the corresponding discrete time RRA algorithm illustrate that the former is more accurate and faster especially for high-demand traffic scenarios.

Mixed Integer Linear formulation of EDAT Problem

To optimally solve the EDAT problem, we develop a MILP formulation that selects road segments, respecting the admissibility conditions, that route a vehicle between

the origin O and the destination D to minimize the arrival time at D. To do so, the optimal path from O to D is described through binary variables χ_{ij}, $(i,j) \in \mathcal{E}$, denoting whether a road segment (i,j) is part of the optimal path p^* ($\chi_{ij} = 1$) or not ($\chi_{ij} = 0$). In addition, the auxiliary binary variables ψ_{ijk}, $(i,j) \in \mathcal{E}$, $k \in \mathcal{K}_{ij}(t_0)$ are introduced to indicate whether a road segment (i,j) satisfies the admissibility condition at the k_{th} time interval ($\psi_{ijk} = 1$) or not ($\psi_{ijk} = 0$).

The mathematical formulation for the EDAT problem can be expressed as follows:

$$(\mathrm{P}_1) \min_{\chi_{ij}, \psi_{ijk}, d_i, \forall i,j,k} d_D \tag{3.27a}$$

$$\text{s.t.} \sum_{(i,j)\in\mathcal{E}} \chi_{ij} - \sum_{(j,i)\in\mathcal{E}} \chi_{ji} = \begin{cases} 1, & \text{if } i = O, \\ -1, & \text{if } i = D, \\ 0, & \text{if } i \in \mathcal{V} \setminus \{O, D\}, \end{cases} \tag{3.27b}$$

$$\sum_{k\in\mathcal{K}_{ij}(t_0)} \psi_{ijk} = \chi_{ij},\ (i,j) \in \mathcal{E}, \tag{3.27c}$$

$$d_i - d_j + \hat{\tau}_{ij}(t_0) \le M_1(1 - \chi_{ij}),\ (i,j) \in \mathcal{E}, \tag{3.27d}$$

$$d_i - d_j + \hat{\tau}_{ij}(t_0) \ge M_2(1 - \chi_{ij}),\ (i,j) \in \mathcal{E}, \tag{3.27e}$$

$$d_i \ge t^l_{ijk}\psi_{ijk},\ (i,j) \in \mathcal{E},\ k \in \mathcal{K}_{ij}(t_0), \tag{3.27f}$$

$$d_j \le t^u_{ijk} + M_3(1 - \psi_{ijk}),\ (i,j) \in \mathcal{E},\ k \in \mathcal{K}_{ij}(t_0), \tag{3.27g}$$

$$d_i \ge t_0,\ i \in \mathcal{V} - \{O\},\ d_i \ge t_0,\ i = O, \tag{3.27h}$$

$$\chi_{ij} \in \{0,1\},\ (i,j) \in \mathcal{E}, \tag{3.27i}$$

$$\psi_{ijk} \in \{0,1\}, (i,j) \in \mathcal{E},\ k \in \mathcal{K}_{ij}(t_0). \tag{3.27j}$$

where M_1, M_2 and M_3 are appropriately selected constants. In the above formulation, equality (3.27b) describes the flow constraints that ensure connectivity of the optimal path from source to destination, while (3.27c) forbids the traversal of link (i,j) at any time if $\chi_{i,j} = 0$. Constraints (3.27d) and (3.27e) ensure the logical condition "if $\chi_{ij} = 1$ then $d_j = d_i + \hat{\tau}_{ij}(t_0)$" which describes the cost increase when traversing road link (i,j). To examine the validity of the condition, notice that for $\chi_{ij} = 1$ constraints (3.27d) and (3.27e) enforce the equality $d_j = d_i + \hat{\tau}_{ij}(t_0)$. For $\chi_{ij} = 0$, constraints (3.27d) and (3.27e) should have no effect on the optimization problem; for this reason, constants M_1 and M_2 are selected to provide tight upper and lower bounds on $d_i - d_j + \hat{\tau}_{ij}(t_0)$,

respectively, such that the resulting inequalities are always true. Letting d^u denote an upper bound to the solution of the EDAT problem (e.g., obtained through a heuristic algorithm), such that $-d^u \leq d_i - d_j \leq d^u$, yields the bounds $M_1 = d^u + \max_{(i,j)\in\mathcal{E}}\{\hat{\tau}_{ij}(t_0)\}$ and $M_2 = -d^u + \min_{(i,j)\in\mathcal{E}}\{\hat{\tau}_{ij}(t_0)\}$. In a similar fashion, we can deduce that constraints (3.27f) and (3.27g) are equivalent to the logical condition "if $\psi_{ijk} = 1$ then $t^l_{ijk} \leq d_i$ and $d_j \leq t^u_{ijk}$" indicating that when $\psi_{ijk} = 1$ then the admissibility condition needs to be satisfied for link (i, j), and time interval k. Note that M_3 needs to provide an upper bound to d_j, hence $M_3 = d^u$. Constraints (3.27h)-(3.27j) simply denote the nature (e.g. continuous, binary) and range of each set of variables. Note that the origin waiting time (i.e., w), is implicitly imposed by letting $d_O \geq t_0$, so that $w = d_O - t_0$.

Problem P_1 is an MILP program that can be solved with standard optimization solvers, yielding the optimal solution to the EDAT problem. Nonetheless, the mixed-integer nature of the formulation implies that in certain cases the MILP solver may need exponentially large time to complete. For this reason, we also develop a close-to-optimal low-complexity heuristic in the next section.

Route Reservation Algorithm Continuous-Time (RRAC)

In this section a polynomial complexity, close-to-optimal algorithm is developed that solves the EDAT problem in continuous-time. As emphasized above, RRAC is a continuous-time adaptation of the discrete-time heuristic proposed in Section 3.5.1 in which route reservations are made for discrete time slots rather than continuous time intervals. RRAC is also an iterative algorithm that solves a series of relaxed problems that provide lower bounds on the optimal solution until it convergence to a feasible solution of the original problem.

The m-th iteration of the RRAC algorithm solves a variation of the EDAT problem with waiting allowed at all nodes while the waiting time at the origin is gradually increased, hereafter referred to as *Relaxed-EDAT*. The Relaxed-EDAT can be optimally solved using a customized version of Dijkstra's algorithm; hence, it provides a lower bound on the optimal solution of the EDAT problem for the specific value of preliminary waiting at the origin, w^p_{m-1}. The preliminary waiting is then updated to $w^p_m = w^p_{m-1} + w^T$, where w^T is the total waiting at all nodes from the solution of the Relaxed-EDAT. RRAC terminates once the solution of the Relaxed-EDAT involves

Algorithm 3 Iterative Dijkstra Algorithm Continuous Time (RRAC).

1: **Input:** $G(\mathcal{V}, \mathcal{E})$, $\mathcal{S}_{ij}(t_0)$, O, D, t_0;
2: **Initialization:** $m = 0$; $w_0^p = 0$;
3: **repeat**
4: $\quad m \leftarrow m + 1$;
5: $\quad [w^T, \mathbf{P}] \leftarrow$ relaxed-EDAT($G(\mathcal{V}, \mathcal{E})$,
6: $\qquad \mathcal{S}_{ij}(t_0)$, O, D, t_0, w_{m-1}^p);
7: $\quad w_m^p \leftarrow w_{m-1}^p + w^T$;
8: **until** $\{w^T = 0\}$
9: **Output:** $\mathbf{P}$ and w_m^p

no waiting. Note that the case of non-zero waiting only at the origin node is included in the above termination condition, because in the next iteration no waiting will be observed at all nodes. Once the algorithm terminates, the path found from the solution of the Relaxed-EDAT, expressed through the predecessor list **P** of the best solution found, is returned with total waiting at the source node equal to w_m^p. RRAC is presented in Algorithm 3.

The customized Dijkstra algorithm for solving the Relaxed-EDAT problem is described in Algorithm 4. First, the algorithm initializes the arrival times d_{v_i} to each junction to infinity except from the arrival time of the origin node which is set equal to the request time plus the preliminary waiting incurred so far from previous iterations of RRAC. Initially all junctions are non-labelled, and hence the set of non-labelled nodes, $\mathcal{Q}$, is set to $\mathcal{V}$, while the predecessor list **P** is set to null (lines 2-3). Then, an iterative procedure is followed until the travel times of all junctions are finalized, i.e. all nodes are labelled. In each iteration, the junction v_l that has the earliest arrival time is labelled (i.e. its travel time is finalized) (line 5) and then it is examined whether the travel time of v_l's neighbours can be improved (lines 6-12). To do so, the smallest waiting time w_{lj} is computed which is needed to go from junction v_l at time d_{v_l} to junction v_j, based on the admissibility of the particular road segment (line 7). If the examined labelled junction (v_l) improves the arrival time at its neighbour (v_j) then the arrival time at v_j, d_{v_j} is updated and v_l is noted as the predecessor of v_j (lines 8-11). In this way, the algorithm calculates and updates the

Algorithm 4 Inner loop of the continuous time RRAC (IL-RRAC).

1: **Input:** $G(\mathcal{V},\mathcal{E})$, $\mathcal{S}_{ij}(t_0)$, O, D, t_0, w^p;
2: **Initialization:** $d_{v_i} = \infty$, $v_i \in \mathcal{V}$, $d_O \leftarrow t_0 + w^p$, $\mathcal{Q} \leftarrow \mathcal{V}$,
3: $P[v_i] \leftarrow NULL$, $v_i \in \mathcal{V}$;
4: **while** $\mathcal{Q} \neq \emptyset$ **do**
5: $\quad v_l \leftarrow \text{argmin}_{v_i \in \mathcal{Q}}\{d_{v_i}\}$;
6: $\quad \mathcal{Q} \leftarrow \mathcal{Q} - \{v_l\}$;
7: $\quad$ **for** $(l, j) \in \mathcal{E}$ **do**
8: $\quad\quad w_{lj} \leftarrow \min_{w \geq 0}\{d_{v_l} + t + w \in \mathcal{S}_{lj}(t_0), 0 \leq t \leq \bar{\tau}_{lj}\}$
9: $\quad\quad c_{lj}(d_{v_l}) \leftarrow \bar{\tau}_{lj} + w_{lj}$;
10: $\quad\quad$ **if** $\{d_{v_j} > d_{v_l} + c_{lj}(d_{v_l})\}$ **then**
11: $\quad\quad\quad d_{v_j} \leftarrow d_{v_l} + c_{lj}(d_{v_l})$, $P[v_j] = v_l$;
12: $\quad\quad$ **end if**
13: $\quad$ **end for**
14: **end while**
15: $w^T \leftarrow d_D$, $v_v \leftarrow D$;
16: **repeat**
17: $\quad w^T \leftarrow w - c_{P[v_v],v_v}(d_{v_{P[v_v]}})$;
18: $\quad v_v \leftarrow P[v_v]$;
19: **until** $\{v_v = O\}$
20: **Output:** w^T, **P**.

earliest arrival times of non-labelled neighbour road junctions in each iteration. The above procedure repeats until all road junctions are characterized as labelled. Finally, the predecessor list which holds the best path from O to D is exploited to calculate the total waiting time at all junctions, w^T (lines 14-17). By allowing waiting at all nodes, Algorithm4 returns a better solution than the proposed solution; nonetheless, it is not applicable to real traffic networks since vehicles cannot stop and wait along arbitrary road junctions.

The complexity of the algorithm is $O(LE^2/V)$, with L denoting the number of reiterations of the Relaxed-EDAT problem that is required. The complexity of RRAC (i.e., the proposed continuous-time variant) is improved to $O(LE^2/\log(V))$ since

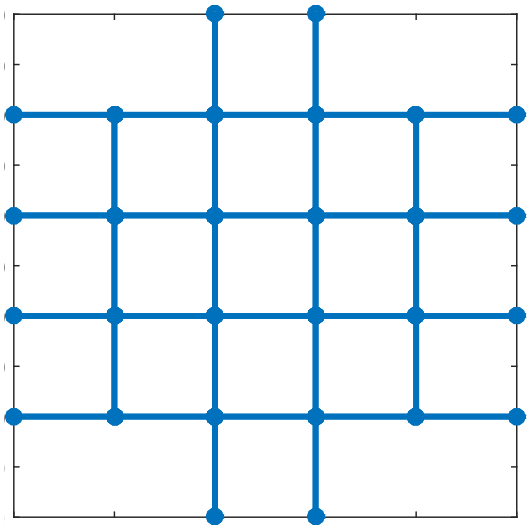


Figure 3.28: Network Topology considered in performance evaluation.

the waiting intervals that may be required at any intermediate road junction can be calculated more efficiently through continuous-time intervals. Additionally, the adapted continuous-time domain solution offers more precise solutions (avoiding quantization errors) in terms of route reservation. Specifically, more vehicles can be scheduled within the same time period while a lower number of Relaxed-EDAT iterations are required (since no time intervals are wasted due to quantization).

3.5.4 Performance Evaluation

Setup

In order to evaluate the performance of the developed algorithms, a Manhattan-style network topology (see Fig. 3.28) is considered consisting of 36 two-way, single-lane road segments and 28 junctions. The Manhattan-style network is selected due to the complexity of the problem and the fact that MILP is able to derive fast solution only in case of small scale networks. Traffic arrives in the network probabilistically, according to Poisson distribution, while for each trip a random $O-D$ pair is selected. All vehicles are assumed to follow the route assigned to them by the RSU using the route reservation scheme. A critical density of $\rho_{ij}^{C} = 30veh/km/lane$ and a free-flow speed of 15 m/s are used. Simulations are performed for varying total flow rates in the range of 1000 to 8000 veh/h. Both RRAC and MILP are evaluated in identical scenarios and reservations initial states for fair comparison. Additionally both are

compared with the results obtained by the Route Reservation Algorithm Discrete-Time (i.e., RRAD). Finally, the MILP formulation is constructed and solved using the Gurobi mathematical programming solver [116], while RRAC and RRAD are implemented in C++.

The performance evaluation is performed in both ideal and micro-simulation environments. In ideal environment it is assumed that there is no uncertainty in the MFD. It is further assumed that the speed under free-flow conditions is constant and equal to the free-flow speed u_f. This means that the topology can be modeled as a graph with edge cost equal to the transit-time of the corresponding road segment with free-flow speed so that the path travel time is equal to the sum of the free-flow transit-times of all traversed links.

On the other hand, the micro-simulation *environment* aims to capture the stochastic nature of traffic mobility within a realistic road network which results in uncertainty in the MFD. In a realistic environment vehicles form queues and follow each other, have acceleration and deceleration times, and experience delays at intersections due to the passing of higher priority vehicles. To capture these effects, the SUMO micro-simulator [112] is used, which employs the Krauss car following model [113] for vehicle mobility. In our simulations the car-following model parameters are set as follows: vehicle length 5 m, maximum speed 15 m/s, acceleration 2.5 m/s^2, deceleration 4.5 m/s^2, and minimum gap distance of 2.5 m. To account for stochastic effects, 10 Monte Carlo simulations are performed for each considered scenario. In addition to comparing the performance of RRAC, RRAD and MILP, comparisons are also conducted against the uncontrolled scenario (referred to as US in the plots) where each vehicle travels from the origin to the destination along the shortest path (based on the distance) and no waiting at the origin is used.

Results (ideal simulation environment)

Fig. 3.29 depicts the mean travel time that vehicles experience within the network for different demand flow rates. Travel time is defined as the time lapse between origin departure time and the destination arrival time. As anticipated, all algorithms experience approximate similar travel times in low demand flow rates (below 3000 veh/h) as there is no congestion. As the flow rate increases, the MILP experience higher

average travel times than the RRAC and RRAD approaches. In terms of path length, MILP produces longer paths as the network travel time increases, while both the RRAC and RRAD paths appear to maintain constant travel times irrespective of the flow demand.

Figs. 3.30 and 3.31 show the origin waiting time and the total time (origin waiting time plus travel time) that vehicles experience for different demand flow rates, respectively. Despite the fact that the MILP approach schedules vehicles through longer paths, the total time is lower compared to both the RRAC and RRAD algorithms, because the former generates solutions with shortest origin waiting times. More specifically, at low flow-rates (1000 - 5000 veh/h) the performance of all algorithms is almost identical as there are no significant restrictions in terms of road segment admissibility. However, at higher demands the MILP approach outperforms the RRAC and RRAD by up to 20% and 30%, respectively, as the admissibility sets become more fragmented and require the examination of a large number of time-interval combinations to find the best path. Fig. 3.30 also illustrates that as the demand flow rate increases, the average waiting time increases exponentially and becomes more than one order of magnitude larger than the mean travel time. This figure also indicates that the origin waiting time for RRAD algorithm is up to 10% larger compared to the RRAC, since RRAD may reserve more time-slots than it actually needs due to inaccuracies in time discretization. Despite the large origin waiting times, we demonstrate that under micro-simulation environment all proposed algorithms lead to significantly better performance compared to the uncontrolled scenario (US) which yields higher travel times due to congestion.

An interesting observation is that the MILP approach may impose virtual waiting within the network by introducing cycles to the vehicle paths. Although such paths do have smaller total times, this often leads to higher fuel consumption, which may be undesirable. On the contrary, the RRAC and RRAD approaches introduce all waiting at the origin, as they are based on Dijkstra's algorithm which prohibits cycles in the produced paths.

Fig. 3.32 examines the average execution time of all algorithms for different flow rates. As expected, higher demand flow rates lead to longer execution times for all algorithms because the admissibility sets become more fragmented. Clearly,

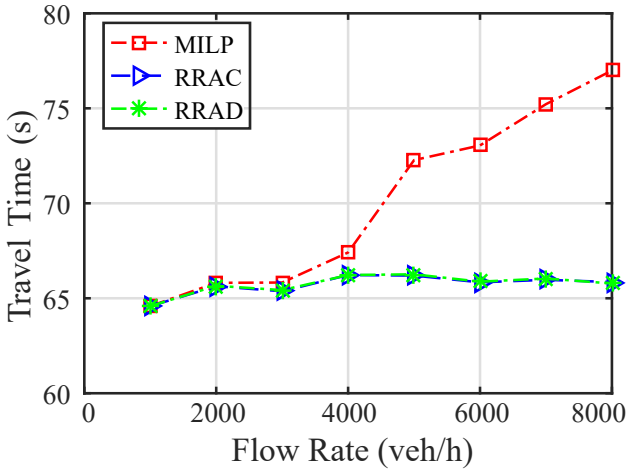


Figure 3.29: Average vehicle travel time from origin to destination for varying demand flow rate.

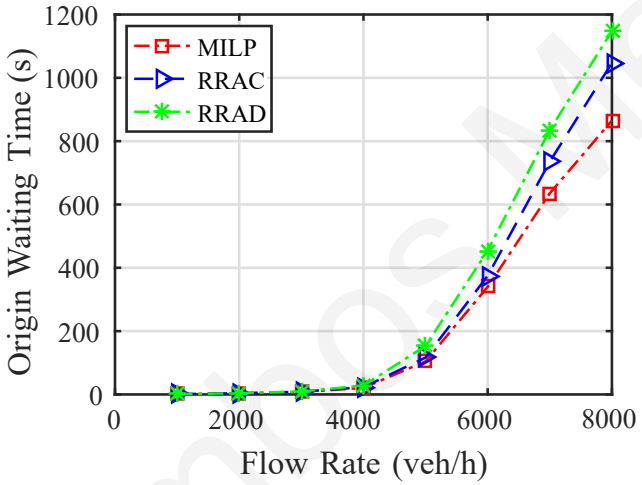


Figure 3.30: Average origin waiting time of all vehicles for varying demand flow rate.

the RRAC and RRAD algorithms are significantly faster than the MILP approach, outperforming the latter by around three orders of magnitude. Another interesting observation regards the execution speed of the two Dijkstra-based algorithms, as no algorithm is dominated by the other. In particular, RRAD appears faster than the RRAC for low flow rates and slower than the RRAC for high flow rates. The increase in the execution time of RRAD at higher flow rates is possibly due to decreased road segment admissibility, as a result of unnecessary reservations in discrete time, which leads to longer horizon problems to be solved. Since, the RRAC algorithm achieves the lowest execution times with near-optimal performance it has also been selected for performance evaluation in micro-simulations.

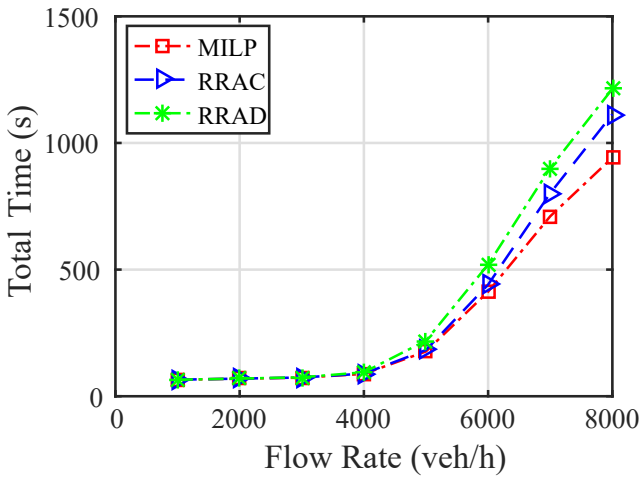


Figure 3.31: Average total time (sum of travel time and origin waiting time) of all vehicles for varying demand flow rate.

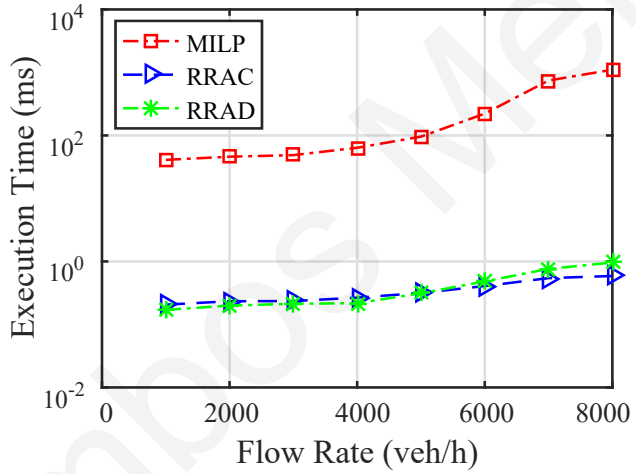


Figure 3.32: Average execution time of different algorithms for varying demand flow rate.

Results (realistic simulation environment)

Figs. 3.33 and 3.34 illustrate respectively, the mean number of vehicles that reach their destination and the average travel time as a function of the different flow rates. Specifically, the dashed lines in Fig. 3.33 represent the average number of vehicles that have finished their journey within the simulation time and the scattered plots are the realizations obtained by each simulation run. Similarly, the dashed lines in Fig. 3.34 represents the value of the average travel time for the different realizations and the scattered plots represent the average travel time across all Monte Carlo simulations.

As illustrated in Figs.3.33 and 3.34, at low demands (flow rate 1000 – 5000) veh/h

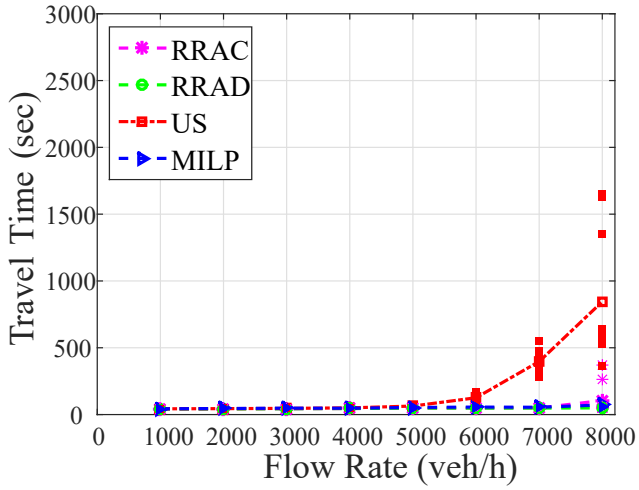


Figure 3.33: Average travel time ($t \rightarrow s$).

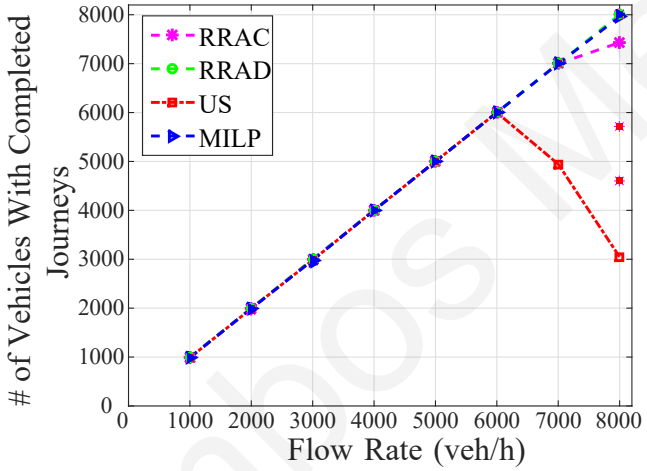


Figure 3.34: Number of vehicles towards to the route end.

all approaches behave similarly in terms of the average number of vehicles that reach their destination and the average travel time. At high demand, the MILP, RRAD and RRAC algorithms outperform by several orders of magnitude the US approach in terms of average travel time. In addition, both algorithms (RRAD and RRAC) allow for the completion of almost all vehicle trips, whereas using the US approach less than 50% of the vehicles arrive at their destination in high congestion scenarios. Comparing the performance of the proposed algorithms, MILP and RRAD results in 6% more vehicles with completed trips and 46% less travel time for the highest congestion scenario compared to RRAC.

Fig. 3.35 illustrates the distribution of the percentage of vehicles that managed to reach their destination as a function of travel time, when a flow rate of 8000 veh/h is

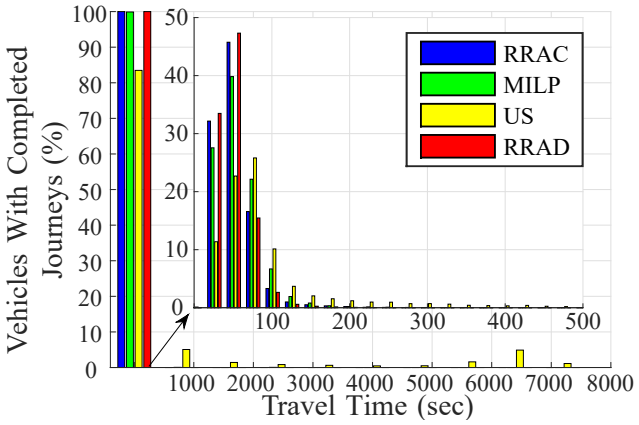


Figure 3.35: Travel time distribution (8000veh/h).

used. From the figure, it is clear that the MILP, RRAC and RRAD approaches provide more robust paths than US, as almost all trips finish within the 500 s, whereas US has a significant portion of vehicles finishing above 1000 s.

Summarizing the results from the realistic environment, it is clear that using the proposed strategy of not admitting vehicles that are in road segments that have reached their critical density, yields significantly improved performance compared to simply routing based on the short-path. In addition, the results indicate that despite the fact that RRAC is a heuristic algorithm, is provides close-to-optimal and fast results.

3.6 Traffic Load Balancing

According to the EDAT problem, vehicles are served in first come first served order, causing the system to be susceptible to fairness issues as some vehicles may be routed through longer routes instead of following the shortest ones or may be forced to wait longer than others. To deal with this issue, a Traffic Load Balancing (TLB) formulation is proposed that seeks to navigate vehicles through congestion-free road segments while minimizing the **variance** of the overall traffic observed in each road segment. Traffic load balancing can lead to better network performance as it improves the homogeneity of the network, alleviating in this way various unwanted phenomena, such as grid-locks. Furthermore TLB, aims also to find a path that provides a good trade-off between early destination arrival time (i.e., EDAT

problem) and traffic load balancing. Although arriving at the destination at the earliest possible time is highly desirable, balancing the traffic creates robustness against travel time estimation inaccuracies due to acceleration/deceleration and queuing at intersections. This section provides a detail mathematically formulation of the TLB problem while two efficient algorithms are derived able to solve EDAT and TLB problem in pseudo-polynomial time considering a discrete time domain.

3.6.1 Traffic Load Balancing (TLB) problem

Given an origin-destination ($O-D$) pair, the time-stamp t_0 of the routing request, the admissibility state $x_{ij}(t)$ of each road segment and the total number of reservations of each road segment $n_{ij}(t)$, the TLB problem requests the path (from O to D) that minimizes the variance of densities across the network. Let p_k denote the k-th path from source O to destination D defined in the same way with the EDAT problem. Let also $T_H = a \cdot d_D^*$ denote the time-horizon for the TLB problem, where d_D^* is the solution to problem of the EDAT problem (Eq. (3.8)) and $a \geq 1$ is a constant that defines the trade-off between achieving the earliest destination arrival for the particular vehicle and load balancing the traffic. Based on the initial conditions of the network, we define the mean μ_0, second moment M_0, and variance σ_0^2 of the densities of the network for the time-horizon considered as:

$$\mu_0 = \frac{1}{N_S} \sum_{(i,j)\in E} \sum_{\tau=t_0}^{T_H} \frac{n_{ij}(\tau)}{b_{ij}} \tag{3.28}$$

$$M_0 = \frac{1}{N_S} \sum_{(i,j)\in E} \sum_{\tau=t_0}^{T_H} \left(\frac{n_{ij}(\tau)}{b_{ij}} \right)^2 \tag{3.29}$$

$$\sigma_0^2 = \frac{1}{N_S} \sum_{(i,j)\in E} \sum_{\tau=t_0}^{T_H} \left(\frac{n_{ij}(\tau)}{b_{ij}} - \mu_0 \right)^2 = M_0 - \mu_0^2 \tag{3.30}$$

where $b_{ij} = l_{ij}\lambda_{ij}$ and $N_S = (T_H - t_0 + 1)N_E$.

When a vehicle with path p_k and waiting time at the origin w enters the network, the number of reservations $n_{ij}(t)$ at the corresponding road segments is increased by one for the occupancy period of each segment. To obtain values for the mean $\mu_k(t)$, the second moment $M_k(t)$, and variance $\sigma_k^2(t)$ of path p_k when the destination is reached at time t, i.e. $t = d_D^k$, we consider the amount of change occurring on

(3.28)-(3.30); this is achieved by increasing the value of $n_{ij}(\tau)$ by one for each road segments included in path p_k, yielding the following expressions:

$$\mu_k(t) = \mu_0 + \frac{1}{N_S} \sum_{(i,j)\in p_k} \sum_{\tau=d_i^k+1}^{d_j^k} \left(\frac{n_{ij}(\tau)+1}{b_{ij}} - \frac{n_{ij}(\tau)}{b_{ij}} \right) = \mu_0 + \frac{1}{N_S} \sum_{(i,j)\in p_k} \frac{c_{i,j}(d_i^k)}{b_{ij}} \tag{3.31}$$

$$M_k(t) = M_0 + \frac{1}{N_S} \sum_{(i,j)\in p_k} \sum_{\tau=d_i^k+1}^{d_j^k} \left(\left(\frac{n_{ij}(\tau)+1}{b_{ij}} \right)^2 - \left(\frac{n_{ij}(\tau)}{b_{ij}} \right)^2 \right) \tag{3.32}$$

$$\sigma_k^2(t) = M_k(t) - (\mu_k(t))^2 \tag{3.33}$$

One important observation is that we can define the mean, second moment and variance of the path p_k up to node v_l^k reached at time t, i.e. $t = d_{v_l^k}^k$, denoted as $\mu_{k,l}(t)$, $M_{k,l}(t)$ and $\sigma_{k,l}(t)$, respectively, based on the associated quantities $\mu_{k,l-1}$, $M_{k,l-1}$ and $\sigma_{k,l-1}^2$ and the incurred increment due to the increase of $n_{ij}(\tau)$. In particular, simple mathematical calculations yield the expressions:

$$\mu_k(v_l^k, t) = \mu_k(v_{l-1}^k, d_{v_{l-1}^k}^k) + \Delta\mu_k(v_l^k, t) = \mu_k(v_{l-1}^k, d_{v_{l-1}^k}^k) + \frac{1}{N_S} \frac{c_{v_{l-1}^k, v_l^k}(d_{v_{l-1}^k}^k)}{b_{v_{l-1}^k, v_l^k}} \tag{3.34}$$

$$\begin{aligned} M_k(v_l^k, t) &= M_k(v_{l-1}^k, d_{v_{l-1}^k}^k) + \Delta M_k(v_l^k, t) \\ &= M_k(v_{l-1}^k, d_{v_{l-1}^k}^k) + \frac{1}{N_S} \sum_{\tau=d_{v_{l-1}^k}^k+1}^{d_{v_l^k}^k} \left(\left(\frac{n_{v_{l-1}^k, v_l^k}(\tau)+1}{b_{v_{l-1}^k, v_l^k}} \right)^2 - \left(\frac{n_{v_{l-1}^k, v_l^k}(\tau)}{b_{v_{l-1}^k, v_l^k}} \right)^2 \right) \end{aligned} \tag{3.35}$$

$$\sigma_k^2(v_l^k, t) = \sigma_k^2(v_{l-1}^k, d_{v_{l-1}^k}^k) + \Delta M_k(v_l^k, t) - 2\mu_k(v_{l-1}^k, d_{v_{l-1}^k}^k)\Delta\mu_k(v_l^k, t) - (\Delta\mu_k(v_l^k, t))^2 \tag{3.36}$$

Based on the above discussion, the TLB problem can be defined as

$$(\Pi_{TLB}) \quad \min_{w\geq 0,\ p_k} \sigma_k^2(D, d_D^k) \tag{3.37a}$$

$$\text{s.t. Constraints (3.1) – (3.3) and (3.7) are satisfied.}$$

$$T_H/a \leq d_D^k \leq T_H. \tag{3.37b}$$

Mathematical formulation (3.37) aims to minimize the spatiotemporal variance of traffic densities in the network provided that the time required to reach the destination is not higher than a percentage $(a-1)100\%$, with respect to the earliest destination arrival time. In this context, other measures can also be considered

such as the weighted sum between mean and variance [117], or some reliability shortest path measure combined with variance, e.g., [118, 119]. Hence, the solution of Mathematical formulation (3.37) requires to optimally solve the EDAT problem (Problem (Π_d)) presented in Section 3.4. In this direction, an optimal solution of the Problem (Π_d)derived in the section that follows.

3.7 Dynamic programming solutions for EDAT and TLB

3.7.1 EDAT problem discrete time optimal solution

To optimally solve the EDAT problem, a directed acyclic graph is build using a space-time network. The space dimension contains indices of the road junctions and the time dimension contains consecutive time slots. Each node replica in the space-time network is identified by the index of the road junction and a specific time slot. Edges on this network represent road segments and the length of each edge reflects the time necessary to travel between adjacent junctions.

To construct a directed graph on this network, edges are assessed based on the reachability of nodes from the origin, and their admissible capacity of edge (i, j), using variables $d_{v_i}(t)$ and $x_{ij}(t)$, respectively. Specifically, variable $d_{v_i}(t)$ determines if node v_i of edge (i, j) is reachable; indicated when $d_{v_i}(t) < \infty$. Thereafter, edge (i, j) is considered admissible when $x_{ij}(t) = 1$. In the process, a topological ordering is imposed for all nodes in the graph. Also, in this directed graph no direct cycles exist, and thus it is easy to indicate reachability from the origin using merely $d_{v_i}(t) < \infty$. In the case when both conditions are satisfied, edge (i, j) is added on the graph. Specifically, if junction v_i is reachable from the originating junction O and $x_{ij}(t) = 1$, then a directed edge from v_i at time t to junction v_j at time $t + c_{ij}(t)$ is added to the graph. The whole process repeats until that time when D becomes reachable, i.e. $d_D(t) < \infty$ for any v_i and t (edge (v_i, D)). It should be noted here that since the earliest destination arrival time route is required, the algorithm stops when D becomes reachable and traces back the nodes in the space-time network that resulted to this route. Algorithm 5 depicts the steps of the aforementioned algorithmic procedure to compute the EDAT.

Algorithm 5 EDAT algorithmic solution.

Input: $G(V, E)$, $n_{ij}(t)$, O, D, t_0, $x_{ij}(t)$;
Initialization:
$t = t_0 - 1$;
$d_{v_i}(t) = \infty \; \forall t, v_i \in V$;
$d_D^* = \infty$;
$d_O(t) = 0, \forall t$;
Algorithm Execution:
while $t < d_D^*$ **do**
    for $(i, j) \in E$ **do**
        $t = t + 1$;
        if $(((v_i == D)$ OR $(v_j == D))$ AND $(d_D(t) < d_D^*))$ **then**
            $d_D^* = d_D(t)$;
        else
            if $((x_{ij}(t) == 1)$ AND $(d_{v_i}(t) < \infty))$ **then**
                $d_{v_j}(t) = d_{v_i}(t) + (t + c_{ij}(t))$;
                $previous[v_j][(t + c_{ij}(t))] = v_i$;
            end if
        end if
    end for
end while
Trace back the optimal path p^* starting from $previous[D][d_D^*]$;
Output: p^* and d_D^*;

Example 1

To better understand the EDAT algorithmic procedure consider the example illustrated in Fig. 3.36 where edge lengths reflect the traversal times for specific road segments. In this example, $t_0 = 0$ and the admissibility along different road segments is given as follows: $x_{OC}(2) = x_{BE}(1) = x_{BE}(2) = x_{CD}(1) = x_{CD}(2) = 0$.

Fig. 3.37 shows the graph constructed by the EDAT algorithm. The space dimension of each node indicates the junction index while the time dimension indicates the replica of the junction created over time. As before, the two variables assess the

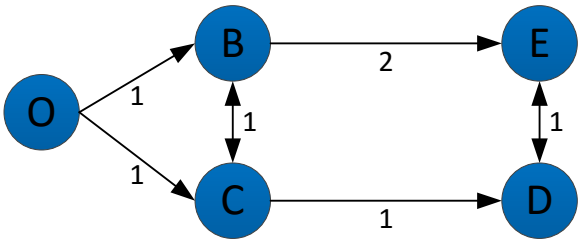


Figure 3.36: Example Network $G(V, E)$

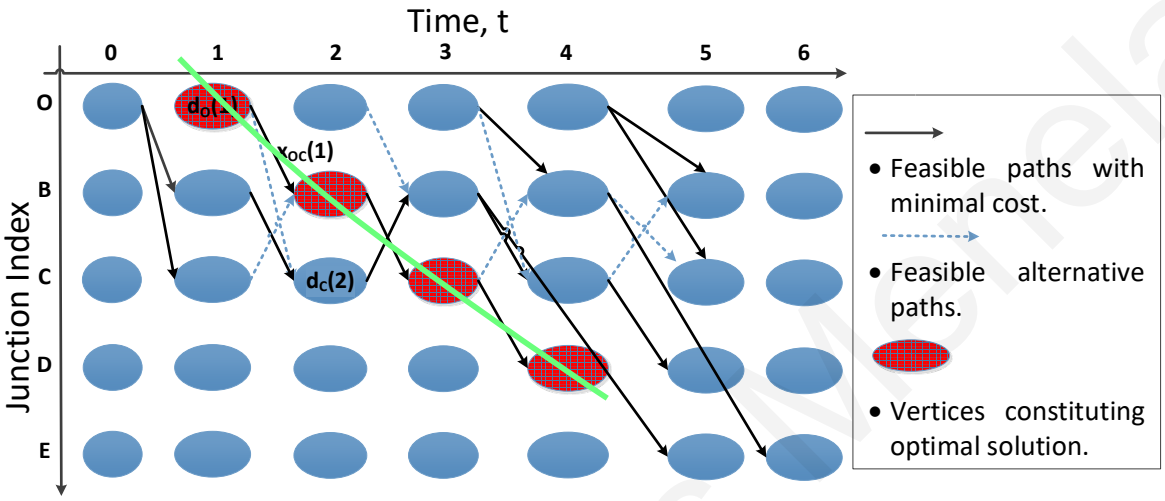


Figure 3.37: EDAT Algorithmic Solution.

reachability of nodes and the admissibility of edges in the graph. As illustrated in the figure, the first column contains only edges emerging from the origin O since all other junctions until time $t = 0$ are not reachable. Similarly, the second column contains edges emerging from junctions O, B, and C since at $t = 1$ these nodes have been reached from origin O.

The dashed-line edges represent road segments that can alternatively be selected without affecting the cost of the solution. The algorithm selects the road segments that were identified first and discard any subsequent arrivals to the specific node. Following the same procedure at the fifth column, the destination is reached and the algorithm terminates. Grid-shaded nodes shown in the graph depict those nodes selected in the route ($O \rightarrow B \rightarrow C \rightarrow D$ and $w = 1$, also denoted with the solid green line).

The devised algorithm results to an optimal solution and executes in pseudo-polynomial time in the discrete time case, due to NP-completeness of the EDAT problem as well since the state space is not known until the algorithm converges

with complexity $O(d_D^* N_E)$. To see why an optimal solution is found, notice that the minimum cost of state (v_i, t) is equal to t if the state is reachable and ∞ otherwise. Hence, a path achieving reachability of state (v_i, t) provides the minimum cost to that state and ensures reachability of all predecessor states forming the path. Therefore, if state $(D, \bar{t})$ is reachable though path p, then all states forming p are also reachable (with minimum cost) and the *optimal substructure property* applies [44]. The reachability of all states is examined for increasing t and thus the optimal solution is found at time d_D^* which is the earliest time at which D is reachable. Next a solution to TLB problem is presented.

3.7.2 TLB algorithmic solution

Solving the TLB problem to optimality is a challenging task because variance does not adhere to the optimal substructure property, i.e. an optimal solution *cannot* be constructed efficiently from optimal solutions of its subproblems [44]. To see why this is true, consider a road network consisting of two paths arriving at junction v_i and a road segment directly connecting junction v_i to the destination D. Let the mean and variance of the two paths p_1 and p_2 up to junction v_i be given by $\mu_1(v_i, t)$, $\sigma_1^2(v_i, t)$ and $\mu_2(v_i, t)$, $\sigma_2^2(v_i, t)$, respectively. Additionally, let $\sigma_1^2(v_i, t) < \sigma_2^2(v_i, t)$ such that the optimal path to v_i is p_1. If the optimal substructure property holds then the minimum variance at the destination must utilize path p_1. However, using Eq. (3.36), it can be easily shown that the optimal path to the destination is through p_2 when $\mu_2(v_i, t) > \mu_1(v_i, t) + (\sigma_2^2(v_i, t) - \sigma_1^2(v_i, t))/(2\Delta\mu(v_i, t))$, where $\Delta\mu(v_i, t) = c_{i,D}(t)/(N_S b_{v_i,D})$ which confirms that the optimal substructure property does not hold for TLB problem.

Based on the above discussion, a dynamic programming algorithm similar to Algorithm 5 cannot be developed to optimally solve TLB. One approach to address this issue is to consider dynamic programming with an additional dimension in the state of the time-expanded graph associated with the origin waiting time so that $\mu_k(v_i, t)$ is constant at one particular state; however, this significantly increases the complexity of the problem which is not desirable. An alternative approach is to approximate the variance with a new metric that has good load balancing performance and satisfies the optimal substructure property. Towards this direction we consider the second moment of the network densities as the cost metric. This

metric provides a good approximation of the variance when the length of the paths reaching the destination are of approximately equal length.

The solution of the TLB problem is outlined in Algorithm 6. Similar to Algorithm 5, a directed acyclic time-expanded graph is constructed with states (v_i, t) indicating that junction v_i is reached at time t and minimum cost of reaching the state equal to $M(v_i, t) = \min_{p_k} M_k(v_i, t)$. The initialization of the algorithm is similar to Algorithm 5, while the main body consists of two blocks. The first examines if the destination has been reached with a better cost than M^*_D, in which case the destination cost is updated; parameter d_D maintains the time-slot with the best cost to the destination. The second block computes the cost of reaching junction v_j through v_i based on Eq. (3.33) and updates it when it is better than the current cost, if v_i is reachable from the origin O and (i, j) is admissible. To backtrack the best path to the destination, a predecessor list is maintained through matrix *previous*, where expression $previous[v_j][t + \bar{\tau}_{ij}] = v_i$ indicates that state $(v_j, t + \bar{\tau}_{ij})$ is reached through state (v_i, t). The complexity of TLB algorithm is equal to $O(T_H \sum_{(i,j)\in E} \bar{\tau}_{ij})$ due to the iteration over time and the summation that appears in the computation of M_{temp}. Note that other research works have also used space-time graphs to solve vehicle routing problems e.g. [120].

Example 2

To illustrate the execution of Algorithm 6, let us revisit Example 1 aiming to solve the TLB problem with $a = 1.25$. In this case, the constructed DAG corresponding to the Algorithm 6 is shown in Fig. 3.38. The dashed green lines indicate edges from candidate paths that have not produced minimal cost, rather than alternative solutions. Comparing Algorithms 5 and 6, the TLB-based algorithm examines solutions up to $T_H = 5 (= 4 \times 1.25)$ rather than $d^*_D = 4$. In addition, the optimal solution provided by Algorithm 6 (illustrated by the red nodes) involves a different path and waiting time compared to the EDAT solution ($O \rightarrow C \rightarrow D$ and $w = 3$ versus $O \rightarrow B \rightarrow C \rightarrow D$ and $w = 1$, also denoted with the solid green line).

Algorithm 6 TLB Algorithmic Solution.

