Thèse n° 9313

Large-scale traffic signal control and multimodal network design

Présentée le 15 décembre 2022

Faculté de l’environnement naturel, architectural et construit
Laboratoire de systèmes de transports urbains
Programme doctoral en génie civil et environnement

pour l’obtention du grade de Docteur ès Sciences

par

Dimitrios TSITSOKAS

Acceptée sur proposition du jury

Dr S. Joost, président du jury
Prof. N. Geroliminis, Dr A. Kouvelas, directeurs de thèse
Prof. L. Leclercq, rapporteur
Prof. H. Yeo, rapporteur
Prof. M. Bierlaire, rapporteur

École
polytechnique
fédérale
de Lausanne

2022

To Giota, Eva and Roxani
my beloved family,
and to all those who struggle with mental health issues.

Δεν έχουμε παρά μια μονάχα στιγμή στη διάθεσή μας. Ας κάνουμε τη στιγμή αυτή αιωνιότητα. Άλλη αθανασία δεν υπάρχει.

Νίκος Καζαντζάκης

“We have but a single moment at our disposal. Let us transform that moment into eternity. No other form of immortality exists”.

Nikos Kazantzakis

Acknowledgements

When I started my PhD studies five years ago, I was highly unaware of the scale of the challenges, the stress and the uncertainty that I was about to take on before reaching this final day today. Luckily though, as knowing could have stopped me from doing it and I can confidently say that looking back in these last five years at EPFL in Switzerland, only beautiful memories and exciting experiences come up to my mind. My doctoral years have been rich in new knowledge and experiences that helped reshape my view of life and the world and redefine myself. During this time, I was lucky enough to meet, collaborate and enjoy life with many kind and interesting people from all corners of the world. This thesis would have been impossible to realize without the support and friendship of many of them, whose contribution I would like to acknowledge here.

First and foremost, I would like to sincerely thank my supervisor Prof. Nikolas Geroliminis for trusting me and giving me the opportunity to join LUTS, for providing ideal conditions for conducting high-quality scientific research while always being friendly and inviting, and for all the guidance, advise and encouragement that he offered me, in both scientific and personal spheres, that helped me succeed in this endeavor and grow as a person. Secondly, I am also grateful to my co-supervisor Dr. Anastasios (Tasos) Kouvelas, for working closely with me in my first year of PhD, for providing support and valuable advise regarding the research presented in this thesis, as well as for hosting me in SVT lab at ETH Zurich for a month during my academic visit. Thank you both, Nikola and Taso.

I would like to cordially thank all the members of my jury, Prof. Michel Bierlaire, Prof. Ludovic Leclercq, Prof. Hwasoo Yeo and Dr. Stéphane Joost, for the time and energy they put in reviewing my thesis, as well as for all the constructive comments and the interesting discussion that we had in the oral exam.

With PhD work being quite stressful and in principle lonely, everyday life in our lab played a significant role in my well-being. I would like to thank our secretary Christine Debossens, for the immense kindness and support in every issue that came up, and her always joyful and inspiring presence. I am also grateful to all my colleagues in LUTS, older and newer, especially to Reza, Raphaël, Martin, Claudia, Işik, Semin, Leonardo, Mikhail, Caio, Patrick, Manos, Sohyeong, Lynn and Pengbo, with whom we overlapped significantly and interacted the most during the last five years. Special thanks to my dearests Claudia and Martin, for all our shared moments also outside of the office, and for their sincere support, kindness and genuine interest, and to dear Patrick, my office mate, for all the kindness with which he often listened to me when I was down and for his encouraging words. I am also thankful to all the colleagues from TRANSP-OR and VITA, for the interesting discussions and evening drinks that we had in the numerous seminars and conferences that we attended together.

A big amount of support also came from many dear friends outside of the lab, that made the off-work time enjoyable. I wish to express my gratitude to all the friends I met in Lausanne and shared many nice moments, in particular to Eirini, Iason, Ioanna, Georgia, Dimitra, Theodora, Sevi, Giannis, Vaios, Andronikos (we first met at a Christmas choir of the Greek army), Kyriaki and Aspa. I am also grateful to Natercio for all the support, the nice moments, the exciting trips and all his help in improving my spoken French, which was not always a pleasant task. A special, big thank you I owe to my dear Federico, who has been actively supporting me during the last year, which was by far the most difficult for me to handle, by listening, advising, analyzing and going deeper in things, by being patient when I kept interrupting him while talking, for all the practical support and emotional encouragement, the nice moments and of course all the delicious cooking. Grazie mille!

I also wish to thank all my friends from Greece that I was lucky to meet and have in my life, in particular my dear friends from University of Patras, Tonia, Giota, Alexandra, Spyros, Leonidas, Vassilis, and Nora, for all the great time that we have had together and all the kindness and support I keep receiving from them, even if we live far away from each other. Also, I wish to thank my closest friends from Aigio, Stella, Lida and Matina, with whom we have been friends for most of our lives and we still care for and support each other in the best way we can, even with big distance among us. Thank you all for being there.

A special thanks I owe to my psychotherapist Dr. Charalampakis for helping me cope with some particularly difficult moments in the last couple of years, for guiding me all the way in understanding and knowing myself and for helping me see the - often disregarded - importance of mental health in our well-being.