Input: $G(V,E)$, $n_{ij}(t)$, O, D, t_0, $x_{ij}(t)$, T_H, b_{ij}, N_S, M_0;
Initialization:
$t = t_0 - 1$;
$M(v_i, t) = \infty$, $\forall t, v_i \in V$;
$M(O, t) = M_0$, $\forall t$;
$M_D^* = \infty$;
Algorithm Execution:
for $t = t_0, t_0 + 1, \ldots, T_H$ **do**
    for $(i, j) \in E$ **do**
        $t = t + 1$;
        if ((v_i == D) OR (v_j == D) AND ($M(D, t) < M_D^*$) **then**
            $d_D = t$;
            $M_D^* = M(D, t)$;
        end if
        if (($x_{ij}(t)$ == 1) AND ($M(v_i, t) < \infty$)) **then**
            $M_{temp}(v_j, t + \bar{\tau}_{ij}) = M(v_i, t) + \frac{1}{N_S} \sum_t^{t+\bar{\tau}_{ij}} \left(\left(\frac{n_{ij}(t)+1}{b_{v_i v_j}} \right)^2 - \left(\frac{n_{ij}(t)}{b_{v_i v_j}} \right)^2 \right)$;
            if ($M(v_j, t) > M_{temp}(v_j, t)$) **then**
                $M(v_j, t) = M_{temp}(v_j, t)$;
                $previous[v_j][t + \bar{\tau}_{ij}] = v_i$
            end if
        end if
    end for
end for
Trace back the optimal path p^* starting from $previous[D][d_D]$;
Output: p^* and M_D^*;

3.7.3 Performance evaluation

Setup

The road network under consideration is identical with the network used in Section 3.5.2. The network is imported into the SUMO micro-simulator, where vehicle

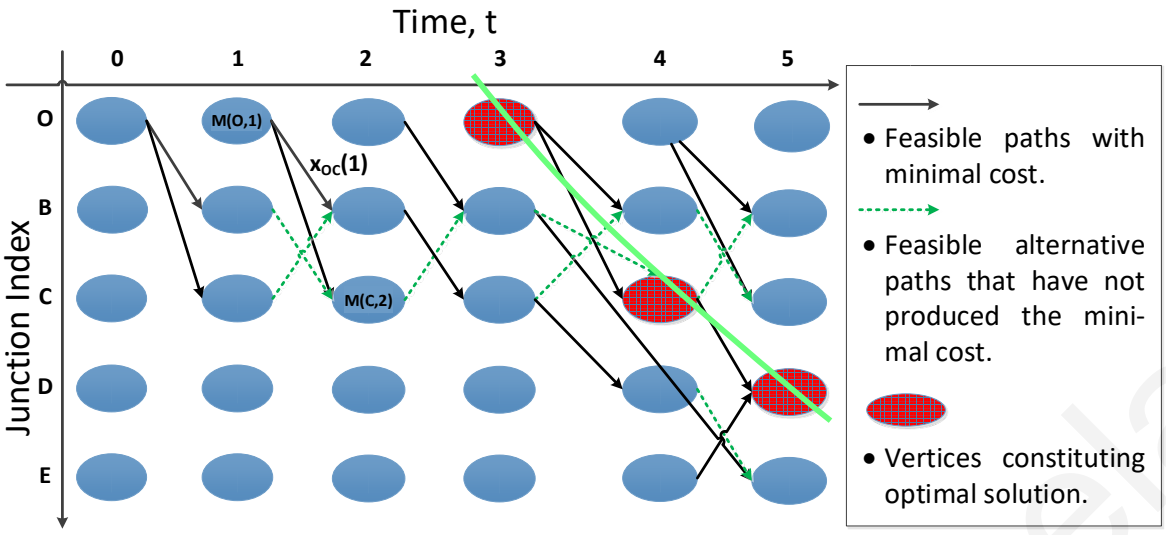


Figure 3.38: TLB Algorithmic Solution.

mobility and car-following model parameters are set as before in Section 3.5.2. Monte Carlo simulations were constructed (10 realizations) whereas both algorithms (EDAT and TLB) were compared against the uncontrolled scenario (US) experienced within the network when no control mechanism is applied and each vehicle travels from the origin to the destination along the shortest (in terms of distance) path. The simulation time-step in SUMO was set to 0.1 s, while the time step of the algorithm is set equal to $T = 1$ s. Finally, as presented in the simulation in Section 3.5.2 vehicles follow strictly their reservation routes, but not their reservation times.

MFD analysis

By injecting flow into the region, identical MFD as depicted in Fig. 3.8 (a) can be derived and accordingly the selected parameters for both algorithms are: $u_f = 47$ km/h, $u_c = 40.5$ km/h, $\rho^J = 1050$ veh and $\rho_{ij}^C = 40$ veh/km/lane (i.e., around 40% of the region's total density). Figs. 3.39 (a) and (b) depict the resulting MFD when the TLB and EDAT algorithms are employed respectively, demonstrating that congestion is alleviated from both algorithms. Additionally, the total volume of flow, density and speed are illustrated in Figs. 3.40, 3.41 and 3.42 as a function of the simulation time, respectively. Comparing these three figures it is demonstrated that as both algorithm operate the density decreases (near 330 veh for TLB and near 410 veh for EDAT compared to more than 700 veh for US as shown in Fig. 3.40). Thus, the

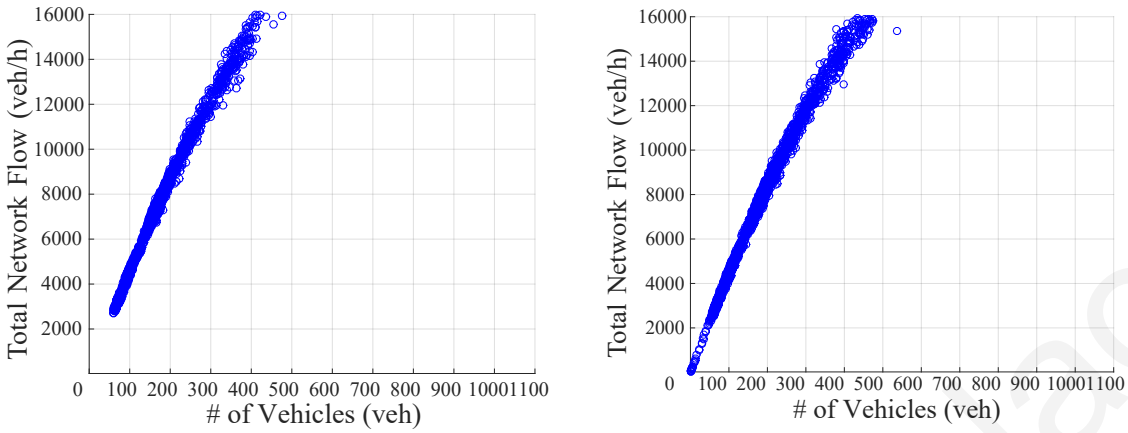


Figure 3.39: Region MFD using algorithm: (a) TLB; (b) EDAT.

link densities are always maintain below the critical capacity (near 350 veh) for both algorithms even when traffic demand is high (e.g., for simulation time 200-240 min). Observing Fig. 3.41 the traveling speed manages to remain high for both algorithms and thus the flow is similar to that of US Fig. 3.42. Indicatively, Fig. 3.41 depicts that in high demand periods TLB outperforms EDAT since the travelling speed is always maintained near u_f.

Results

For the results presented hereafter, ten Monte Carlo simulations were executed with random $O-D$ pairs and for flow rates varying between $1000-8000$ veh/h over a duration of two hours. Figs. 3.44 and 3.43(b) show the average number of vehicles that reach their destination and the average vehicle travel time as a function of the different flow rates, respectively. Specifically, the dashed lines in Fig. 3.44 represent the average number of vehicles that have finished their journey within the simulation time and the scattered plots are the realizations obtained by each simulation run. Similarly, the dashed lines in Fig. 3.43 (b) represent the value of the average travel time for the different realizations and the scattered plots represent the average travel time across all Monte Carlo simulations.

As illustrated in Figs. 3.44 and 3.43(b), at low flow rates EDAT and US behave similarly while TLB slightly lags in terms of the average number of vehicles that reach

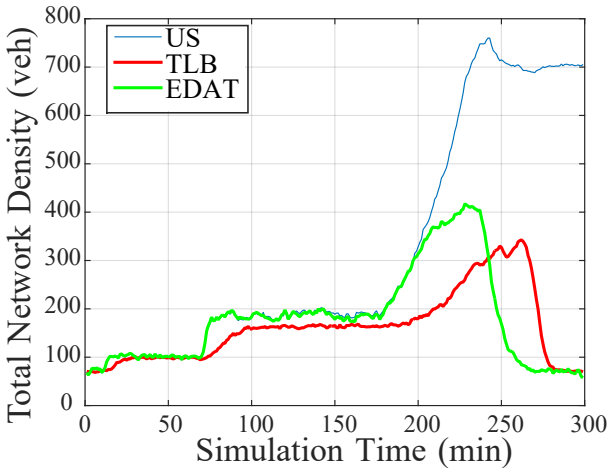


Figure 3.40: Total network density over time.

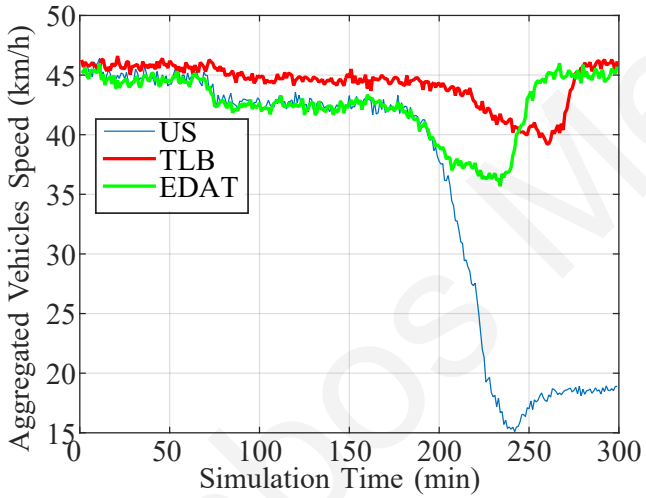


Figure 3.41: Aggregated network speed over time.

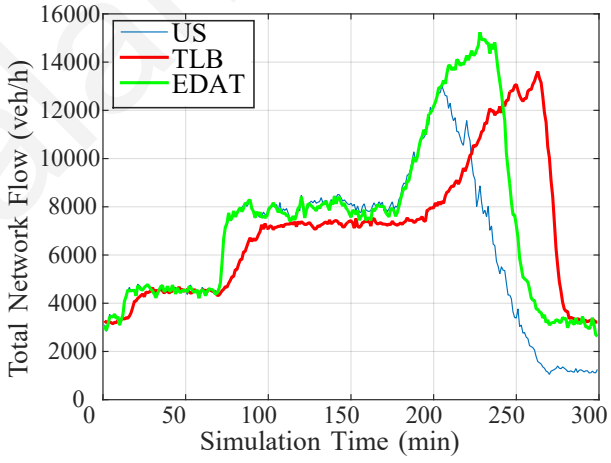
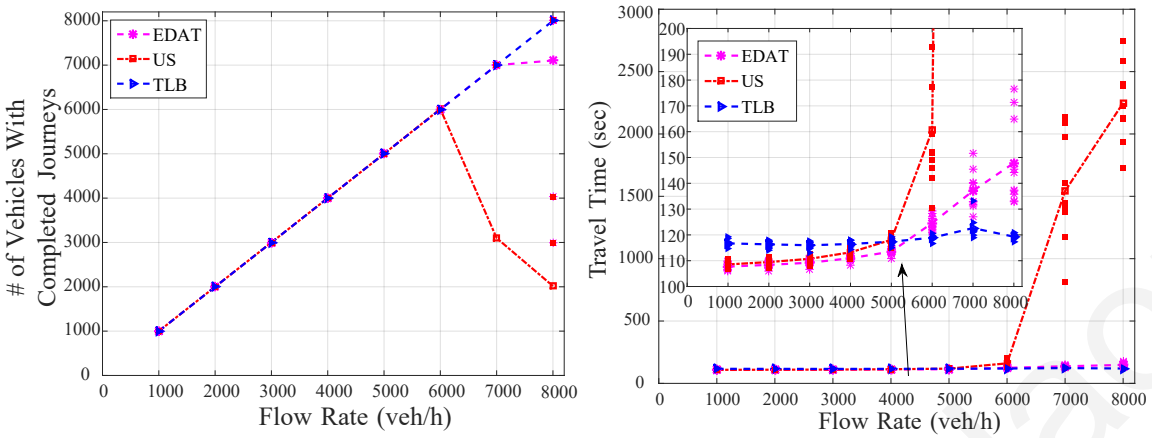


Figure 3.42: Total network flow over time.

their destination within the simulation time. On the other hand, TLB outperforms EDAT and US in high flow rates, where a larger number of vehicles reach their



(a) (b)

Figure 3.43: (a) Number of vehicles towards to the route end; (b) Average travel time ($t \rightarrow s$).

destination and results in a more robust and shorter travel times. In either case, both algorithms perform much better than US. Comparing EDAT against TLB it is clear that better network utilization can be achieved in the case where EDAT is applied at low flow rates while TLB employed at higher flow rates.

Figs. 3.44 illustrates the travel time distribution for all vehicles that managed to reach their destination during the simulation time when a flow rate of 8000 veh/hour is used. As clearly shown, EDAT and TLB greatly outperform US. In numbers, the mean travel time for EDAT is 139.7 s, for TLB is 118.9 s and for US is 2160.8 s. The standard deviation for EDAT is 86.65, for TLB is 58.2 and for US is 2762 demonstrating that as congestion of the road segments increases, TLB is more stable and accurate than the other two solutions. Additionally as shown in the figure, TLB is highly resilient to the increase in flow rate since travel times do not significantly deviate. Finally Fig. 3.44 clearly indicates that TLB achieves better travel times while eliminating spillbacks.

3.8 Summary

This chapter proposes a new route-reservation architecture which aims to prevent congestion by restricting the traffic density in different road segments within a homogeneous region. The key advantage of this architecture is that it considers both the spatial and temporal density of regions, and it exploits more reliable future traffic

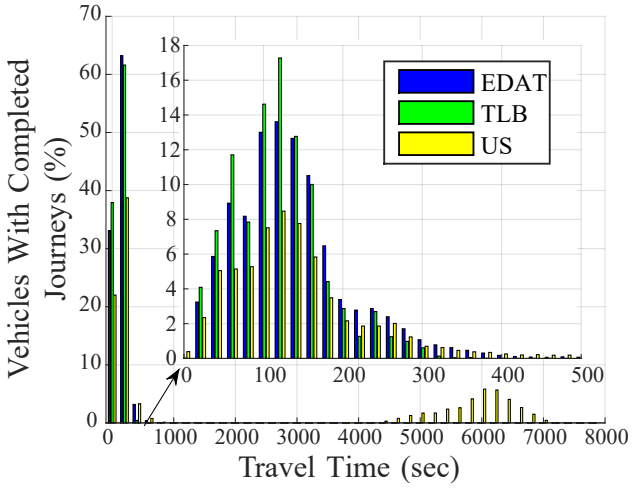


Figure 3.44: Travel time distribution (8000veh/h).

estimates through reservations. For this scheme, the earliest destination arrival time problem and the traffic load balancing problem are examined which have a complementary nature as the earliest destination arrival time algorithm provides better results for low congestion conditions and the traffic load balancing algorithm for high congestion conditions. Both problems are solved by developing several algorithms that can provide different solution qualities and execution times. Simulation results demonstrate the superiority of the route reservation architecture compared to the traditional traffic behavior (where no reservations are made), achieving substantial improvements in terms of road utilization and travel times, especially during high demand.

Chapter 4

Improved Route Reservations with Travel Time Prediction

4.1 Introduction

According to the proposed route-reservation architecture, the future state of each road segment is estimated based on the received reservations and assuming that all vehicles travel at a constant speed which is set equal with the free-flow speed (or the speed at capacity). Despite the fact route reservation architecture can achieve free-flow conditions for the entire road network, speed variations may be observed due to various factors such as the stochastic nature of human driving which leads to speed fluctuations between interacting vehicles, and the crossing of intersections with same or different road segment priorities. Hence even within a homogeneous region, the assumption of traveling in constant speed is not always valid which can lead to wrong predictions of the time that a vehicle occupies a road segment which in turn affects the accuracy of the overall reservation architecture. Therefore, the key objective of this chapter is to investigate a better method for estimating the time that vehicles take to traverse each road segment.

The remainder of the chapter is organized as follows. Section 4.2 extends the proposed route reservation architecture presented in Chapter 3 by enabling it to predict the transit-time of each road segment while Section 4.3 mathematically describes the continuous-time EDAT problem that takes into consideration these transit-time predictions. Section 4.4 proposes the Time-Varying Multiple Linear Re-

gression (TVMLR) method, which makes the transit-time predictions. The required modification of the considered continuous-time route reservation algorithm (RRAC) is presented in Section 4.5 with simulation results in Section 4.6 indicating the benefits that can be gained through the proposed prediction methods. Finally Section 4.7 summarizes this chapter.

4.2 Route Reservation Architecture extension

Within the context of connected vehicles, previous Chapter 3 proposes a route reservation architecture that routes vehicles through non-congested road-segments. In the proposed architecture, a road side unit (RSU) is responsible for computing vehicles routes while it also ensures that each vehicle will arrive at its destination at the earliest possible time. At the same time, the RSU also reserves each road segment at the time that the vehicle is expected to traverse them. Each road segment transit-times are calculated assuming constant vehicles speed (typically the free-flow speed). Nonetheless, considering constant speed is not always a valid assumption in practice.

In this chapter, an extension of the route reservation architecture is proposed to address the above shortcoming. To do so, instead of considering constant transit-times, we investigate the use of predicted transit-times by introducing feedback to the RSU capabilities. In this context, when a vehicle exits a road segment, it sends to the RSU the time it took to traverse this particular road segment. The RSU utilizes this information to make near-future predictions of the transit-times of all road segments without requiring to know the average vehicle's speeds in the network or any other information about the network's state. Therefore, as time goes by, it is expected that the road segment transit estimates will be closer to the actual values and thus the future state predictions will be more accurate, and therefore the reservations will be more effective. Recent advances in connected and autonomous vehicle technologies support the development of such a scheme.

According to the proposed extension, the transit-times of each road segment can be interpreted as a time-series where samples are affected by the segment's instantaneous density. As the density of the segment increases the speed decreases

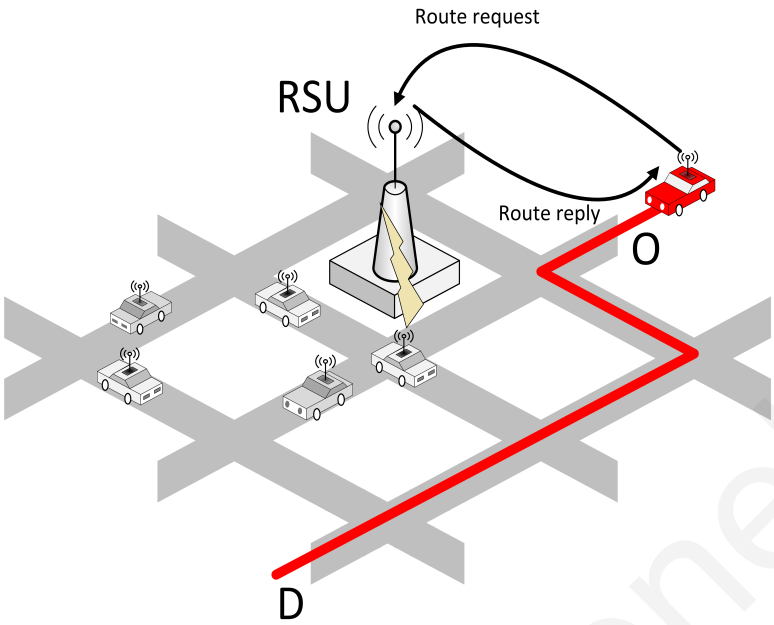


Figure 4.1: Extended of the Route Reservations Architecture. When a vehicle traverse a road segment sends to the RSU the time it took to traverse this particular road segment.

and hence, the transit-time increases, [121] exhibiting seasonal changes (depending on density). Considering such behavior, a simple prediction method is proposed to address the accuracy problem under route reservation architecture. The proposed method based on the Time-Varying Multiple Linear Regression (TVMLR) technique that uses several explanatory variables to predict the outcome of a response variable as the segment's transit-time depends on more than one factor. In this chapter, one factor is the segment's current density while the second factor is the density of all adjacent segments that share the same exiting junction [122]. Hence, there is seasonality on the transit-time response, meaning that there is a pattern of transit-times that is related to these two dependent variables (factors).

4.3 Continuous time formulation modification

Considering the continuous time mathematical formulation presented in Section 3.3.2, transit-times predictions can be included by replacing the constant transit-time assumption (i.e., $\bar{\tau}_{ij} = l_{ij}/u_f$) with the variable $\hat{\tau}_{ij}(t)$ that denotes the travel time prediction for each road segment (i, j), as predicted at time t. In this way, the admissibility intervals of each road segment are based on the assigned vehicle path,

the past reservations status and the predicted travel times for each road segment (i.e., $\hat{\tau}_{ij}(t)$). Hence, the cost of traversing each road segment (i.e., $c_{ij}(t)$ is replaced by the variable $c_{ij}(d_{b_{v_i}}, t_0)$, mathematically defined as follows:

$$c_{ij}(d_{v_i}, t_0) = \begin{cases} \hat{\tau}_{ij}(t_0), \text{ if } d_i + t \in \mathcal{S}_{ij}(t_0), \forall\ 0 \leq t \leq \hat{\tau}_{ij}(t_0),\ i \neq O \\ \hat{\tau}_{ij}(t_0) + w, \text{ if } t_0 + w + t \in \mathcal{S}_{ij}(t_0), \\ \qquad\qquad \forall\ 0 \leq t \leq \hat{\tau}_{ij},\ i = O \\ \infty, \text{ otherwise,} \end{cases} \tag{4.1}$$

where d_{v_i} denotes the arrival time at node v_i, t_0 the vehicle's request time, O the origin node, and $w \in \mathbb{R}$ the *origin waiting time*. Note that, in this case, the cost of traversing each road segment is time varying and calculate at the reservation request time. Furthermore, let the variables $M(t)$ and $N(t)$ denote the actual and estimated (e.g., through reservations) instantaneous network density. Hence, the variable $N(t)$ defines the network's accumulated number of reservations (*i.e.*, $N(t) = \sum_{\forall (i,j) \in \mathcal{E}} n_{ij}(t)$) while similarly, $M(t)$ mathematically is defined as ($M(t) = \sum_{\forall (i,j) \in \mathcal{E}} \rho_{ij}(t) * l_{ij}$), at t.

Earliest destination arrival Time (EDAT) problem considering travel time predictions

Given the vehicle routing request time (t_0), the travel time prediction for each segment until t_0 (i.e., $\hat{\tau}_{ij}(t_0)$), the origin-destination ($O - D$) pair and the admissible sets $\mathcal{S}_{ij}(t_0)$, $\forall\ (i,j) \in \mathcal{E}$, the EDAT problem requests the earliest-arrival-time-at-destination (from O to D) subject to avoiding links that are at their critical density. Let p_h denote the h-th path from source O to destination D denoted as $p_h = (v_0^h, v_1^h),\ (v_1^h, v_2^h),\ (v_2^h, v_3^h), \cdots (v_{L_h-1}^h, v_{L_h}^h)$, where $v_j^h \in \mathcal{V}$ is the j-th visited node in the h-th path, with $v_0^h = O$ and $v_{L_k}^h = D$. Additionally, as before let d_j^h denote the arrival time at junction v_j then, the arrival time to each node of the path can be expressed as:

$$\begin{aligned} d^h_{v^h_0} &= t_0, \\ d^h_{v^h_1} &= d^h_{v^h_0} + c_{v^h_0, v^h_1}(d^h_{v^h_0}, t_0) \\ &\vdots \\ d^h_{v^h_{L_h}} &= d^h_{v^h_{L_h-1}} + c_{v^h_{L_h-1}, v^h_{L_h}}(d^h_{v^h_{L_h}}, t_0) \end{aligned} \tag{4.2}$$

with, $w \geq 0$ see Equation (4.1). Accordingly, the EATD problem can be expressed in compact form as:

$$(\Pi_{c_{tc}})\ d_D^* = \min_{w \geq 0,\ p_h} d_D^h \tag{4.3}$$

$$\text{s.t. Constraints (3.4)} - \text{(3.5) and (4.1)} - \text{(4.2) are satisfied.}$$

In the next section we develop a method for predicting the travel-time of road segments.

4.4 Time-Varying Multiple Linear Regression (TVMLR) method

This section describes the proposed Time-Varying Multiple Linear Regression predictor, that is utilized to predict the current transit-time of each road section for improving the accuracy of the reservation scheme. Both predictors are based on statistical techniques where the k-th observation (*i.e.*, $\hat{\tau}_{ij}^k(t_0)$)) is interpreted as a time-series that utilizes the most recent transit-time observations in order to forecast the near future transit-times.

Linear regression is a statistical technique that considers historical observations to forecast near-future states. In this chapter, we employ a simple prediction method based on the Multiple Linear Regression technique to predict the transit-times of road segment using the most recent transit-time observations in each road segment. Evidently, as the density of a particular road segment increases, the speed of vehicles traversing the road segment decreases [121] which further results in increased travel times. Nonetheless, other factors also affect travel times such as the network traffic state, road junction priorities, the road geometry and the weather conditions. In this chapter, in addition to the density of the traversed road segment we explicitly consider the densities of neighboring segments (i.e., the road segments forming the down-link intersection). Other factors affecting travel times are considered implicitly by adopting a time-varying prediction approach that employs the most "recent" collected observations to capture the short-term traffic dynamics achieving more accurate predictions.

The Time-Varying Multiple Linear Regression (TVMLR) prediction method uses the H most recent measurements of the *response variable*, i.e., the transit-time of the road segment of interest, and the *predictor variables*, i.e., the observed densities of the road segment of interest and its neighbors, to construct the best linear relationship between the two variable sets using the vector of *regression coefficients*.

To simplify notation, the TVMLR prediction method is described for a single road segment, say (i, j). Let (i, j), $i \in \mathcal{P}_j$ denote the segments whose down-link intersection is j, and $|\mathcal{P}_j|$ the number of such links. Let also y_k and $\hat{y}_k$ denote the observed and predicted transit-time of the k-th vehicle traversing (p, j), respectively; while y is the response variable of the TVMLR prediction method. Assuming that the k-th measurement is collected at time t_k, the number of vehicles in link (i, j) at time t_k is denoted by the variable $m_{ik} = \rho_{ij}(t_k) * l_{ij}$, $i \in \mathcal{P}_j$, $i = 1, \cdots, |\mathcal{P}_j|$, value that can also derived from the reports obtained from each vehicle. Inasmuch as, each vehicle reports to the RSU whenever it exits a certain road segment, and considering that, vehicles' routes are known from the RSU, an accurate calculation of the number of vehicles in each segment can be maintained by simply increasing/decreasing the density value whenever a vehicle enters/exits a segment, i.e.,

$$m_{lk} \leftarrow \begin{cases} \rho_{ij}(t_k) * l_{ij} + 1, \text{ if vehicle enters } (i, j), \\ \rho_{ij}(t_k) * l_{ij} - 1, \text{ if vehicle exits } (i, j), \\ \rho_{ij}(t_k) * l_{ij}, \qquad \text{otherwise,} \end{cases}$$

assuming that $\rho_{ij}(0) = 0$.

In this way one obtains the measurement vector $\mathbf{y}_H^K = [y_K, \cdots, y_{K-H+1}]^\top$, and density matrix $\mathbf{M}_H^K = [\mathbf{m}_K, \cdots, \mathbf{m}_{K-H+1}]^\top$, where K is the latest vehicle that traversed link (p, j), and $\mathbf{m}_K = [m_{lk}, \cdots, m_{|\mathcal{P}_j|k}]^\top$. Using this information, the TVMLR prediction method builds a linear model of the form

$$\mathbf{y}_H^K = [\mathbf{1}_{K-H}, \mathbf{M}_H^K]\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.4}$$

where, $\mathbf{1}_{K-H}$ is a vector consisted of $K - H$ ones, $\boldsymbol{\beta} = [\beta_0, ..., \beta_{|\mathcal{P}_j|}]^\top$ is the regression vector and $\boldsymbol{\epsilon}$ are the residual terms of the model.

Because (4.4) is an over-determined linear system of equations, the TVMLR prediction method finds the regression vector $\hat{\boldsymbol{\beta}}$ that minimizes the least square residual

error by solving the optimization problem

$$\min_{\hat{\beta}} \sum_{k=K-H+1}^{K} \epsilon_k^2 = (\mathbf{y}_H^K - \mathbf{M}_H^K \boldsymbol{\beta})^T (\mathbf{y}_H^K - \mathbf{M}_H^K \boldsymbol{\beta}) \tag{4.5}$$

Whenever a new route reservation request arrives, the method identifies for each road segment $(i, j) \in \mathcal{E}$ the best regression vector in the least square sense, $\hat{\boldsymbol{\beta}}_{ij}$. Using the resulting regression vectors one can obtain the predicted travel time of road segment (p, j) for the $K + 1$-th vehicle traversing the segment, given by

$$\hat{\tau}_{pj} = \begin{bmatrix} 1 & m_{1K} & \dots & m_{|\mathcal{P}_j|K} \end{bmatrix} \hat{\boldsymbol{\beta}}_{pj}. \tag{4.6}$$

Note that, to estimate future travel times reservation counts can be potentially utilized instead to using the number of current vehicles. Furthermore, the initial values of the predicted road segment travel-times are set equal to those corresponding to free-flow speed, i.e., $(\hat{\tau}_{pj} = l_{pj}/u^f)$.

Next, we present how the low-complexity heuristic algorithm RRAC can be modified to obtain a solution to the problem in real-time.

4.5 EDAT solution considering travel time predictions

EDAT solution can be provided by modifying the aforementioned Route Reservation Algorithm (continuous time, RRAC) as introduced in Section 3.5.3 (Algorithms 3 and 4). RRAC is a heuristic algorithm that solves the EDAT problem over a sequence of iterations by employing a variant of Dijkstra's shortest path algorithm. In the previous version of the algorithm, reservations are made assuming constant speed equal with the free-flow speed. This assumption is removed in the proposed variant as presented in Algorithm 7.

The algorithmic procedure is identical with the procedure presented in Section 3.5.3. In the same manner Algorithm 3 is the outer loop that is responsible to re-iterate until a feasible solution is found. On the contrary the inner loop changes where at the beginning of each iteration (as depicted in line (3) of the Algorithm 7) the regression coefficients have to be derived. Furthermore, in lines (8-11) coefficients are used to estimated the transit-times of each road segment (i.e., $\hat{\tau_{ij}}(t_0)$). Finally, the complexity of algorithm remains the same and thus is equal with $O(LE^2/V)$, with L denoting the number of reiterations of the Relaxed-EDAT problem that is required.

Algorithm 7 Inner loop of the continuous time RRAC with travel time predictions (IL-RRACP).

Input: $G(\mathcal{V}, \mathcal{E})$, $\mathcal{S}_{ij}(t_0)$, O, D, t_0, w^p;
Initialization: $d_{v_i} = \infty$, $v_i \in \mathcal{V}$, $d_O \leftarrow t_0 + w^p$, $\mathcal{Q} \leftarrow \mathcal{V}$,
Calculate regression coefficients (i.e., $\beta_{ij} \; \forall \, (i, j) \in \mathcal{E}$)
$P[v_i] \leftarrow NULL$, $v_i \in \mathcal{V}$;
while $\mathcal{Q} \neq \emptyset$ **do**
    $v_l \leftarrow \text{argmin}_{v_i \in \mathcal{Q}}\{d_{v_i}\}$;
    $\mathcal{Q} \leftarrow \mathcal{Q} - \{v_l\}$;
    for $(l, j) \in \mathcal{E}$ **do**
        $w_{lj} \leftarrow \min_{w \geq 0}\{d_{v_l} + t + w \in \mathcal{S}_{lj}(t_0), 0 \leq t \leq \hat{\tau}_{ij}(t_0)\}$
        $c_{lj} \leftarrow \hat{\tau}_{ij}(t_0) + w_{lj}$;
        if $\{d_{v_j} > d_{v_l} + c_{lj}\}$ **then**
            $d_{v_j} \leftarrow d_{v_l} + c_{lj}$, $P[v_j] = v_l$;
        end if
    end for
end while
$w^T \leftarrow d_D$, $v_v \leftarrow D$;
repeat
    $w^T \leftarrow w^T - c_{P[v_v], v_v}$;
    $v_v \leftarrow P[v_v]$;
until $\{v_v = O\}$
Output: w^T, **P**.

4.6 Performance evaluation

4.6.1 Setup

Performance evaluation is performed in a micro-simulation environment to capture the stochastic nature of traffic mobility within a real road network which results in uncertainty in the MFD and thus predictions of road segment's travel-times are required to increase reservation accuracy. Micro-simulations are performed using SUMO micro-simulation software [112] within which in total 10 Monte Carlo simulations are conducted.

In SUMO the traffic mobility characteristics are determined by the Krauss car-following model [113] with the selected parameters: vehicle length 5 m, maximum speed 15 m/s, acceleration 2.5 m/s^2, deceleration 4.5 m/s^2, driver imperfection 5%, driver reaction time 0.5 s, minimum gap distance 2.5 m, and simulation time-step 0.5 s. As before all vehicles assumed follow RSU's instructions (route, origin waiting time) without any deviation.

The network under consideration is the un-signalized homogeneous region of down-town San Francisco depicted in Fig. 3.7 in which three approaches are examined:

1. **RRAC**: The heuristic approach proposed in Section 4.5 that is combined with the TVMLR prediction method proposed in Section 4.4. The RRAC algorithm without the TVMLR prediction method proposed in Section 3.5.3 (Algorithms 3 and 4). In this case, the RRAC algorithm to calculate each road segment travel time assumes that all vehicles are traverse constantly with the free-flow speed (i.e., u_f).

2. **DOT**: The Decreasing Order of Time (DOT) algorithm [42]. The DOT algorithm seeks to find the time-dependent travel-time path that minimizes the user's arrival time at the destination within a user-specified departure time window.

3. **US**: The uncontrolled scenario where each vehicle travels along its shortest-distance path without any waiting time at the origin.

Note that in case of the RRAC without predictions (as presented in Section 3.5.3) in this simulation results denoted as RRACNP meaning RRACNP with no predictions.

Furthermore, in the case of both RRAC and RRACNP, new route reservations are computed using only information from previous reservations without any consideration of the actual network state.

4.6.2 MFD Analysis

Considering that the proposed algorithm rely on critical density of the region of interest, we first generate the MFD of the considered road network to identify the region's critical density and free-flow speed (ρ_{ij}^{C} and v_f) parameters. To do so, a 6 hours scenario was simulated within which for the first hour the input flow was set to 2000 veh/h and incrementally increased by 2000 veh/h for the next three hours. Thereafter, for the last two hours the input flow was set to 4000 veh/h and 2000 veh/h in order to allow a network discharge. In the loading procedure, both endogenous and exogenous flows are considered meaning that vehicles start and finish within the imported area (i.e., endogenous) either can start and end their journey from/at the region's boundaries or, a combination of the two.

Fig. 4.2 shows the total network flow as a function of the network's density (i.e. the total number of vehicles within the region) for the US, which depicts the MFD diagram of the region. In the figure, each point corresponds to the sliding mean of 5 measurements that are calculated every 15 s. To calibrate the model, the *Van-Aerde* automated calibration method proposed in [114] is employed. The procedure produced the following calibrated parameters v_f = 42.5 km/h, v_c = 27.5 km/h and ρ^J = 675 vehicles, which correspond to the red line depicted in the figure. Having obtained the calibrated model, one can analytically obtain the critical density which in our case is equal to ρ_{ij}^{C} = 33 veh/km/lane (i.e., around 33% of the region's jam density); accordingly, the vehicle speed that will be used for the RRACNP is v_f = 42.5 km/h. Note that, the shape and scatters of Fig. 4.2 differs form Fig. 3.8 (a) since in that case the generate flows are both endogenous and exogenous.

Figs. 4.3 (a) and (b) compare the accuracy of reservations for the RRACNP and RRAC algorithms over the 6-hour scenario. In particular, the figures depict the actual and predicted instantaneous network density over time with red and green lines, $M(t)$ and $N(t)$, respectively, while the blue dotted line represents the instantaneous residual density, $E(t) = |M(t) - N(t)|$. During the first two hours of the considered

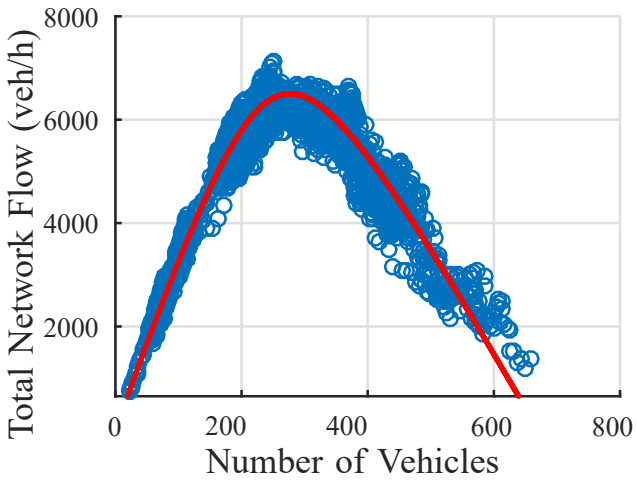


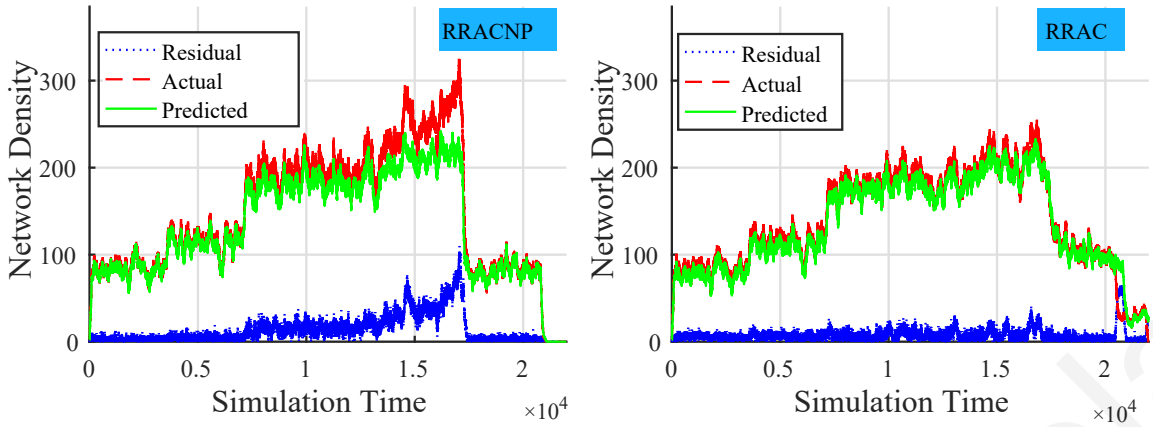
Figure 4.2: Regional MFD for the uncontrolled scenario.

scenario, the predicted density, through reservations, closely follows the actual network density, as the demand flow rate that enters the network is low. As the demand increases over time, the network becomes more congested which results in a significant growth of the residual density when using the RRACNP; on the contrary, the RRAC manages to maintain an excellent prediction of the network density over the entire 6-hour scenario, due to the integration of the TVMLR prediction method.

Figs. 4.4 (a) and (b) depict the maximum instantaneous residual density of individual road segments for the RRACNP and RRAC, respectively. From the figures it is clear that more than one third of the network's road segments (around 70 segments) exhibit high maximum instantaneous residual density for the RRACNP (larger than the 10 vehicles). On the contrary, when the RRAC approach is employed only six of the segments exhibits high maximum instantaneous residual density at any simulation step. Even for these six segments the value of the residual density is significantly smaller compared to the road segments with high residual density in the RRACNP approach.

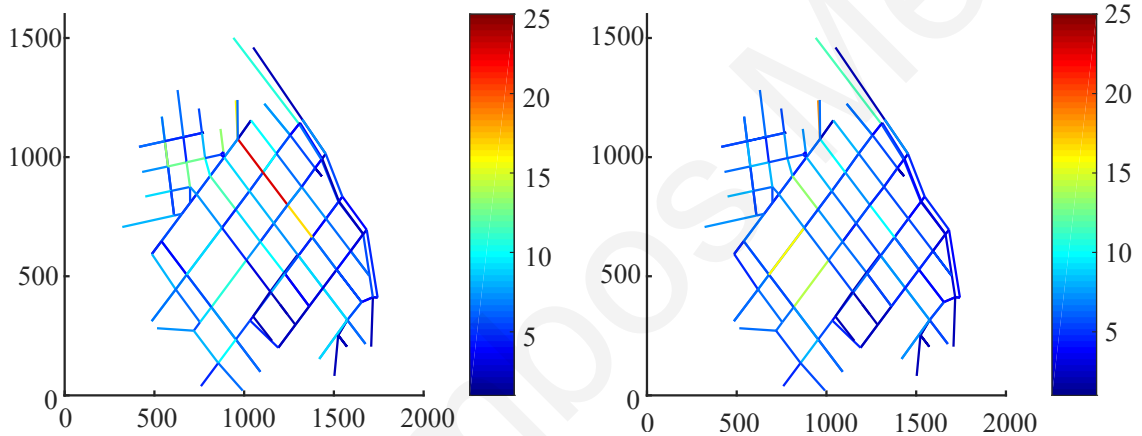
4.6.3 Results

To further evaluate the efficiency of the TVMLR prediction method in terms of network operation, the resulting MFD diagram of the RRACNP and RRAC methods are demonstrated in Figs. 4.5 (a) and (b), respectively. The figures illustrate acceptable performance, as both methods maintain the network operation within the non-



(a) (b)

Figure 4.3: Instantaneous network density over time for the (a) RRACNP and (b) RRAC.



(a) (b)

Figure 4.4: Maximum instantaneous residual density of individual road segments for the (a) RRACNP and (b) RRAC.

congested region. Comparing both approaches, the RRACNP is slightly better than the RRAC method as it can achieve 2% higher flow rate (7100 veh/h compared to 6980 veh/h). This behavior is anticipated since the RRACNP approach makes reservations that under-estimate the transit time of each road segment allowing a little bit higher rate to pass through the network.

The four aforementioned methods are evaluated for various scenarios and varying demand flow rates (3000 – 8000 veh/h) over a 2 hours simulation period where the demand is constant for the first hour and equal to zero for the second hour. For a fair comparison, the origin waiting time for all approaches is not considered in the

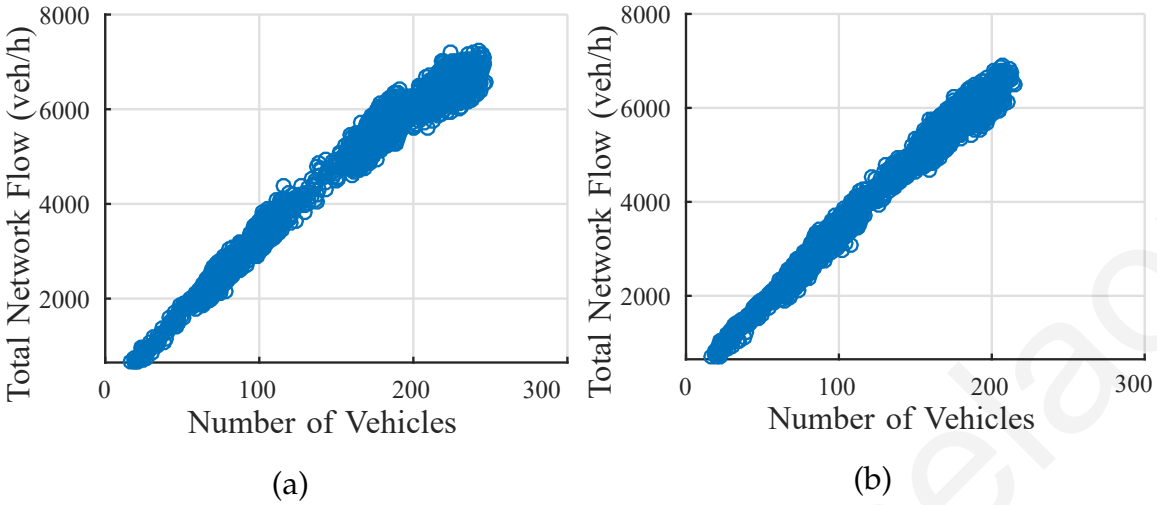


Figure 4.5: Regional MFD for the (a) RRACNP and (b) RRAC.

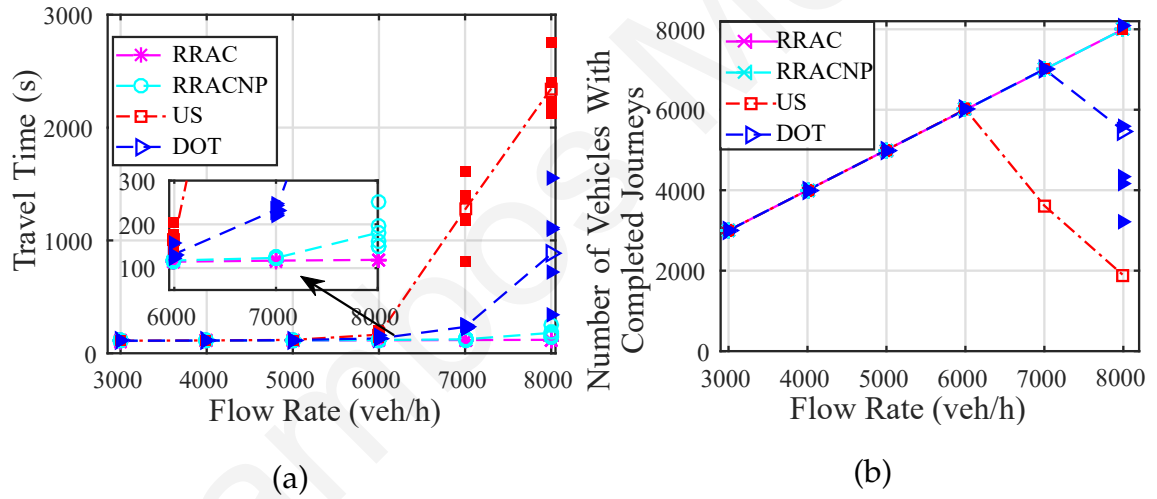


Figure 4.6: (a) Travel time and (b) Number of vehicles with completed journeys for different simulation scenarios with varying demand flow rate.

travel time and for DOT the maximum allowed origin waiting time is set to 1 minute (i.e., half the average trip length for the considered network).

Figs 4.6 (a) and (b) depict the average travel time of all vehicles and the total number of vehicles that have managed to finish their journey within the simulation time, respectively. In the figures dashed lines indicate the average values over the five simulations run, while the markers indicate each individual simulation run value. From the figures it is clear that for low demand flow rates (under 6000 veh/h)

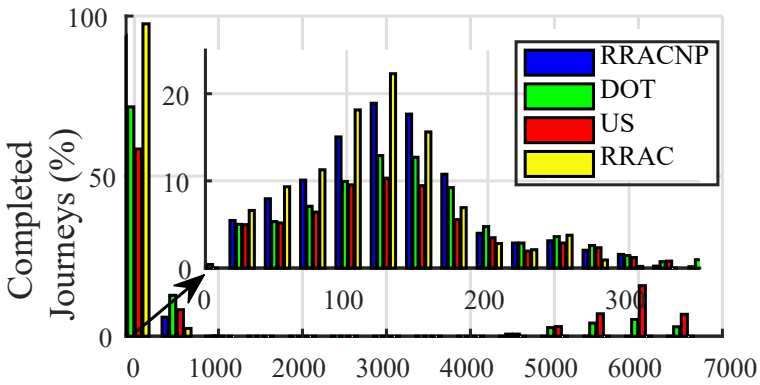


Figure 4.7: Per vehicle travel time distribution for the highest demand flow rate (i.e. 8000 veh/h).

all approaches have similar behavior with the US because no congestion occurs in the network. Nonetheless, under congested conditions (demand flow rates larger or equal to 6000 veh/h) the behaviour of the different algorithms varies significantly. With regards to travel time, Fig. 4.6 (a) indicates that both the RRACNP and RRAC significantly outperform both DOT and US approaches. As congestion intensifies the travel time of both DOT and US grows exponentially, while the travel time of the RRACNP increases only for the highest demand scenario. Interestingly, the RRAC exhibits constant travel time in all scenarios considered which indicates that it is robust to different demand levels. With respect to the total number of vehicles that have managed to finish their journey within the simulation time, Fig. 4.6 (b) demonstrates that for all demands the RRACNP and RRAC manage to complete all the assigned trips within the simulation time, while in the case of US and DOT, vehicles experience major delays due to congestion and as a result a large percentage of vehicles do not complete their journeys.

This observation is also supported from the distribution of the vehicle travel times for the highest demand flow rate (8000 veh/h) shown in Fig. 4.7. As can be seen, the RRAC manages to complete more vehicles than all the other algorithms for all bins that correspond to travel times smaller than 125 s.

Figs 4.8 (a) and (b) compare the accuracy of reservations for the RRACNP and RRAC algorithms over the 2 hours scenario. The figure supports the observations made in Figs. 4.3, as the instantaneous residual network density over time grows significantly for the RRACNP, while for the RRAC the instantaneous residual network

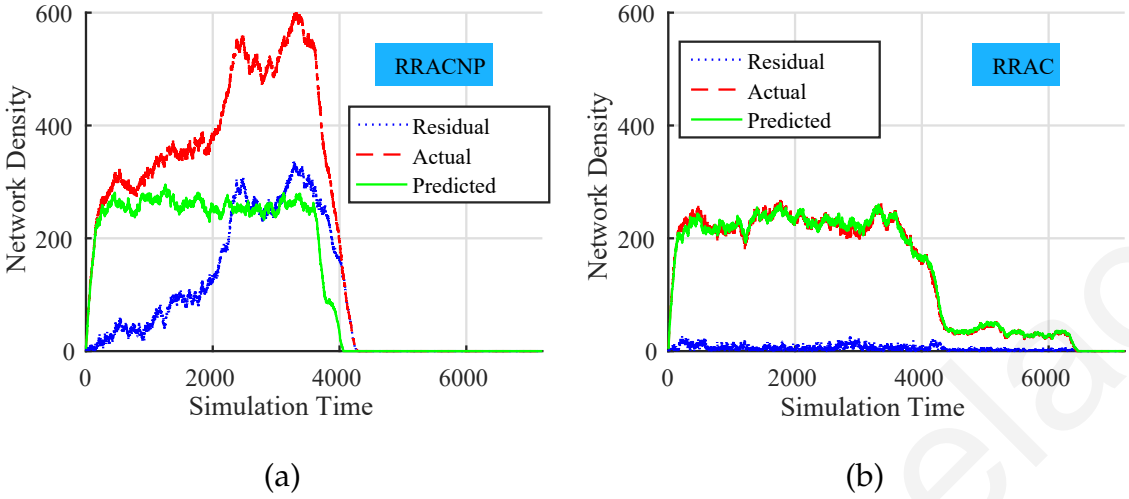


Figure 4.8: Performance evaluation of the RRACNP (a) and RRAC (b), with respect to the instantaneous network density over time, for the highest demand flow rate (i.e., 8000 veh/h).

density is very close to zero. The superior reservation accuracy of the RRAC over RRACNP is also shown from Figs. 4.9 (a) and (c), as well as Figs 4.9 (b) and (d) which depict the instantaneous residual density of individual links and the maximum residual density of individual links over time, respectively. More specifically, the average residual density of the RRAC is approximately three times smaller compared to the RRACNP (8 vehicles compared 25 vehicles per road segment, respectively). Figs 4.9 (b) and (d) further indicate that more than one third of network's segments exhibit high maximum residual density (more than 20 vehicles) when using the RRACNP, while for the RRAC only two road segment exhibit this behaviour.

Figs 4.10 (a) and (b) illustrate the origin waiting time assigned to vehicles before commencing their trip for the RRACNP and RRAC solutions in the form of a boxplot. For both cases, as demand grows, the origin waiting time increases since more vehicles request to enter the network, with the RRAC responding accordingly at higher input rates as opposed to the RRACNP solution. More specifically, the origin waiting time assigned by the RRAC is about 4 times higher than the RRACNP approach. Nonetheless, the average origin waiting is within acceptable levels (14 min for the highest demand approach), while origin waiting times observed at the origin do not impact travel times. This behavior is expected because the RRACNP allows more vehicles to enter the network compared to the critical capacity (see Fig. 4.8 (a)) so that the origin waiting time of the vehicles is smaller.

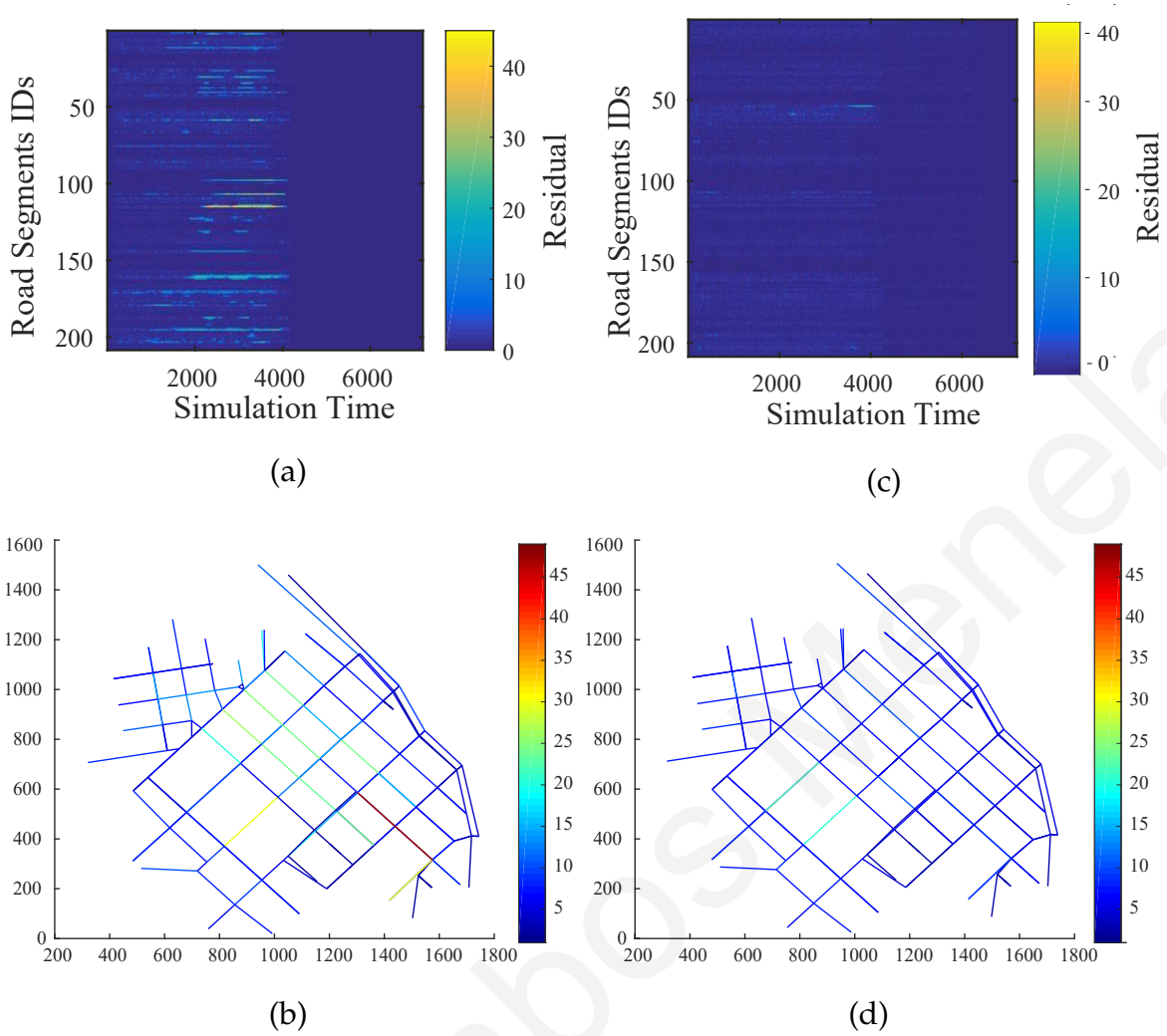
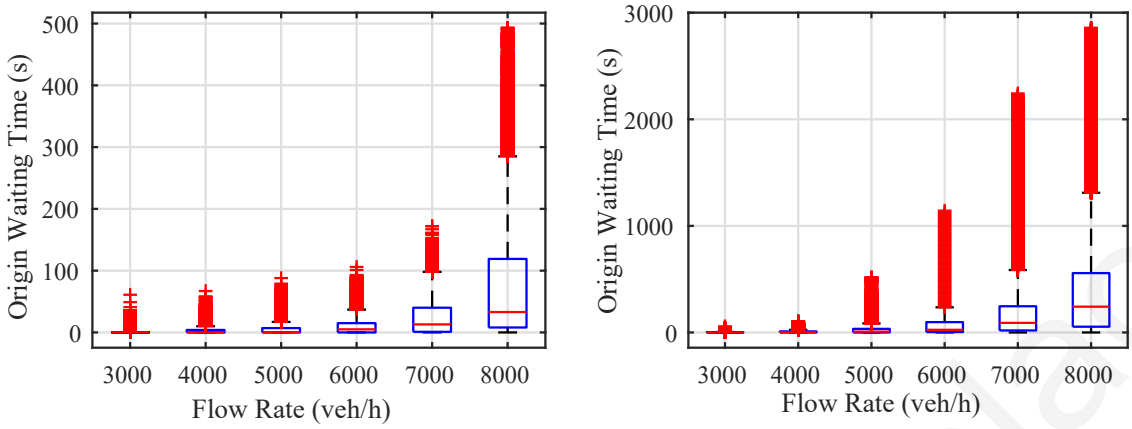


Figure 4.9: Performance evaluation of the RRACNP (a)-(b) and RRAC (c)-(d) with respect to the residual density of individual road segments over time, and the maximum instantaneous residual density of individual road segments, respectively, for the highest demand flow rate (i.e., 8000 veh/h).

4.7 Summary

This chapter extends the route reservation architecture proposed to investigate the EDAT problem in continuous time and develop prediction methods for more accurate estimation of the time needed to traverse different road segments for better route reservations. In this context, a time-varying multiple linear regression method is developed that takes into account the densities of neighboring road segments affecting the exit of a vehicle from a particular road segment. For the solution of the EDAT problem in continuous time, a heuristic approach is also developed based



(a) (b)

Figure 4.10: Origin waiting time for the (a) RRACNP and (b) RRAC.

on a customized version of the RRAC algorithm that provides fast and close-to-optimal solutions. Extensive performance evaluation confirms the usefulness of the continuous-time approach to solving the EDAT problem as it leads to better results compared to algorithms proposed previously.

Chapter 5

Effective Multi-region Traffic Control and Demand Management Using an Overlay Route-Reservation Scheme

5.1 Introduction

According to the route reservation architecture (presented in Chapters 3 and 4), vehicles are assigned to traverse the network only through non-congested paths. In solving this problem, the earliest arrival time at the destination can be achieved while network utilization remains within the congestion-free regime. In the proposed architecture, it is assumed that there is a detailed reservation plan along the exact path that all vehicles should follow from their origin to their destination. This approach has two possible drawbacks. First, it requires a large amount of information to be stored, fact that makes it difficult to scale. Second, a vehicle movement is affected by the actions of other vehicles which introduces randomness, meaning that a vehicle may not be able to follow precisely the plan assigned by the RSU.

To resolve these problems, this chapter proposes an aggregation scheme where vehicles are forced to follow a regional-level path aiming to make the Route Reservation architecture more scalable and efficient. Assuming that an urban area is partitioned into several homogeneous[1] regions in which an overlay graph is con-

[1]Homogeneous regions are those where their road segments exhibit similar traffic characteristics such as small variance of link densities and traffic demand distribution; the existence of homogeneous

structed to aggregate all road segments that are in each region. In this approach, a vehicle is allowed to enter a region only in the case that a region will not exceed its critical density during the interval that the vehicle will be expected to traverse it. In this setup, a variation of the route reservation architecture is proposed. However, the routing now is done based on the overlay network, that aggregates the actual network.

The remaining of the chapter is organized as follows. Section 5.2 introduces the system model of the proposed aggregated reservation architecture. Section 5.3 mathematically describes the proposed reservation scheme and formulates the Earliest Destination Arrival Time (EDAT) problem over the overlay graph. Section 5.4 presents an algorithm for solving the EDAT problem and a variation that aims to balance the utilization of the boundary nodes. Simulation results are included in Section 5.5 that demonstrate the potential benefits that can be achieved from applying the proposed approach. Finally, the chapter concludes with Section 5.6.

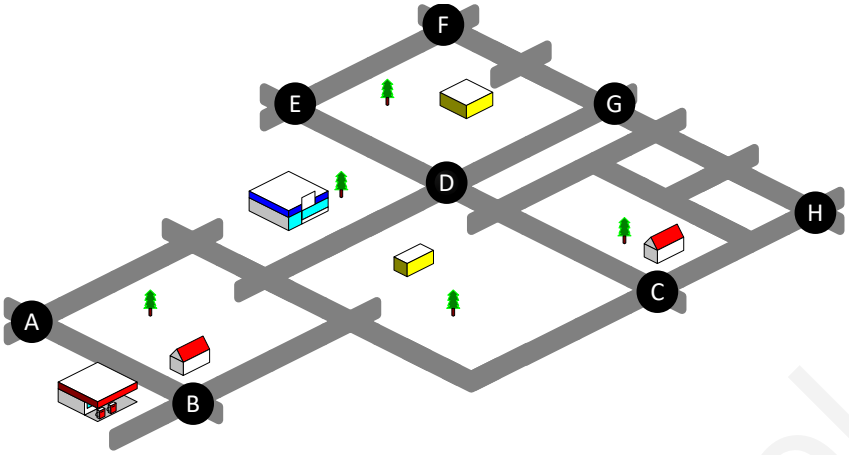
5.2 Overlay Route-Reservation Architecture

An overview of the proposed architecture is presented in Fig. 5.1. Schematic (a) illustrates a heterogeneous urban area that consisted of 3 homogeneous regions (each region is separated with the colored ellipsoid) whereas, each identified region has a number of boundary nodes shared by adjacent regions[2]. In the sequel, schematic (b) depicts the overlay graph formed by connecting all boundary nodes of the same region together. In the example illustrated in the schematic (b) each boundary node is represented by junctions $(A, \ldots, H)$ with the coloring denoting the regions that each one is a member of. For instance, junction D is shared among three neighboring regions (region 1, 2 and 3). Note that, the overlay graph consists solely of boundary nodes that are shared between at least two adjacent regions.

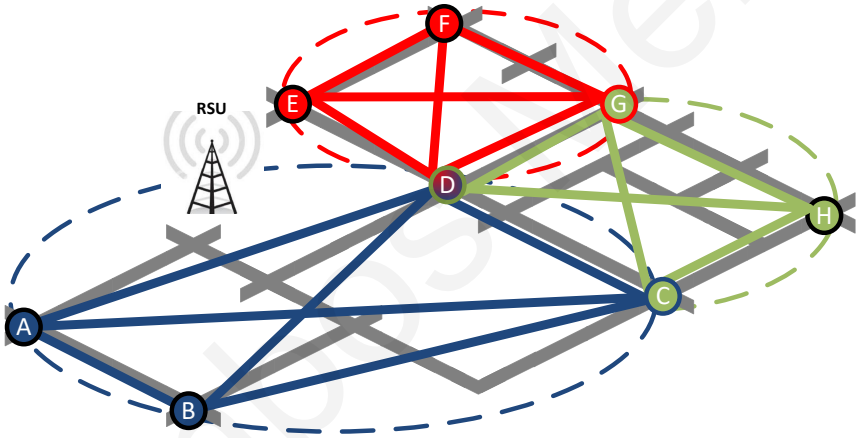
In the overlay graph (which can be created off-line), each boundary node is connected to the other boundary nodes of the same region whenever a physical path exists between them. For each of these links, the traveling time cost is set to

regions has been recently studied using empirical data by [19] and [25].

[2]In this chapter region clustering is based on [23]



(a) Road network



(b) Overlay network

Figure 5.1: Proposed Architecture.

be the time required to traverse the shortest path in the underlying road network considering free-flow conditions. Doing so enables accurate travel-time estimates to be made on the expected time necessary to traverse a region from one boundary node to another.

Given the origin-destination ($O-D$) pair requests, and the past path assignments, an RSU keeps an estimate of the number of vehicles that are expected to traverse each region at each time slot. When a vehicle is about to start its journey, it sends a request to the RSU with its ($O-D$) pair and its preferred starting time. The RSU responds with the actual starting time and the sequence of way-points that consist of the

boundary nodes that the vehicle should follow along its journey. The RSU decisions have as primary objective to minimize the vehicle's arrival time to its destination while ensures that all regions always remain below their critical capacity. Between way-points, a vehicle is free to follow any desired route to reach the next boundary node or its destination.

If during a time-slot an overlay link (which connects two boundary nodes) reaches its critical density, then it becomes non-admissible and hence the RSU refrains from using it in its route-reservation algorithm until the region is cleared. Additionally, the RSU may impose a waiting period at the origin if the arrival time at the destination will be minimized or when no feasible path exists (as required regions exceed their critical density). Subsequently, the architecture computes a "high-level" path and instructs the vehicle to navigate through specific regions while allowing to the vehicle to navigate within each region freely. The provided route avoids regions that are at their capacity, and the RSU may also instruct the vehicle to delay its departure until a region is cleared.

5.3 EDAT problem formulation

Consider a network partitioned in $r \in \mathcal{R} = \{1, \ldots, R\}$ regions. Each in similar manner with previous chapters each region exhibits specific MFD characteristics that determine ρ_r^C, $\rho_r{}^J$, L_r, and u_r^C, representing the critical density (corresponding to the maximum flow), the jam density, the total region length and the speed-at-capacity. Let parameter $N_r(t)$ denote a region's accumulated number of vehicles at time-slot t.

Denote the overlay graph $\mathcal{G} = (\mathcal{B}, \mathcal{E})$ where $\mathcal{B}$, is the set of boundary nodes and $\mathcal{E}$ the set of all overlay links between boundary nodes of each region. Also when vehicle h requests to determine its path, a "temporary" graph is formed $\mathcal{G}^h = (\mathcal{B}^h, \mathcal{E}^h)$ where $\mathcal{B}^h = \mathcal{B} \bigcup \{O, D\}$, i.e., it includes the boundary nodes together with the origin and destination nodes of vehicle h and $\mathcal{E}^h$ includes all links of $\mathcal{E}$ plus the links that connect the $\{O, D\}$ nodes to $\mathcal{B}$.

The traffic dynamics for each link $(i, j) \in \mathcal{E}$ which is within region r is characterized the link's jam density $\rho_{ij}^J = \rho_r^J l_{ij}$ where $l_{i,j}$ is the length of link (i, j). Furthermore, we define $\rho_{ij}^C = (\rho_r^J / \rho_r^C)\rho_{ij}^J$ whereas, the parameter $\rho_{ij}(t)$ denotes the instantaneous density

of link (i, j) at time-slot t and the parameter ρ_{ij}^C denotes the maximum allowable density that link (i, j) is expected to accommodate in order to operate at the region's speed-at-capacity, u_r^C, i.e., $\rho_{ij}(t) \leq \rho_{ij}^C$. The traversal time for each link in the overlay graph is denoted by the parameter $\bar{\tau}_{ij}$ i.e., $\bar{\tau}_{ij} = \left\lfloor l_{ij}/u_r^C/T \right\rceil$, where $\lfloor z \rceil$, is the nearest integer to z and T is the sampling interval. Since all regions are assumed to operate at the free-flow regime we assume that the vehicle speed within a region is constant and equal to u_r^C.

As described above, the RSU keeps track of the cumulative number of arrivals and departures for each region up to time-slot t, with parameters $\alpha_r(t)$ and $\beta_r(t)$, respectively. The actual number of reservations within each region is then $N_r(t) = \alpha_r(t) - \beta_r(t)$. Similarly, the RSU keeps track of the accumulated number of vehicles across each link of the overlay graph, $n_{ij}(t) = \alpha_r ij(t) - \beta_{ij}(t)$ for time-slot t.

The accumulated number of vehicles that pass through a specific boundary node can be computed by

$$h_{v_i}(t) = \sum_{\tau=t_0}^{t} \sum_{j \in \mathcal{B}, j \neq i} n_{ji}(\tau) \tag{5.1}$$

Interestingly, the parameters $N_r(t)$ and $n_{ij}(t)$ are used to track the admissibility of each region and each link in the overlay graph respectively. A road segment (i, j) is considered as *admissible* at time-slot t if the number of reservations at time-slot t is not larger than the number of vehicles corresponding to the link's critical density, i.e. $n_{ij}(t_0) \leq \rho_{ij}^C l_{ij}$ while at same time $N_r(t_0) \leq \rho_r^C L_r$.

Formally, the *admissibility* of road segment (i, j) at time-slot t is denoted by parameter $x_{ij}(t)$ given as $x_{ij}(t) = 1$ if a a link is admissible, and $x_{ij}(t) = 0$ otherwise. That is

$$x_{ij}(t) = \begin{cases} 1, \text{ if } n_{ij}(\tau)/l_{ij} \leq \rho_{ij}^C \; AND \; N_r(\tau)/L_r \leq \rho_r^C, \\ \qquad \forall\, \tau = t, \ldots, t + \bar{\tau}_{ij}, \; v_i \neq O, \; v_j \neq D \\ 1, \text{ if } N_r(\tau)/L_r \leq \rho_r^C, \\ \qquad \forall\, \tau = t, \ldots, t + \bar{\tau}_{ij}, \; v_i = O \; OR \; v_j = D \\ 0, \text{ otherwise} \end{cases} \tag{5.2}$$

where the quantity $n_{ij}(\tau)/l_{ij}$ and $N_r(\tau)/L_r$ denotes the estimated densities of link (i, j) and region r at time τ according to the reservations estimates, respectively. Let t_0

and d_{v_i} denote the vehicle request time and the vehicle arrival time at node v_i of link (i, j), respectively. Considering the above notation, the traversing cost $c_{ij}(t)$ can be expressed as follows:

$$c_{ij}(t) = \begin{cases} \bar{\tau}_{ij}, \text{ if } x_{ij}(t) = 1 \\ \infty, \text{ if } x_{ij}(t) = 0 \text{ and } i \neq O \\ \bar{\tau}_{ij} + w, \text{ if } x_{ij}(t) = 0 \text{ and } i = O \end{cases} \tag{5.3}$$

where, w denotes the number of time-slots that a vehicle may wait at the originating junction O.

The resulting route-reservation problem that arises requires the computation of a path using only admissible links and regions (of the overlay network), with waiting allowed at the originating node if deemed beneficial. In this computation two alternative options arise (as before in Chapter 3) when the shortest path from origin to destination includes non-admissible links. The first forces vehicles to wait at their origin until links become admissible. The second chooses an alternative route (instead of the shortest-time path) consisting of admissible links. Note that in some cases a combination of the two options may yield a better solution (i.e., wait for a certain time at the origin and then take an alternative path).

Earliest Destination Arrival Time (EDAT) Problem

The EDAT problem, aims to schedule vehicles across congestion-free regions while minimizing the expected arrival time to the destination. Given the overlay graph $\mathcal{G}$ and the origin-destination $(O - D)$ pair, the vehicle scheduling request time t_0 and the reservation states $x_{ij}(t)$, $(i, j) \in \mathcal{E}$, $\forall t \geq t_0$, then the EDAT problem determines the path that would allow the vehicle to arrive at its destination at the earliest possible time avoiding non-admissible links). Let p_k denote the k-th overlay path from source O to destination D denoted as $p_k = (v_0^k, v_1^k)$, (v_1^k, v_2^k), (v_2^k, v_3^k), ...$(v_{L_k-1}^k, v_{L_k}^k)$, where $v_j^k \in V$ is the j-th visited boundary node in the k-th path, with $v_0^k = O$ and $v_{L_k}^k = D$, and L_k is the length of the overlay path p_k. Additionally, let w and $d_{v_j^k}$ denote the waiting time at the origin junction and the earliest arrival time at boundary node v_j, respectively. Then, the earliest arrival time to each node of the path can be expressed as:

$$
\begin{aligned}
d^k_{v^k_0} &= t_0, \\
d^k_{v^k_1} &= d^k_{v^k_0} + c_{v^k_0, v^k_1}(d^k_{v^k_0}) \\
&\vdots \\
d^k_{v^k_{L_k}} &= d^k_{v^k_{L_k-1}} + c_{v^k_{L_k-1}, v^k_{L_k}}(d^k_{v^k_{L_k}})
\end{aligned} \tag{5.4}
$$

In compact form, the EDAT problem can be expressed as follows:

$$
(\Pi_o g)\ d^*_D = \min_{w,\ p_h} d^h_D \tag{5.5}
$$

$$
\text{s.t. Constraints (5.2) } - \text{ (5.4) are satisfied}
$$

5.4 Solution approaches

A heuristic solution to the EDAT problem over the overlay network is derived through a modification of the Route Reservation Algorithm (RRA) introduced in Chapter 3 refereed as Reservation-Based Routing Algorithm (RBRA) algorithm. To help the reader, a brief description of the RBRA algorithm is presented next. However, when the routing is done based on the overlay network, it is possible that two links of the overlay network, though they appear disjoint (they start from the same boundary node and connect two different boundary nodes), may share one or more road segments of the actual network. Reason of this is that, these road segments may be part of the shortest paths that lead to several other boundary nodes that theirs shortest paths contained shared links. Under such scenarios, the reservation scheme may allow the number of vehicles that will traverse the road segments that are in multiple shortest paths to exceed their capacity which may cause congestion phenomena to appear. To limit the probability of this happening, a balancing scheme is also introduced that aims to balance the traffic between the boundary nodes. This approach is referred to as Boundary nOde Load Balancing (BOLB). In this way, a possible solution of BOLB algorithm may lead the vehicle to follow a slightly longer path in order to distribute traffic more evenly across the network (social optimum).

5.4.1 Reservation-Based Routing Algorithm (RBRA)

The RBRA employs the well known Dijkstra's algorithm which uses the label setting-property and the relaxation technique [44] to find the shortest path from the origin to a destination. During the relaxation process of each iteration, the earliest-arrival-time to each boundary junction is updated (i.e., $d_{v_j} = \min\{d_{v_j},\ d_{v_i}+c_{ij}(t)\}$) and junction v_j is set as the shortest path from the origin to node v_j.

RBRA solves the EDAT problem in two loops, the inner and outer loop. The inner loop is responsible to identify the path that will enable each vehicle to arrive at its destination at the earliest time, assuming that a vehicle is allowed to wait at any intermediate node until a link's state changes from non-admissible to admissible (see Equation (5.2)). The RBRA initialization is identical to that of Dijkstra's algorithm, where the traversing cost of each link is dynamically calculated from every labeled junction to its neighbors. In the case that a link is non-admissible, then a vehicle is assumed to wait at node v_i of link (i, j), until its admissibility state changes. The minimum number of time-slots that may be required, denoted by variable $w_{ij}(t)$, can be calculated based on both the reservation status of the concerned link (i, j) and the earliest-arrival-time at junction v_i (i.e., $d^*_{v_i}$). Hence, each link traversal time is calculated using the constant travel time cost and the estimated waiting time duration if required (i.e., $c_{ij}(d^*_{v_i}) = \bar{\tau}_{ij}+w_{ij}(t)$). Hence, on every iteration the algorithm identifies the new labeled junction which is set to be the one that has the minimum earliest-arrival-time ($d^*_{v_i}$) (label-setting property). Finally, the inner loop returns to the outer loop the identified path and the minimum time ($w_{min} = \min(w_{ij}(t))$) that the vehicle needs to wait at an intermediate junctions or 0 if it does not need to wait at *any* intermediate node. In the outer loop, if $w_{min} > 0$, then the minimum waiting across all intermediate nodes is transferred at the origin and the vehicle's starting time is updated accordingly (i.e., $t_0 = t_0 + w_{min}$) and the problem is resolved, or if $w_{min} = 0$ the algorithm terminates.

5.4.2 Boundary nOde Load Balancing Algorithm (BOLB)

BOLB employs RBRA to compute the earliest-arrival-time path while tries balance the utilization observed at each boundary node. To do so, the link cost includes both

the travel time cost and the node utilization cost. Doing so, each link's traversal cost (5.3) is modified to also account the node's utilization as follows:

$$c_{ij}(d^*_{v_i}) \;=\; (\bar{\tau}_{ij} + w_{ij}(t)) \times \frac{h_{v_j}(t)}{\max_{q \in \mathcal{B}} h_{v_q}(t)} \tag{5.6}$$

where $h_{v_i}(t)$ is given by (5.1). The normalized utilization expressed by the second term in the latter cost function guides reservations through boundary nodes that are not highly utilized, effectively increasing each boundary node utilization.

Both RBRA and BOLB have a complexity in the order of $O(ME^2V)$, where $M < \infty$ denotes the number of iterations that the outer loop of both algorithms requires to converge. Note that both algorithms identify the starting time (demand management) and the boundary nodes that vehicles should follow but not the exact road segments to traverse. In this chapter, vehicles are free to follow their desired paths, passing through the given boundary nodes.

5.5 Performance evaluation

5.5.1 Setup

The road network under consideration is a 2.5 square miles non-signalized area of Downtown San Francisco, as illustrated in Fig. 5.2. A similar area has been used in [24] which also provides a detailed breakdown of the homogeneous regions that exist. Specifically, the selected area consists of 143 road junctions and 319 single-lane road segments with lengths varying from 100 m to 400 m. The SUMO micro-simulator [112] has been used to create traffic across this area, considering the Krauss' car-following model [113] with model parameters set as follows: vehicle length of 5 m, maximum speed 15 m/s, acceleration 2.5 m/s^2, deceleration 4.5 m/s^2, driver imperfection 5%, driver reaction time 0.5 s, and minimum gap distance 2.5 m. The simulation time step was set to 0.1 s, while the discretization of the algorithm's time-slots was set equal to $T = 1$ s. For the application of the proposed route-reservation scheme, the area is partitioned into three homogeneous regions ($R = 3$) (separated with different colors as depicted in Fig. 5.2).

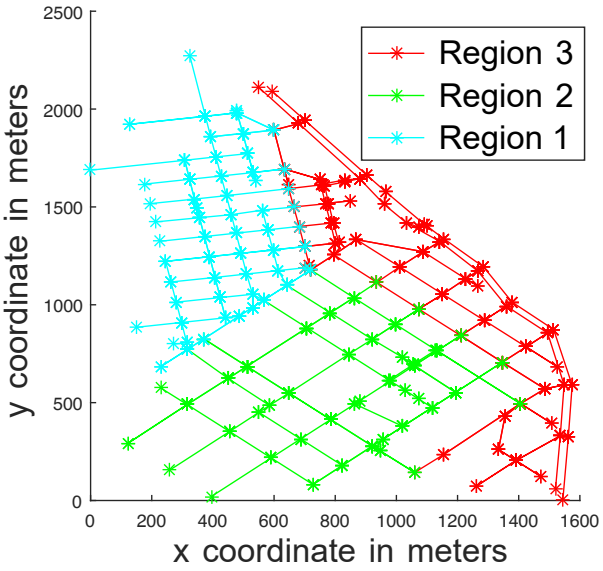
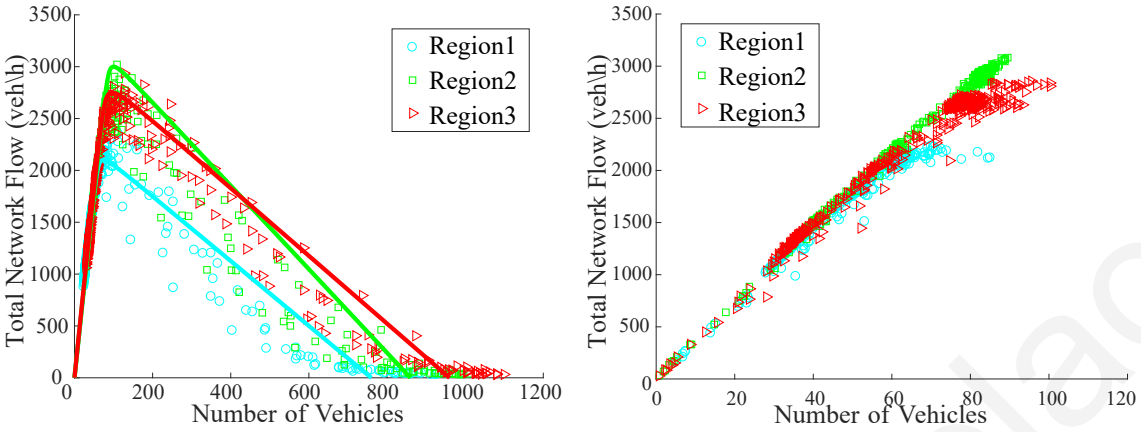


Figure 5.2: The simulated network (a segment of Downtown San Francisco).

5.5.2 MFD analysis

To derive and use the MFD of each region, simulations were performed for the uncontrolled scenario (US) (i.e., where vehicles follow their desired shortest paths without any waiting or re-routing). To do so, a 4 hours scenario was simulated within which the input flow was initially set to 4000 veh/h and incrementally increased by 2500 veh/h for the next three hours considering measurements every 5 minutes. Also, for the results presented hereafter, ten Monte Carlo simulations were conducted where vehicles arrive according to a Poisson process. Doing so, illustration in Fig. 5.3 (a), shows that in the US scenario, all three regions experience moderate scattering along varying densities. Accordingly, each region's MFD is calibrated through the automated calibration method proposed by [114] for the single-regime *Van Aerde* model [115] as indicated by the colored solid lines. From the calibrated model in the figure, the each region's model parameters are obtained as follows: 1) the speed-at-capacity u_1^C = 36 km/h, u_2^C = 35.5 km/h, u_3^C = 37 km/h, 2) the per region jam density as ρ_1^J = 740 veh, ρ_1^J = 880 veh and ρ_1^J = 960 veh and 3) the per region critical density as ρ_1^C = 85 veh, ρ_2^C = 120 veh and, ρ_3^C = 105 veh. Notably, for both RBRA and BOLB algorithms the travel time calculations are estimated using the respective speed-at-capacity u_r^C value. Fig. 5.3 (b) depicts the resulting MFD when the BOLB algorithm is employed demonstrating the absence of the congested regime. This is achieved as



(a) (b)

Figure 5.3: Each region's MFD of the simulated topology: (a) Uncontrolled scenario; (b) BOLB algorithm.

BOLB algorithm restricts the number of vehicles allowed to simultaneously traverse the network.

5.5.3 Results

In the results that follow, the performance of both RBRA and BOLB algorithms is compared with the uncontrolled scenario (US)[3]. It is emphasized that even though route reservations are computed solely based on information from previous reservations made, the result presented here reflect the actual paths of the vehicles which may be different from what the reservation approach has predicted due to the randomness and other uncertainties involved. Finally, similarly as before 10 Monte Carlo simulations were contacted with random $O - D$ pairs across the whole area with flow rates varying between $2000 - 10000$ veh/h over a period of 2 hours.

Fig. 5.4 and Fig. 5.5 show the average vehicle travel times and, the average number of vehicles that completed their journeys as a function of the different flow rates. The scattered plots in Fig. 5.4 illustrate the mean travel time of each realization, while the dashed lines represent the mean travel time for all realizations. Similarly, the dashed lines in Fig. 5.5 illustrate the average number of vehicles that managed to

[3]Methods like [52, 55, 60] can not be directly compared since they require other parameters e.g., split rates among the boundaries of the examined region which can arbitrarily affect the results.

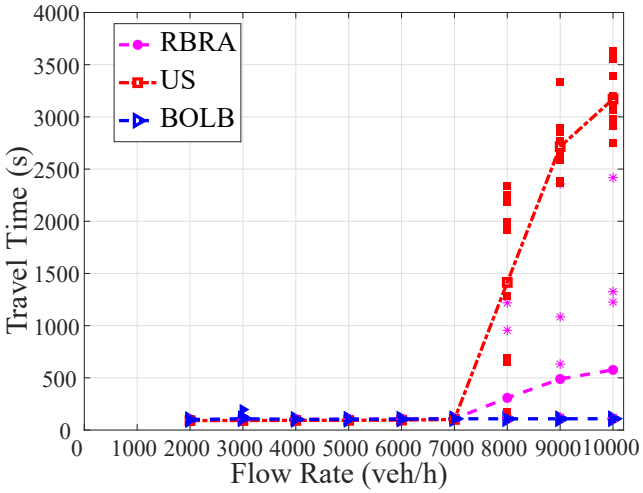


Figure 5.4: Average vehicle travel time from origin to destination for varying demand flow rate.

complete their journey within the simulation time while, the scattered plots represent the realizations obtained by each simulation run.

Fig. 5.4 and Fig. 5.5 depict the network performance according to different flow rates. As demonstrated, in low flow rates (ranging from (2000 veh/h – 7000 veh/h), no congestion is observed and both approaches behave similar to US. At higher flow rates congestion begins to form, demonstrating the superior performance of BOLB as it greatly outperforms RBRA and US by managing to maintain smaller travel times, that are also mostly unaffected by the increasing demand. It is also important to note that, at high flow rates RBRA behaves somewhat unpredictably due to over-utilization of some boundaries nodes that may produced spill-backs and queues at the perimeter of neighboring regions.

Fig. 5.6 illustrates the travel time distribution for all vehicles that manage to reach their destination during the simulation time at flow rates of 10000 veh/h. As depicted, both RBRA and BOLB greatly improved travel time compared to US.

Fig. 5.7 and Fig. 5.8 illustrate the variance of all boundary nodes and the evolution of utilization of each boundary node for both RBRA and BOLB as a function of consecutive vehicle requests. Fig. 5.7 depicts the variance of utilization between all boundary nodes demonstrating that BOLB manages to maintain a significantly lower variance compared to RBRA. The increased variance of RBRA is due to the fact that particular boundary nodes may reside in locations that intersect a higher

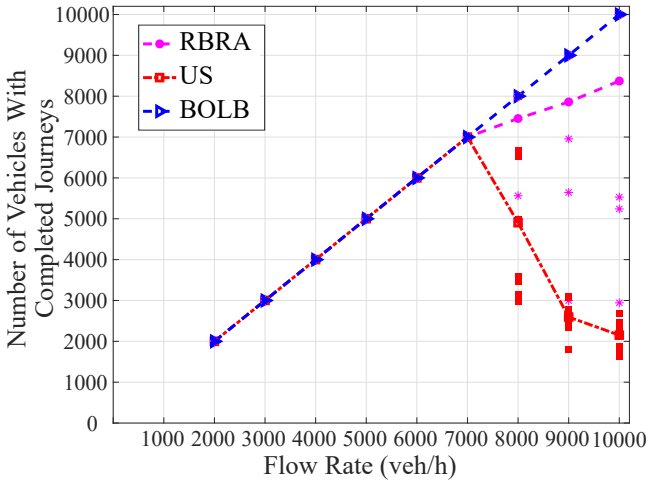


Figure 5.5: Number of vehicles with completed journeys for different simulation scenarios with varying demand flow rate.

number of regions increasing their attractiveness. This is more clearly evident in Fig. 5.8 where in the case of RBRA 4 nodes are significantly over-utilized compared to the rest.

Fig. 5.9 (a) and Fig. 5.9 (b) depicts the admissibility state of each region as it evolves over time for traffic flow of 10000 veh/h for BOLB and RBRA algorithms, respectively. The non-admissible state is represented by the value 1 and the admissible state is represented by the value 0. As shown in the figure, all regions frequently become non-admissible due to the high traffic demand considered in this scenario. Also compared to RBRA, BOLB manages to maximize the number of times each region enters the non-admissible state mainly due to the fact that it strives to equalize the utilization across all boundary nodes.

As indicated above the demand management is perform by restricting the allowed number of vehicles that simultaneously traverse each region below region's critical density result in the overall network efficiency improvement. However, with an increase of traffic demand the waiting time at the origin increases as well with this behavior can be observed in Fig. 5.10 which shows the waiting-time that vehicles require to wait before depart for their journeys for different flow rates (increasing demand). Fig. 5.10 shows that as flow rate increases vehicles prefer to wait at their origin until an admissible path becomes feasible for both algorithms. This behavior is desirable since in high-demand scenarios, significant waiting needs to be incurred

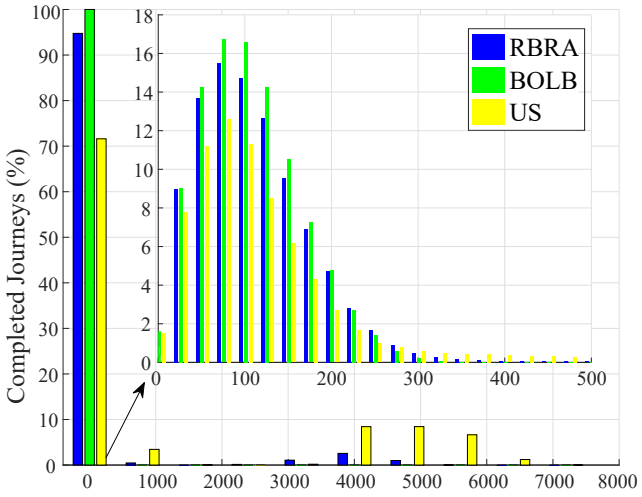


Figure 5.6: Per vehicle travel time distribution for the highest demand flow rate (i.e. 10000 veh/h).

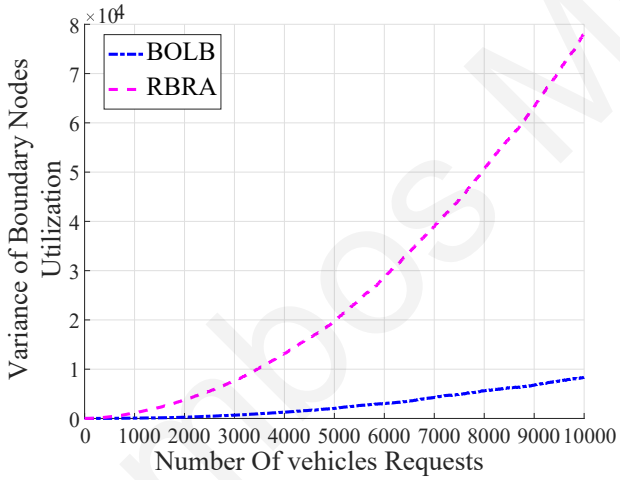


Figure 5.7: Evolution of the variance of boundaries utilization.

to maintain high network flows and speeds. Even so, for both algorithms the average waiting is within acceptable levels (10 minutes).

Finally, the sensitivity of the BOLB performance to the changes in drivers compliance levels is examined in Fig. 5.11 in which the heaviest loaded demand scenario of 10000 veh/h is evaluated considering seven different drivers percentages of drivers' compliance level (i.e., 70%, 75%, 80%, 85%, 90%, 95% and 100%). Note that in the ideal scenario (i.e., the compliance rate is 100 %) all drivers will opt to follow the waiting intervals and routes provided by the RSU. The figure depicts the boxplot of travel times in which the solid red line representing the measured median while the solid red dot represents the mean travel time of all vehicles in each considered

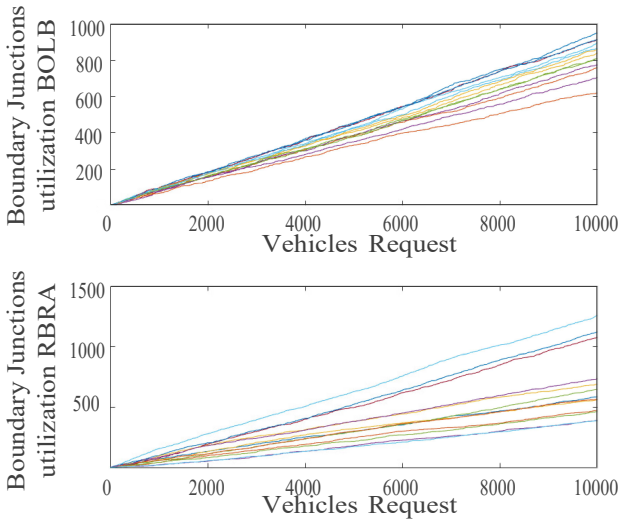


Figure 5.8: Evolution of the boundaries utilization.

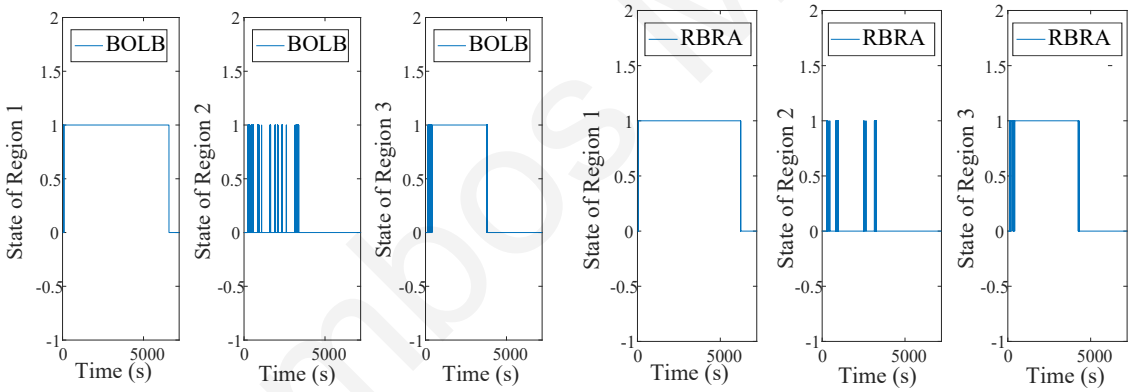


Figure 5.9: Region admissibility over time for: (a) the BOLB algorithm and (b) the RBRA algorithm.

simulation. As expected as the compliance level is reduced then the performance of the BOLB method decreases. More specific with a decrease in compliance level the scattering increases while the average experience travel time increases as well. Furthermore, it's evident that for compliance level higher than 80%, the BOLB method slightly affected by drivers that did not opt to the RSU instructions. On the other hand, at compliance level of 70%, BOLB behaves similarly with ordinary RRA (in the sense of average travel time) with an increasing possibility of gridlock.

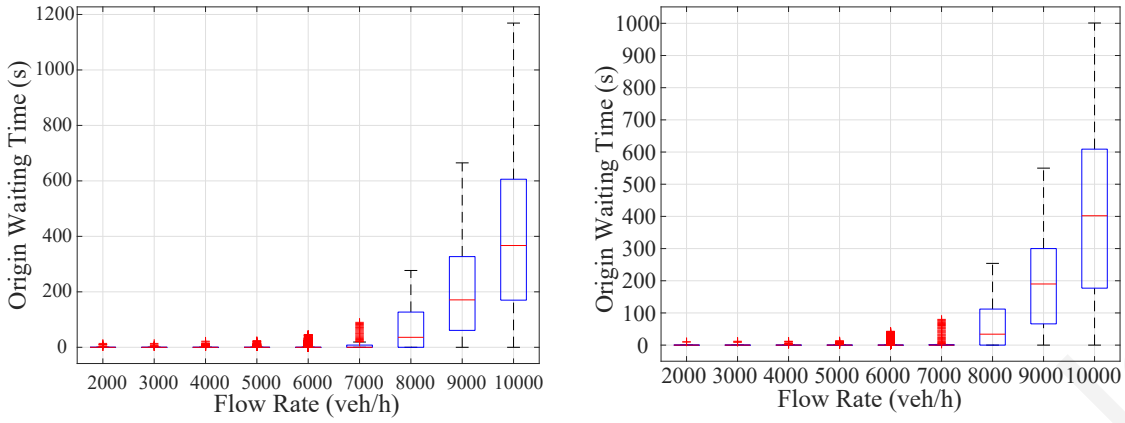


Figure 5.10: Origin waiting time for the (a) the BOLB algorithm and (b) the RBRA algorithm.

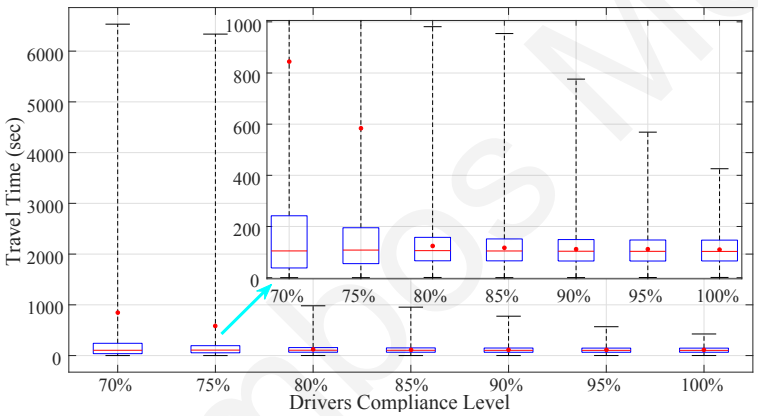


Figure 5.11: The sensitivity of RBRA performance to changes in the percentage of drivers' compliance level considering the heaviest loaded demand scenario of 10000 veh/h.

5.6 Summary

This chapter proposes an aggregation of the route-reservation scheme that controls vehicles route in a multi-region urban area. The advantage of this scheme is that the resulting algorithm that utilizes an overlay graph to control the traffic ensures effective, scalable, and congestion free routing solutions in large-scale multi-region networks. Simulation results demonstrate the significant gains achieved by the proposed route-reservation scheme compared to the uncontrolled traffic behavior resulting in many-fold gains in serving traveling requests and reductions in travel times, especially during high demand flows.

Chapter 6

Scheduling Vehicles for On-Time Arrival using Route-Reservations

6.1 Introduction

This chapter proposed an alternative route reservation architecture that aims to compute the vehicles departure times and reserves their route to reach their destination at the desired time. In this way, the objective of this chapter is to minimize the difference between the desired and the actual destination arrival times. For this problem, vehicles transmit to RSU their origin and destination pair and the desired time that they require to arrive at the destination. The RSU determines each vehicle's departure time as well as the path to be followed while making the appropriate route reservations on the selected path such that all scheduled vehicles avoid congested road segments. Due to the reservations, the RSU can guarantee the on-time arrival at the destination for each vehicle request. This can be done by coordinating the departure times for each vehicle (i.e., apply demand management) which can significantly improve the traffic flows and sustain travel times around those achieved assuming free-flow speed conditions.

The remaining of the chapter is organized as follows. Section 6.2 mathematically describes the proposed scheme and defines the on-time arrival problem (OTA). Section 6.3 derives an algorithmic solution for the OTA problem which utilizes a backward route-reservation scheme that schedules vehicles through road-segments that are below their critical density. The performance of the proposed solution is

investigated in Section 6.4, demonstrating the gains achieved for several different metrics while also indicating that vehicles almost always arrive at their destination on time. Concluding remarks are provided in Section 6.5.

6.2 Problem formulation

Similar to previous chapter an urban area is modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the sets $\mathcal{V}$ and $\mathcal{E}$ represent the road junctions (i.e., $\{v_i, v_j\} \in \mathcal{V}$)) and the road-segments (i.e., $(i, j) \in \mathcal{E}$), respectively. Each road-segment $(i, j) \in \mathcal{E}$ is described by the parameters λ_{ij} and l_{ij} (km), denoting the number of lanes and its length, respectively. In addition, we assume that the road network under investigation constitutes an urban area with a well defined MFD. Hence, the segment's $(i, j) \in \mathcal{E}$ traffic dynamics are characterized by the parameters ρ_{ij}^C and ρ_{ij}^J, denoting the critical and jam densities, respectively. Note that to obtain the critical density of the road-segment (i, j) we use:

$$\rho_{ij}^C = (\rho^C / \rho^J) \rho_{ij}^J. \tag{6.1}$$

Hence, the critical density of each segment is proportional to the region's critical density. Furthermore, let variable $\rho_{ij}(t)$ veh/km denote the instantaneous density of each road segment at each time-slot $t \in \mathcal{T}$, where $\mathcal{T}$ defines the time horizon of the problem. Considering Eq. (6.1) it is true that, for all $\rho_{ij}(t) \leq \rho_{ij}^C$ vehicles can be assumed to travel with free-flow speed u_f. On this premises, the number of time-slots that a vehicle is require to traverse road segment (i, j) can be expressed as:

$$\bar{\tau}_{ij} = \left\lfloor l_{ij} / u_f / T \right\rceil, \tag{6.2}$$

where T is the sampling interval and $\lfloor z \rceil$ denotes the nearest integer to z.

The proposed methods utilizes route reservation to keep track of the accumulated number of vehicles reservations (i.e., $n_{ij}(t)$) of each road-segment (i, j) for time-slot t). As i previous chapters, a road-segment is assumed to be *admissible* at the discrete time-slot t if a vehicle starting from road junction v_i at time-slot t can traverse road segment (i, j) without making the accumulated reserved density larger than the critical density at any point within the transit time. Let variable $x_{ij}(t)$ denote

the admissibility state taking the value $x_{ij}(t) = 1$ if segment $(i, j) \in \mathcal{E}$ is admissible and $x_{ij}(t) = 0$, otherwise. Mathematically the admissibility state can be defined as follows:

$$x_{ij}(t) = \begin{cases} 1, \text{ if } n_{ij}(\tau)/(\lambda_{ij} l_{ij}) \leq \rho_{ij}^{C}, \ \forall \ \tau = t, \ldots, t + \bar{\tau}_{ij} \\ 0, \text{ otherwise} \end{cases} \tag{6.3}$$

where the quantity $n_{ij}(\tau)/(\lambda_{ij} l_{ij})$ is the accumulated reserved density of road segment (i, j) at time $\tau \in [t, \ldots, t + \bar{\tau}_{ij}]$. Given the admissibility state, the cost of traversing a road segment (i.e., $c_{ij}(t)$) can be defined as follows:

$$c_{ij}(t) = \begin{cases} \bar{\tau}_{ij}, \text{ if } x_{ij}(t) = 1 \\ \infty, \text{ if } x_{ij}(t) = 0 \end{cases} \tag{6.4}$$

On-Time Arrival (OTA) problem:

Given the origin-destination pair of the m-th vehicle (i.e., $O_m - D_m$, with $O_m, D_m \in \mathcal{V}$), the desirable destination arrival time $d_{D_m}^{des}$, and the reservation states $x_{ij}(k)$, $(i, j) \in \mathcal{E}$, $\forall k \in \mathcal{T}$, then, the OTA problem seeks to find the starting time s_m^* and the path p_m^* that minimize the difference between $d_{D_m}^{des} - s_m^*$. In other words, OTA finds the latest time that the m-th vehicle should start from its origin such that it will arrive at the destination on or before the desired arrival time.

To complete the problem formulation, let p_h, denoting the h-th path from source O_m to destination D_m, be defined as $p_h = (v_0^h, v_1^h), (v_1^h, v_2^h), (v_2^h, v_3^h), \ldots, (v_{L_h-1}^k, v_{L_h}^h)$, where $v_j^h \in \mathcal{V}$ is the j-th visited node in the h-th path, with $v_0^h = O_m$, $v_{L_h}^h = D_m$, and L_h is the length of p_h. Additionally, let variable $d_{v_j}^h(s)$ be the arrival time at junction $v_j \in \mathcal{V}$ if a vehicle departs form its origin at $s \in \mathcal{T}$. Then, the arrival time to each node of the h-th path can be expressed as:

$$\begin{aligned} d_{v_0^h}^h(s) &= s, \\ d_{v_1^h}^h(s) &= d_{v_0^h}^h(s) + c_{v_0^h, v_1^h}(d_{v_0^h}^h(s)) \\ &\vdots \\ d_{v_{L_h}^h}^h(s) &= d_{v_{L_h-1}^h}^h(s) + c_{v_{L_h-1}^h, v_{L_h}^h}(d_{v_{L_h}^h}^h(s)) \end{aligned} \tag{6.5}$$

Thus, for the m-th scheduled vehicle, the central-controller has to compute s_m^* and p_m^* that solve the problem (P_1) below:

$$(P_1) \quad \min_{s,p^h} \; J_T = d_{D_m}^{des} - d_{v_0}^h(s) \tag{6.6a}$$

$$\text{s.t. Model Dynamics (6.1) – (6.5),}$$

$$d_{D_m}^h(s) \leq d_{D_m}^{des}. \tag{6.6b}$$

Constraint eq. (6.6b) is added to ensure that vehicle m will not arrive after the desired time to the destination. Furthermore, the constraints in (6.1) - (6.5) define the model dynamics which consider each road segment's admissibility state.

Clearly, if at a given time there are no road segments that are at their capacity, the path that the m-th vehicle should follow is the shortest path from O_m to D_m and it should start at time $s_m^* = d_{D_m}^{des} - c_m^*$, where $c_m^* = l_m^*/u_f/T$ and l_m^* is the length of the shortest path. On the other hand, if there are links of the shortest path that are at their capacity, then the vehicle may have two options, either depart much earlier when all links of the shortest path are admissible (and arrive earlier and wait at the destination) or start a little earlier and take a longer path and arrive at the destination on time. Out of these possibilities, (P_1) will select the one that will allow the vehicle to depart as late as possible from the origin and still make it to the destination on time.

6.3 On-Time Arrival problem algorithmic solution

A solution to the OTA problem is obtained, based on dynamic programming [44], by constructing a time-space Graph (TSG). Algorithm 8 obtains an OTA solution taking into account the m-th request (O_m, D_m and $d_{D_m}^{des}$), the current number of reservations (i.e., $n_{ij}(t)$), and the current admissibility state (i.e., $x_{ij}(t)$) of each edge $(i, j) \in \mathcal{E}$) over the $t \in \mathcal{T}$ (Note, that the reservations and admissibility states can be easily expressed in the form of 2-D matrices with the columns representing link indices while the rows represent time indices). The constructed TSG is a directed acyclic graph where the space dimension contains all the indices of the nodes in $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and the time dimension includes consecutive time slots in descending order (starting from the desired destination arrival time, $d_{D_m}^{des}$ and going backwards in time). In this

way, each node in the space-time network represents the node where a vehicle arrives at the specific time-slot t. Once we have all nodes we can construct the TSG by inserting the edges that connect two nodes in reverse direction of the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and backwards in time on the TSG, with edges inserted only if there is a physical connection between the two adjacent nodes on $\mathcal{G}$, with its associated travel time cost reflecting the node on TSG that the vehicle will arrive.

The edge insertion procedure is accomplished based on two states. The first is the admissibility state, where an edge is considered as admissible if $x_{ij}(t) = 1$ according to Eq. (6.3). The second is the reachability state which defines if the newly inserted edge of TSG is reachable from the destination or not, meaning that there is a path that connects the destination node with the starting node of the related edge. Both states can be determined simply using variable, $d_{v_i}(t)\, v_i \in \mathcal{V}$, which denotes the arrival time at each node since the constructed graph does not contain a cycle. Therefore, in case that $d_{v_i}(t) = \infty$ then the edge is both not reachable and not admissible, while in the case that $d_{v_i}(t) = k$, the edge is both reachable and admissible. In case that both conditions are satisfied, (line 13 of Algorithm 8), an edge (i, j) is added on TSG, (lines 14-15). The whole process repeats until the time-slot that O_m becomes reachable, i.e. $d_{O_m}(k) < \infty$ for any v_i and t (edge (i, O_m)), (lines 8-20). In that case, the algorithm converges and returns the identified path by tracing back the nodes from O_m to D_m with the vehicle's departure time be equal with $s^*_m = d_{O_m}(t)$, (line 11).

Therefore, the solution of Algorithm 8 provides the m-th vehicle's departure time (i.e., s^*_m) and its route to follow (i.e., p^*_m). This information is utilized to make the appropriate route reservations on each road segment at the expected traversal times. We emphasize that, to compute route reservations for each vehicle; we use the constant parameter (i.e.,$\bar{\tau}_{ij}$) which defines the number of time-slots that a vehicle is required to traverse a road segment. Hence, by knowing the m-th vehicle path to follow and its departure time the expected traversal time for each road-segment can be calculated assuming each segment requires $\bar{\tau}_{ij}$ time slots to be traversed. Hence, the reservation status is updated during the expected transit times. In the same manner, from $n_{ij}(t)$ we also update the admissibility state of each road segment (i.e., $(i, j) \in \mathcal{E}$).

Initially, $n_{ij}(t) = 0$ and $x_{i,j}(t) = 1$ for all $(i, j) \in \mathcal{E}$ and $t \in \mathcal{T}$. Furthermore, it

Algorithm 8 TSG Alogrithm

Input: $\mathcal{G}(\mathcal{V},\mathcal{E})$, $n_{ij}(t)$, O_m, D_m, $x_{ij}(t)$, $d^{des}_{D_m}$, $c_{ij}(t)\ \forall\, t \in \mathcal{T}$;
Initialization:
$k = d^{des}_{D_m}$;
$d_{v_i}(k) = \infty,\ \forall\, k \in \mathcal{T},\ v_i \in \mathcal{V}$;
$d_{D_m}(k) = k,\ \forall\, k \in \mathcal{T}$
$s^*_m = -\infty$
Algorithm Execution:
while $k > s^*_m$ **do**
    for $(i, j) \in \mathcal{E}$ **do**
        if $((i == O_m)$ **OR** $(j == O_m))$ **AND** $(d_{O_m}(k) > s^*_m)$ **then**
            $s^*_m = d_{O_m}(k)$;
        else
            if $(x_{ij}(k) == 1)$ **and** $(d_{v_i}(k) < \infty)$ **then**
                $d_{v_j}(k) = d_{v_i}(k) - c_{ij}$;
                $previous[v_j][d_{v_j}(k)] = v_i$;
            end if
        end if
    end for
    $k = k - 1$;
end while
Trace back p^*_m and $previous[O_m][s^*_m]$;
Reservations-Admissibility status Update:
Update Reservations(p^*_m, s^*_m);
Update Admissibility(p^*_m, s^*_m);
Output: p^*_m, s^*_m;

is assumed that vehicle requests are collected over an interval and are sorted in descending order based on the desired arrival time. Then, they are processed by the TSG algorithm sequentially starting from the latest desired arrival time to the earliest.

The algorithm 8 results in an optimal solution in the discretized space-time do-

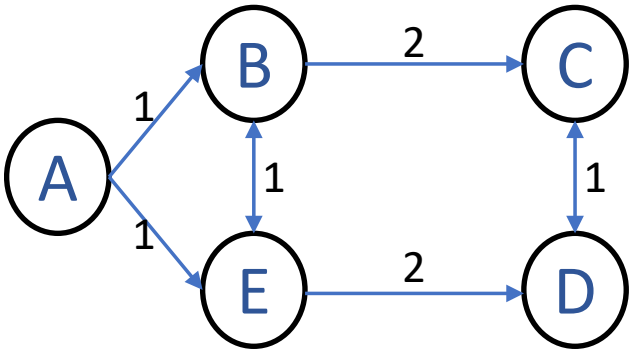


Figure 6.1: An example network of $\mathcal{G}(\mathcal{V},\mathcal{E})$.

main that it operates and executes in pseudo-polynomial time since the state space for the m-th request to be solved is not known until the first time that the algorithm visits the origin node in which $d_{O_m}(t) < \infty$, meaning that the algorithm converges with complexity $O((d^{des}_{D_m} - s^*_m)|\mathcal{E}|)$. Note that, in this algorithm we have to find a solution within a fixed interval $d^{des}_{D_m} T_{sub}$ where T_{sub} is the time a vehicle has submitted its request. If no such solution is found the algorithm returns failure. The optimum solution can be derived considering that each node in TSG is reachable only if the reachability state of all predecessor nodes forming the minimum path from destination to that particular node starting form the destination at the corresponding time. Hence, if a node is reachable through path p, then all nodes forming p are also reachable (with the minimum cost) and the *optimal substructure property* applies [44]. The reachability of all states examined for decreasing t and thus the optimal solution is found at time $d^*_{O_m}$ which in turn represent the latest time at which a vehicle should start from O_m to reach D_m on-time or on earlier time considering the admissibility states of the edges in $\mathcal{G}(\mathcal{V},\mathcal{E})$.

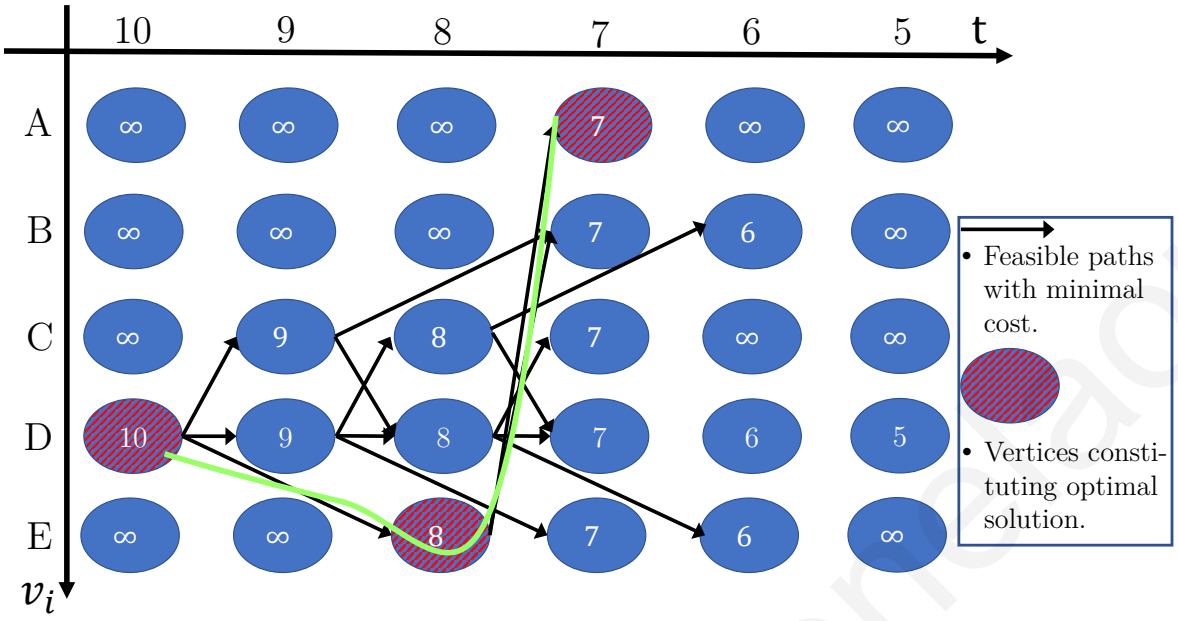
Illustrative Example

To better understand the proposed procedure, consider the example illustrated in Figure 6.1 (a) where edge lengths reflect the traversal times for specific road segment while the critical density of all edges in the graph is equal with 1 veh/edge. In this example, initially, no reservations are made and two vehicles request a path from A to D desiring to arrive at D at time slots $d^{des}_{D_1} = 9$ and $d^{des}_{D_2} = 10$, respectively. These requests are first sorted in descending order and thus the second request will be

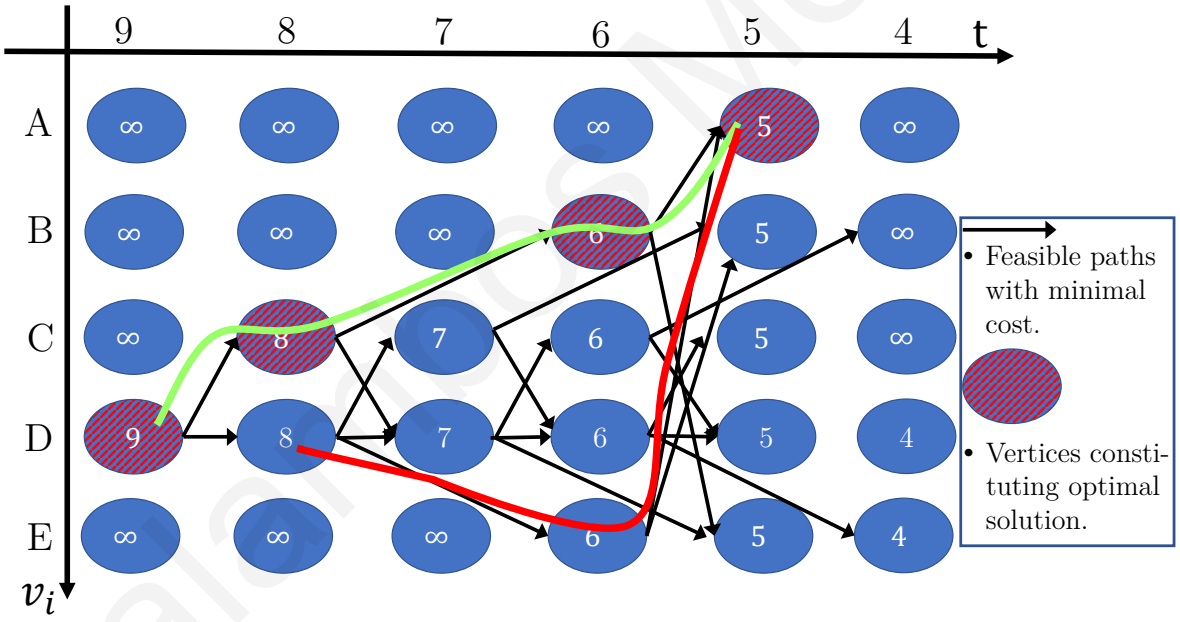
executed first by Algorithm 8.

Figure 6.2 (a) shows the TSG graph that is constructed by executing the first sorted request. The space dimension of each node indicates the junction index while the time dimension indicates the node created over time (with the time index starting from $d_{D_2}^{des} = 10$). The reachability of each node is assess from variable $d_{v_i}(t)$ where for the case of D_2 for all time-slots is reachable and thus $d_{D_2}(t) < \infty \, \forall \, t \in \mathcal{T}$. As illustrated in the figure, in the first column edges emerge only from the destination node (e.g., D_2) since all other nodes are not reachable at time-slot $t = 10$. Similarly, in the second column edges emerge from nodes D_2 and C since at time-slot $t = 9$ they have been reached from the destination (e.g., D_2). Note that the black solid-line edges are those that are added to construct the TSG which has a feasible path form the destination to the specific node. As Fig. 6.2 (a) shows, at the fourth column is the first time index that the origin is reached with the algorithm converging at this time index. In that way, the grid-shaded nodes represent the nodes consisting of the path p_2^* (i.e., $A \rightarrow E \rightarrow D_2$, also denoted with the solid green line) where the latest departure time is $s_2^* = 7$. Next, the algorithm updates the reservations based on the obtained solution and the admissibility state of those edges changes as follows $x_{EA}(7) = x_{DE}(8) = x_{DE}(9) = 0$.

Subsequently, the algorithm re-executes the TSG procedure for the other request (e.g., $d_{D_1}^{des} = 9$) where the associated TSG is depicted in Figure 6.2 (b). For that case, the time index begins at the 9-th time-slot (according to vehicle's request) while due to reservations made from the first vehicle we can observe that the shortest path is not a feasible solution due to the non-admissible states that emerge for some particular time-slots. Hence, the first time that the originating junction is reached at the 5-th time-slot (fifth column) where two alternative solutions exist (denoted with green and red solid lines, respectively) and the algorithm selects as a solution the p_1^* (i.e., $A \rightarrow B \rightarrow C \rightarrow D$, green line) with the $s_1^* = 5$. Note that, both solutions have equivalent objective value $J_T = 4$ but, their length differs. More specifically, if the vehicle follows the green path, then the duration of its travel time will be 4 time-slots and will arrive at the destination exactly on time. Otherwise, if vehicle follows the red path, the duration of its travel time will be 3 time-slots and will arrive at the destination 1 time-slot earlier, thus the vehicle will wait at the destination for 1



(a)



(b)

Figure 6.2: The direct acyclic graph that generated from TSG procedure to solve (a) the first vehicles request (b) the second vehicle request.

time-slot. In other words, if the second vehicle would like to arrive at the destination at $t = 9$, it cannot leave from its origin at $t = 6$ because will produce congestion at the edge (E, D) at the time-slot $t = 8$.

6.4 Performance evaluation

6.4.1 Setup

To evaluate the performance of the proposed solution we consider an 1.8 km^2 non-signalized urban region of the downtown San Francisco, as illustrated in Fig. 3.7 (the same are used in Chapter 3, The network was imported in the SUMO micro-simulator [112], and the Krauss car following model [113] was used. The car-following model parameters are set as follows: vehicle length 5 m, maximum speed 15 m/s, acceleration 2.5 m/s^2, deceleration 4.5 m/s^2, and minimum-gap-distance 2.5 m while no vehicle overtaking is allowed. The simulation time-step in SUMO was set to 0.1s while the time step of the algorithm was set equal to T = 1s with all simulations were performed for 2 hours. The vehicle desired arrival times are requested only during the first simulation hour in which requests are uniformly distributed during 6 time intervals. Specifically, all desired arrival times were distributed uniformly in the time intervals of 8:00, 8:10,. . . ,9:00 am. Hence, the network is loaded only during the first simulation hour while the second is used to empty the network and record some measurements. A critical density of ρ_{ij}^C = 33 veh/km/lane and a free-flow speed of u_f = 10.0 m/s is used to calculate each segment's travel time. Note, the ρ_{ij}^C and u_f values changes compared with values selected in Chapter 3 as the distribution of requests times differs. Finally, a total of ten Monte Carlos simulations are constructed (10 realizations) for varying flow rates from $1000 - 8000 veh/h$.

The proposed algorithm described in Section 6.3 is compared against the case where no control mechanism is applied (i.e., uncontrolled scenario US) where vehicles traverse from their origin to their destination along the shortest distance path. The travel time for each path is calculated assuming free-flow speed conditions while the departure time for each vehicle is assigned assuming that vehicles will take their shortest time path plus a uniformly distributed time budget (between 0-3 min) that is allocated to each vehicle to depart in advance to its shortest time path. Note that within the simulated network, the average trip length is around 1.5 min. Also, for the OTA results, it is assumed that all vehicles comply to the derived schedule. Furthermore, note that all vehicles schedules are obtained based only on the reservation estimates, rather than the actual state of each road segment, however, deviations

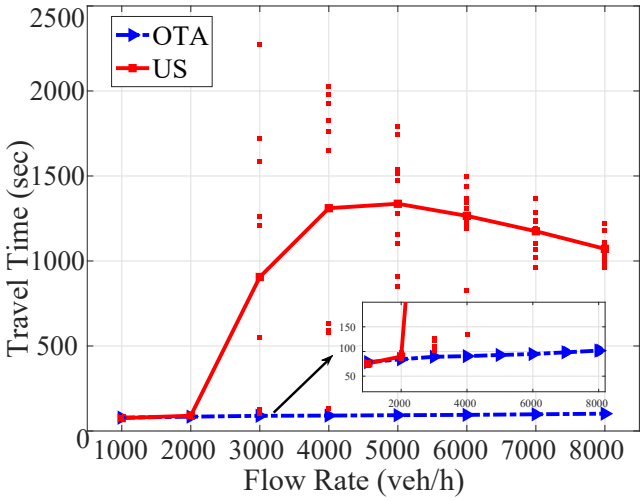


Figure 6.3: Travel time for different simulation scenarios with varying demand flow rate.

between the reservations and actual state exists due to the underlying uncertainty involved in the microscopic simulation. Finally, in the results presented hereafter only vehicles that have completed their journeys during the 2-hour simulation time are considered.

6.4.2 Results

Figs. 6.3, and 6.4 show the vehicle average travel time and the number of vehicles that manage to reach their destination during the simulation time, respectively. The scattered plots in Figs. 6.3 depict the mean travel time of each realization, while the lines represent the mean travel time for all realizations. Similarly, the scattered plots in Fig. 6.4 depicts the number of vehicles that manage to reach their destination within the simulation time for each realization, while the lines represent the average number of vehicles that completed their journey. As illustrated in Figures 6.3 and 6.4 (as expected), at low flow rates both OTA and US approaches perform equally well. However, at higher flow rates it is evident that OTA outperforms US in terms of travel times since OTA avoids congestion. Additionally, Fig. 6.4 shows that for the case of OTA all vehicles manage to reach their destination, unlike the case of the non-controlled case where a significant number of vehicles cannot manage to reach their destination due to the formation of severe traffic congestion.

The scatterplot of Fig. 6.5 represents the number of vehicles that have reached their destination after their desired arrival time for all realizations obtained by each

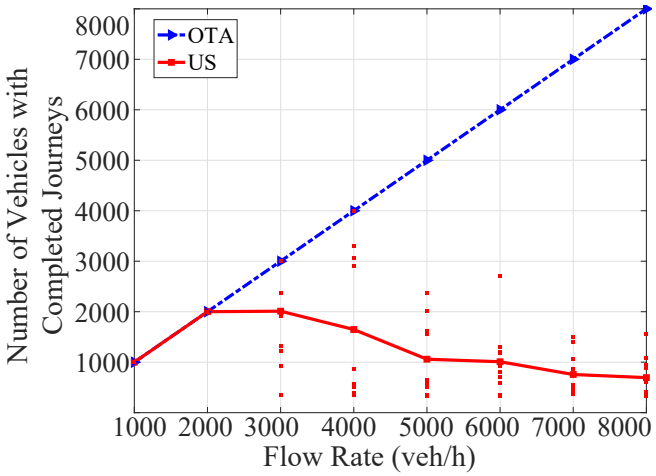


Figure 6.4: Number of vehicles with completed journeys for different simulation scenarios with varying demand flow rate.

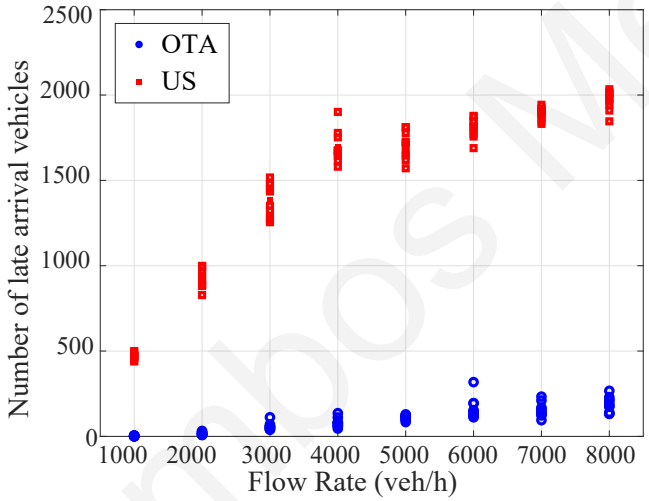


Figure 6.5: Number of late arrival vehicles.

Monte Carlo run. Similarly, the scatterplot in Fig. 6.6 shows the average time by which vehicles exceeded their desired arrival time for each realization while the line represents the mean value for all realizations. Both figures clearly show that the OTA approach manages to schedule most of the vehicles on-time with those exceeding their desired arrival time have negligible delays. Note that, those delays occur due to the uncertainty in the micro-simulation environment. On the other hand, for the non-controlled case, it is observed that at the highest flow-rates the congestion is unavoidable with many vehicle arriving late while a large number of them cannot reach the destination within the simulation period.

Fig. 6.7 illustrates the average waiting observed at the destination. In this figure,

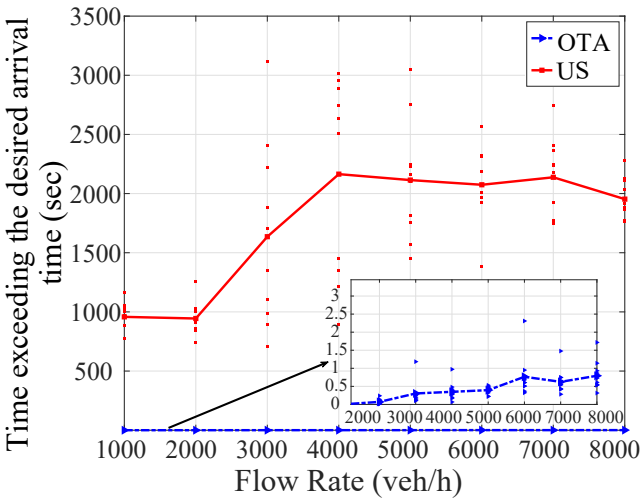


Figure 6.6: The time that vehicles exceeding their desired arrival time.

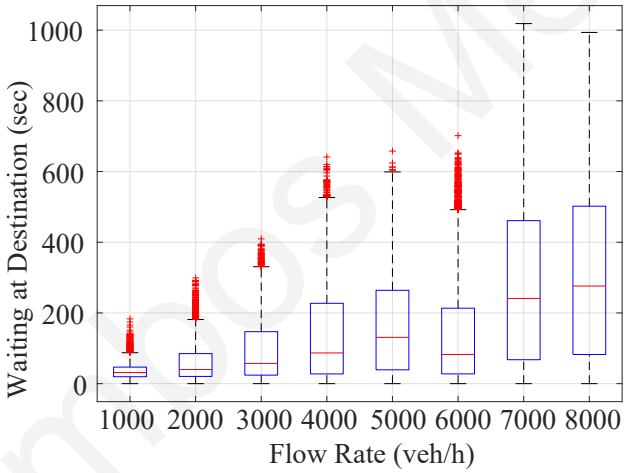


Figure 6.7: The waiting time at destination for all considered flow rates measured as the difference between the derided and the actual arrival time.

we measure the time that vehicles arrive earlier than their desired arrival time. As expected, with higher flow rates, the waiting time tends to be higher than with lower flow rates, meaning that travelers arrive much earlier than their desired time. This phenomenon occurs since an increase in demand results in more vehicles having to traverse the network in the presence of non-admissible segments. Even so, the vehicles arrive earlier than the desired time, while the waiting time at the destination is sustained within acceptable levels (around 3 min on average for the case of the highest flow rate scenario).

6.5 Summary

This chapter proposes a route-reservation approach that aims to schedules vehicles to arrive at their destination at their desired time while at the same time, traffic congestion is eliminated by restricting the density of all road segments below their critical value. In this framework, the on-time arrival (OTA) problem is examined and solved by developing a dynamic programming algorithm that solves the OTA problem in pseudo-polynomial time. Simulation results demonstrate that under the proposed solution, the on-time arrival for all vehicle requests is guaranteed, while also demonstrate the substantial improvements gained in terms of network operation and the experienced travel times, especially during high flow rates.

Chapter 7

Joint route guidance and demand management for real-time control of multi-regional networks

7.1 Introduction

In this thesis, demand management occurs by managing traffic inflow inside a region, e.g., through the route reservation scheme. Against this background, this chapter proposes a Model Predictive Control (MPC) framework that joints the multi-regional route guidance scheme with a novel demand management method. Route guidance is used to minimize network's density imbalances, while demand management is utilized to reduce the conditions that cause congestion. This can be achieved by manipulating vehicle routes (i.e., using route guidance) and/or by instructing a portion of the vehicles to wait at their origin before commencing their journey (demand management).

On these premises, this chapter develops a regional-level (macroscopic) non-linear non-convex formulation to solve the joint route guidance and demand management MPC problem. In general, this chapter proposes several formulations that have been designed to provide accurate and efficient solutions to the original problem with varying properties in solution quality and execution time. One approach to do so, is to involve the development of a Mixed Integer Linear Program (MILP) that yields to tight lower bounds to the optimal solution. Nonetheless, the resulting MILP

formulation is computationally hard to be implemented in practice under real-time constraints due to the MILP complexity. In this direction, this chapter also proposes a novel Linear Programming (LP) MPC formulation that also offers tight lower bounds to the optimal solution. Bearing in mind that through demand management, each region can be operated only with the free-flow regime of the macroscopic fundamental diagram a second LP formulation is derived which provides a feasible but a sub-optimal solution to the original non-linear non-convex MPC problem. The key benefit of both LP formulations is they can be solved accurately and fast with the use of standard LP solvers.

The remainder of this chapter is organized as follows. Section 7.2 describes the regional level system model and Section 7.3 derives the nonlinear MPC formulation of the multi-region RGDM problem, while Sections 7.4 and 7.5 relaxes the problem into a Mixed Integer Linear Program (MILP) and a Linear Program (LP), respectively. Section 7.6 exploits demand management to allow regions to operate only within the free-flow regime, which results in a linear MPC formulation. Section 7.7 presents simulation results to illustrate how the linear MPC formulation produces competitive results compared to other state-of-the-art solutions of higher complexity. Finally, Section 7.8 concludes this chapter.

7.2 System model

7.2.1 Traffic Flow Model

An urban area is partitioned into R homogeneous regions [25], denoted by $r \in \mathcal{R} = \{1, \ldots, R\}$, with traffic dynamics for each region modelled according to the region's MFD as depicted in Fig 7.1 [123]. The traffic parameters of each region r are: the *jam density*, ρ_r^J, the *capacity*, $q_r^C = \rho_r^C u_r^f$, which denotes the maximum outflow of region r (observed at the critical density ρ_r^C), the *free-flow speed* u_r^f, and the *backward congestion propagation speed* $w_r = q_r^C/(\rho_r^J - \rho_r^C)$ [121]. Fig. 7.1 depicts the triangular flow-density MFD diagram which is complemented by the fundamental relationship that the *intended outflow*[1] $q_r(\rho_r(k))$ (veh/h) is equal to the product of density $\rho_r(k)$ (veh/km)

[1]The word "intended" represents the total flow that region r is ready to transfer to its neighboring regions and/or the outside world, if no flow/storage capacity restrictions where applicable from other

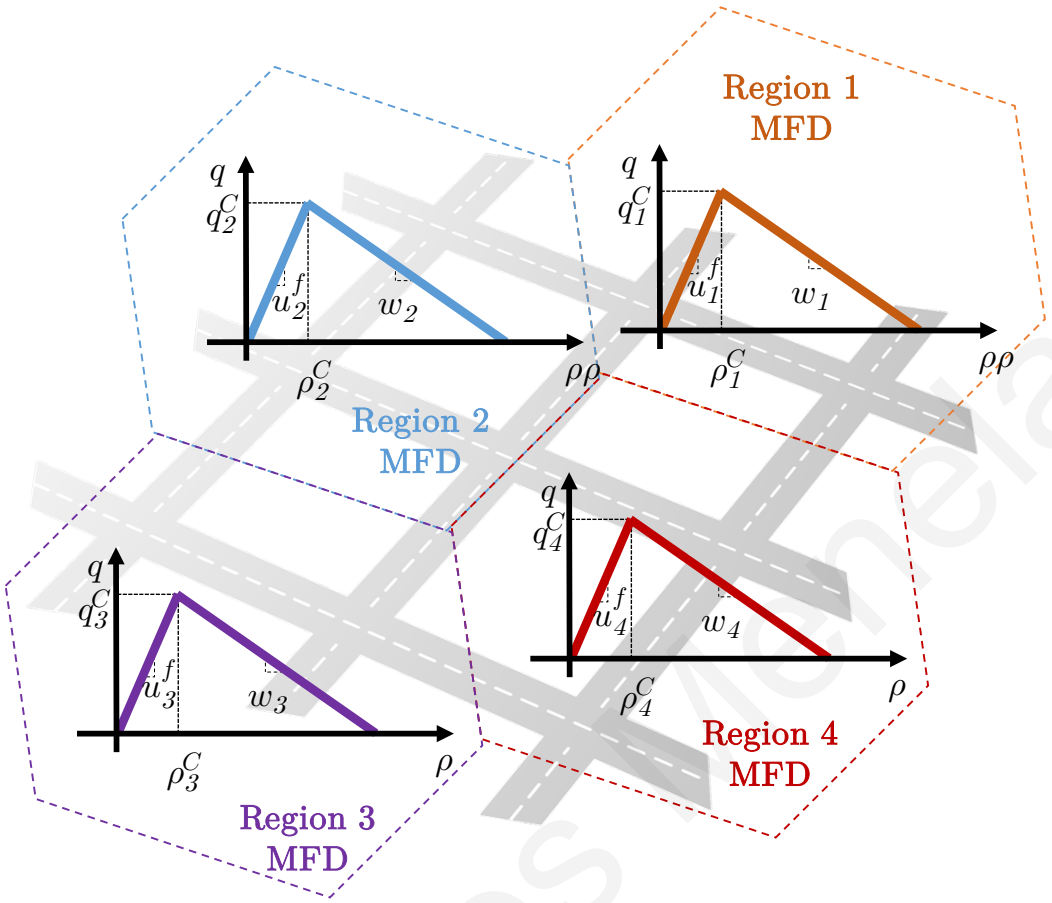


Figure 7.1: A four region network where the outflow traffic dynamics are captured through the regional triangular flow-density MFDs.

and speed $u_r(\rho_r(k))$ (km/h) at each time-step k, i.e., $q_r(\rho_r(k)) = \rho_r(k)u_r(\rho_r(k))$. Note that the variables of the intended outflow and speed are functions of density; henceforth, we suppress the dependency on $\rho_r(k)$ for the sake of simplicity. According to the MFD theory [123], the intended outflow of each region $q_r(k)$, $r \in \mathcal{R}$, can be approximated using the asymmetric unimodal curve of the triangular MFD [121], shown in Fig. 7.1, which is defined as

$$q_r(k) = \begin{cases} \dfrac{q_r^C}{\rho_r^C}\rho_r(k), \text{ if } 0 \le \rho_r(k) \le \rho_r^C, \\ w_r(\rho_r^J - \rho_r(k)), \text{ otherwise.} \end{cases} \tag{7.1}$$

In this work, we assume that the distance travelled by a vehicle inside each region is independent of the origin-destination pair and the drivers' route choice (similar to [64] and [124]) with the parameter L_r (km) denoting the total length of all roads in

regions.

region $r \in \mathcal{R}$.

Let sets $\mathcal{O} \subseteq \mathcal{R}$ and $\mathcal{D} \subseteq \mathcal{R}$ determine the regions considered as the origins and destinations of flows, respectively. Let also $\mathcal{J}_r^- \subseteq \mathcal{R}$ be the set of neighboring regions directly accessible from region $r \in \mathcal{R}$ (i.e., the immediately next region of $r \in \mathcal{R}$) and similarly let $\mathcal{J}_r^+ = \mathcal{J}_r^- \cup \{r\}$, such that:

$$\mathcal{J}_r = \begin{cases} \mathcal{J}_r^+, \text{ if } r \in \mathcal{D} \\ \mathcal{J}_r^-, \text{ otherwise.} \end{cases} \tag{7.2}$$

The *instantaneous external demand* and *admitted external demand*, from region $o \in \mathcal{O}$ to $d \in \mathcal{D}$, during time-step k are denoted by $d_{od}(k)$ and $\tilde{d}_{od}(k)$, respectively. The instantaneous external demand captures the number of new vehicles that request to enter region o towards d. The admitted external demand indicates the number of vehicles that actually enter region o towards d and is restricted by three factors:

1. The physical ability of the region to accommodate more vehicles.

2. The maximum possible demand that can physically enter region $o \in \mathcal{O}$ denoted by D_{od}^{MAX}.

3. Demand management which allows only a portion of the requested vehicles to enter the network; the remaining vehicles wait at their origins (outside the network) until they are admitted.

To keep track of the remaining flows to be served, $D_{od}(k)$ represents the *total external demand* at time-step k defined as follows:

$$D_{od}(k+1) = D_{od}(k) - \tilde{d}_{od}(k) + d_{od}(k), \;\; D_{od}(0) = 0, \tag{7.3}$$

for $k = 1, 2, \ldots$. Furthermore, let the variable $\rho_{rd}(k)$ denote the portion of $\rho_r(k)$, $r \in \mathcal{R}$, that is destined to $d \in \mathcal{D}$ such that

$$\rho_r(k) = \sum_{d \in \mathcal{D}} \rho_{rd}(k). \tag{7.4}$$

Accordingly, let variables $q_{rd}(k)$ and $q_{rjd}(k)$ denote the *intended transfer flow* from region $r \in \mathcal{R}$ to destination region $d \in \mathcal{D}$ and the corresponding flow in region $r \in \mathcal{R}$ destined to region $d \in \mathcal{D}$ that passes through neighbouring region $j \in J_r$, respectively,

defined as

$$q_{rd}(k) = \frac{q_r(k)}{\rho_r(k)}\rho_{rd}(k) = u_r(k)\rho_{rd}(k), \tag{7.5}$$

$$q_{rd}(k) = \sum_{j\in\mathcal{J}_r} q_{rjd}(k), \tag{7.6}$$

$$q_r(k) = \sum_{d\in\mathcal{D}} q_{rd}(k). \tag{7.7}$$

Note that in the case that $r = j = d$, $d \in \mathcal{D}$, then variable $q_{ddd}(k)$ denotes the number of vehicles that exit the network from their destination, termed *exiting vehicles*. Note also that $q_{djd}(k) = 0$, $j \in \{\mathcal{J}_r \setminus d\}$, and hence $q_{dd} = q_{ddd}$.

The intended transfer flow between neighbouring regions $r \in \mathcal{R}$ and $j \in \mathcal{J}_r^-$ is restricted by their inter-boundary capacity, $C_{rj}(\rho_j(k))$, which is the maximum flow that can be exchanged between the two neighbouring regions, for a specific value of $\rho_j(k)$. According to [124], $C_{rj}(\rho_j(k))$ can be defined as:

$$C_{rj}(\rho_j(k)) = \begin{cases} C_{rj}^{\text{MAX}}, \text{ if } \rho_j(k) \le \alpha\rho_j^J, \\ \dfrac{C_{rj}^{\text{MAX}}}{1-\alpha}(1 - \dfrac{\rho_j(k)}{\rho_j^J}), \text{ otherwise,} \end{cases} \tag{7.8}$$

where C_{rj}^{MAX} is the maximum inter-boundary capacity and $\alpha\rho_j^J$ is the point where the inter-boundary capacity starts to decrease with $0 < \alpha < 1$. Considering Eq. (7.8), the value of $q_{rjd}(k)$ depends on the total number of vehicles in region $r \in \mathcal{R}$, while the transfer flow of neighbouring region j relies on its remaining storage capacity; which also depends on the transfer flows from all other regions $s \in \{\mathcal{J}_j \setminus r\}$. Hence, the *actual transfer flow* from $r \in \mathcal{R}$ to $j \in \mathcal{J}_r$, denoted by variable $\tilde{q}_{rjd}(k)$, is defined as

$$\tilde{q}_{rjd}(k) = \min\left(q_{rjd}(k), C_{rj}(\rho_j(k))\frac{q_{rjd}(k)}{\sum_{y\in\mathcal{D}} q_{rjy}(k)}\right). \tag{7.9}$$

The dynamics of the number of vehicles in region $r \in \mathcal{R}$ towards destination $d \in \mathcal{D}$, can be defined as

$$\rho_{rd}(k+1) = \rho_{rd}(k) + \frac{1}{L_r}\tilde{d}_{rd}(k) + \frac{T_s}{L_r}\sum_{j\in\mathcal{J}_r}(\tilde{q}_{jrd}(k) - \tilde{q}_{rjd}(k)), \tag{7.10}$$

where, T_s denotes the simulation time-step that governs the evolution of the regional dynamics as described in (7.10).

7.3 Joint Route Guidance and Demand Management Problem

In this section, we formulate the optimal joint route guidance and demand management problem utilizing the MFD of each region alongside with the regional model dynamics as described in Section 7.2.

7.3.1 Objective function

In order to define our objective function, let variables $S^a(k)$ and $S^b(k)$ be the cumulative number of vehicles that request to enter the network and successfully arrive at their destination, respectively, defined as

$$S^a(k+1) = S^a(k) + \sum_{o \in \mathcal{O}} \sum_{d \in \mathcal{D}} d_{od}(k), \quad S^a(0) = 0, \tag{7.11}$$

$$S^b(k+1) = S^b(k) + T_s \sum_{d \in \mathcal{D}} \tilde{q}_{ddd}(k), \quad S^b(0) = 0. \tag{7.12}$$

for $k = 1, 2, \ldots$. Summing over all time-steps, yields the *Total Time Spent* (TTS) in the system of all vehicles J_{TTS} (veh·h)

$$J_{TTS} = T_s \sum_k (S^a(k) - S^b(k)). \tag{7.13}$$

Note that, the total time spent (TTS) is the sum of the *Total Waiting Time* (TWT) and the *Total Travel Time* (TTT) of all vehicles (TTS=TTT+TWT). The TWT and TTT are defined as the sum of the time that individual vehicles spent waiting at their origin outside the network and travelling inside the network, respectively.

7.3.2 Problem Formulation

To formulate and solve our problem, a Model Predictive Control framework is considered where the control time-step is set equal to the simulation time-step, such that a distinct control action can be taken every T_s time units. We consider that the control and prediction horizons are both equal to N_p, while a new MPC problem is solved every m time-steps. Hence, we solve the l-th MPC problem, $l = 1, 2, \ldots$, for the time horizon $\mathcal{K}_l = \{m(l-1)+1, \ldots, m(l-1)+N_p\}$ and apply to the traffic network

the control actions corresponding to time-steps $\{m(l-1)+1, ..., ml\}$. In our case, the l-th MPC problem aims to select the best values for the indented transfer flows $q_{rjd}(k)$ and admitted external flows $\tilde{d}_{od}(k)$ to minimize the total time spent over the time horizon $\mathcal{K}_l$. The mathematical formulation of the l-th MPC problem is given in (7.14).

$$(\mathrm{P}_1) \quad \min \; J_{TTS}^{MPC}(l) = T_s \sum_{k \in \mathcal{K}_l} (S^a(k) - S^b(k)) \tag{7.14a}$$

s.t. Traffic Dynamics (7.1) – (7.12),

$$\tilde{d}_{od}(k) \leq D_{od}^{MAX}, \; k \in \mathcal{K}_l, o \in \mathcal{O}, d \in \mathcal{D}, \tag{7.14b}$$

$$\tilde{d}_{od}(k) \leq D_{od}(k), \; k \in \mathcal{K}_l, o \in \mathcal{O}, d \in \mathcal{D}, \tag{7.14c}$$

$$0 \leq \rho_r(k) \leq \rho_r^J, k \in \mathcal{K}_l, r \in \mathcal{R}, \tag{7.14d}$$

Variables: $\rho_r(k)$, $\rho_{rd}(k)$, $\tilde{d}_{od}(k)$, $D_{rd}(k)$, $q_r(k)$,

$q_{rd}(k)$, $q_{rjd}(k)$, $\tilde{q}_{rjd}(k)$, $u_r(k)$, $S^a(k)$, $S^b(k)$

In problem P_1, constraints (7.1) - (7.12) define the traffic dynamics modelled according to the triangular MFD. The physical constraints of the external demand inflows are ensured through (7.14b) and (7.14c), such that the external demand inflow is always smaller than the maximum possible external inflow, D_{od}^{MAX}, and the total external demand, $D_{od}(k)$. Constraint (7.14d) simply ensures that the density of each region is within physical limits. The mathematical optimization problem P_1 is a nonconvex NonLinear Program (NLP) due to the presence of the non-affine functions (7.1) and (7.8), and the product of variables in (7.5) and (7.9).

7.3.3 General MPC Framework

The block diagram in Fig. 7.2 describes the general operation of an arbitrary MPC scheme for the solution of Problem P_1. Every m time-steps the external demands for the prediction horizon and the current state of the network are inputted into the MPC controller which computes the best values for the control variables (admitted demands and indented transfer flows) according to a specific MPC scheme for the entire control horizon. Due to possible modelling approximation errors between the considered MPC scheme and the physical plant, the indented transfer flows q_{rjd}

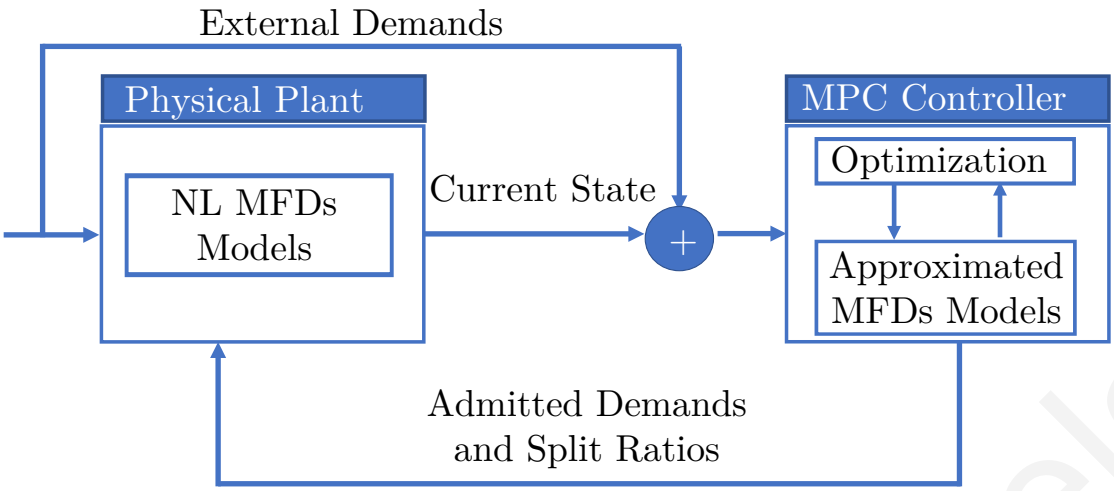


Figure 7.2: Block diagram describing the general operation of an arbitrary MPC scheme for the solution of Problem P_1.

are converted into *split ratios* $a_{rjd}=q_{rjd}/\sum_{d\in\mathcal{D}} q_{rjd}$ which denote the percentage of the indented transfer flow to each of the destination regions. The values of the admitted demands and split ratios for the first m time-steps are used as input to the physical plant which updates the state of the traffic network for the next m time-steps using the nonlinear multi-regional model dynamics given by Eqs. (7.1)-(7.10). Note that the physical plant uses as indented transfer flows the values $a_{rjd}q_{rd}$ instead of the values q_{rjd} produced by the corresponding MPC scheme. In this way the initial state of the next MPC iteration is computed and the procedure is repeated again until the end of the simulation. In practice, the indented transfer flows/split ratios resulting from the considered MPC scheme can be realized using local controllers located at the boundary of each region through traffic signal control, as discussed in [60] and [64].

7.4 MILP reformulation

In this section, the NLP Problem (7.14) shown in Section 7.3 is approximated with a Mixed Integer Linear Program (MILP) that can be solved optimally using standard mathematical programming solvers. To do this, we have to replace all the non-linear constraints (e.g., (7.1), (7.5), (7.8) and (7.9)) with equivalent MILP constraints.

To do so, we first approximate the product of variables in constraint (7.5) with a

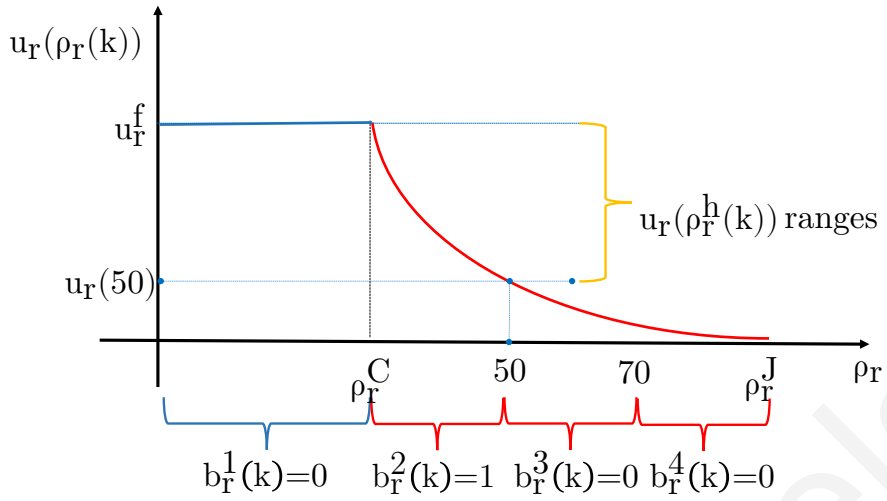


Figure 7.3: The speed function $u_r(k) = q_r(k)/\rho_r(k)$ that is produced when considering a triangular MFD.

set of linear inequalities. This can be achieve by considering segments of the density for the function $u_r(k)$, each of which is defined over a lower and upper bound of speed. Hence, for each region we introduce a set of binary variables $b_r^h(k) = \{0, 1\}$, $h \in \mathcal{H} = \{1, \ldots, |\mathcal{H}|\}$, $r \in \mathcal{R}$ and $k \in \mathcal{K}_l$ which indicate whether $\rho_r(k) \in [\rho_r^{h-}, \rho_r^{h+})$, where ρ_r^{h-} and ρ_r^{h+} is the lower and upper bound of density segment h, as shown in Fig. 7.3. Given that the MFD is composed of two regimes, the spacing is not uniform: $b_r^1(k)$ indicates whether region r is in the free-flow so that $\rho_r^{1-} = 0$ and $\rho_r^{1+} = \rho_r^C$, while the rest indicate the corresponding segment in the congested regime. In total, no more than one set member may be non-zero and positive hence:

$$\sum_{h \in \mathcal{H}} b_r^h(k) = 1, \; r \in \mathcal{R}, \; k \in \mathcal{K}_l \tag{7.15}$$

Subsequently, we introduce a set of new continuous variables $\rho_r^h(k) \in [0, \rho_r^J]$ defined over the constraints derived below:

$$\sum_{h \in \mathcal{H}} \rho_r^h(k) = \rho_r(k), \; r \in \mathcal{R}, \; k \in \mathcal{K}_l \tag{7.16}$$

$$b_r^h(k)\rho_r^{h^-} \leq \rho_r^h(k) \leq b_r^h(k)\rho_r^{h^+}, h \in \mathcal{H}, r \in \mathcal{R}, k \in \mathcal{K}_l \tag{7.17}$$

For each time-step each region can only have one variable $\rho_r^h(k)$ set to be non-zero and equal to $\rho_r(k)$.

Utilizing the above segments of density, we can obtain the lower and upper bounds of the transfer flows (e.g., $q_{rd}^{h^-}(k)$ and $q_{rd}^{h^+}(k)$, $h \in \mathcal{H}$, $k \in \mathcal{K}_l$, $r \in \mathcal{R}$, $j \in \mathcal{J}_r$) as

follows:

$$q_{rd}^{h^-}(k) = \rho_{rd}(k)u_r(\rho_r^{h^+})b_r^h(k), \tag{7.18}$$

$$q_{rd}^{h^+}(k) = \rho_{rd}(k)u_r(\rho_r^{h^-})b_r^h(k), \tag{7.19}$$

where $u_r(\rho_r^{h^+})$ and $u_r(\rho_r^{h^-})$ are the corresponding lower and upper bounds on the speed for density segment h.

Fig. 7.3 depicts a toy example of how the transfer flows can be approximated, with the following parameters: $\rho_r^J = 90$, $\rho_r^C = 30$, $q_r^c = 1800$ and $|\mathcal{H}| = 4$. Hence, we can separate $u_r(k)$ in the following four density segments: $[0, \rho_r^C]$, $(\rho_r^C, 50]$, $(50, 70]$ and $(70, 90]$. For instance, at an arbitrary time-step k, $\rho_r(k) = 40$ veh/km with $u_r^f = 60$ km/h and $u_r(40) = 37.5$ km/h. Then $b_r^2(k) = 1$ and $b_r^1(k) = b_r^3(k) = b_r^4(k) = 0$, with $\rho_r^2(k) = \rho_r(k) = 40$ veh/km and $\rho_r^1(k) = \rho_r^3(k) = \rho_r^4(k) = 0$ with $u_r(\rho_r^{2^+}) = 60$ km/h and $u_r(\rho_r^{2^-}) = 30$ km/h.

In this regard, constraint (7.5) is approximated by constraints Eqs. (7.18) and (7.19) which contain a product of a continuous and a binary variable. Considering that the continuous variable (e.g., $\rho_{rd}(k)u_r(\rho_r^{h^+})$ or $\rho_{rd}(k)u_r(\rho_r^{h^-})$) is bounded below by zero and above by C_{rj}^{MAX} then, equality constraints (7.18) and (7.19) can be equivalently transformed to a set of MILP inequalities using the big "M" notation [125] with $M = C_{rj}^{MAX}$. This can be done considering that the transfer flows are upper bounded by the maximum inter-boundary capacity. In view of the above, these constraints are equivalent to (7.20) and (7.21) that comprise of four MILP constraints as follows:

$$q_{rd}^{h^-}(k) \leq Mb_r^h(k) \tag{7.20a}$$

$$q_{rd}^{h^-}(k) \leq \rho_{rd}(k)u_r(\rho_r^{h^-}) \tag{7.20b}$$

$$q_{rd}^{h^-}(k) \geq 0 \tag{7.20c}$$

$$q_{rd}^{h^-}(k) \geq \rho_{rd}(k)u_r(\rho_r^{h^-}) - (1 - b_r^h(k))M. \tag{7.20d}$$

$$q_{rd}^{h^+}(k) \leq Mb_r^h(k) \tag{7.21a}$$

$$q_{rd}^{h^+}(k) \leq \rho_{rd}(k)u_r(\rho_r^{h^+}) \tag{7.21b}$$

$$q_{rd}^{h^+}(k) \geq 0 \tag{7.21c}$$

$$q_{rd}^{h^+}(k) \geq \rho_{rd}(k)u_r(\rho_r^{h^+}) - (1 - b_r^h(k))M. \tag{7.21d}$$

Considering all the above, constraint (7.5) can be approximate with the following lower and upper bounds on $q_{rd}(k)$:

$$\sum_{h \in \mathcal{H}} q_{rd}^{h^-}(k) \leq q_{rd}(k) \leq \sum_{h \in \mathcal{H}} q_{rd}^{h^+}(k) \tag{7.22}$$

Subsequently, constraint eq. (7.1) can be transformed into a MILP equality considering the new variables $b_r^1(k)$ and $\rho_r^h(k)$ as depicted in constraint eq (7.23).

$$q_r(k) = \Big(\frac{q_r^C}{\rho_r^C} + w_r\Big)\rho_r^1(k) + w_r\rho_r^J(1 - b_r^1(k)) - w_r\rho_r(k). \tag{7.23}$$

Similar to [124] for the case of MILP formulation we omit the inter-boundary capacity constraints (7.8) and (7.9) from the prediction model used in the developed MPC optimization approach described in mathematical Program P_1 , as the effect of the critical capacity is significantly larger than that of the inter-boundary capacity. Furthermore, the work presented in [124] has extensively studied the sensitivity to changes of the inter-boundary capacity value, indicating that MPC schemes are insensitive to the inter-boundary capacities.

Given that we only provide bounds for the flows that passes through neighboring regions, in our model we add the following flow conservation equation within region $r \in \mathcal{R}$:

$$\hat{q}_r(k) = \sum_{d \in \mathcal{D}} q_{rd}(k) \tag{7.24}$$

By doing this, better approximation can be achieved to the final values of the flows through the developed MILP formulation.

In summary, problem P_1 can be transformed into an MILP by replacing Eqs. (7.1) and (7.5) with Eqs. (7.15)-(7.17) and Eqs. (7.20)-(7.24) while omitting Eqs. (7.8) and (7.9). yielding formulation (7.25).

$$\min J_{TTS}^{MPC}(l) = T_s \sum_{k \in \mathcal{K}_l} (S^a(k) - S^b(k)) \tag{7.25}$$

s.t. Constraints: (7.2) – (7.4), (7.6) – (7.7), (7.10) – (7.12), (7.14b) – (7.14d), (7.15) – (7.17) and (7.20) – (7.24).

Variables: $\rho_r(k)$, $\rho_{rd}(k)$, $\tilde{d}_{od}(k)$, $D_{rd}(k)$, $q_r(k), q_{rd}(k)$, $q_{rjd}(k)$, $\tilde{q}_{rjd}(k)$, $u_r(k)$, $S^a(k)$, $S^b(k)$.

Baring in mind that the range of function $u_r(\rho_r)$ is the set of all values given by $u_r(\rho_r)$ for all possible $\rho_r(k)$ which are defined over the close set $[0, \rho_r^J]$ (i.e., $u_r(\rho_r) \in$

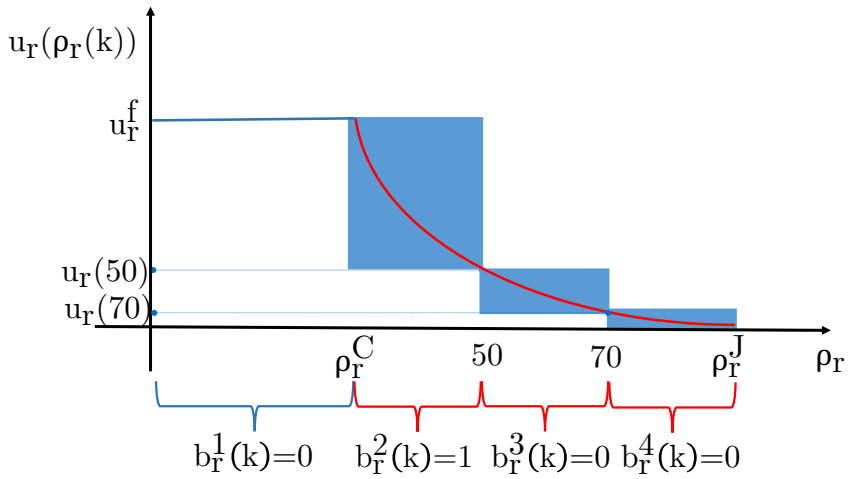


Figure 7.4: The relaxed feasible domain of the speed Function.

$\mathcal{X} = \{0, \ldots, u_r^f\}$). Now by taking segments of density for the function $u_r(k)$, the segments of the range of function $u_r(\rho_r)$ are part of the set (i.e., $\mathcal{X}^{\langle} \subseteq \mathcal{X}$) of all values given by $u_r(\rho_r)$ for all possible $\rho_r(k)$ defined over the set $\rho_r(k) \in [\rho_r^{h-}, \rho_r^{h+})$ (i.e., $u_r(\rho_r) \in \mathcal{X}^{\langle} = \{u_r(\rho_r^{h^+}), \ldots, u_r(\rho_r^{h^-})\}$). Considering the relaxed constraint in (7.22), the range of function $u_r(\rho_r)$ is the set defined within the plane for each considered segment (blue shaded rectangles depicted in Fig. 7.4). In addition, let the optimum solution of the program (P_1) denoted as (P_1^*) be the optimum solution of the MILP reformulation (i.e., $(\hat{P}_1)$) denoted as $\hat{P}_1^*$. Hence, any feasible solution to problem (P_1) is also a feasible solution to its corresponding relaxed MILP program (e.g., $(\hat{P}_1)$) thereby, $\hat{P}_1$ gives a lower bound of the (P_1) program (i.e., $\hat{P}_1^* \leq \hat{P}_1$). Note that, for the first segment, the speed for both cases is equal to u_r^f and thus $P_1^* = \hat{P}_1^*$. Finally, it is worth mentioning that with an increase of the number of segments $|\mathcal{H}|$ that we are use to approximate $u_r(\rho_r)$, tighter bounds on the transfer flows are attained.

7.5 Linear relaxation

In this section, the NLP Problem P_1 is relaxed to a linear programming formulation that can easily be solved using standard mathematical optimization solvers. The developed formulation relaxes all the nonconvex constraints with linear constraints that lie in convex domains that are supersets of the corresponding nonconvex constraint domains. As a result, the obtained solution from this formulation yields lower

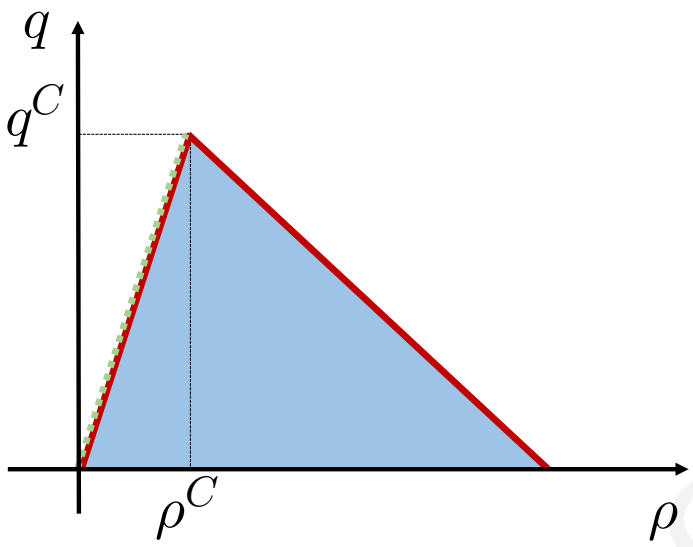


Figure 7.5: The relaxed feasible domain of the triangular MFD.

bounds to the optimal objective value and hence, the particular linear relaxation can be used to derive the optimality gap of any developed solution approach. Next, we derive superset linear constraints for the four nonconvex constraints of Problem P_1, namely, (7.1), (7.5), (7.8) and (7.9).

First, let us consider the triangular MFD relationship between flow, speed and density (7.1). It is true that this constraint can be equivalently written as

$$q_r(k) = \min\left(\frac{q_r^C}{\rho_r^C}\rho_r(k), w_r(\rho_r^J - \rho_r(k))\right). \tag{7.26}$$

Constraint (7.26) can be relaxed by substituting the equality sign "=" with the inequality sign "≤" yielding

$$q_r(k) \leq \frac{q_r^C}{\rho_r^C}\rho_r(k), \tag{7.27}$$

$$q_r(k) \leq w_r(\rho_r^J - \rho_r(k)). \tag{7.28}$$

Notice that constraints (7.27) and (7.28) produce a convex feasibility domain for $\{q_r(k), \rho_r(k)\}$ (shown with the blue shaded area in Fig. 7.5) which is a superset of the nonconvex feasibility domain produced by constraint (7.26) (shown with the red solid line in Fig. 7.5).

Second, let us consider constraint (7.5) which involves the product of two variables. Since $u_r(k) \leq u_r^f$ for all densities $\rho_r(k)$ the constraint (7.5) can be relaxed into

$$q_{rd}(k) \leq u_r^f \rho_{rd}(k). \tag{7.29}$$

which produces a convex feasibility domain for $\{q_{rd}(k), u_r(k), \rho_{rd}(k)\}$ that is a superset of the nonconvex feasibility domain of (7.5).

Finally, constraints (7.8) and (7.9) are handled together. Similar to (7.1), constraint (7.9) can be relaxed into the following two inequalities

$$\tilde{q}_{rjd}(k) \leq q_{rjd}(k) \tag{7.30}$$

$$\tilde{q}_{rjd}(k) \leq C_{rj}(\rho_j(k)) \frac{q_{rjd}(k)}{\sum_{y \in \mathcal{D}} q_{rjy}(k)}. \tag{7.31}$$

Although constraint (7.30) is linear, constraint (7.31) is nonconvex and further relaxation is needed. Summing (7.31) over all $\tilde{q}_{rjd}(k)$ for $d \in \mathcal{D}$ yields

$$\sum_{d \in \mathcal{D}} \tilde{q}_{rjd}(k) \leq C_{rj}(\rho_j(k)), \tag{7.32}$$

which is a relaxed version of (7.31) as individual constraints are always at least as tight as the sum of the associated constraints. In constraint (7.32), $C_{rj}(\rho_j(k))$, defined in Eq. (7.8), can be rewritten as

$$C_{rj}(\rho_j(k)) = \min\left(C_{rj}^{\text{MAX}}, \frac{C_{rj}^{\text{MAX}}}{1-\alpha}(1 - \frac{\rho_j(k)}{\rho_j^J})\right),$$

which has the same form with (7.1). Thus, Eq. (7.32) can further be relaxed into the following two linear constraints

$$\sum_{d \in \mathcal{D}} \tilde{q}_{rjd}(k) \leq C_{rj}^{MAX}, \tag{7.33}$$

$$\sum_{d \in \mathcal{D}} \tilde{q}_{rjd}(k) \leq \frac{C_{rj}^{\text{MAX}}}{1-\alpha}(1 - \frac{\rho_j(k)}{\rho_j^J}), \tag{7.34}$$

for all $k \in \mathcal{K}_l$, $r \in \mathcal{R}$, $j \in \mathcal{J}_r$. Therefore, eqs. (7.8) and (7.9) are relaxed into the linear constraints (7.30), (7.33) and (7.34).

In summary, problem P_1 can be transformed into an LP by replacing Eqs. (7.1), (7.5), (7.8) and (7.9) with Eqs. (7.27)-(7.30) and (7.33)-(7.34), yielding formulation (7.35).

$$\min J_{TTS}^{MPC}(l) = T_s \sum_{k \in \mathcal{K}_l} (S^a(k) - S^b(k)) \tag{7.35}$$

s.t. Constraints: (7.2) – (7.4), (7.6) – (7.7), (7.10) – (7.12), (7.14b) – (7.14d), (7.27) – (7.30) and (7.33) – (7.34).

Variables: $\rho_r(k)$, $\rho_{rd}(k)$, $\tilde{d}_{od}(k)$, $D_{rd}(k)$, $q_r(k), q_{rd}(k)$, $q_{rjd}(k)$, $\tilde{q}_{rjd}(k)$, $u_r(k)$, $S^a(k)$, $S^b(k)$.

The resulting LP relaxation, provides a lower bound to the optimal objective value which can be used to assess the optimality gap of any solution approach for Problem P_1. Although, formulation (7.35) may lead to infeasible solutions due to possible non-satisfaction of the relaxed constraints, a feasible solution can be obtained through the procedure outlined in Section 7.3.3 for the use of the split ratios instead of the indented transfer flows.

7.6 LP Feasible Solution to Problem

In this section, we develop an LP formulation that provides a feasible solution to Problem P_1. Towards this direction, we capitalize on the flexibility offered by demand management to enforce operation of the traffic network in non-congested conditions at all times. This can be achieved because demand management can control the number of vehicles entering the network irrespective of the demand quantity and profile.

Contrary to the previous section where the four nonconvex constraints were relaxed, in this section these constraints are tightened to yield a feasible solution to P_1 which is achieved by enforcing free-flow conditions. Starting with constraint (7.1), operation in the free-flow regime can be guaranteed if we consider the constraint

$$0 \leq \rho_r(k) \leq \rho_r^C, \tag{7.36}$$

which ensures that the density of a region does not exceed its critical density. As a result, (7.1), can be simplified to

$$q_r(k) = (q_r^C/\rho_r^C)\rho_r(k) = u_r^f \rho_r(k), \tag{7.37}$$

which is denoted by the green dashed line in Fig. 7.5. Notice that the green dashed line is a subset of the red solid line which indicates that a potential solution to the problem will yield a feasible solution which, however, may not be optimal because part of the feasibility domain in not used (the part of the red line in the congested regime).

By not allowing any region of the network to enter the congested regime, implies that the vehicles travel with free-flow speed in all regions, i.e., $u_r(k) = u_r^f$. As a result,

constraint (7.5) is simplified to

$$q_{rd}(k) = \rho_{rd}(k) u_r^f. \tag{7.38}$$

To linearise the third nonconvex constraint (7.8) while maintaining feasibility of the solution, we enforce the inter-boundary capacity to always maintain its maximum value (i.e., $C_{rj}^{MAX} \, \forall j \in \mathcal{J}_r$). To achieve this, we further tighten constraint (7.14d) by replacing it with the constraint

$$0 \leq \rho_r(k) \leq \min(\rho_r^C, \alpha \rho_r^J), \tag{7.39}$$

i.e., the region's density should never exceed the critical density and also the point of density where its region's inter-boundary capacity starts to decrease. As a result, constraint (7.8) is simplified to

$$C_{rj}(\rho_j(k)) = C_{rj}^{\text{MAX}}. \tag{7.40}$$

Finally, given the already defined constraints, the nonconvex constraint (7.9) can be simplified to

$$\tilde{q}_{rjd}(k) = q_{rjd}(k), \tag{7.41}$$

$$\sum_{d \in \mathcal{D}} \tilde{q}_{rjd}(k) \leq C_{rj}^{MAX}. \tag{7.42}$$

In this manner, constraint (7.42) enforces demand management to admit lower external demands in order to satisfy both (7.41) and (7.42). Taking all linearisations into account yield the LP formulation (7.43).

$$\min J_{TTS}^{MPC}(l) = T_s \sum_{k \in \mathcal{K}_l} (S^a(k) - S^b(k)) \tag{7.43}$$

s.t. Constraints: (7.2) − (7.4), (7.6) − (7.7), (7.10) − (7.12),

(7.14b) − (7.14c), and (7.37) − (7.42).

Variables: $\rho_r(k)$, $\rho_{rd}(k)$, $\tilde{d}_{od}(k)$, $D_{rd}(k)$, $q_r(k)$, $q_{rd}(k)$, $q_{rjd}(k)$, $\tilde{q}_{rjd}(k)$, $u_r(k)$, $S^a(k)$, $S^b(k)$.

Formulation (7.43) is a linear program that minimizes the total time spent in the system under the enforcement of non-congested conditions. Its solution offers a feasible solution to Problem (P_1). The extensive simulation results that follow indicate that the above formulation can lead to remarkable travel time reductions since formulation (7.43) guarantees that the network will always operate below or at its critical capacity.

Figure 7.6: The simulated urban area consists of 16 regions, four origin regions (1, 4, 11 and 16) and four destination region (2, 8, 9 and 14).

7.7 Performance evaluation

7.7.1 Setup

For the evaluation of the proposed methodologies, a Manhattan-style network topology, shown in Fig. 7.6, is considered as a case study network (i.e., the physical plant) consisting of 16 regions. Region are assumed to have identical well-defined triangular MFDs [123] with parameters: $\rho_r^C = 30$ veh/km, $\rho_r^J = 130$ veh/km, $L_r = 1$ km, $u_r^f = 60$ km/h, $q_r^C = 1800$ veh/h, $C_{rj}^{\text{MAX}} = 2000$ veh/h and $\alpha = 0.25$. The simulation time-step is set equal to $T_s = 60$ s and the duration of the whole simulation experiment is set to $T = 120$ min. For the considered MPC schemes we set $m = 5$ and $N_p = 20$ time-steps, while the corresponding optimization problems are solved using the Gurobi mathematical programming solver [116].

All schemes are evaluated across three scenarios: (i) *light* with average demand 2700 veh/h and range [1500, 5800] veh/h, (ii) *moderate* with average demand 3600 veh/h and range [2000, 7800] veh/h and (iii) *heavy* with average demand 4000 veh/h and range [2300, 8500] veh/h. The demand loading procedure lasts for one hour and varies for different O-D pairs. For each scenario we consider four origin regions (1, 4, 11 and 16) and four destination regions (2, 8, 9 and 14). It is also assumed that the compliance rate of drivers is equal to 100%.

In this setting the performance of the following solution approaches is examined:

- **SP**: In this scheme all vehicles follow the shortest distance path from their

origin to their destination.

- **RG**: An ordinary route guidance MPC scheme with no demand management.
- **LRDM**: The linear relaxation MPC scheme of the joint demand management and route guidance approach based on input obtained using formulation (7.35) in conjunction with the general MPC procedure described in Section 7.3.3.
- **NCDM**: The non-congested feasible LP MPC scheme of the joint demand management and route guidance approach based on input obtained using formulation (7.43) in conjunction with the general MPC procedure described in Section 7.3.3.

To formulate the ordinary route guidance scheme we have to replace the MILP constraints (7.14b) and (7.14c) with the constraint:

$$\tilde{d}_{od}(k) = \min\Big(\frac{(\rho_o^J - \rho_r(k))L_r}{|\mathcal{D}|}, D_{od}(k), D_{od}^{MAX}\Big) \tag{7.44}$$

By doing this, the controller's ability to regulate the external inflows is removed. Hence, constraint (7.44) allows all of the requesting demands to enter unless they are physically restricted by the flow/storage capacity of the region. Note that constraint (7.44) is also a non-linear function that needs to be relaxed to solve by the standard solver. Nonetheless, the current state-of-the-art solvers (e.g., Gurobi [116]) can transform general type of constraints (e.g., min constraint) with built-in functions that utilizes binary variables (MILP programs) as discussed in [126].

7.7.2 Results

Table 7.1 presents the performance of the different solution schemes in terms of the Average Time Spent (ATS), Average Travel Time (ATT) and Average Waiting Time (AWT) at the origin (waiting occurs outside the network) for three demand scenarios. From the table, it is clear that the SP scheme yields poor results as it leads to very large travel times even for the light demand scenario (82.72 min). On the contrary, the other approaches yield optimal performance for the light demand scenario, as they have the same performance with the Ideal Shortest Path (ISP) solution. Nonetheless, as the demand increases the performance of the different schemes diversifies. For the

		Demand Level		
		Light	Moderate	Heavy
ATS (min)	SP	82.72	171.83	431.59
	RG	3.84	4.61	8.11
	LRDM	3.84	3.96	4.58
	NCDM	3.84	3.96	4.09
	ISP	3.84	3.82	3.81
ATT (min)	SP	54.74	95.1	159.1
	RG	3.84	4.6	7.48
	LRDM	3.84	3.82	3.81
	NCDM	3.84	3.82	3.81
	ISP	3.84	3.82	3.81
AWT (min)	SP	27.4	75.71	270.96
	RG	0	0.01	0.63
	LRDM	0	0.13	0.76
	NCDM	0	0.13	0.28
	ISP	0	0	0

Table 7.1: Performance evaluation of different solution approaches for three demand levels: light, moderate and heavy. **ISP** indicates the ideal case where all vehicles follow their shortest distance path with free-flow speed.

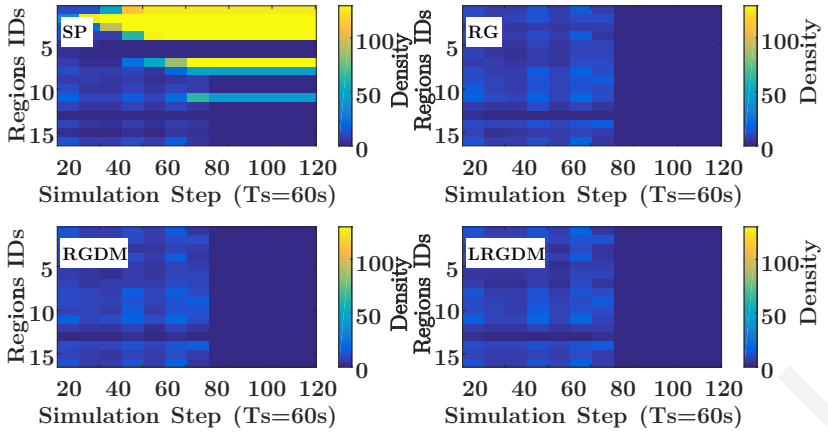
moderate demand scenario, notice that although the RG scheme has only 0.01 min of average waiting time, compared to 0.13 min for the LRDM and NCDM schemes, its average time spent is approximately 15% higher compared to the LRDM and NCDM schemes. Note that although waiting at the origin is not explicitly imposed in the RG scheme, waiting occurs implicitly for vehicles that want to enter a region that is full. For the heavy demand scenario NCDM is the clear winner achieving roughly 10% and 100% better performance compared to the LRDM and RG schemes. Another important observation is that the two demand management schemes achieve this excellent performance by imposing on average less than 20% waiting time at the

origin compared to the travel time. Furthermore, it is interesting to observe that the waiting time of the RG scheme due to congestion is almost identical to the enforced waiting from the LRDM and NCDM schemes, despite having significantly worse performance.

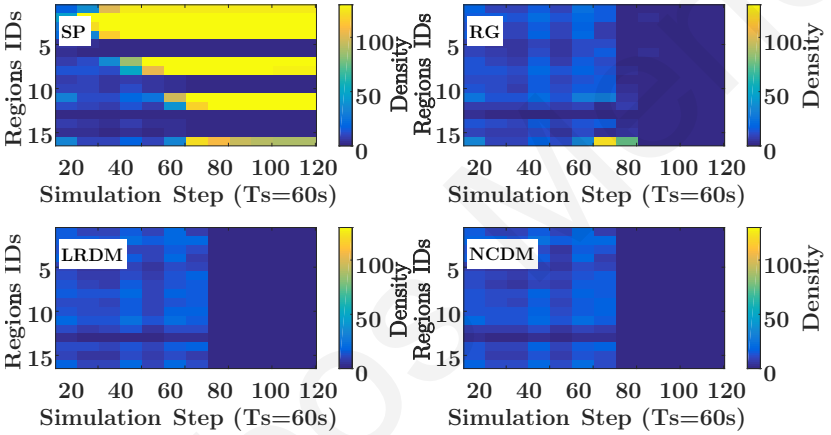
Fig. 7.7 depicts the space-time density diagram for the three demand scenarios. The performance of the three schemes (RG, LRDM and NCDM) for the light demand scenario is almost identical, while it is clear that the SP scheme suffers from severe congestion. On the contrary, for the moderate and heavy demand levels, congestion appears also for the RG scheme. These findings support the notion that the RG scheme can delay the emergence of congestion but cannot eventually avoid it when the demand keeps increasing. On the contrary, it is evident that both LRDM and NCDM schemes can successfully optimize the network's efficiency, as the density of all regions is sustained below the critical density even for the heavy demand scenario. Moreover, it can be observed that despite the fact that we selected only four distinct origin - destination pairs, the optimization problem routes traffic through various paths to utilize all available regions and hence maximize performance.

Figs. 7.8 illustrate the cumulative number of vehicles that request to enter the network (i.e., GeneratedVehs) compared with the number of vehicles that have completed their trip (i.e., ExitVehs) considering all methodologies (i.e., SP, RG, LRDM and NCDM). Interestingly the three route guidance schemes outperform SP in all considered scenarios. Indicatively, with SP more than the two thirds of the vehicles were unable to arrive at their destination within the simulation time. Moreover, during the light demand scenario the three route guidance methods performed equally well as no congestion occurred, but in both moderate and heavy scenarios the demand management schemes outperformed RG. Thus, it is clear that demand management can offer significant reductions in travel times and can enhance the overall network performance.

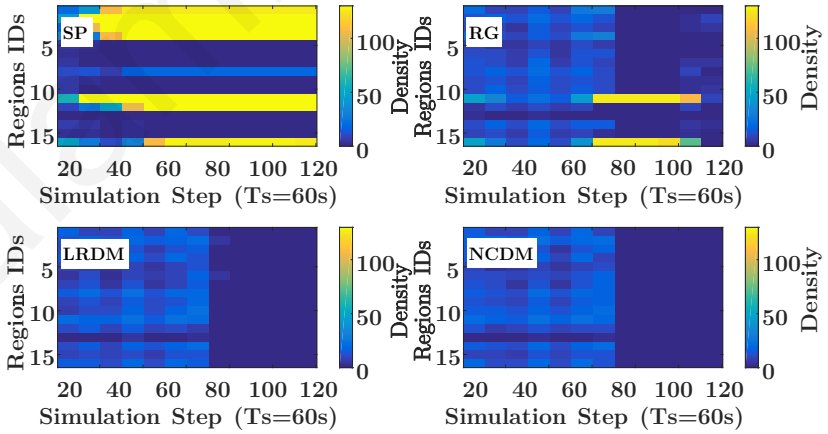
Similar results are obtained in Fig. 7.9 which illustrates the cumulative number of instantaneous external demand (i.e. VehsRequests) compared to the admitted external demand (i.e., VehsEnter). In all three scenarios, the demand management schemes manage to serve the vehicular flows in higher rates than the no demand management schemes. The reason is that, the no demand management schemes



(a)

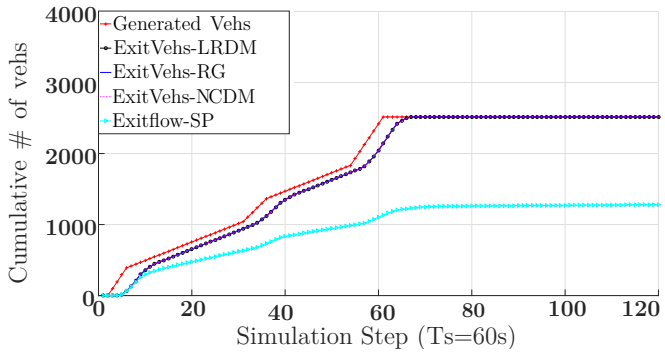


(b)

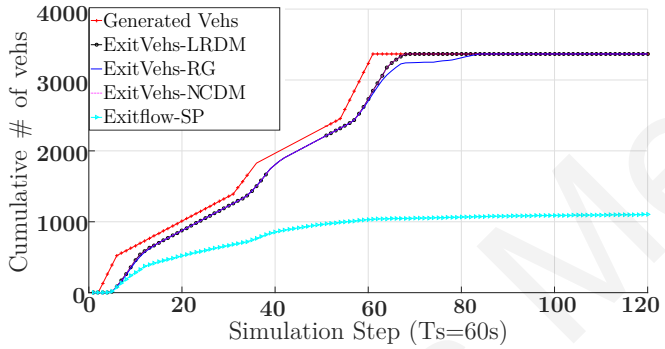


(c)

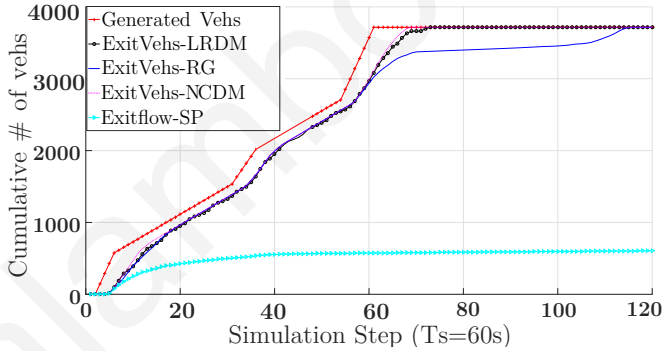
Figure 7.7: The instantaneous density of each region observed at each simulation step (Ts) considering (a) light, (b) moderate and (c) heavy loaded demand scenarios.



(a)



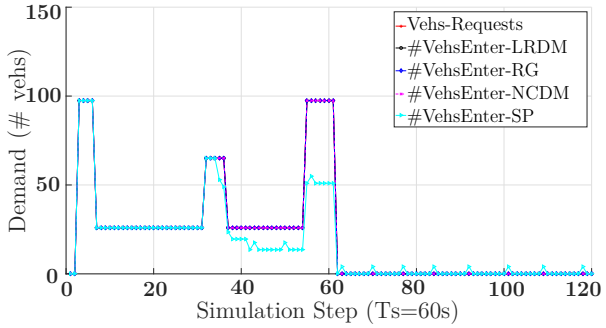
(b)



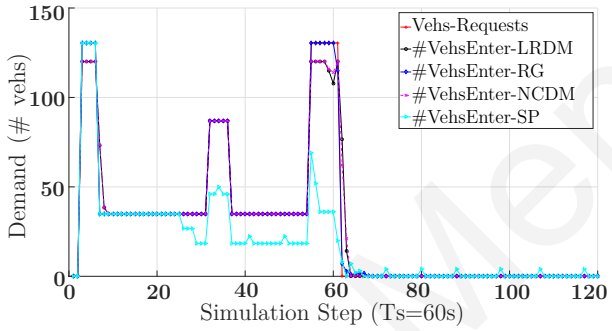
(c)

Figure 7.8: The cumulative summation of the vehicles number that request to enter the network (Generated), that exit the network (Outflow) and their difference (residual) up to each time slot (T_s), considering (a) light, (b) moderate and (c) heavy loaded demand scenarios.

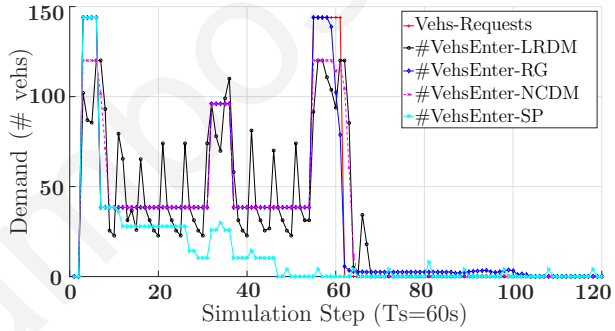
allow all requesting flows to enter unless the originating regions are full. On the other hand, both demand management schemes allow vehicles to enter only in case that regions to be traversed are below the critical density and by doing this, the



(a)



(b)

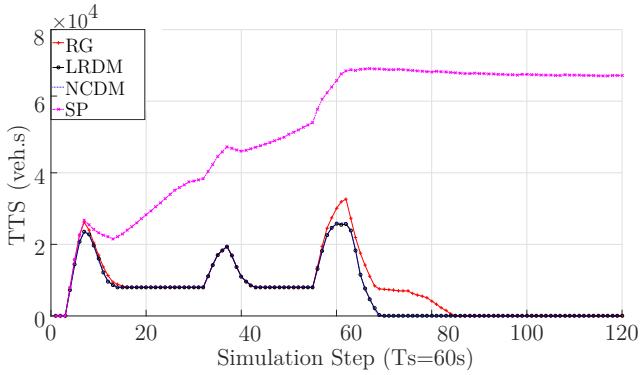


(c)

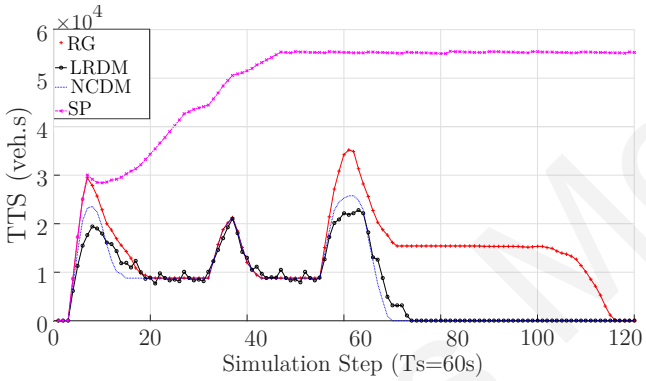
Figure 7.9: The cumulative summation of vehicles number that request to enter the network (Generated) and those that actually entered (granted) for each time-slot (T_s) considering (a) light, (b) moderate and (c) heavy loaded demand scenarios.

outflow is maximized while travel time is minimized.

Fig. 7.10 examines the Total Time Spend in the network (TTS) for the moderate Fig. 7.10 (a) and heavy Fig. 7.10 (b) demand scenarios. From this figure, its clear that the route guidance schemes can improve the TTS metric compared to the levels achieved by the SP scheme as vehicular flows can be re-routed through paths that



(a)



(b)

Figure 7.10: TTS in network for (a) the moderate and (b) the heavy demand levels.

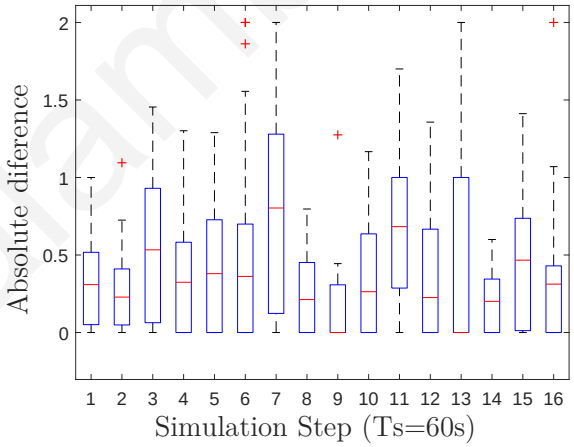


Figure 7.11: The absolute value of the difference between the values of the selected control inputs and their values at their first prediction for each region

minimize the overall time spend. The necessity for this is that, SP tries to develop solutions close to the user optimum in which case congestion occurs and hence the TTS metric is increased exponentially. Similar behavior can be observed from the

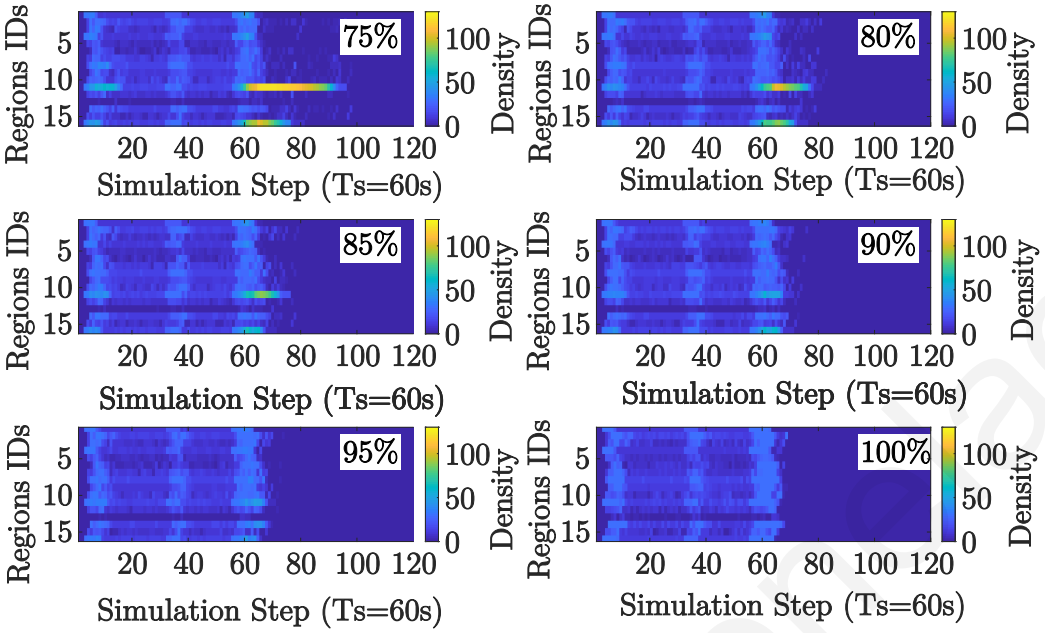


Figure 7.12: The sensitivity of LRDM performance to changes in the percentage of drivers' compliance level considering the heavy loaded demand scenario.

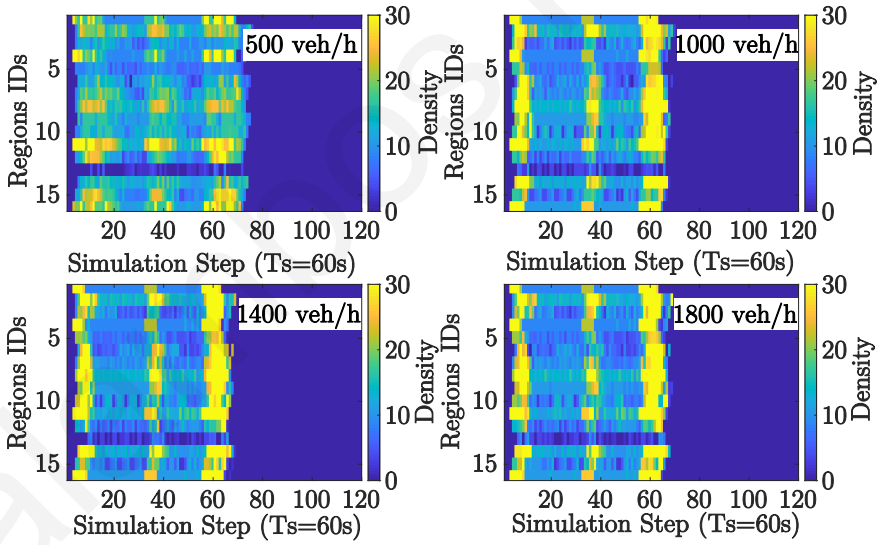


Figure 7.13: The sensitivity of NCDM performance to changes in inter-boundary capacity for the heavy loaded demand scenario.

RG scheme as it tries to force more drivers to use the physical shortest path however, comparing RG with SP it is true that RG can potentially delay the emerge of congestion as it tries to distribute the load evidently through the network. On the other hand, both demand management methodologies try to the optimize social optimum which produces significant reductions in the TTS metric irrespective of the considered demand level, always diminishing the possibility of congestion emer-

gence. Comparing LRDM with NCDM in moderate demand level both demand management strategies can offer similar results, but in heavy demand levels NCDM can offer slightly better results in terms of TTS due to the fact that the LRDM method approximates the optimum solution while the NCDM can offers the optimal solution through a linear program. As shown in Fig. 7.10, (b) for the heaviest demand level the average TTS for SP is 5.775610^6 veh.s for RG is 1.668410^6 veh.s, for LRDM is 8.512910^5 veh.s, and for NCDM is 8.511210^5 veh.s, this mean that the NCDM scheme can reduce by 85% the TTS metric compared with the SP scheme. It is worth to mentioning that for the case of SP scheme in both demand levels due to congestion a significant large portion of vehicular flows can not manage to enter in the network with the above results considering only flows that have managed to enter in the network.

The importance of the selected prediction horizon (i.e., $N_p = 20$) is illustrated in Fig. 7.11 which shows the change in control inputs (i.e,. $n_{rjd}(k)$ and $\tilde{d}_{od}(k)\ \forall\ k = mN_p$) comparing the first predictions with what is actually selected for control action at the current time-step. In the figure, each boxplot determines the absolute value of the difference between selected control inputs (i.e,. $n_{rjd}(k)$ and $\tilde{d}_{od}(k) \forall k = mN_p$) and their predictions made at time-step $k = mN_p - N_p$. The majority of regions have median larger that 0.3, meaning that each applied control input changes on average up to 30% compared with the first predicted value.

The sensitivity of the LRDM performance to the changes in drivers compliance levels is examined in Fig. 7.12 in which the heavy loaded demand scenario is evaluated considering six different drivers percentages of drivers' compliance level (i.e., 70%, 75%, 80%, 85%, 90%, 95% and 100%). Note that in the ideal scenario (i.e., the compliance rate is 100 %) all drivers will opt to follow the waiting intervals provided by the LRDM scheme. Furthermore, the compliance rate is examined only for the case of the LRDM scheme, as a potential 90% disobedience may lead to densities higher than the region's critical value. As expected as compliance levels are reduced then the performance of the LRDM method decreases. Furthermore, it's evident that for compliance level higher than 75%, the LRDM method still outperforms the ordinary route guidance method. On the other hand, in compliance level lower than 70%, LRDM behaves almost similarly with the ordinary route guidance method.

Scenario Number	Average Demand	Optimality Gap LRDM	Optimality Gap NCDM
1	2600 veh/h	0.0%	0.0%
2	2900 veh/h	0.0%	0.04%
3	3200 veh/h	0.0%	0.14%
4	3500 veh/h	0.0%	0.11%
5	3700 veh/h	0.0%	0.57%
6	4000 veh/h	0.0%	0.13%
7	5300 veh/h	0.0%	0.0059%
8	8000 veh/h	0.0005%	3.19%

Table 7.2: The optimality gap of NCDM and LRDM compared to a lower bound of the optimal solution for different demand scenarios.

The sensitivity of the NCDM scheme to changes in the inter-boundary capacity (i.e, $C_{rj}^{\text{MAX}}(\rho_j(k))$) value is examined in Fig. 7.13 in which four different simulations are conducted considering various values of $C_{rj}^{\text{MAX}}(\rho_j(k))$ (i.e., 500, 1000, 1400, and 1800 veh/h) for the heavy loaded demand scenario. Fig. 7.13 illustrates the space-time diagram of the volume of density for the four different simulation runs. Looking into the results presented in Fig. 7.13, it is clear that for inter-boundary capacity rates higher than 1000 veh/h, the performance is almost identical with the highest rate at 1800 veh/h. On the other hand, for the lowest rates of 500 veh/h, the NCDM scheme is fairly insensitive to the value of inter-boundary capacity. Interestingly, as boundary capacity gets lower the performance slightly decays as the loaded demand served in a negligibly longer time.

7.7.3 Optimality Gap

To investigate the optimality of the LRDM and NCDM MPC schemes we have evaluated their performance in comparison to a Lower Bound (LB) of the optimal objective value that can be obtained using formulation (7.35) when the problem is solved once for the entire time horizon, i.e., $\mathcal{K} = \{1, \ldots, T + N_p\}$. For the LRDM and NCDM schemes we consider $m = 5$ and $N_p = 120$ time-steps, similar to previous

experiments. The optimality criterion of choice is the *optimality gap* defined as

$$\text{Optimality Gap} = \frac{J_{\text{TTS}}^{\text{Alg}} - J_{\text{TTS}}^{\text{LB}}}{J_{\text{TTS}}^{\text{LB}}} \times 100\%$$

where $J_{\text{TTS}}^{\text{LB}}$ and $J_{\text{TTS}}^{\text{Alg}}$, Alg = {LRDM, NCDM}, denote the TTS values, according to Eq. (7.13), obtained from the LB solution and the solution from the LRDM and NCDM schemes, respectively.

Table 7.2 illustrates the optimality gap of the NCDM and the LRDM MPC schemes for eight demand scenarios of increasing average value for the same simulation time-step and duration (i.e., T_s = 1 min and T = 120 min). From the results, three important observations can be made. First, the NCDM scheme practically provides optimal results in all considered scenarios. This is a very important result which highlights the fact that a traffic network operating under congestion free conditions will yield excellent results. Second, the obtained lower bounds are tight as indicated by the fact that the optimality gap for the NCDM scheme is equal to zero for all considered cases. Third, the LRDM scheme has excellent performance as the optimality gap is less than 0.6% in almost all cases except from the scenario with the largest demand. In fact it appears that for the LRDM scheme the performance tends to reduce for increasing congestion.

7.8 Summary

This chapter extends the multi-regional route guidance scheme by jointly optimizing the route calculated with demand management. The proposed scheme seeks solutions that schedule flows through shortest travel time paths while at the same time preventing traffic congestion by imposing waiting at the origin nodes when this action benefits the overall network's operation. An MPC formulation is developed which leads to a highly complex non-linear problem while a relaxed linear reformulation is also derived that offers a lower bound solution to the original non-linear problem.

This chapter extends the multi-regional route guidance scheme by jointly optimizing the route calculated with demand management. The proposed scheme seeks solutions that schedule flows through shortest travel time paths while at the same

time preventing traffic congestion by imposing waiting at the origin nodes when this action benefits the overall network's operation. An MPC formulation is developed which leads to a highly complex non-linear problem while a relaxed linear reformulation is also derived that offers a lower bound solution to the original non-linear problem.

Provided that by sustaining each region's density below the critical value its outflow is maximized, the proposed demand management methodology permits us to transform the highly complex non-linear problem to a linear one, without affecting the performance of the proposed scheme. Also, the latter approach can offer a feasible solution to the original problem in real-time in contrasts with the state-of-the-art NLP formulations. Extensive simulation results confirm that the joint route guidance and demand management scheme can lead to substantially better performance compared to the on-demand management option, verifying that the linear case outperforms all other approaches.

Chapter 8

Path-based joint demand management and route guidance for multi-region traffic networks

8.1 Introduction

The non-linear non-convex joint route guidance and demand management problem formulated in Chapter 7 assumes that all paths pass through each region have a constant length independently from their origin-destination pair. However, this assumption is often violated in practice; thus, this chapter reformulates the problem and explicitly define the paths followed for each origin-destination pair. In particular, the road network is partitioned into a number of regions with well-defined Macroscopic Fundamental Diagrams (MFDs) [106] within which a set of predefined paths exists. A path is an ordered sequence of region indices that guide vehicular flows from their origin to their destination. For example, Fig. 8.1 depicts a 4-region network with two origin-destination pairs (O1-D1 and O2-D2) served by five paths.

Given the demand profile of each origin-destination pair, the proposed scheme regulates the inflow rate by which external demands enter the network and splits the inflow among multiple predefined paths associated with each origin-destination pair. The amount of flow allocated to a specific path aims to minimize the total travel time of all vehicles, while the control of the rate by which external flows enter the network (through demand management) aims to alleviate congestion by restricting a

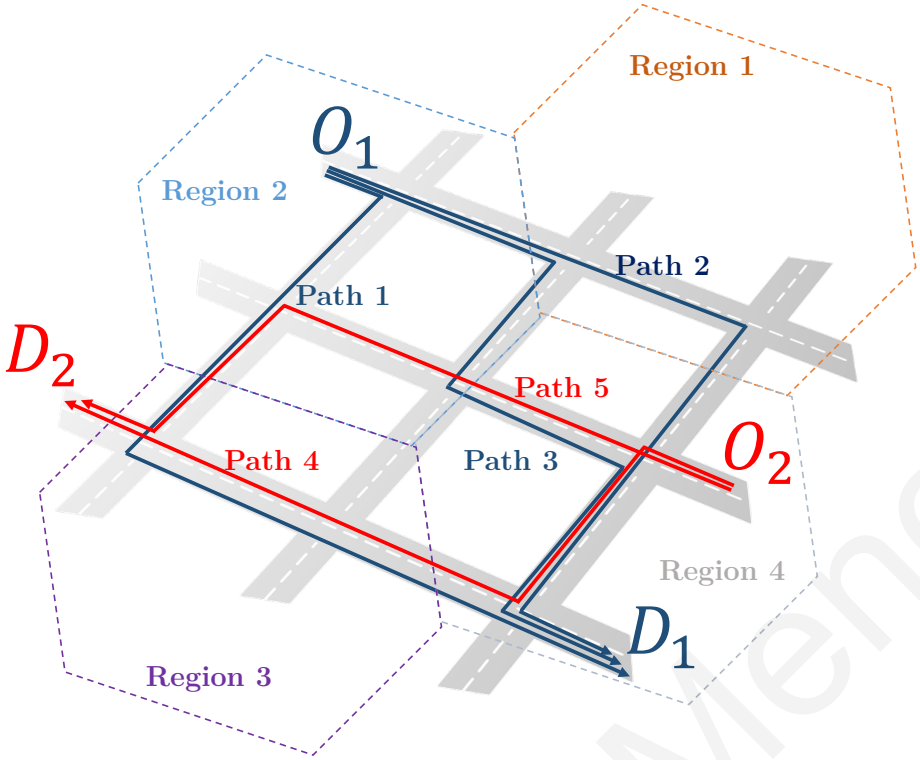


Figure 8.1: The proposed path-based framework.

portion of the inflows at their origins, before entering the network. The novelty of this chapter lies both in modeling and solving the resulting problem under these control measures. In this way, a non-linear and non-convex MPC is formulated to address the path-based joint route guidance and demand management problem. Similarly, as in Chapter 7, an LP formulation for the non-convex MPC problem is developed by relaxing the non-convex constraints. The resulting formulation provides a lower bound to the original problem, which helps in assessing the optimality gap of the developed non-congested solution technique. However, by restricting the density of each region within the non-congested regime, then a second LP formulation can be developed to provide a feasible but sub-optimal solution to the original non-convex problem.

The remaining of this chapter is organized as follows. Section 8.2 presents the path-based regional level model. Section 8.3 derives the non-linear MPC formulation of the problem, while Section 8.4 presents an LP relaxation to the original problem that provides a lower bound to the optimal solution. Section 8.5 develops the proposed solution approach to the original problem which constructs an LP formulation of the original problem by restricting the density of each region to lie

within the non-congested (free-flow) regime at all times. Subsequently, Section 8.6 presents a performance evaluation of the proposed solution approach and finally, Section 8.7 concludes this chapter.

8.2 Traffic flow model

Let an urban area be partitioned into R homogeneous regions, denoted by $r \in \mathcal{R} = \{1, \ldots, R\}$. For instance, the network in Fig. 8.1 is partitioned into four regions. Furthermore, we assumed that all traffic dynamics in each region are defined according to a triangular NFD such that

$$q_r^{out}(\rho_r(k)) = \begin{cases} \dfrac{q_r^C}{\rho_r^C}\rho_r(k), \text{ if } 0 \leq \rho_r(k) \leq \rho_r^C \\ w_r(\rho_r^J - \rho_r(k)), \text{ otherwise,} \end{cases} \tag{8.1}$$

where $\rho_r(k)$ (veh/km) is the vehicle density of region r at time-step k and $q_r^{out}(\rho_r(k))$ (veh/h) is the region's *intended* outflow[1]. Furthermore, ρ_r^C and ρ_r^J are the region's critical and jam densities respectively, while q_r^C is the region's maximum outflow observed at the critical density. In addition, it is true that

$$q_r^{out}(\rho_r(k)) = \rho_r(k)u_r(k)$$

where $u_r(k)$ is the average speed in region $r \in \mathcal{R}$. To complete the model, we assume, the *capacity* $q_r^C = \rho_r^C u_r^f$ where u_r^f denotes the region's free-flow speed and w_r the region's *backward congestion propagation speed* such that $w_r = q_r^C/(\rho_r^J - \rho_r^C)$ [121].

Let sets $\mathcal{O} \subseteq \mathcal{R}$ and $\mathcal{D} \subseteq \mathcal{R}$ denote the regions considered as the origins and destinations of flows, respectively. Also, let $\mathcal{P}$ denote the set of indices of all region-level paths in the network. A path $p \in \mathcal{P}$ is defined by the set containing the sequence of regions to be followed from the origin $o \in \mathcal{O}$ to the destination $d \in \mathcal{D}$ of the particular path, i.e., $P_p = (o, \cdots, r, s, \cdots, d)$; regions $\{r, s\} \in \mathcal{R}$ are intermediate regions of path p such that s is the downstream region of r. The set $\mathcal{P}_r \subseteq \mathcal{P}$ denotes the subset of all paths that pass through region $r \in \mathcal{R}$ and the set $\mathcal{B}_{rs} \subseteq \mathcal{P}$ denotes the subset of paths that pass through region $r \in \mathcal{R}$ and their immediate downstream

[1]By *intended* outflow we mean the region's outflow, assuming that the downstream region(s) have enough capacity to accommodate it.

region is $s \in \mathcal{R}$. Likewise, the set $\mathcal{P}_{od} \subseteq \mathcal{P}$, denotes the set of all paths that start from $o \in \mathcal{O}$ and end at $d \in \mathcal{D}$. The length of the section of path p that passes through region r is denoted by L_{pr}; thus, the total length of paths that pass through region r is given by $L_r = \sum_{p \in \mathcal{P}_r} L_{pr}$.

During every time step k, a flow (external demand) $d_{od}(k)$ (veh/h) requests to enter the network at region $o \in \mathcal{O}$ with destination $d \in \mathcal{D}$ and $\tilde{d}_p(k)$ (veh/h) is admitted in the origin region of path $p \in \mathcal{P}_{od}$. If no demand management is applied, $\tilde{d}_p(k)$ will be limited by the network's physical constraints as:

1. The physical ability of the origin region to accommodate more vehicles.

2. The physical ability of each path in the particular region to accommodate more vehicles.

3. The maximum possible demand that can physically enter path p denoted by D_p^{MAX}.

On the other hand, if demand management is applied, then $\tilde{d}_p(k)$ becomes a decision variable and can take values lower than the network's physical constraints to prevent the occurrence of congestion.

Given an origin-destination pair, the remaining demand that still needs to enter the network is given by

$$D_{od}(k+1) = D_{od}(k) - \sum_{p \in \mathcal{P}_{od}} \tilde{d}_p(k) + d_{od}(k), \tag{8.2}$$

for $k = 0, 1, \cdots$, and with initial condition $D_{od}(0) = 0$.

The density of the section of path p in region r is denoted by $\rho_{pr}(k)$ (veh/km), $p \in \mathcal{P}_r$ such that the density of region r at time step k is given by

$$\rho_r(k) = \frac{\sum_{p \in \mathcal{P}_r} \rho_{pr}(k) L_{pr}}{\sum_{p \in \mathcal{P}_r} L_{pr}}. \tag{8.3}$$

Eq. (8.3) emanates from the fact that the density of the region of an NFD is equal to the total number of vehicles divided by the total length of all roads in the region.

From the NFD of region r, given the density of r, we can obtain the intended outflow of the region, $q_r^{out}(\rho_r(k))$ and also the average speed $u_r(k)$ in region r. Using

this information, and spreading the outflow proportionally among paths, we can also obtain the per path outflow of each region,

$$q_{pr}^{out}(k) = \frac{\rho_{pr}(k)}{\rho_r(k)} q_r^{out}(\rho_r(k)) = \rho_{pr}(k)\, u_r(k) \tag{8.4}$$

From the per-path outflow of region r, we can obtain again the total intended outflow of the region as:

$$q_r^{out}(\rho_r(k)) = \frac{\sum_{p\in\mathcal{P}_r} q_{pr}^{out}(k) L_{pr}}{\sum_{p\in\mathcal{P}_r} L_{pr}} \tag{8.5}$$

On a given path p passing from region r to s, the intended outflow is restricted by the inter-boundary capacity, $C_{rs}(\rho_s(k))$, that determines the maximum flow from region r to downstream region s,

$$C_{rs}(\rho_s(k)) = \begin{cases} C_{rs}^{\text{MAX}}, \text{ if } \rho_s(k) \le \alpha\rho_s^J \\ \dfrac{C_{rs}^{\text{MAX}}}{1-\alpha}(1 - \dfrac{\rho_s(k)}{\rho_s^J}), \text{ otherwise,} \end{cases} \tag{8.6}$$

where C_{rs}^{MAX} is the maximum inter-boundary capacity and $\alpha\rho_s^J$ denotes the point where the inter-boundary capacity starts to decrease with $0 < \alpha < 1$. The inter-boundary capacity will be shared proportionally among all paths that pass from r to s, denoted by $\mathcal{B}_{rs}$; thus, the actual transfer flow from r to s on path p is given by

$$\tilde{q}_{pr}^{out}(k) = \min\left(q_{pr}^{out}(k), C_{rs}(\rho_s(k)) \frac{q_{pr}^{out}(k)}{\sum\limits_{p\in\mathcal{B}_{rs}} q_{pr}^{out}(k)} \right). \tag{8.7}$$

Considering all the above, the vehicle density of each path $p \in \mathcal{P}_r$ in region r is given by

$$\rho_{pr}(k+1) = \rho_{pr}(k) + \frac{T_s}{L_{pr}} \left(\tilde{q}_{pr}^{in}(k) - \tilde{q}_{pr}^{out}(k) \right), \tag{8.8}$$

where T_s is the sampling time step and $\tilde{q}_{pr}^{in}(k)$ is the inflow in region r on path p from the region immediately upstream from r (this can also be obtained from (8.7) by appropriately adjusting the indices). If r is the originating node of path p, then $\tilde{q}_{pr}^{in}(k) = \tilde{d}_p(k)$ which is the new flow admitted into the network on path p during time-step k.

8.3 Path-based Joint Demand Management and Route Guidance formulation

In this section, we present the formulation of the joint Route Guidance and Demand Management MPC scheme that takes into consideration all traffic dynamics presented in the previous section.

Objective function

Let $S^a(k)$ be the cumulative number of vehicles that request to enter the network,

$$S^a(k+1) = S^a(k) + T_s \sum_{o \in \mathcal{O}} \sum_{d \in \mathcal{D}} d_{od}(k). \tag{8.9}$$

Similarly, let $S^b(k)$ be the cumulative number of vehicles that successfully arrive at their destination

$$S^b(k+1) = S^b(k) + T_s \sum_{p \in \mathcal{P}} \tilde{q}_{pd}^{out}(k), \tag{8.10}$$

where index $d \in \mathcal{D}$ denotes the destination region of path p and assuming that the initial values are $S^a(0) = 0$ and $S^b(0) = 0$. Then, our objective function can be defined as the *Total Time Spent* (TTS) in the system of all vehicles J_{TTS} (veh·h) $J_{TTS} = T_s \sum_k (S^a(k) - S^b(k))$ (veh·h).

Note that, the total time spent (TTS) is the sum of the Total Waiting Time (TWT) and the Total Travel Time (TTT) of all vehicles (TTS=TTT+TWT). The TWT and TTT are defined as the sum of the time that individual vehicles spent waiting at their origin outside the network and travelling inside the network, respectively.

To formulate the corresponding MPC problem we assume that a new problem instance is solved every m time-steps. The control time-step (i.e., N_c) is set equal to simulation step and the prediction horizon equal to mN_p time-steps. Then, for the l-th MPC problem solution $l = 1, 2, \ldots$, we define the time horizon $\mathcal{K}_l = \{m(l-1)+1, \ldots, m((l-1)+N_p)\}$. Hence, the l-th formulated problem chooses the per path admitted external flows $\tilde{d}_p(k)$ that minimize the total time spent. The complete model

is derived in problem (P_2) below:

$$(P_2) \quad \min \; J_{TTS}^{MPC}(l) = T_s \sum_{k \in \mathcal{K}_l} (S^a(k) - S^b(k)) \tag{8.11a}$$

s.t. Traffic Dynamics (8.1) − (8.10),

$$\tilde{d}_p(k) \leq D_p^{MAX}, \; p \in \mathcal{P}, k \in \mathcal{K}_l, \tag{8.11b}$$

$$\sum_{p \in \mathcal{P}_{od}} \tilde{d}_p(k) \leq D_{od}(k), \; k \in \mathcal{K}_l, o \in \mathcal{O}, d \in \mathcal{D}, \tag{8.11c}$$

$$0 \leq \rho_r(k) \leq \rho_r^J, k \in \mathcal{K}_l, r \in \mathcal{R}, \tag{8.11d}$$

$$S^a(0) = 0, S^b(0) = 0, \tag{8.11e}$$

Variables: $\rho_r(k)$, $\rho_{pr}(k)$, $\tilde{d}_{od}(k)$, $D_{rd}(k)$, $q_r^{out}(k)$, $q_{pr}^{out}(k)$, $\tilde{q}_{pr}^{out}(k)$, $u_r(k)$, $S^a(k)$, $S^b(k)$

The constraints of (P_2) are due to the traffic dynamics defined in Eq. (8.1)-(8.8) while constraints (8.11b)-(8.11c) ensure that all admitted demands satisfy the physical capacity constraints of each region. Constraint (8.11d) maintains the density of each region within its physical limits and constraint (8.11e) sets the initial conditions of variables $S^a(k)$ and $S^b(k)$. Problem (P_1) is a non-convex Non-Linear Program (NLP) due to constraints (8.1), (8.4), (8.6) and (8.7) that is hard to solve to global optimality by standard mathematical programming solvers. Hence, in the next section we investigate a linear relaxation of the problem that yields in high-quality lower bounds to the optimal solution of Problem (P_2).

8.4 Linear Relaxation to Problem (P_2)

In this section, we present how the non-linear constraints of (P_2) (8.1), (8.4), (8.6) and (8.7) are relaxed to linear constraints.

Due to the triangular NFD form, the intended outflow of a region (8.1) can equivalently be written as

$$q_r^{out}(k) = \min \left\{ \frac{q_r^C}{\rho_r^C} \rho_r(k), w_r(\rho_r^J - \rho_r(k)) \right\}. \tag{8.12}$$

Hence, the intended outflow can be relaxed by bounding $q_r^{out}(k)$ to be smaller than

the two linear terms of the *min* operator in Eq. (8.12), i.e.,

$$q_r^{out}(k) \leq \frac{q_r^C}{\rho_r^C}\rho_r(k) \tag{8.13}$$

$$q_r^{out}(k) \leq w_r(\rho_r^J - \rho_r(k)). \tag{8.14}$$

In this way, it is ensured that

$$q_r^{out}(k) \leq \min\left\{\frac{q_r^C}{\rho_r^C}\rho_r(k), w_r(\rho_r^J - \rho_r(k))\right\}.$$

Constraint (8.4) involves the product of two variables in (P_1): $\rho_{pr}(k)$ and $u_r(k)$. To eliminate this product of variables, notice that the speed of all vehicles is always below the free flow speed, i.e., $u_r(k) \leq u_r^f$. Thus, (8.4) can be relaxed by

$$q_{pr}^{out}(k) \leq u_r^f \rho_{pr}(k), \tag{8.15}$$

which is a linear equality constraint.

In a similar way, the constraint Eq. (8.7) can be written as:

$$\tilde{q}_{pr}^{out}(k) \leq q_{pr}^{out}(k), \tag{8.16}$$

$$\tilde{q}_{pr}^{out}(k) \leq C_{rs}(\rho_s(k))\frac{q_{pr}^{out}(k)}{\sum\limits_{p\in\mathcal{B}_{rs}} q_{pr}^{out}(k)}, \tag{8.17}$$

Taking the sum of $\tilde{q}_{pr}^{out}(k)$ over $p \in \mathcal{B}_{rs}$ in eq. (8.17) yields

$$\sum_{p\in\mathcal{B}_{rs}} \tilde{q}_{pr}^{out}(k) \leq C_{rs}(\rho_s(k)), \tag{8.18}$$

which can be further relaxed (using (8.6)) into

$$\sum_{p\in\mathcal{B}_{rs}} \tilde{q}_{pr}^{out}(k) \leq C_{rs}^{MAX}, \tag{8.19}$$

$$\sum_{p\in\mathcal{B}_{rs}} \tilde{q}_{pr}^{out}(k) \leq \frac{C_{rs}^{\mathrm{MAX}}}{1-\alpha}(1 - \frac{\rho_s(k)}{\rho_s^J}) \tag{8.20}$$

Constraints (8.16), (8.19) and (8.20) are used to relax the model constraints (8.6) and (8.7).

The transformations presented above relax all the non-linear constraints of problem (P_1) to linear ones, resulting in an LP formulation which can be solved very efficiently. Hence, the Problem (P_1) can be transformed into a relaxed LP problem

by relaxing the non-linear constraints (8.1), (8.4), (8.6) and (8.7) with the constraints (8.13)-(8.16) and (8.19)-(8.20). However, the relaxed constraints have a larger feasible constraint set which implies that the solution of the associated LP formulation will produce a *lower bound* to the optimal solution of (P_2). Although a derived solution from this problem may not be a feasible solution to (P_2), it helps in assessing the optimality gap of a feasible solution to problem (P_2). Such a feasible solution is derived in the next section.

8.5 Linear solution Approach to Problem (P_2)

Due to the demand management, it is possible to impose tighter constraints on the inflow to the network and as a result maintain all regions in the non-congested (free flow) regime. Under such conditions, the network's outflow is maximized [16] and the solution of the relaxed problem becomes feasible. To guarantee operation in the non-congestion region, constraint (8.11d) is replaced with the constraint $0 \leq \rho_r(k) \leq \rho_r^C$, i.e., the region's density should never exceed the critical density. As a result, constraint (8.1) is simplified to

$$q_r^{out}(k) = \frac{q_r^C}{\rho_r^C}\rho_r(k) = u_r^f \rho_{pr}(k) \tag{8.21}$$

Similarly, constraint (8.4) can be simplified to

$$q_{pr}^{out}(k) = \rho_{pr}(k)\frac{q_r^C}{\rho_r^C} = \rho_{pr}(k)u_r^f \tag{8.22}$$

Similarly, to guarantee a feasible solution, we should ensure that the inter-boundary capacity constraint (8.6) performs at its maximum value (i.e., C_{rs}^{MAX}). To achieve this, we further tighten constraint (8.11d) by replacing it with the constraint $0 \leq \rho_r(k) \leq \min(\rho_r^C, \alpha\rho_r^J)$, i.e., the region's density should never exceed the critical density and also the point of density where its region's inter-boundary capacity starts to decrease. As a result, constraint (8.6) is simplified to

$$C_{rs}(\rho_s(k)) = C_{rs}^{\mathrm{MAX}} \tag{8.23}$$

Therefore, the constraint (8.7) can be written as:

$$\tilde{q}_{pr}^{out}(k) = q_{pr}^{out}(k), \tag{8.24}$$

$$\sum_{p \in \mathcal{B}_{rs}} \tilde{q}_{pr}^{out}(k) \leq C_{rs}^{MAX} \tag{8.25}$$

Constraint (8.25) involves an inequality which seems that it may violate the feasibility of the Problem (8.11). However, this is not the case as constraint (8.25) ensures that the constraint (8.24) would be never violated. Furthermore, constraint (8.25) enforces demand management to admit lower external demands in order to satisfy both (8.24) and (8.25). Considering the above simplifications, the mathematical formulation of the linear problem can be written as:

$$\min \; J_{TTS}^{MPC}(l) = T_s \sum_{k \in \mathcal{K}_l} (S^a(k) - S^b(k)) \tag{8.26a}$$

s.t. Dynamics (8.2) – (8.3), (8.5), (8.8) – (8.10), (8.11b) – (8.11c)

(8.11e) and (8.21) – (8.25)

$$0 \leq \rho_r(k) \leq \min(\rho_r^C, \alpha \rho_r^J), k \in \mathcal{K}_l, r \in \mathcal{R}, \tag{8.26b}$$

Variables: $\rho_r(k)$, $\rho_{rp}(k)$, $\tilde{d}_{od}(k)$, $D_{rd}(k)$, $q_r^{out}(k)$,

$q_{rp}^{out}(k)$, $\tilde{q}_{rp}^{out}(k)$, $u_r(k)$, $S^a(k)$, $S^b(k)$

Problem (P_2) is a linear programming problem with the objective of minimizing the total time spent considering the linear traffic dynamic constraints in (8.2)-(8.3), (8.5), (8.8), (8.11e), (8.21)-(8.25), (8.26b), and the physical capacity limitations of each path (8.11b)-(8.11c). An important property of Problem (8.26) is that despite the simplifications that are made its solution is also a feasible solution to Problem (8.11). Furthermore, the performance evaluation that follows demonstrates that the above formulation can lead to significant travel time reductions since (P_2) guarantees that the network will always operate below or at its critical capacity.

8.6 Performance evaluation

8.6.1 Setup

This section provides detailed simulation results that investigate the performance of the proposed joint demand-management and route guidance scheme. For the simulations, we consider the case study network as shown in Fig. 8.2), consists of 7 regions, with their traffic dynamics assumed to have identical triangular MFD [123] with parameters: ρ_r^C = 30 veh/km, ρ_r^J = 130 veh/km, v_r^f = 60 km/h and q_r^C = 1800 veh/h. In total 15 paths are used through all regions where at least two paths exist that connect each origin (i.e., regions: 1, 2 and 6) to every destination (i.e., regions: 4, 5 and 7). In addition, α = 0.25 and C_{rs}^{MAX} = 2000 veh/h. The prediction horizon is selected as mN_p = 30 while m = 2 for all particular MPC schemes while, the simulation time-step is T_s = 60 s. In this setting, the performance of the following MPC schemes is examined:

- **RG** The ordinary Route Guidance without demand management considering the MILP formulation.
- **LRDM** The Linear Relaxation of the joint route guidance and Demand Management formulation as presented in Section 8.4.
- **NCDM** The non-congested joint route guidance Demand Management solution as presented in the Problem (8.26).

Note that, to implement the ordinary demand management method we have to replace constraints (8.11b) and (8.11c) of the problem presented in Section 8.4 with the constraint given by:

$$\sum_{p\in\mathcal{P}_{od}} \tilde{d}_p(k) = \min\Big(\sum_{p\in\mathcal{P}_{od}} \frac{(\rho_{op}^J - \rho_{op}(k))L_{op}}{|\mathcal{D}|},\, D_{od}(k),\, D_{od}^{MAX} \Big) \tag{8.27}$$

By incorporating Eq. (8.27), we remove the ability of the MPC to regulate the external inflows, with all requesting demands allowed to enter unless they are physically restricted by each region's flow/storage capacity (jam density). Note that the above constraint is also a non-linear function. Nonetheless, the problem is solved by employing standard solvers (e.g., Gurobi [116]).

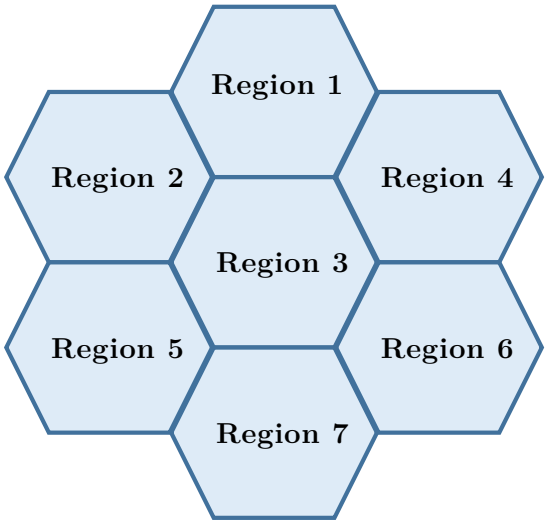


Figure 8.2: Simulated urban area consisted of 7 regions (origin regions: 1, 2 and 6, destination regions: 4, 5 and 7).

The results presented hereafter illustrates the traffic state measurements obtained from a simulated environment in which the control inputs (i.e., $\tilde{d}_p(k)$) of RG, LRDM, and NCDM schemes are applied to the case study network. In the simulated environment, the path-based regional model dynamics given by Eqs. (8.1)-(8.8) are used to represented reality in which traffic states updated every $T_s = 60$ s. All formulated MPC schemes constructed and solved using the Gurobi mathematical programming solver [116], where the simulation environment developed in Matlab. Finally, the drivers' compliance rate equal to 100%, with all simulations performed for 2 hours across 2 loading demand level scenarios (e.g., light and heavy).

8.6.2 Results

In the topmost part of Table 8.1 we depict the Average Time Spent (ATS) of all vehicles in the system while the lower part of Table 8.1 illustrates the Average Waiting Time (AWT) of vehicles before commencing their trips. Under light demand, LRDM and NCDM perform equally well while both of them outperform the RG scheme in terms of ATS. Observing the AWT for light demand scenario, it is obvious that RG incurs no waiting at the origin while for the schemes that include demand management it is clear that some of the vehicles are forced to wait at their origin to achieve lower travel times. In the heavy demand scenario, the average time spent of the RG grows

		Demand Scenarios	
		Light	Heavy
TTT	RG	6.86	255.51
	LRDM	3.52	6.36
	NCDM	3.51	6.32
AWT	RG	0	203.77
	LRDM	0.13	2.97
	NCDM	0.14	2.95

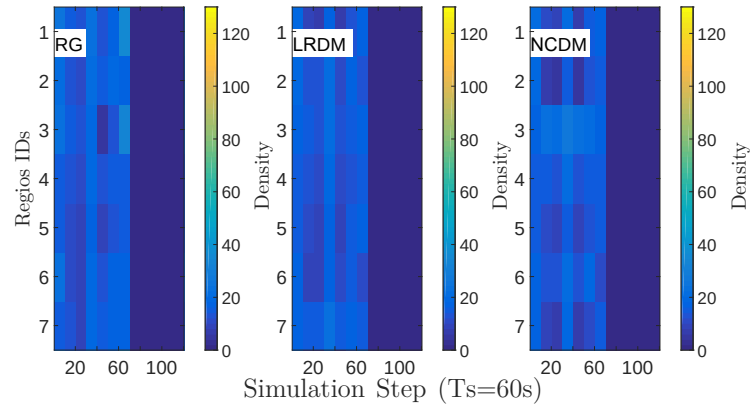
Table 8.1: Total Travel Time (TTT) and Average Waiting Time (AWT) for different demand scenarios.

exponentially since higher demand causes congestion to emerge. Furthermore, it is interesting to observe that the waiting time of the RG method is substantially higher than the waiting enforced by the demand management methods due to congestion that is caused by the high demand scenario.

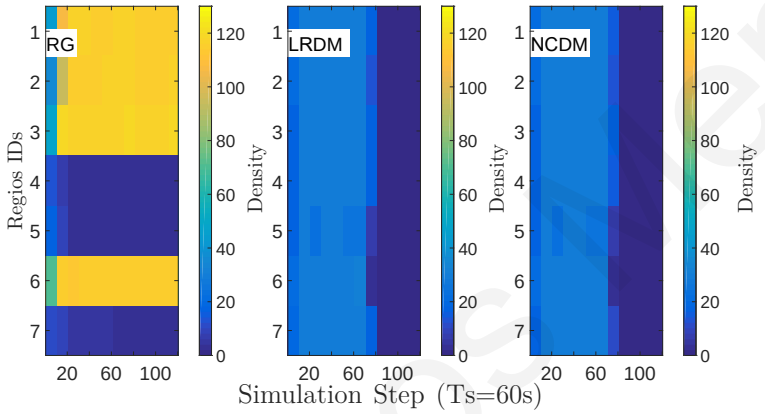
Fig. 8.3 illustrates the space-time diagram of density for the two considered scenarios. In the case of light demand Fig. 8.3 (a) the performance of all three approaches is almost identical as light congestion occurs. On the other hand, in the case of heavy demand scenario both demand management approaches outperform the ordinary route guidance as both of them sustain each region's density around the critical points.

Figs. 8.4 illustrates the cumulative number of vehicles that request to enter the network (generated) with the number of vehicles that have completed their trip (exiting vehicles). As expected, in the light demand scenario all methods work equally well; while the demand management methods achieve slightly better performance. In heavy demand scenarios however, it is clear that both LRDM and NCDM greatly outperform RG as vehicles can be served with higher speeds and implicitly offering substantial travel time reductions.

Similar results are obtained in Fig. 8.5 which illustrates the cumulative number of instantaneous external demand (i.e. VehsRequests) compared to the admitted external demand (i.e., VehsEnter). For the light demand scenario (Fig. 8.5 (a)), the



(a)

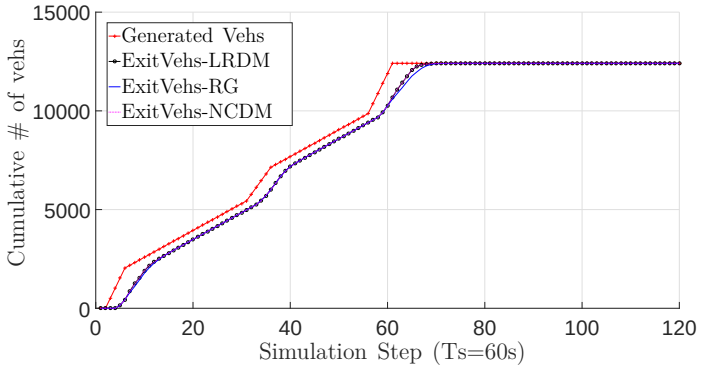


(b)

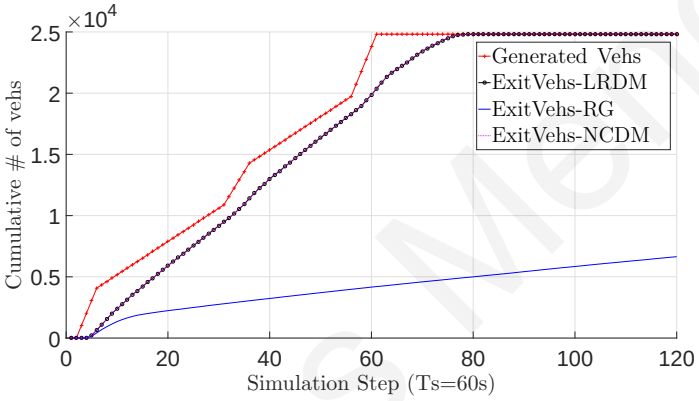
Figure 8.3: The instantaneous density of each region observed at each simulation step (*Ts*) considering (a) light and (b) heavy loaded demand scenarios.

three approaches serve vehicle requests around the same time, with the RG scheme admitting almost instantaneously all vehicles requests. For the heavy demand scenario (Fig. 8.5 (b)), it is evident that demand management is able to server all vehicle requests faster than RG. The reason is that, the no demand management schemes allow all requesting flows to enter unless the originating paths are full, something that produces heavy congestion and hence vehicles can not manage to enter the network.

Fig. 8.6 depicts the Total Time Spent in the network (TTS) for both considered scenarios. From both figures it is clear that the demand management approach can improve the TTS metric compared to RG. The reason is that RG tries to select paths that improve the user optimum in which case the possibility of congestion



(a)

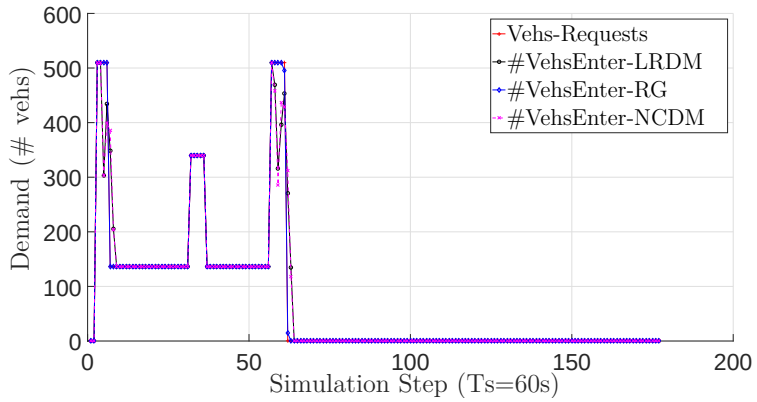


(b)

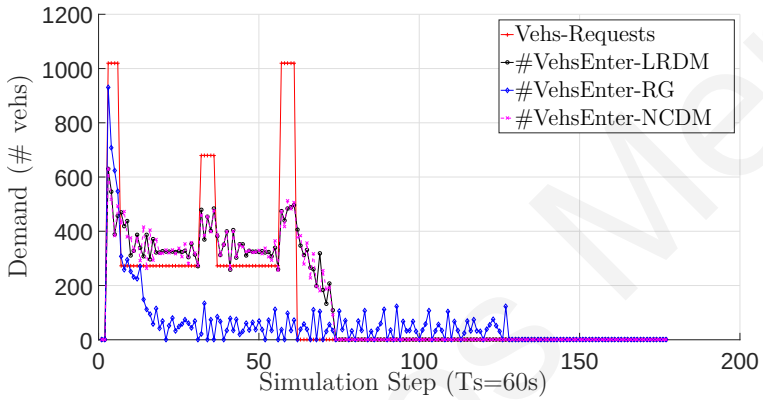
Figure 8.4: The cumulative number of the vehicles that request to enter the network (Generated), that exit the network (Outflow) and their difference (residual) up to each time slot (T_s), considering (a) light and (b) heavy loaded demand scenarios.

to occur is high and thus the TTS metric increases exponentially. On the contrary, when demand management is used, the aim is to improve the system's optimum and produce significant reductions in TTS as no congestion emerges. Comparing LRDM with NCDM in heavy demand scenario, it is evident that NCDM can offer better results in terms of TTS due to the fact that the LRDM solution may results in control actions that maybe not be adequate in the physical model.

To further investigate the performance of LRDM and NCDM formulations we evaluated their objective function under various demand scenarios in which the prediction and control horizons are set to be $mN_p = 120$ and $m = 120$, respectively. For comparison purposes, we consider a total of six demand scenarios considering the same simulation time-step and simulation duration (e.g., $T_s = 60$ min $T = 120$



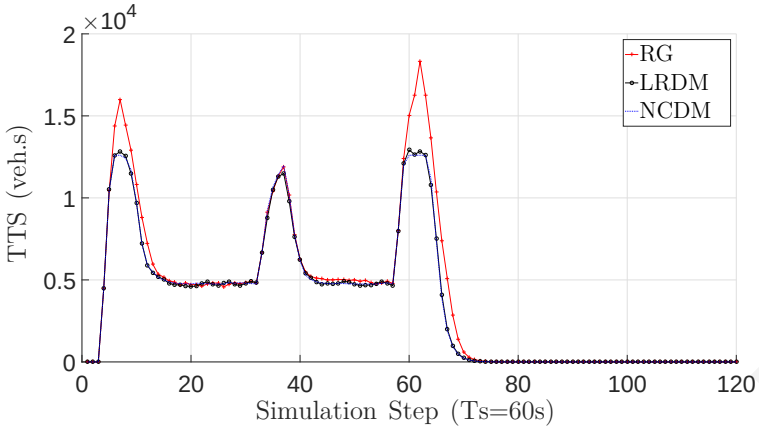
(a)



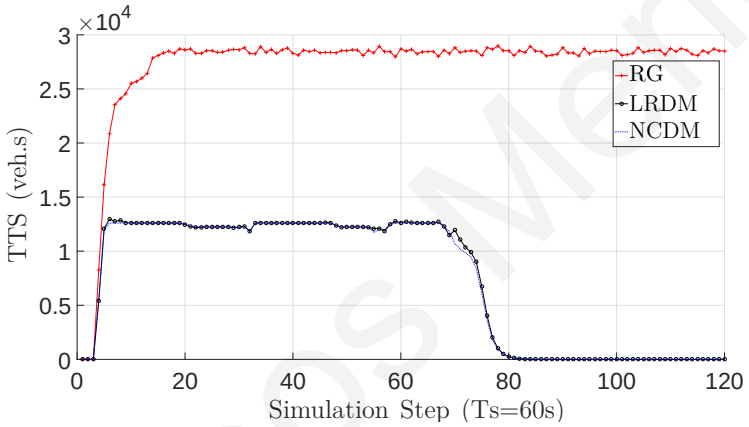
(b)

Figure 8.5: The cumulative sum of vehicles requesting to enter the network (Generated) and those that actually entered (granted) for each time-slot (T_s) considering (a) light and (b) heavy loaded demand scenarios.

min). Note that in this case, we compare the objective functions as obtained from Gurobi and not from the simulated environment. Table 8.2 shows the percentage that the objective functions (Total travel time) of RG and NCDM solutions result in higher values compared to the objective function of the lower bound solution (i.e. LRDM scheme). In table, we consider only solutions that were found in less than 48 hours with, "NSF" meaning that feasible solution is not found within the 48 hours. First it's evident that for the case of the ordinary route guidance a feasible solution hardly can be obtained due to the MILP constraint incorporated form Eq. (8.27) making it impractical for real case scenarios while, on the contrary, both lower bound (i.e., LRDM) and NCDM solutions can converge very fast as they are linear programs.



(a)



(b)

Figure 8.6: TTS in network for (a) the moderate and (b) the heavy demand levels.

Furthermore, Table 8.2 indicates that the objective function of NCDM formulation is almost equal to the objective function obtained by the lower bound solution (i.e., LRDM) no matter the size of the demand that requests to enter the network. These results offer a good indication that the proposed Linear Optimal formulation (as presented in Section.8.5) can provide near-optimal feasible solutions.

8.7 Summary

This chapter extends the multi-regional to path-based route guidance scheme by jointly optimizing the route calculated with demand management method that aims to prevent traffic congestion. Under this model, the assumption of constant trip lengths within each region is removed, making this approach more realistic and able

Scenario Number	Optimality Gap NCDM	RG
1	0.0463%	0.8%
2	0.0211%	6.09%
3	0.0526%	435.1684%
4	0.0857%	NSF
5	0.0955%	NSF
6	0.0587%	NSF

Table 8.2: The optimality gap of NCDM and RG compared to a lower bound of the optimal solution for different demand scenarios.

to handle realistic scenarios in real-time. Furthermore, the proposed demand management methodology defers vehicle departures to avoid congestion and improve overall travel times.

Simulation results confirm the significant gains in travel times that can be achieved through demand management compared to the ordinary route guidance methodology. Besides, the linear relaxation can optimally make decisions at a fraction of the time required by state-of-the-art NLP formulations.

Chapter 9

Critical Density Estimation

9.1 Introduction

In this chapter, we relax the assumption on the requirement of knowing apriori the critical density of a particular region of the road network. The aim is to estimate that value in an online fashion by employing stochastic fluid modeling (SFM). Therefore, this chapter, a single region of the road network is abstracted as a single queue, and the gradient estimator obtained through Infinitesimal Perturbation Analysis (IPA) is employed to **estimate** the critical density of the region (i.e., buffer content). This value can then be used by the RRA algorithms to compute congestion-free routes over *O-D* pairs in the specific region of the road network.

The remainder of this chapter is organized as follows: Section 9.2 presents the system model and the basic flow control problem for the SFM setting of the route-reservation architecture while the performance metrics of the related problem are also mathematically formulated. Section 9.3 derives the IPA estimators for the region's throughput gradients based on the SFM setting. Section 9.4 includes simulation results demonstrating how the SFM-based gradient estimators can be used for the on-line estimation of the critical capacity, showing an approximation method which can be on-line applied to the actual system (not the SFM). Finally, Section 9.5 concludes this chapter and discusses future research directions motivated by this chapter.

9.2 System model and problem statement

9.2.1 Traffic flow model

Consider a *homogeneous* urban road region [25] defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the sets $\mathcal{V}$ and $\mathcal{E}$ represent the road junctions (i.e., $\{v_i, v_j\} \in \mathcal{V}$) and the road-segments (i.e., $(i, j) \in \mathcal{E}$), respectively. Due to the homogeneity of the region, the Macroscopic Fundamental Diagram [121] can describe the macroscopic traffic behaviour using three fundamental parameters: *speed*, $u(t)$ (km/h), *flow* $q(t)$ (veh/h), and *density* $\rho(t)$ (veh/km). Fig. 7.1 depicts a typical flow-density relationship which is comprised of two distinct regimes separated from the *critical density*, ρ_C: 1) the *free-flow regime* where traffic flows at free-flow speed u_f, and 2) the *congested regime* where traffic experiences a speed reduction due to congestion. The flow-density diagram is complemented by the fundamental relationship that the flow is equal to the product of density and speed, i.e., $q(t) = \rho(t)u(t)$. Using this information, one can define other important parameters of the MFD depicted in Fig. 7.1 such as the *capacity* $q_C = \rho_C u_f$ which is the maximum possible flow of the region observed at the critical density, the *jam density*, ρ_J, and the *backward congestion propagation speed* $w = q_C/(\rho_J - \rho_C)$ [121]. Notice that above the critical density ρ_C the outflow of the region decreases [16].

To maximize the flow through the region, current literature controls traffic to regulate the density of the network below or equal to ρ_C, assuming that the parameters of the MFD are known. Such control mechanisms include perimeter control that regulates exogenous traffic entering the network [5] and all exogenous and endogenous traffic control by the aforementioned demand management schemes that are presented in previous chapters.

In this chapter we consider the use of route-reservations to maintain the traffic density of the region below ρ_C. Route-reservations are used to keep track of the cumulative number of arrivals and departures within the region. Let variable $r(t)$ denote the accumulated number of vehicle reservations within the region, and L the total length of all roads in the region. Then, the quantity $n(t)/L$ approximates $\rho(t)$ at time t. Considering the MFD of Fig. 7.1 and the fact that the route-reservation scheme operates within the free-flow regime, it is true that vehicles traverse the entire region with a constant speed equal to u_f. Hence, vehicle l entering the region

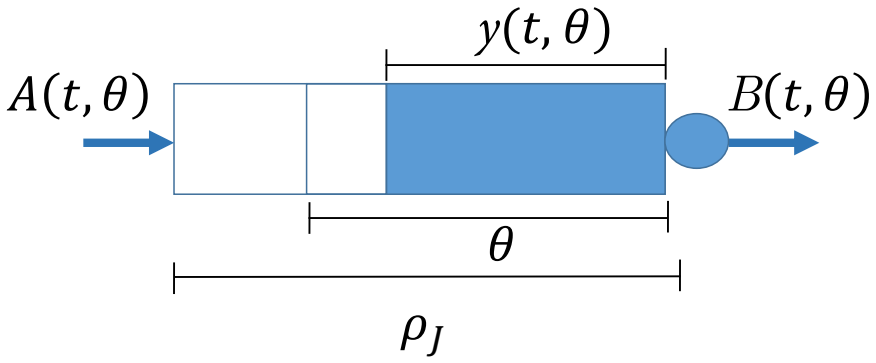


Figure 9.1: The corresponding Stochastic Fluid Model (SFM) of the considered network.

at time t remains within the region up to time $t + t_l$ where $t_l = L_l/u_f$ denotes the travel time and L_l the route length of vehicle l. Hence, the region is denoted as *admissible* if a vehicle l entering at time t can traverse the region without making the accumulated reserved density larger than the critical density for the entire traversing period. Hence, the *admissibility state* $x(t)$ can be defined as:

$$x(t) = \begin{cases} 1, \text{ if } n(t+k)/L \leq \rho_C, \ \forall\, k \in [0,\ t_l] \\ 0, \text{ otherwise} \end{cases} \tag{9.1}$$

Therefore, under the route-reservation scheme vehicles are allowed to reserve routes only during time periods where $x(t) = 1$ to ensure that the region never enters the congested regime.

Contrary to previous literature assuming known MFD parameters, this chapter aims to estimate the critical density by maximizing the outflow of the region. Next, it is described how a Stochastic Fluid Model can be used to represent the traffic network and formulate the investigated problem.

9.2.2 Stochastic fluid model representation

The traffic flow model under consideration can be represented as a Stochastic Fluid Model (SFM) based on continuous fluid-flow dynamics characterized by a set of stochastic processes defined on a common probability space $(\Omega, \mathcal{F}, \mathcal{P})$ [127, 128].

As shown in Fig. 9.1, the road network can be represented by a fluid-storage queue with finite density (content) ρ_J with a single-server to determine the traversal time (service time) of vehicles within the region. The control parameter of interest is θ which denotes the maximum queue size allowed within the queue. Parameter θ is

regulated using some control mechanism (in our case route reservations). Parameter θ aims to estimate the critical density in order to maximize the outflow of the queue. According to the MFD, for $\theta > \rho_C$ the region is over-utilized resulting in a reduction of the outflow as the region experiences congestion. On the contrary, for values of $\theta < \rho_C$ the region is underutilized, also resulting in a reduction of outflow. Hence, the aim is to define a strategy that changes online the value of θ in order to operate as close as possible to the critical density of the system that maximizes the outflow.

Let $y(t, \theta)$, $A(t, \theta)$ and $B(t, y(t, \theta))$ denote the SFM state (queue content), arrival rate[1] (inflow) and departure rate (outflow) at time t, respectively.

The arrival rate of vehicles depends on θ and is given by

$$A(t, \theta) = \begin{cases} a(t), & \text{if } n(t) < \theta \\ 0, & \text{if } n(t) \geq \theta \end{cases} \tag{9.2}$$

where, the variable $a(t)$ denotes the vehicle arrival process which is a time-varying and unknown function independent of θ. According to (9.2), when the number of reservations reaches the parameter θ (which should approximate the critical density of the region) the inflow is set to zero so that no more vehicles to enter, until $n(t) < \theta$. Here, it is assumed that the reservations are consistent with the actual state of the region ($y(t, \theta) = n(t)$). Although, this is not generally true due to the stochastic nature of traffic, it is a reasonable assumption in light of the emergence of connected and automated vehicles.

The departure rate $B(t, y(t, \theta))$ depends on the MFD; when the density exceeds ρ_C the function $B(t, y(t, \theta))$ changes from a linear increasing function (free-flow regime) to a linear decreasing function (congested regime). Hence, $B(t, y(t, \theta))$ is defined as

$$B(t, y(t, \theta)) = \begin{cases} u_f y(t, \theta), & \text{if } \ y(t, \theta) < \rho_C \\ w(\rho_J - y(t, \theta)), & \text{if } y(t, \theta) \geq \rho_C \end{cases} \tag{9.3}$$

Notice from Eq. (9.3) that the departure rate is significantly affected by the instantaneous density in two ways: (a) when parameter θ overestimates ρ_C, undesirable vehicle delays are produced that further exacerbate congestion conditions, and (b) when θ underestimates ρ_C the region is underutilized leading to lower outflow rates.

[1]Consistent with the proposed architecture a central entity is responsible to schedule vehicles according to the described route-reservation scheme. In this way, the arrival rate is controlled to ensure that $y(t, \theta) \leq \theta$.

The queue content is determined by the following differential equation:

$$\dot{y}(t,\theta) = \begin{cases} 0, \text{ if } y(t,\theta) = 0 \,\&\, A(t,\theta) = 0, \\ 0, \text{ if } y(t,\theta) = \theta, \\ A(t,\theta) - B(t, y(t,\theta)), \text{ otherwise,} \end{cases} \tag{9.4}$$

with the initial condition that $y(0) = y_0$, with y_0 known. Here, it is assumed for simplicity that $y(0) = 0$. Note that according to Eq. (9.4) whenever $y(t,\theta) > 0$ a non-zero flow rate should be observed. Moreover, for the case $y(t,\theta) = \theta$, it may be true that $A(t,\theta) = B(t, y(t,\theta)) \neq 0$ such that $\dot{y}(t,\theta) = 0$. In addition, we make the technical assumption that $a(t) \geq -\epsilon$ where ϵ is a small positive number. This assumption is needed to make sure that the queue becomes empty at a finite time (and does not go to zero asymptotically). For practical systems, this assumption does not have any impact, since $a(t) \geq 0$ and empty periods are always observed.

The above SFM setting can be viewed as a hybrid system, with the time-driven dynamics described by Eq. (9.4) and with event-driven dynamics denoted by the region's full and empty periods. Hence, the region's operation can be determined with a Stochastic Hybrid Automaton (SHA) as depicted in Fig. 9.2 which consists of three (3) modes. This model is similar to the one used in [99] (single buffer case), but different as the inflow and outflow rates depend on the parameter θ. Let the time interval $[0, T]$; the region operation can be determined by the set of events $E = \{e_1, e_2, e_3, e_4\}$ defined as:

$e_1 : y(t,\theta) = \theta$, queue reaches capacity.

$e_2 : y(t,\theta) = 0$, queue becomes empty.

e_3 : the sign of $A(t,\theta) - B(t, y(t,\theta))$ changes from positive to negative and queue content ceases to be full.

e_4 : the sign of $A(t,\theta) - B(t, y(t,\theta))$ changes from negative to positive and the queue content ceases to be empty.

All events whose occurrence time depends on the parameter θ are called *endogenous events*, while all other that are independent of the parameter θ are referred to as *exogenous events*.

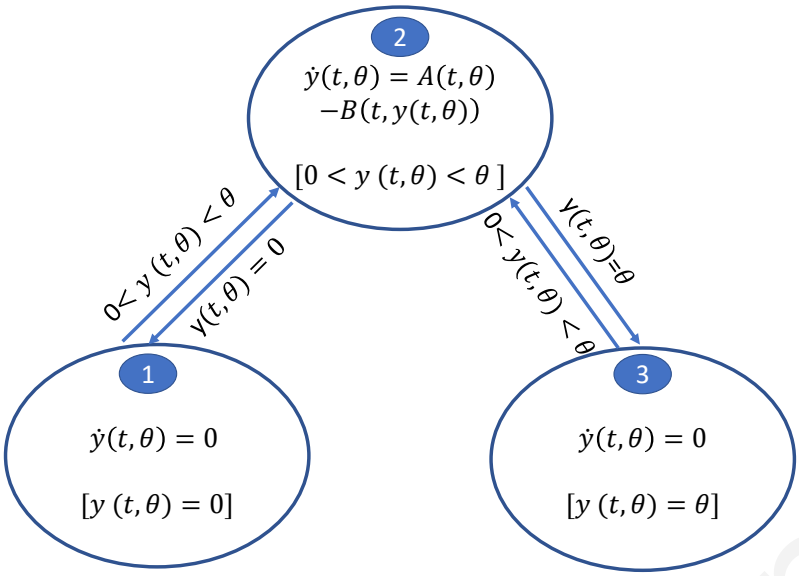


Figure 9.2: The Stochastic Hybrid Automaton model.

9.2.3 Problem statement

As mentioned earlier, we seek to estimate the critical density which by definition, is the density that maximizes the average outflow of the region $W(t,\theta)$ over the interval $[0,T]$, defined as follows:

$$W_T(t,\theta) = \frac{1}{T}\int_0^T B(t,y(t,\theta))dt \tag{9.5}$$

Thus the critical density will be approximated by the control parameter θ that will maximize the outflow following the solution of the optimization problem:

$$\max_{\theta} J(t,\theta) = E[W_T(t,\theta)] \tag{9.6}$$

In the next section we employ the IPA method to determine the best value for the control parameter θ^* in an online fashion.

9.3 Infinitesimal Perturbation Analysis

Infinitesimal Perturbation Analysis review

Let $v_k(\theta)$ denote the occurrence times of k-th event, then the time derivative of queue content (i.e., $y(t,\theta)$) and event occurrence times (i.e., $v_k(\theta)$) with respect to θ can be expressed as:

$$y'(t,\theta) = \frac{dy(t,\theta)}{d\theta} \quad v'_k(\theta) = \frac{dv_k(\theta)}{d\theta} \tag{9.7}$$

Let k denote the k-th interval $[v_k, v_{k+1}) \in T$ within which the dynamics of $y(t,\theta)$ are fixed representing the right-hand-side expression of Eq. (9.4). If the SHA is in mode 2, then the queue content at time $t \ \forall\, t \in [v_k, v_{k+1})$ is formulated as:

$$\begin{aligned} y(t,\theta) &= y(v_k,\theta) + \int_{v_k}^{t} \dot{y}(\tau,\theta)\, d\tau \\ &= y(v_k,\theta) + \int_{v_k}^{t} (A(\tau,\theta) - B(\tau, y(t,\theta)))\, d\tau \end{aligned} \tag{9.8}$$

If the SHA is in modes 1 or 3, then

$$y(t,\theta) = y(v_k,\theta) \quad \forall t \in [v_k, v_{k+1})$$

As above, taking the derivatives with respect to θ and let $t = v_k^+$ the boundary initial condition can be obtained as:

$$y'(v_k^+) = y'(v_k^-) + [\dot{y}(v_k^-,\theta) - \dot{y}(v_k^+,\theta)]v'_k \tag{9.9}$$

Furthermore, taking the derivatives with respect to t in Eq. (9.4) for all $t \in [v_k, v_{k+1})$:

$$\frac{\partial}{\partial t} y'(t,\theta) = \frac{\partial \dot{y}(t,\theta)}{\partial y} y'(t,\theta) + \frac{\partial \dot{y}(t,\theta)}{\partial \theta} \tag{9.10}$$

As mentioned earlier, the derivative with respect to θ of each event occurrence time (i.e., v'_k) depends on the type of event that occurs. Hence, a discrete time transition that is independent from θ is an exogenous event with $v'_k = 0$. Otherwise, if event depends on the control parameter θ, a continuously differentiable function $g_k : \mathcal{R}^n \times \Theta \longrightarrow \mathcal{R}$ exist such that $v_k = \min\{t > v_{k-1} : g_k(y(t,\theta),\theta) = 0\}$ (this function constitutes the guard function [128]). Now, taking the derivatives with respect of θ we obtain

$$v'_k(\theta) = -\Big[\frac{\partial g_k}{\partial y}\dot{y}(v_k^-,\theta)\Big]^{-1}\Big(\frac{\partial g_k}{\partial \theta} + \frac{\partial g_k}{\partial y} y'(v_k^-)\Big) \tag{9.11}$$

Proof of the above expressions (Eq. (9.9) - (9.11)) can be found in [128].

Infinitesimal Perturbation Analysis

Our solution approach is based on the aforementioned IPA analysis that is used to estimate the gradient of our performance metric e.g., throughput Eq. (9.13) which

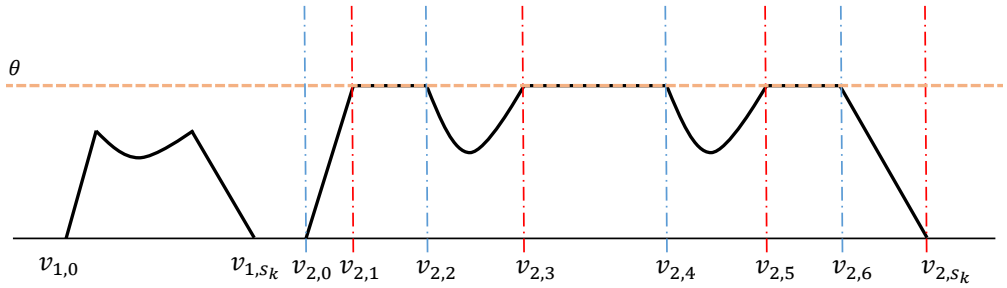


Figure 9.3: A typical sample path of the queue's content.

in turn is employed within a stochastic approximation based algorithm in order to converge towards to the maximum throughput. In this manner we are interested in estimating the $\frac{d}{d\theta}J(t,\theta)$ through an iterative scheme of the form

$$\theta_{k+1} = \theta_k - hH_k(\theta_k, \omega^{SFM}) \tag{9.12}$$

where h is a constant value step-size and $H_k(\theta_k, \omega^{SFM})$ is an estimate of $\frac{d}{d\theta}J(t,\theta)$ derived on-line and is based on information of its sample path as depicted in Fig. 9.3.

The related sample path consists of time intervals over which $y(t,\theta) > 0$, called Non-Empty Periods (NEPs), followed by intervals where $y(t,\theta) = 0$, called Empty-Periods (EPs). The k-th NEP period starts at $v_{k,0}$ and ends at v_{k,s_k} where $k = 1, 2, ..., N_T$ and $|N_T|$ denotes the number of NEPs in time-interval T. On that premise some of NEPs may also contain some periods that the system is full at its capacity (FPs) that attain during the interval $[v_{k,2j-1}, v_{k,2j}]$ i.e., $j = 1, \ldots, \frac{s_k-1}{2}$. Note that, the even index $(2j)$ represents the ending time of each particular FP.

Even though the start of a Non-Empty Period looks like an endogenous event (rates $A(t,\theta)$ and $B(t, y(t,\theta))$ generally depend on θ, at the specific time, they are independent of θ. This can be justified combining equations Eq. (9.3) and (9.4) as the $y(t,\theta) = 0$ only if $A(t,\theta) = B(t, y(t,\theta)) = 0$ and according Eq. (9.3) the queue content switches to $y(t,\theta) > 0$ only when the inflow changes from $A(t,\theta) = 0$ to $A(t,\theta) = a(t) > 0$ which is independent of θ. Considering, the SHA in Fig. 9.2 the transition from mode 2 to 3 is the result of event e_1 which is dependent on θ (endogenous) while the opposite direction (modes 3 to 2) is due to an e_3 event which is independent from θ (exogenous). Considering that during mode 3 $B(t, y(t,\theta))$ is maintained constant and thus the e_3 event occurs only with a decrease of the inflow

rate $A(t,\theta) = a(t)$ which is independent from θ. The transition from 2 to 1 (e_2) is considered an endogenous event as its dependent on the queue content. Note that, the times that endogenous events occur indicated with red dashed lines in Figure 9.3.

Using the above notation, the network's outflow (see Eq. (9.5)) can be rewritten as:

$$\Omega_T(t,\theta) = \frac{1}{T}\sum_{k=1}^{N_t}\omega_k = \frac{1}{T}\sum_{k=1}^{N_t}\int_{v_{k-1,s_k}}^{v_{k,s_k}} B(t, y(t,\theta))dt \tag{9.13}$$

where ω_k is the outflow during the kth NEP. Taking derivatives with respect to θ and observing that all EPs are independent from θ then the required IPA gradient derivative $\frac{d}{d\theta}\Omega_T(t,\theta)$ of the Eq. (9.5)

$$\frac{\partial\Omega_T(t,\theta)}{\partial\theta} = \frac{1}{T}\sum_{k=1}^{N_t}\frac{d\omega_k}{d\theta} \tag{9.14}$$

The IPA tries to evaluate these derivatives as a function of the observable sample path quantities using similar framework establish in [99, 128]. In this way, the derivation of the IPA derivatives requires some mild assumptions in order to guarantee the existence of derivatives as follows:

1. $A(t,\theta) < \infty$ and $B(t, y(t,\theta)) < \infty$ for all $t \in [0,T]$.
2. For all $\theta \in \Theta$, w.p.1, no two events occur at the same time.

Time derivatives

Taking all the possible transition events for a single NEP we have:

At event e_2, a transition from mode 2 to 1 takes place. This is an endogenous event with $g_2(y(t,\theta),\theta) = y(v_{s_k},\theta) = 0$. Applying Eq. (9.11) we have

$$v'_{s_k} = \frac{-y'(v^-_{s_k},\theta)}{A(v^-_{s_k},\theta) - B(v^-_{s_k},\theta)} \tag{9.15}$$

In addition applying Eq. (9.9) we get

$$y'(v^+_{s_k},\theta) = y'(v^-_{s_k},\theta) + [\dot{y}(v^-_{s_k},\theta)]v'_{s_k} \tag{9.16}$$

and combining the two equations above we have

$$y'(v^+_{s_k},\theta) = 0 \tag{9.17}$$

The e_1 event is an endogenous and thus, there exists a continuous differentiable function denoted as $g_1(y(t,\theta),\theta) = y(v_{2j-1},\theta) - \theta = 0 \;\forall\; j = 1,\ldots,\frac{s_k-1}{2}$ and applying Eq. (9.11) we get

$$v'_{2j-1} = \frac{1 - y'(v^-_{2j-1},\theta)}{A(v^-_{2j-1},\theta) - B(v^-_{2j-1},\theta)} \tag{9.18}$$

in the sequel combining Eq. (9.18) with Eq. (9.9) we have

$$y'(v^+_{2j-1},\theta) = y'(v^-_{2j-1},\theta) + [\dot{y}(v^-_{2j-1},\theta)]v'_{2j-1} \tag{9.19}$$

and combining the two equations above Eqs. (9.19)-(9.18) we get

$$y'(v^+_{2j-1},\theta) = 1 \tag{9.20}$$

The e_3 is an exogenous event with $y'(v_{2j},\theta) = 0 \;\forall\; j = 1,\ldots,\frac{s_k-1}{2}$.

Finally from Eqs. (9.17)-(9.20) it follows that $y'(t,\theta)$ always starts from 0 and at every FP switches to 1 and always at the end of the NEP reset back again to 0 value.

Infinitesimal Perturbation Analysis for throughput

Lemma 3. *Eq.* (9.13) *measures the total outflow as the summation of region's NEPs starting from the beginning of an EP until the beginning of the next EP. Therefore, considering that the region's outflow rate is determined by Eq.* (9.3) *then taking the derivatives with respect to* θ *we get*

$$\frac{d\omega}{d\theta} = \frac{1}{T}\Big[\sum_{j=1}^{\frac{s_k-1}{2}} C(v_{2j-1} - v_{2j})\Big] \tag{9.21}$$

where the parameter C is obtained from $B(t, y(t,\theta))$ *defined by Eq.* (9.3) *and thus*

$$C = \frac{\partial}{\partial y}B(t, y(t,\theta)) = \begin{cases} u_f, & \text{if } y(t,\theta) \le \theta \\ -w, & \text{otherwise} \end{cases} \tag{9.22}$$

Proof. Considering that Eq. (9.13) can be re-stated as

$$\Omega_T(t,\theta) = \frac{1}{T}\sum_{k=1}^{N_t}\Big[\int_{v_{k-1,s_k}}^{v_{k,0}} B(t, y(t,\theta))dt + \int_{v_{k,0}}^{v_{k,s_k}} B(t, y(t,\theta))dt\Big] \tag{9.23}$$

then, taking the derivative with respect to θ we get

$$\frac{d\omega_k}{d\theta} = \frac{1}{T}\frac{d}{d\theta}\int_{v_{k,0}}^{v_{k,s_k}} B(t, y(t,\theta))dt \tag{9.24}$$

since, $\frac{d}{d\theta}\int_{v_{k-1,s_k}}^{v_{k,0}} B(t,y(t,\theta))dt$ is zero as $B(t,y(t,\theta)) = 0$ during an EP. In the sequel, considering the Leibniz rule the above derivative can be computed as

$$\frac{d}{d\theta}\int_{v_{k,0}}^{v_{k,s_k}} B(t,y(t,\theta))dt = B(v_{k,s_k})v'_{k,s_k} + \int_{v_{k,0}}^{v_{k,s_k}} \Big[\frac{\partial}{\partial y}B(t,y(t,\theta))\frac{\partial y}{\partial \theta} + \frac{dB(t,y(t,\theta))}{d\theta}\Big]dt \tag{9.25}$$

considering the Eq. (9.25) we can observe that the term $\frac{dB(t,y(t,\theta))}{d\theta} = 0$ as is not dependent on θ while the term $\int_{v_{k,0}}^{v_{k,s_k}} \frac{\partial y}{\partial \theta}$ can be computed as follows:

Considering a single NEP, the term $\int_{v_0}^{v_{s_k}} \frac{\partial y}{\partial \theta}dt$ can be expressed as

$$\int_{v_0}^{v_{s_k}} \frac{\partial y}{\partial \theta}dt = \int_{v_0}^{v_1} y'(t,\theta)dt + \sum_{j=1}^{\frac{s_k-1}{2}}\int_{v_{2j-1}}^{v_{2j}} y'(t,\theta)dt + \sum_{j=1}^{\frac{s_k-3}{2}}\int_{v_{2j}}^{v_{2j+1}} y'(t,\theta)dt + \int_{v_{s_k}}^{v_{s_k-1}} y'(t,\theta)dt \tag{9.26}$$

Taking one term at time then, during the all the FPs the queue content $y(t,\theta) = \theta$ and thus

$$\sum_{j=1}^{\frac{s_k-1}{2}}\int_{v_{2j-1}}^{v_{2j}} y'(t,\theta)dt = \sum_{j=1}^{\frac{s_k-1}{2}}\int_{v_{2j-1}}^{v_{2j}} 1dt \tag{9.27}$$

According to Eq. (9.25) and considering the interval in-between two consecutive FPs the buffer content can be calculated as

$$y(t,v_{2j+1}) = y(v_{2j},\theta) + \int_{v_{2j}}^{v_{2j+1}} \dot{y}(\tau,\theta) \tag{9.28}$$

for all $j = 1,\ldots,\frac{s_k-3}{2}$. Then, taking the derivatives with respect to θ and considering that $y(t,v_{2j+1}) = y(v_{2j},\theta) = \theta$ then we get that

$$v'_{2j+1}\dot{y}(v_{2j+1},\theta) - v'_{2j}\dot{y}(v_{2j},\theta) + \int_{v_{2j}}^{v_{2j+1}} y'(\tau,\theta) = 0 \tag{9.29}$$

for all $j = 1,\ldots,\frac{s_k-3}{2}$. However, considering that $\dot{y}(v_{2j+1}) = v'_{2j} = 0$ then we have

$$\sum_{j=1}^{\frac{s_k-3}{2}}\int_{v_{2j}}^{v_{2j+1}} y'(t,\theta)dt = 0 \tag{9.30}$$

In similar way, during the interval $[v_0,v_1]$ we get

$$\int_{v_0}^{v_1} y'(t,\theta)dt = 1 \tag{9.31}$$

while during the interval $[v_{s_k-1}, v_{s_k}]$ we have

$$\int_{v_{s_k}}^{v_{s_k-1}} y'(t,\theta)dt = -1 \tag{9.32}$$

Therefore, according to Eq. (9.25) the $\int_{v_0}^{v_{s_k}} \frac{\partial y}{\partial \theta}$ has a unit value only during each FPs while it first and last terms are cancel then its follows that

$$\frac{d}{d\theta}\int_{v_{k,0}}^{v_{k,s_k}} B(t, y(t,\theta))dt = +\sum_{j=0}^{\frac{s_k-1}{2}} \int_{v_{2j-1}}^{v_{2j}} \frac{d}{dy} B(t, y(t,\theta))dt \tag{9.33}$$

then combining Eqs. (9.24)-(9.33) then Eq. (9.21) follows.

□

9.4 Performance evaluation

9.4.1 Setup

The area under consideration is an 1 km^2 homogeneous [25] region with the following MFD parameters: ρ_C =300 veh/km, ρ_J = 1000 veh/km and u_f =15 m/s all defined over the triangular macroscopic fundamental diagram as denoted by eq. (9.3) [121].

The actual system is simulated along side the route-reservation algorithm awhere each vehicle arrives to the simulated region with a Poisson arrival process. The RRA reschedules the vehicle departure times from their origin according to its objective (that is, maintain each road-segment's density below the critical density). Furthermore, the RRA determines each vehicle's route such that congested links are avoided. To achieve this, the RRA assumes that it knows every link's critical capacity and can determine the exact path of each vehicle assuming that it will traverse its path using the free flow speed u_f. In all earlier Chapters 3-8, the critical density was measured (through extensive simulation) a priori and it was assumed known by the RRA. In this Chapter, the RRA utilizes an estimate of the critical density θ, which is continuously updated such that RRA is able to learn on-line the true value of the critical density.

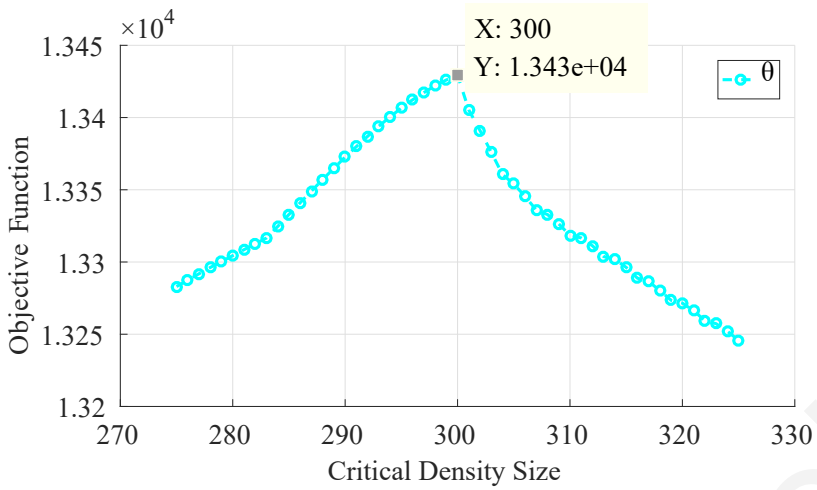


Figure 9.4: Region's outflow rate as a function of θ.(Brute-force method)

9.4.2 Results

For the assumed network, Fig. 9.4 depict the region's outflow as a function of the critical density assumed by the RRA. This result is obtained by running long simulations with varying θ within the range of $[275, \ldots, 325]$ in a brute-force manner. As observed by the Fig. 9.4 the maximum outflow-rate is obtained when $\theta = 300$veh/km while, as expected, for all other values lower flow-rates are observed since using these values imply that the region is under/over utilized.

The critical density estimated by the RRA is updated through the stochastic approximation rule

$$\theta_{i_{k+1}} = \theta_{i_k} - h\frac{\partial \Omega_T(t,\theta)}{\partial \theta} \tag{9.34}$$

with h denoting the step size while $\frac{\partial \Omega_T(t,\theta)}{\partial \theta}$ denotes the sensitivity of the region's outflow with respect to the parameter θ as computed by the IPA (eq. (9.21)). At this point it is worth pointing out that despite the fact that the IPA algorithm was derived based on an SFM, the underlying system model used for the simulations is a more realistic discrete event model.

As mentioned above, as the region's density is maintained within the free-flow regime, the outflow has different rate compared to that of the congested regime, fact that can be justified by findings in Fig. 9.4. Therefore, the derived gradient estimator of Eq. (9.21) requires the value of the parameter c (see eq. (9.22)) which is not know since the true state of the network is also not known. In such manner, Eq. (9.22)

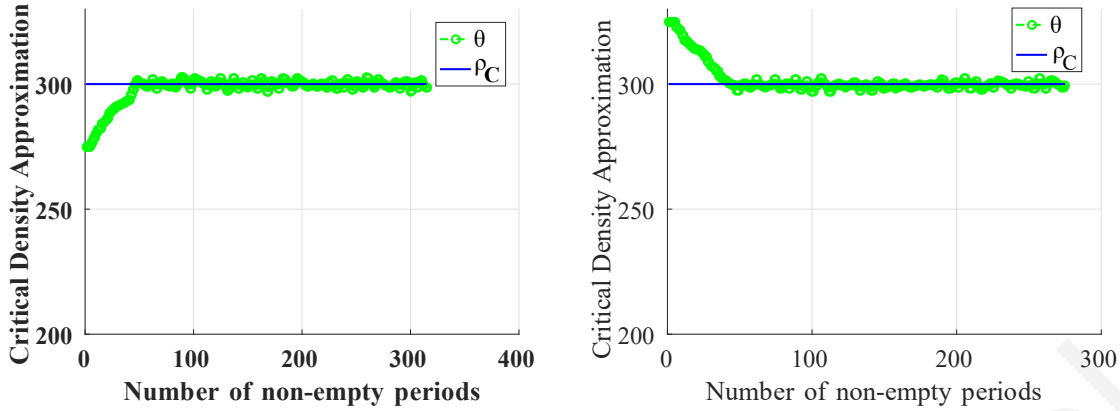


Figure 9.5: IPA estimators starting from different initial values: (a) θ = 275veh/km (b) θ = 300veh/km as a function of the number of NEP observed within the simulation time (iterations).

approximates all the unknown parameters and the direction of our stochastic approximation algorithm. Notably, considering Eq. (9.22) the sign of parameter C is highly correlated with the region's state as is depends on whether the estimated parameter θ over/under estimates the actual ρ_C. Since the region's true critical density is not known, it is a challenge to determine the true state of the network.

To address this challenge, we utilize on-line measurements of the outflows of the actual discrete event system. In this way, the parameter C is approximated with the parameter $\hat{C}$, by sampling the region's outflow rate every time that the control parameter θ updates. More specific, we compute the average change of the outflow rate by taking real time measurements across the current and the previous state θ updates (i.e., θ_k and θ_{k-1}) as follows

$$\hat{C} = \frac{\bar{q}(\theta_k) - \bar{q}(\theta_{k-1})}{\Delta L} \tag{9.35}$$

with the parameter $\bar{q}(\theta)$ denotes the real time measurement of region's outflow as a function of θ and the parameter $\Delta L = |\theta_k - \theta_{k-1}|$ denotes the difference of θ values of the two measurements (this approximation is called as the "Euler's" backward derivative approximation method). In this manner, every time that we are going to update the new θ_{k+1} value we compare the previous measured outflow with the current observation in order to drive the estimated $\hat{C}$. Notably, to further improve

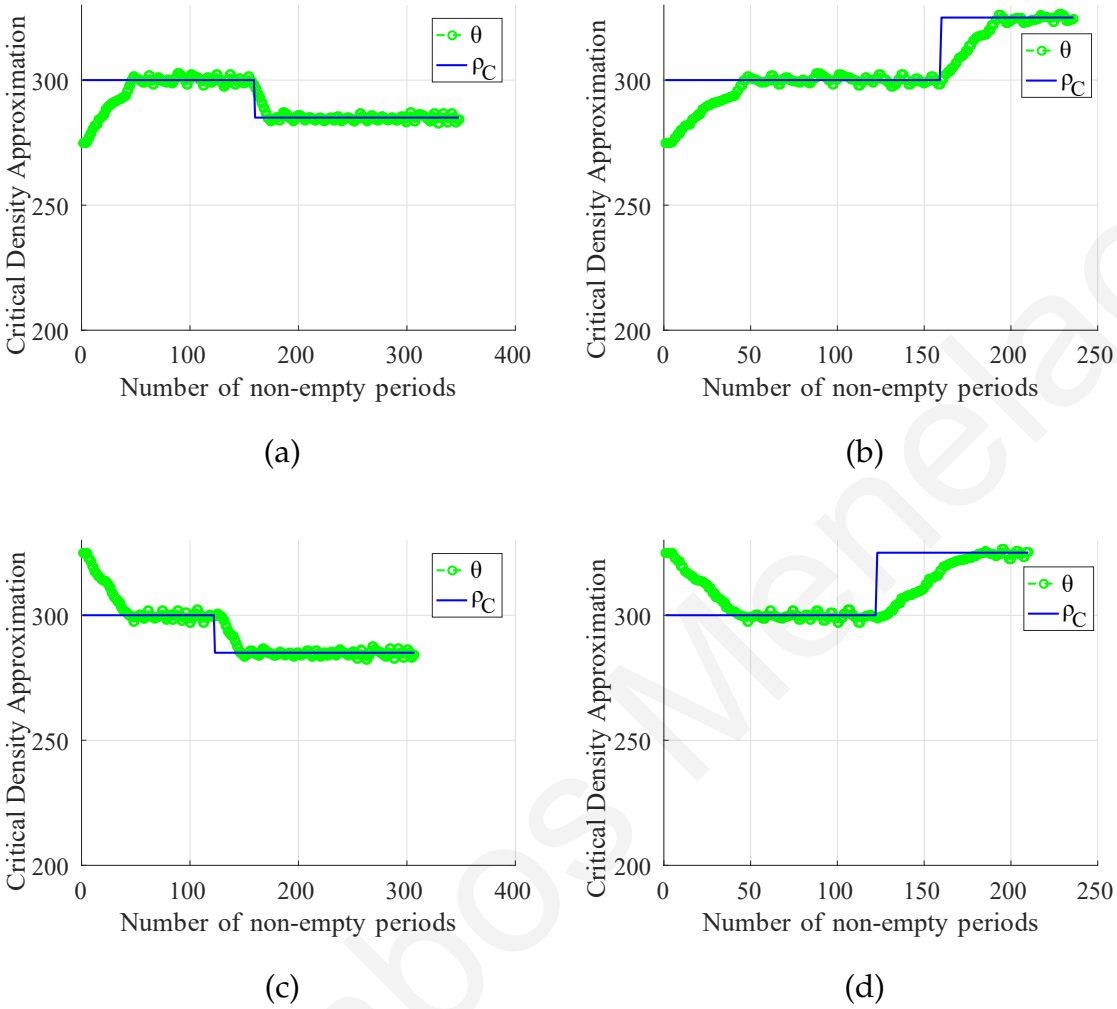


Figure 9.6: IPA estimators starting from different initial values with a suddenly change of ρ_C value: (a) and (b) with initial value starting from 275veh/km to 285veh/km and 315veh/km, respectively while (c) and (d) with initial value starting from 325veh/km to 285veh/km and 315veh/km, respectively as a functions of the number of NEP observed within the simulation time (iterations).

the approximation accuracy measurements are taken only during the FPs where the outflow has its maximum possible rate according to the current θ value.

The obtained results of the on-line estimation of $\frac{\partial \Omega_T(t,\theta)}{\partial \theta}$ are depicted in the Fig. 9.5 which indicate how θ is updated assuming different initial values of θ. In this figure with the green color scatters we denote the updates θ_{k+1} observed on each NEP while with the solid blue line represents the true value of ρ_C =300veh/km. According to the figure, it is clear that for both of these cases the IPA estimates can be used to learn the true critical density irrespective of the initial values. Note that, in the first case

of Fig. 9.5(a) the initial value under-estimated while, in the second case Fig. 9.5(b) over-estimate the true critical density.

Figs. 9.6 illustrates scenarios where the actual critical density starts with the initial value of ρ_C =300veh/km while at some point during the simulation time ρ_C suddenly changes due to external factors (e.g., weather conditions) either increases or decreases. Likewise, in Fig. 9.5, with the green color scatters we denote the θ_{k+1} as they are updated on every NEP while with the solid blue line we depict the true value of ρ_C. The first two figures Figs. 9.6 (a) and (b) start with a parameter θ that under-estimates ρ_C and as time progresses it can efficiently approximates the initial critical density value. Subsequently, when ρ_C suddenly changes, it automatically learns that and quickly it converges to the new ρ_C. The same behavior is observed when the critical density value increases or decreases Figs. 9.6 (a) and (b), respectively.

9.5 Summary

This chapter investigates a stochastic fluid model with switching dynamics that can be utilized for the on-line estimation of the critical density of an urban area. The approach utilizes a stochastic approximation based algorithm that seeks to learn the region's critical density. The stochastic approximation algorithm is driven by sensitivity estimates that are obtained through IPA on stochastic fluid models. The IPA estimate requires minimal information (e.g., timers and average speeds and flow rates). An important challenge of the derived IPA estimator is that it requires knowledge of the state of the network (free flow or congested), which is information not directly observable; however, it is information that can be inferred from the average speed. Hence, the major advantage of this approach in that is mainly the simple implementation and its online execution.

Chapter 10

Conclusions, implementation challenges and future work

10.1 Conclusions

This Ph.D. thesis proposes the use of demand management in road traffic networks via a novel route reservation architecture which aims to maximize the efficiency of the urban transport systems while eliminate the emergence of traffic congestion altogether. Demand management at a microscopic level is realised by reserving individual vehicle routes while at a macroscopic level it is realised by managing traffic inflow at a region level. Through the proposed route reservation architecture, several approaches are investigated that either delay drivers' departure times or suggest routes to minimize travel times. In this thesis, the route reservation problems that arise in both microscopic and macroscopic levels are modeled and solved. Their performance evaluation confirms the usefulness of the proposed route reservation architecture as it leads to substantial improvements in terms of network operation and the overall travel time experienced as compared with the ordinary route guidance frameworks.

The key advantage of the proposed demand management schemes is that the density of vehicles in the system under investigation never exceed the road network's critical value which ensures that a congestion-free operation exhibiting the social awareness of the proposed methodologies. Hence, the primary objective of this thesis is to improve the system/social optimum by managing each driver (or a

group of drivers) departure times. Other than that, the extensive simulations results demonstrate that this socially-oriented behavior can further enhance the efficiency of the current state-of-the-art traffic management schemes, e.g., route guidance, ensuring their proper operation. The reason for this is that by delaying departure times, vehicles will be kept outside the network and thus will not affect the network's operation and will not interact with other vehicles which may produce unwanted delays. Apart from this, by forcing drivers to wait at their origin, they will not be contributing to congestion and wasting time in traffic jams. Furthermore, the proposed demand management schemes may force drivers to take longer congestion-free paths if that action minimizes the arrival time at the destination. Broadly speaking, these actions seem to produce fairness issues to the proposed demand management schemes since some of the drivers may have increased travel distances or may incur some waiting at their origin. On the other hand however, during congestion, drivers would have to wait for much longer in traffic jams. Hence, demand management actions can potentially benefit the overall user base compared to those that are sacrificing some of their time to alleviate congestion. Extensive simulation results validate the superior performance of the proposed demand management approaches in terms of the total travel time and the congestion caused. The benefits of the proposed methods are not limited to travel time reductions but also can offer numerous improvements in the terms of environmental impact and economic growth in future city operations.

Of course, these improvements can be achieved only when considering high driver compliance. Interestingly, with connected autonomous vehicles the necessary compliance can effortlessly be supported. Therefore, the proposed route reservation architecture and the related demand management approaches can stand as an alternative innovative solution to manage access to the road network by autonomous connected vehicles based on route reservations.

More specifically, this thesis begins with the proposition that the proposed route reservation architecture, aims to prevent congestion by restricting the traffic density in different road segments. Accordingly, the Earliest Destination Arrival Time (EDAT) problem is formulated (considering both continuous and discrete time domains), which is shown to be an NP-complete problem. In the sequel, this thesis proposes the Route Reservation Algorithm that produces low-complexity, close-to-

optimal solutions. Furthermore, the Reservation Algorithm is also customized to adopt continuous time, whereas, the optimal solution to this problem is obtained by constructing an appropriate MILP formulation. Importantly, the included simulation results provide evidence for the superior performance of both algorithms compared to the traditional traffic behavior (where no route reservations are made), achieving substantial improvements in terms of road utilization and the travel times experienced by vehicles.

To resolve any fairness problems, this Ph.D. thesis specifically formulates the Traffic Load Balancing problem, that aims to find a path that provides a good trade-off between the destination arrival time and the number of reservations that are made on each road segment. Importantly, the included results provide evidence that the TLB formulation can offer a more robust and fair solution that seems to be highly resilient on an increase of flow rates.

The proposed architecture assumes that each segment's transit-time is calculated assuming a constant speed, (i.e., the free-flow speed). However, this assumption is quite restrictive. Unfortunately, inaccurate transit-time predictions may lead to unstable solutions as long queues may be observed. Hence, the application of reservations on bigger subareas rather than individual segments is prohibited as reservation inaccuracies may lead to congestion. This Ph.D. thesis proposes a simple predictor that estimates the transit-times of each road segment, aiming to improve the route reservations accuracy. This predictor can result in complexity reductions that can enable more scalable routing solutions. Extensive performance evaluation confirms the usefulness of the predicted transit times that ensure proper utilization of the infrastructure's capacity leading to shorter trips. This thesis shows that a performance trade-off exists between the two proposed prediction methods where long waiting times at the origin can impact to more accurate reservations.

In this thesis an aggregated route-reservation scheme is also developed, which is more scalable compared to the original route reservation architecture. The significant advantage of this scheme is that it utilizes an overlay graph to control the traffic in a large-scale multi-region network ensuring effective, scalable, and congestion free routing solutions. Simulation results demonstrate the superiority of the proposed aggregated scheme compared to the uncontrolled traffic behavior resulting in many-

fold gains in serving traveling requests and reductions in travel times, especially during high demand flows. Additionally, an investigation is performed on how network operation is impacted if only a small percentage of the vehicles follows the RSU's schedules, demonstrating that the proposed approach is robust even if at least 80% of drives adhere to the RSU schedules (with similar gains achieved to the ideal case of 100% driver adherence).

Another extension of the route reservation architecture is proposed that address the problem of scheduling vehicle departures from their origin such that they will arrive at their destination on the desired time. For this problem, vehicles transmit to the RSU their origin and destination pair and their desire time to arrive at their destination. In return, the RSU response to each vehicle its departure time as well as the path to be followed while making the appropriate route reservations on the selected path such that all scheduled vehicles avoid congested road segments. Due to the reservations, the RSU can guarantee an on-time arrival at the destination for each vehicle request. This thesis presents a mathematical formulation to model this problem is provided while an efficient algorithmic solution is derived. Microscopic simulation results demonstrate the proposed algorithm's effectiveness in realistic simulation scenarios.

Joint demand management and multi-regional route guidance is also investigated with the aim to minimize the total travel time of vehicles in road networks that are characterized by a well-defined MFD. To solve this joint problem, a mathematical formulation is suggested that aims towards to minimize the total time that vehicles spend in the network while maintaining non-congested conditions at all times through demand management. A relaxed LP and MILP formulations of the original non-convex, the non-linear problem is also proposed which provide tight lower bound to the optimal solution. Besides, as the obtained solutions (from both relaxed formulations) prefer to sustain the density of each region below the critical density, this thesis further approximated the LP program by proposing another LP that provides a feasible solution to the actual non-convex program. The proposed LP constitutes the major contribution of this thesis as it can be solved in real time in contrasts with the state-of-the-art NLP formulations. An essential assumption of the multi-region MPC scheme is that all paths in each region have a constant length

regardless form their origin-destination pairs. This assumption is often violated in practice; thus, this thesis reformulates the problem to a path-based formulation which enables the routing of vehicles through multiple paths. Similar LP algorithms are extracted for the path-based formulation while, extensive simulation results show the importance of demand management in minimizing the total travel time and demonstrate the effectiveness of all proposed MPC approaches, as they can provide fast and very close to optimal results for all various demand scenarios.

Finally, this Ph.D. thesis adobes the framework of the Stochastic Fluid Modeling framework to model the critical density of a homogeneous region where the route-reservation architecture is employed to control traffic demand and thus to achieve a congestion-free operation. On that account, an Infinitesimal Perturbation Analysis (IPA) is applied in an online fashion to capture the dynamic changes in the critical density value as a consequence of different incidents.

Currently, demand management methods remain mostly unexplored, and hopefully, the recent advances in autonomous-connected vehicles capabilities push future ITS solutions towards this promising direction that elaborates on intelligent demand management schemes. As a conclusion, this Ph.D. thesis significantly contributes towards this research direction with the proposed demand management schemes achieving a better network utilization, travel times reduction, congestion-free operation and better energy savings making them the best proposition for the development of future ITS solutions.

10.2 Implementation challenges

At the same time, the proposed demand management strategies (especially the route reservation architecture) admittedly faces certain challenges need to be addressed prior its real-life implementation as listed below.

Communication issues:

The real-life implementation of such approaches require the direct communication between vehicles and the infrastructure, a fact that may lead to various communication and computation issues due to the size and the complexity of transportation

networks. Given the recent developments in the information and communication domain, the Internet of Things (IoT) technology and the proliferation of connected vehicles, these challenges will most definitely be addressed in the near future.

Drivers adhering:

It is evident that within the proposed schemes, vehicles are driven by selfish drivers that are only interested in optimizing their own travel time. Hence, several drivers may not follow either the suggested departure time or suggested route (i.e., the non-compliant drivers). This act can potentially reduce the performance of the proposed schemes and inhibit their real-life implementation. Nonetheless, the impact and the number of non-compliant drivers can be reduced by introducing innovative policies that affect drivers route choices by providing incentives to compliant drivers.

One possible solution to the compliance issue involves the development of pricing mechanisms that incentivize drivers to follow the suggested departure times and routes. Such pricing mechanisms have been explored in recent literature, indicating that pricing policies can potentially influence drivers routing decisions aiming to improve the social optimum [129]. In this fashion, a pricing mechanism can be introduced to prompt drivers to participate in the demand management framework. In doing so, the proposed schemes will be combined with a dynamic pricing mechanism that will aim at identifying the optimal tolling prices that would discourage drivers from disregarding the controller schedules [130]. Thereby, non-compliant drivers will pay a tolling fee for using the road network [131]. An alternative direction may include the utilization of a time-dependent pricing scheme (similarly with the work in [132]). In this setting, tolling prices will be adjusted based on the levels of congestion in combination with the adaptability of drivers to the route guidance suggestions.

Another possible solution to anticipate this issue is to restrict some lanes for exclusive use from drivers that adhere to demand management suggestions. Thereby, a part of the network will be accessible only to compliant drivers with the non-compliant ones restricted to use the non-prioritized busy lanes. Hence, drivers that are willing to participate in the demand management schedules will be prioritized, an act that can significantly benefit the social optimum. On the other hand, drivers

that violate their scheduling will be charged with a penalty cost to avoid occasions where non-compliant drivers use the prioritized lanes. Note that restrictions should only be imposed during rush hours where demand surpasses the network's available capacity.

Furthermore, a supplementary solution to the compliance issue is to combine the proposed demand management scheme with a perimeter control approach. This combined scheme will be useful during peak periods where non-compliant drivers can considerably affect the performance of the proposed schemes, as a perimeter control mechanism may further mitigate their impact. Therefore, in case of congestion, perimeter control could restrict the inflow demands at the periphery of the affected regions, protecting compliant drivers from experiencing congestion.

Centralized implementation:

All the proposed schemes are implemented in a centralized manner a fact that increases their computation complexity and reduces the reliability of the solution as there is a single point of failure. Therefore, an interesting topic for investigation in future work is how the proposed methodologies can be transformed to operate in a distributed manner. A distributed framework will result in substantial complexity reduction enabling scalable routing solutions that can handle large-scale networks. Having a distributed scheme reduces the dimension of the considered optimization problem as each region can be partitioned into a set of subregions of smaller dimension, each with its dedicated control unit that will manage traffic in a similar fashion with the proposed demand management solutions. Nonetheless, such a distributed scheme may result in suboptimal solutions as each control unit will operate under partial information. In such a case, control units should cooperate and communicate information between them to obtain information regarding the reservation status of their nearby regions. Moreover, a load balancing framework could be applied to balance the load of each subregion and to manage the load at the boundaries of neighboring regions.

Privacy:

Under the proposed architecture, drivers are asked to provide private and sensitive information to the RSU, such as their origin-destination (*OD*) pair. A possible solution to this problem is that the drivers can use an OD pair from a pre-specified set of OD pairs such that users will not reveal their identity.

Fairness:

In this thesis, vehicles are served in first-come-first-served order, making the system susceptible to fairness problems as some vehicles may be instructed to follow longer routes instead of following the shortest ones, or some of them may be forced to wait longer than others. A possible solution to this problem is to balance the number of reservations across each road segment provided that the time required to reach the destination is not higher than a percentage of the earliest destination arrival time.

Evidently, today's communication technologies can effortlessly support such demand management schemes. The proposed schemes can also bring significant added value to the network upon the emergence of connected and automated vehicles which can be fully compliant to any instruction provided.

10.3 Future Work

Exploring the route reservation architecture over the years has generated a plethora of interesting research questions that still remain unaddressed.

A good starting point for further research is to investigate the use of the reservation architecture in larger networks where vehicles will travel from one city to another. In that case, a dedicated controller will be responsible for controlling vehicular flows in each particular city. The problem that arises is that the arrival time at the periphery of each visited city may be uncertain so that its controller may be unable to identify a congestion-free path. One way to tackle this issue is to employ the Route Reservation Protocol (RSVP), designed to reserve resources across a computer network for quality of service, that can be used to coordinate the communication between the city controllers. This coordination aims to enable them to have access to the reservations of their neighboring regions. In doing so, several controllers can

exchange information among them, to ensure that vehicles will traverse through each city only following congestion-free road segments. Hence, the drivers will be informed about a multiregional path to follow while the RVSP protocol will identify the expected traversal time in each city.

An alternative solution is to develop a hierarchical control framework in which the optimization problem will be formulated and solved at two different control layers, the upper, and the lower layer. The upper control layer will be responsible for finding the departure time of each vehicle and the regional-path that the vehicle will follow. The upper-layer controller aims to balance the traffic demand between regions and avoid the occurrence of congestion in each region. The lower control layer will operate at a regional level, where a dedicated regional controller will manage all the traffic movements. This hierarchical framework can also be combined with a perimeter control strategy aiming to regulate the inflow at the boundaries of each distinct region. In doing so, we can ensure that a reservation will be assigned to all vehicles at the times that they wait at the periphery of the controlled region.

As mention in the implementation challenges, the proposed methodologies investigate only the ideal case with all drivers adhere to the route guidance instructions. However, some drivers may prefer to follow their own routes and/or to start their journey immediately, a fact that evidently can affect the performance of the proposed scheme. Hence, future research should examine strategically how driver compliance affects the performance of the proposed scheme and investigate new designs that encourage drivers to use reservation architecture. Additionally, future implementations should investigate scenarios where the considered network is utilized by manual and autonomous vehicles, and examine how the route reservation architecture's performance can be affected by increasing or reducing the percentage of ordinary vehicles. In this setting, one could analyze the performance as a percentage of ordinary vehicles and their effect on traffic state.

Another future direction is to investigate scenarios where instead of having an explicit starting time for each vehicle to have a time range for its departure, a feature that might be essential to increase the applicability of the proposed demand management schemes. In such a case, robust formulations for the proposed demand management schemes should be investigated that take into account the distribution

of the driver choices according to the provided departure time range.

Future work will also investigate how vehicle-to-vehicle (V2V) communications can help the proposed demand management schemes. Primarily V2V communications can be used as an advanced monitoring system that provides useful information regarding the state of the network. For instance, when the number of non-compliant drivers is high, then the V2V communications can be employed to detect and track the occurrence of congestion. In this case, V2V communications can provide real-time state information, that in turn can be used to implement a rescheduling mechanism that eliminates overcrowding conditions. Furthermore, V2V communications can improve the accuracy of reservations by providing real-time state information, which in turn can be used to detect any inaccuracies that may arise. Eliminating predicted state and model uncertainty inaccuracies offer the opportunity to re-optimize vehicle departure times and routes in real-time to obtain better performance.

Besides, a limitation of the proposed route reservation architecture is that it does not consider any incidents or any vehicle reservations cancellations. Therefore, future formulations should allow the on-demand rescheduling of vehicles in order to account for early or late departures and even anticipate any reservation cancellations. Unfortunately, the rescheduling of vehicles may negatively affect compliant drivers because either they will be routed through longer paths or they will be asked to wait longer at their origins. This issue can be alleviated by providing rewards to incentivize both compliant and non-compliant drivers remain compliant or become compliant, respectively. Another issue that may arise and needs to be tackled in future work is that V2V communications may expose vehicles to malicious cyber-attacks that indent to affect the performance of the network.

Further studies should investigate the introduction of uncertainty into the generated demands and to further analyze if the LP MPC formulation can be affected by uncertainties in demand or by measurement noise observed in the parameters that estimate the actual density of each region. It will be important that future research also investigates the use of a generalized-shape MFD in the proposed MPC frameworks. In considering those limitations, future work can result in more robust formulations that address robustness issues. Future work will also examine how the

driver compliance rate can affect the performance of the proposed MPC formulations and investigate new schemes that incentivize drivers to use demand management towards optimal system performance.

Future work includes the proof of the unbiasedness of the derived IPA estimators which constitute a more difficult task compared to earlier works in IPA due to unobservant switching dynamics. Future avenues also include the introduction of uncertainty to the route-reservation estimates, something that allows the formation of a queue that is longer than the region's actual critical density. This will constitute a more realistic approach as in real application inaccuracies may be observed within the reservation plan. Finally, future work should also examine how the perturbations are propagated between neighboring regions.

Bibliography

[1] E. Commission, *Roadmap to a Single European Transport Area: Towards a Competitive and Resource Efficient Transport System: White Paper*, 2011.

[2] R. Arnott and K. Small, "The economics of traffic congestion," *American scientist*, vol. 82, no. 5, pp. 446–455, 1994.

[3] K. Luten, B. Katherine, D. Deborah, H. Tanisha, and S. Eric, *Mitigating traffic congestion: The role of demand-side strategies*, 2004.

[4] K. T. Geurs and B. van Wee, "Accessibility evaluation of land-use and transport strategies: review and research directions," *Journal of Transport Geography*, vol. 12, no. 2, pp. 127 – 140, 2004.

[5] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2043–2067, 2003.

[6] F. T. Seik, "An effective demand management instrument in urban transport: the area licensing scheme in singapore," *Cities*, vol. 14, no. 3, pp. 155–164, 1997.

[7] C. Chen, Z. Jia, and P. Varaiya, "Causes and cures of highway congestion," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 26–32, 2001.

[8] M. Papageorgiou, "Dynamic modeling, assignment, and route guidance in traffic networks," *Transportation Research Part B: Methodological*, vol. 24, no. 6, pp. 471 – 495, 1990.

[9] H. Mahmassani, S. Peeta, T.-Y. Hu, and A. Ziliaskopoulos, "Dynamic traffic assignment with multiple user classes for real-time atis/atms applications," in *Large Urban Systems. Proceedings of the Advanced Traffic Management ConferenceFederal Highway Administration*, 1993.

[10] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature communications*, vol. 7, no. 10793, 03 2016.

[11] J. G. Wardrop, "Some theoretical aspects of road traffic research," in *Inst Civil Engineers Proc London/UK/*, 1952.

[12] J. R. Correa, A. S. Schulz, and N. E. Stier-Moses, "Selfish routing in capacitated networks," *Mathematics of Operations Research*, vol. 29, no. 4, pp. 961–976, 2004.

[13] A. K. Ziliaskopoulos, "A linear programming model for the single destination system optimum dynamic traffic assignment problem," *Transportation science*, vol. 34, no. 1, pp. 37–49, 2000.

[14] M. D. Meyer, "Demand management as an element of transportation policy: using carrots and sticks to influence travel behavior," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 7-8, pp. 575–599, 1999.

[15] M. D. Meyer *et al.*, "A toolbox for alleviating traffic congestion and enhancing mobility," *Transportation research record*, 1997.

[16] C. F. Daganzo, "Urban gridlock: macroscopic modeling and mitigation approaches," *Transportation Research Part B: Methodological*, vol. 41, no. 1, pp. 49–62, 2007.

[17] R. Landman, S. Hoogendoorn, M. Westerman, S. Hoogendoorn-Lanser, and J. Van Kooten, "Design and implementation of integrated network management in the netherlands," in *TRB 89th Annual Meeting Compendium of Papers DVD*, 2010.

[18] N. Geroliminis and C. F. Daganzo, "Macroscopic modeling of traffic in cities," in *TRB 86th annual meeting*, no. 07-0413, 2007.

[19] ——, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 759–770, 2008.

[20] C. Roncoli, M. Papageorgiou, and I. Papamichail, "Traffic flow optimisation in presence of vehicle automation and communication systems–part i: A first-order multi-lane model for motorway traffic," *Transportation Research Part C: Emerging Technologies*, vol. 57, pp. 241–259, 2015.

[21] Y.-C. L. Ho and X.-R. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems*. The Springer International Series in Engineering and Computer Science, 1991.

[22] K. Ampountolas and A. Kouvelas, "Real-time estimation of critical vehicle accumulation for maximum network throughput," in *2015 American Control Conference (ACC)*, 2015, pp. 2057–2062.

[23] M. Keyvan-Ekbatani, M. Yildirimoglu, N. Geroliminis, and M. Papageorgiou, "Multiple concentric gating traffic control in large-scale urban networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2141–2154, 2015.

[24] K. Aboudolas and N. Geroliminis, "Perimeter and boundary flow control in multi-reservoir heterogeneous networks," *Transportation Research Part B: Methodological*, vol. 55, pp. 265–281, 2013.

[25] A. Mazloumian, N. Geroliminis, and D. Helbing, "The spatial variability of vehicle densities as determinant of urban network capacity," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1928, pp. 4627–4647, 2010.

[26] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1393 – 1403, 2012.

[27] J. Haddad and N. Geroliminis, "On the stability of traffic perimeter control in two-region urban cities," *Transportation Research Part B: Methodological*, vol. 46, no. 9, pp. 1159–1176, 2012.

[28] M. Keyvan-Ekbatani, M. Papageorgiou, and I. Papamichail, "Urban congestion gating control based on reduced operational network fundamental diagrams," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 74–87, 2013.

[29] Y. Ji and N. Geroliminis, "On the spatial partitioning of urban transportation networks," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1639–1656, 2012.

[30] 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.

[31] M. Ramezani, J. Haddad, and N. Geroliminis, "Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control," *Transportation Research Part B: Methodological*, vol. 74, pp. 1–19, 2015.

[32] R. C. Carlson, I. Papamichail, M. Papageorgiou, and A. Messmer, "Optimal motorway traffic flow control involving variable speed limits and ramp metering," *Transportation Science*, vol. 44, no. 2, 2010.

[33] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: An overview," *IEEE transactions on intelligent transportation systems*, vol. 3, no. 4, pp. 271–281, 2002.

[34] I. Papamichail and M. Papageorgiou, "Traffic-responsive linked ramp-metering control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 111–121, 2008.

[35] I. Papamichail, M. Papageorgiou, V. Vong, and J. Gaffney, "Heuristic ramp-metering coordination strategy implemented at monash freeway, australia," *Transportation Research Record*, vol. 2178, no. 1, pp. 10–20, 2010.

[36] M. Keyvan-Ekbatani, R. C. Carlson, V. L. Knoop, S. P. Hoogendoorn, and M. Papageorgiou, "Queuing under perimeter control: Analysis and control strategy," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016, pp. 1502–1507.

[37] A. Kouvelas, M. Saeedmanesh, and N. Geroliminis, "Enhancing model-based feedback perimeter control with data-driven online adaptive optimization," *Transportation Research Part B: Methodological*, vol. 96, pp. 26–45, 2017.

[38] A. Kouvelas, D. Triantafyllos, and N. Geroliminis, "Hierarchical control design for large-scale urban road traffic networks," in *Transportation Research Board 97th Annual Meeting, January 7–11, 2018, Washington, D.C.*, 2018.

[39] L. Xiao and H. K. Lo, "Adaptive vehicle navigation with en route stochastic traffic information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1900–1912, 2014.

[40] D. E. Kaufman and R. L. Smith, "Fastest paths in time-dependent networks for intelligent vehicle-highway systems applications," *Journal of Intelligent Transportation Systems*, vol. 1, no. 1, pp. 1–11, 1993.

[41] I. Kaysi, M. Ben-Akiva, and H. Koutsopoulos, "An integrated approach to vehicle routing and congestion prediction for real-time driver guidance," Ph.D. dissertation, Doctoral dissertation, Massachusetts Institute of Technology Cambridge, Mass, 1993.

[42] I. Chabini, "Discrete dynamic shortest path problems in transportation applications: Complexity and algorithms with optimal run time," *Transportation Research Records*, vol. 1645, pp. 170–175, 1998.

[43] A. Ziliaskopoulos and H. Mahmassani, "Time dependent, shortest-path algorithm for real-time intelligent vehicle highway system applications," *Transportation research record*, no. 1408, pp. 94–100, 1993.

[44] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *et al.*, *Introduction to algorithms*. MIT press Cambridge, 2001, vol. 2.

[45] L. Du, S. Chen, and L. Han, "Coordinated online in-vehicle navigation guidance based on routing game theory," in *Transportation Research Board 94th Annual Meeting*, no. 15-3613, 2015.

[46] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transportation Research Part B: Methodological*, vol. 53, pp. 64–81, 2013.

[47] K. E. Wunderlich, D. E. Kaufman, and R. L. Smith, "Link travel time prediction for decentralized route guidance architectures," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 1, pp. 4–14, 2000.

[48] H.-E. Lin, R. Zito, M. Taylor, *et al.*, "A review of travel-time prediction in transport and logistics," in *In proceedings of the Eastern Asia Society for transportation studies*, 2005, pp. 1433–1448.

[49] L. Foschini, J. Hershberger, and S. Suri, "On the complexity of time-dependent shortest paths," *Algorithmica*, vol. 68, no. 4, pp. 1075–1097, 2014.

[50] Y. Jiang and X. Li, "Travel time prediction based on historical trajectory data," *Annals of GIS*, vol. 19, no. 1, pp. 27–35, 2013.

[51] F. Sun, X. Hu, Y. Zou, and S. Li, "Adaptive unscented kalman filtering for state of charge estimation of a lithium-ion battery for electric vehicles," *Energy*, vol. 36, no. 5, pp. 3531 – 3540, 2011.

[52] V. Knoop, S. Hoogendoorn, and J. Van Lint, "Routing strategies based on macroscopic fundamental diagram," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2315, pp. 1–10, 2012.

[53] A. H. Chow, "Properties of system optimal traffic assignment with departure time choice and its solution method," *Transportation Research Part B: Methodological*, vol. 43, no. 3, pp. 325–344, 2009.

[54] M. Yildirimoglu and N. Geroliminis, "Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams," *Transportation Research Part B: Methodological*, vol. 70, pp. 186–200, 2014.

[55] M. Yildirimoglu, M. Ramezani, and N. Geroliminis, "Equilibrium analysis and route guidance in large-scale networks with mfd dynamics," *Transportation Research Part C: Emerging Technologies*, vol. 59, pp. 404–420, 2015.

[56] L. Leclercq and N. Geroliminis, "Estimating mfds in simple networks with route choice," *Procedia-Social and Behavioral Sciences*, vol. 80, pp. 99–118, 2013.

[57] R. Zhang, F. Rossi, and M. Pavone, "Routing autonomous vehicles in congested transportation networks: Structural properties and coordination algorithms," *Autonomous Robots*, vol. 42, no. 7, pp. 1427–1442, 2018.

[58] M. Yildirimoglu, I. I. Sirmatel, and N. Geroliminis, "Hierarchical control of heterogeneous large-scale urban road networks via path assignment and regional route guidance," *Transportation Research Part B: Methodological*, vol. 118, pp. 106 – 123, 2018.

[59] S. Batista, L. Leclercq, and N. Geroliminis, "Estimation of regional trip length distributions for the calibration of the aggregated network traffic models," *Transportation Research Part B: Methodological*, vol. 122, pp. 192 – 217, 2019.

[60] M. Hajiahmadi, V. L. Knoop, B. D. Schutter, and H. Hellendoorn, "Optimal dynamic route guidance: A model predictive approach using the macroscopic fundamental diagram," in *16th International IEEE Conference on Intelligent Transportation Systems*, 2013, pp. 1022–1028.

[61] E. F. Camacho and C. B. Alba, *Model predictive control*. Springer Science & Business Media, 2013.

[62] L. D. Baskar, B. D. Schutter, and J. Hellendoorn, "Hierarchical model-based predictive control for intelligent vehicle highway systems: Regional controllers," in *13th International IEEE Conference on Intelligent Transportation Systems*, Sept 2010, pp. 249–254.

[63] I. Papamichail, A. Kotsialos, I. Margonis, and M. Papageorgiou, "Coordinated ramp metering for freeway networks–a model-predictive hierarchical control approach," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 3, pp. 311–331, 2010.

[64] M. Hajiahmadi, J. Haddad, B. D. Schutter, and N. Geroliminis, "Optimal hybrid perimeter and switching plans control for urban traffic networks," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 2, pp. 464–478, March 2015.

[65] A. Kouvelas, M. Saeedmanesh, and N. Geroliminis, "Real-time estimation of aggregated traffic states of multi-region urban networks," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[66] ——, "Linear parameter varying model predictive control for multi-region traffic systems," in *98th Annual Meeting of the Transportation Research Board (TRB 2019)*, 2019, pp. 19–03 933.

[67] A. Jamshidnejad, I. Papamichail, M. Papageorgiou, and B. D. Schutter, "Sustainable model-predictive control in urban traffic networks: Efficient solution based on general smoothening methods," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 3, pp. 813–827, May 2018.

[68] A. C. Pigou, "Wealth and welfare," *MacMillan, UIK*, 1920.

[69] M. Mirshahi, J. Obenberger, C. A. Fuhs, C. E. Howard, R. A. Krammes, B. T. Kuhn, R. M. Mayhew, M. A. Moore, K. Sahebjam, C. J. Stone, *et al.*, "Active traffic management: the next step in congestion management," United States. Federal Highway Administration, Tech. Rep., 2007.

[70] J. Schade and B. Schlag, "Acceptability of urban transport pricing strategies," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 6, no. 1, pp. 45–61, 2003.

[71] K. M. Kockelman and S. Kalmanje, "Credit-based congestion pricing: a policy proposal and the public's response," *Transportation Research Part A: Policy and Practice*, vol. 39, no. 7, pp. 671 – 690, 2005.

[72] E. T. Verhoef, "Second-best congestion pricing in general networks. heuristic algorithms for finding second-best optimal toll levels and toll points," *Transportation Research Part B: Methodological*, vol. 36, no. 8, pp. 707 – 729, 2002.

[73] H. Mirzahossein and S. A. Zargari, "A combined model of congestion toll pricing based on system optimization with minimum toll," *Tehnički vjesnik*, vol. 25, no. 4, pp. 1162–1168, 2018.

[74] J. Dahlgren, "High occupancy vehicle lanes: Not always more effective than general purpose lanes," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 2, pp. 99–114, 1998.

[75] D. J. Fagnant and K. M. Kockelman, "Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in austin, texas," *Transportation*, vol. 45, no. 1, pp. 143–158, 2018.

[76] M. Barth and S. A. Shaheen, "Shared-use vehicle systems: Framework for classifying carsharing, station cars, and combined approaches," *Transportation Research Record*, vol. 1791, no. 1, pp. 105–112, 2002.

[77] Y. Li, A. K. Ziliaskopoulos, and S. T. Waller, "Linear programming formulations for system optimum dynamic traffic assignment with arrival time-based and departure time-based demands," *Transportation Research Record*, vol. 1667, no. 1, pp. 52–59, 1999.

[78] R. B. Noland and K. A. Small, "Travel-time uncertainty, departure time choice, and the cost of morning commutes," *Transportation Research Record*, no. 1493, 1995.

[79] E. Fekpe, S. Collins, *et al.*, *Evaluation of intelligent transportation infrastructure program (ITIP) in Pittsburgh and Philadelphia, Pennsylvania*, 2003.

[80] R. De Neufville and A. Odoni, *Airport Systems. Planning, Design and Management*, Eurocontrol, 12 2004.

[81] D. Condorelli, "Efficient and equitable airport slot allocation," *Rivista di politica economica*, vol. 1, pp. 81–104, 2007.

[82] C. G. Panayiotou and C. G. Cassandras, "A sample path approach for solving the ground-holding policy problem in air traffic control," *IEEE Transactions on Control Systems Technology*, vol. 9, no. 3, pp. 510–523, 2001.

[83] H. Akahane and M. Kuwahara, "A basic study on trip reservation systems for recreational trips on motorways," *Proc. 3rd World Congr. Intelligent Transportation Systems*, pp. 1–7, 1996.

[84] J.-T. Wong, "Basic concepts for a system for advance booking for highway use," *Transport Policy*, 1997.

[85] R. De Feijter, J. J. Evers, and G. Lodewijks, "Improving travel time reliability by the use of trip booking," in *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, vol. 1. IEEE, 2003, pp. 205–210.

[86] Y. Zhao, K. Triantis, D. Teodorović, and P. Edara, "A travel demand management strategy: The downtown space reservation system," *European Journal of Operational Research*, vol. 205, no. 3, pp. 584–594, 2010.

[87] P. Edara and D. Teodorović, "Model of an advance-booking system for highway trips," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 1, pp. 36–53, 2008.

[88] P. Su and B. B. Park, "Auction-based highway reservation system an agent-based simulation study," *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 211–226, 2015.

[89] K. Liu, E. Chan, V. Lee, K. Kapitanova, and S. H. Son, "Design and evaluation of token-based reservation for a roadway system," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 184–202, 2013.

[90] W. Liu, H. Yang, and Y. Yin, "Efficiency of a highway use reservation system for morning commute," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 293–308, 2015.

[91] E. D. Miller-Hooks and H. S. Mahmassani, "Least expected time paths in stochastic, time-varying transportation networks," *Transportation Science*, vol. 34, no. 2, pp. 198–215, 2000.

[92] Y. M. Nie and X. Wu, "Shortest path problem considering on-time arrival probability," *Transportation Research Part B: Methodological*, vol. 43, no. 6, pp. 597–613, 2009.

[93] S. Samaranayake, S. Blandin, and A. Bayen, "A tractable class of algorithms for reliable routing in stochastic networks," *Procedia-Social and Behavioral Sciences*, vol. 17, pp. 341–363, 2011.

[94] L. Yang and X. Zhou, "Optimizing on-time arrival probability and percentile travel time for elementary path finding in time-dependent transportation networks: Linear mixed integer programming reformulations," *Transportation Research Part B: Methodological*, vol. 96, pp. 68–91, 2017.

[95] H. Frank, "Shortest paths in probabilistic graphs," *Operations Research*, vol. 17, no. 4, pp. 583–599, 1969.

[96] C. F. Daganzo, "The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck," *Transportation science*, vol. 19, no. 1, pp. 29–37, 1985.

[97] E. J. Gonzales and C. F. Daganzo, "The evening commute with cars and transit: Duality results and user equilibrium for the combined morning and evening peaks," *Procedia-Social and Behavioral Sciences*, vol. 80, pp. 249–265, 2013.

[98] C. G. Panayiotou, "Infinitesimal perturbation analysis for a single stochastic fluid model node with a class of feedback controlled traffic," in *American Control Conference, 2004. Proceedings of the 2004*, vol. 3. IEEE, 2004, pp. 2308–2313.

[99] C. G. Cassandras, Y. Wardi, B. Melamed, G. Sun, and C. G. Panayiotou, "Perturbation analysis for online control and optimization of stochastic fluid models," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, pp. 1234–1248, 2002.

[100] H. Yu and C. G. Cassandras, "Perturbation analysis for production control and optimization of manufacturing systems," *Automatica*, vol. 40, no. 6, pp. 945–956, 2004.

[101] Y. Wardi, C. Seatzu, X. Chen, and S. Yalamanchili, "Performance regulation of event-driven dynamical systems using infinitesimal perturbation analysis," *Nonlinear Analysis: Hybrid Systems*, vol. 22, pp. 116–136, 2016.

[102] C. G. Panayiotou, W. C. Howell, and M. Fu, "Online traffic light control through gradient estimation using stochastic fluid models," *In Proceedings of IFAC Volumes*, vol. 38, no. 1, pp. 90–95, 2005.

[103] R. Chen and C. G. Cassandras, "Stochastic flow models with delays and applications to multi-intersection traffic light control," *Systems and Control*, 2017.

[104] Y. Geng and C. G. Cassandras, "Multi-intersection traffic light control with blocking," *Discrete Event Dynamic Systems*, vol. 25, no. 1-2, pp. 7–30, 2015.

[105] Y. Wardi and C. Seatzu, "Infinitesimal perturbation analysis of stochastic hybrid systems: Application to congestion management in traffic-light intersections," in *Proceedings Decision and Control (CDC), 2014 IEEE 53rd Annual Conference*. IEEE, 2014, pp. 6752–6757.

[106] N. Geroliminis and J. Sun, "Properties of a well-defined macroscopic fundamental diagram for urban traffic," *Transportation Research Part B: Methodological*, vol. 45, no. 3, pp. 605–617, 2011.

[107] A. Orda and R. Rom, "Shortest-path and minimum-delay algorithms in networks with time-dependent edge-length," *Journal of the ACM (JACM)*, vol. 37, no. 3, pp. 607–625, 1990.

[108] R. G. Michael and S. J. David, "Computers and intractability: a guide to the theory of np-completeness," *WH Freeman & Co., San Francisco*, 1979.

[109] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[110] H. Bast, D. Delling, A. Goldberg, M. Müller-Hannemann, T. Pajor, P. Sanders, D. Wagner, and R. Werneck, "Route planning in transportation networks," Microsoft Research Silicon Valley, Tech. Rep. MSR-TR-2014-4, 2015.

[111] M. Saeedmanesh and N. Geroliminis, "Clustering of heterogeneous networks with directional flows based on 'snake' similarities," *Transportation Research Part B: Methodological*, vol. 91, pp. 250–269, 2016.

[112] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo-simulation of urban mobility-an overview," in *SIMUL 2011, The Third International Conference on Advances in System Simulation*, 2011, pp. 55–60.

[113] S. Krauss, P. Wagner, and C. Gawron, "Metastable states in a microscopic model of traffic flow," *Physical Review E*, vol. 55, no. 5, p. 5597, 1997.

[114] M. Van Aerde and H. Rakha, "Multivariate calibration of single regime speed-flow-density relationships," in *Proceedings of the 6th 1995 Vehicle Navigation and Information Systems Conference*, vol. 334, 1995, p. 341.

[115] M. V. Aerde, "Single regime speed-flow-density relationship for congested and uncongested highways," in *Presented at the 74th TRB Annual Conference, Washington, D.C. Paper No. 950802*, 1995.

[116] Gurobi Optimization Inc., "Gurobi Optimizer Reference Manual," 2016.

[117] A. Khani and S. D. Boyles, "An exact algorithm for the mean–standard deviation shortest path problem," *Transportation Research Part B: Methodological*, vol. 81, pp. 252–266, 2015.

[118] T. Xing and X. Zhou, "Reformulation and solution algorithms for absolute and percentile robust shortest path problems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 943–954, 2013.

[119] ——, "Finding the most reliable path with and without link travel time correlation: A lagrangian substitution based approach," *Transportation Research Part B: Methodological*, vol. 45, no. 10, pp. 1660–1679, 2011.

[120] M. Mahmoudi and X. Zhou, "Finding optimal solutions for vehicle routing problem with pickup and delivery services with time windows: A dynamic programming approach based on state–space–time network representations," *Transportation Research Part B: Methodological*, vol. 89, pp. 19–42, 2016.

[121] L. Immers and S. Logghe, *Traffic flow theory*, Department of Civil engineering Section Traffic and Infrastructure, Belgium, 05 2003, course H 111.

[122] A. Bryman and D. Cramer, *Quantitative Data Analysis for Social Scientists*. New York, NY, 10001: Routledge, 1994.

[123] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

[124] I. I. Sirmatel and N. Geroliminis, "Economic model predictive control of large-scale urban road networks via perimeter control and regional route guidance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1112–1121, April 2018.

[125] A. Modeling, *Integer Programming Tricks*, 2012.

[126] F. X. O. Suite, *MILP formulations and linearizations: A quick reference*, 2009.

[127] C. G. Cassandras and S. Lafortune, *Introduction to discrete event systems*. Springer Science & Business Media, 2009.

[128] Y. Wardi, R. Adams, and B. Melamed, "A unified approach to infinitesimal perturbation analysis in stochastic flow models: the single-stage case," *IEEE Transactions on Automatic Control*, vol. 55, no. 1, pp. 89–103, 2010.

[129] P. N. Brown and J. R. Marden, "Optimal mechanisms for robust coordination in congestion games," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2437–2448, 2017.

[130] D. A. Hensher, "The valuation of commuter travel time savings for car drivers: evaluating alternative model specifications," *Transportation*, vol. 28, no. 2, pp. 101–118, 2001.

[131] K. Yang, M. Menendez, and N. Zheng, "Heterogeneity aware urban traffic control in a connected vehicle environment: A joint framework for congestion pricing and perimeter control," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 439–455, 2019.

[132] N. Zheng, G. Rérat, and N. Geroliminis, "Time-dependent area-based pricing for multimodal systems with heterogeneous users in an agent-based environment," *Transportation Research Part C: Emerging Technologies*, vol. 62, pp. 133–148, 2016.