Last but not least, I am deeply grateful to my family, my mother Giota, my sister Eva and my grandmother Roxani, who have always been there for me and contributed the most in making me who I am today, who have been unconditionally supporting me in multiple ways in all periods of my life and have encouraged me to pursue my goals, even when this meant that I needed to be far away from them. I would not have made it without them.

Lausanne, 1/12/2022 Dimitris Tsitsokas

Abstract

Traffic congestion with its multi-dimensional implications constitutes one of the most frequent, yet challenging, problems to address in the urban space. Caused by the concentration of population, whose mobility needs surpass the serving capacity of urban networks, congestion cannot be resolved in the long term by creating more road space. Instead, network capacity can be increased by maximizing efficiency of traffic operations and road space use. Several methods can be employed for this purpose, including optimization of public transportation systems operations and intelligent, adaptive traffic signal control. This thesis contributes to the existing research in these directions, by developing and evaluating new modeling, optimization and adaptive signal control approaches, aiming at improving mobility in highly-congested, large-scale networks, while considering the dynamic characteristics of congestion propagation.

The problem of optimal Dedicated Bus Lanes (DBL) location assignment in large networks with existing bus systems of fixed operational characteristics is addressed in chapter 2. A combinatorial optimization problem is formulated on the basis of an enhanced version of Store-and-Forward paradigm, a dynamic, queue-based macroscopic traffic model, able to properly capture the dynamics of backwards propagation of congestion due to queue spill-backs. Changes in mode choice equilibrium are considered in the evaluation of candidate solutions. An algorithmic scheme based on Local Search, problem-specific heuristics and Large Neighborhood Search (LNS) metaheuristic is developed to address the complex problem. Various destroy and repair operators for LNS are proposed, together with a learning process for assessing the importance of links in terms of receiving DBL, and a network decomposition strategy for accelerating the solution process for very large networks.

A two-layer hierarchical traffic-responsive signal control framework is proposed in chapter 3, combining aggregated multi-region perimeter control (PC) with distributed Max Pressure (MP) control in isolated intersections. Partial deployment of MP in subsets of network nodes is performed and a methodology for identifying critical nodes for MP control based on node traffic characteristics is developed, in the scope of reducing MP implementation cost. Various node layouts of different network penetration rates are evaluated, both in independent MP application and as part of the combined framework. Simulation experiments are performed for a large-scale network of more than 1500 links and 900 intersections for two scenarios resulting in moderately and highly congested states, respectively. Results provide meaningful insights in terms of both independent and combined application of PC and efficient MP control. Under congested conditions, a properly selected subset of critical intersections with MP produces better performance than installing MP everywhere. Adding PC with MP creates even more significant improvements.

Finally, detailed analysis of total remaining travel distance in multi-region networks is performed via microscopic simulation, in the scope of evaluating the benefits of utilizing the recently proposed M-Model, which is disconnected from the steady-state approximation of conventional PL model, in aggregated MFD-based network control applications. Results indicate significant potential improvement in terms of accuracy of prediction, especially in cases of highly-dynamical traffic evolution patterns.

Keywords: dedicated bus lanes, store-and-forward, large neighborhood search, adaptive signal control, perimeter control, Max Pressure, hierarchical control, macroscopic fundamental diagram (MFD), remaining travel distance.

Résumé

La congestion routière, dont les impacts sont multiples, est l'un des problèmes les plus courants et les plus difficiles à résoudre dans l'espace urbain. Due à la concentration de la population, dont les besoins en mobilité dépassent la capacité des réseaux urbains, elle ne peut être résolue à long terme en créant davantage d'espace routier. Cependant, la capacité du réseau peut être augmentée en maximisant l'efficacité des activités liées à la circulation routière et l'utilisation de l'espace qui leur est dédié. Cette thèse vise à contribuer à la recherche existante dans ces directions, en développant et en évaluant de nouvelles approches de modélisation, d'optimisation et de contrôle adaptatif pour améliorer la mobilité dans les réseaux routiers à grande échelle fortement congestionnés, en considérant les caractéristiques dynamiques de la propagation de la congestion.

Le chapitre 2 aborde le problème de localisation optimale des voies réservés aux transports en commun dans les grands réseaux où circulent des services de bus aux caractéristiques opérationnelles fixes. Un problème d'optimisation combinatoire est formulé sur la base d'une version améliorée du paradigme Store-and-Forward, capable de capturer correctement la dynamique de la propagation de la congestion vers l'arrière. Un ensemble algorithmique basé sur la recherche locale combinant des heuristiques spécifiques au problème et la métaheuristique Large Neighbohrhood Search (LNS) est développé pour ce problème complexe. Divers opérateurs de destruction et de réparation de solution pour la LNS sont proposés, ainsi qu'un processus d'apprentissage pour sélectionner les tronçons les plus prometteurs pour l' installation des voies de bus réservées, et une stratégie de décomposition du réseau pour accélérer le processus de recherche de solution pour les très grands réseaux.

Le chapitre 3 propose un cadre de contrôle adaptatif de la signalisation hiérarchique à deux couches, combinant le contrôle au cordon (perimeter control PC) multi-régional avec le contrôle Max-Pressure (MP) distribué dans des intersections isolées. Le déploiement partiel de MP dans des sous-ensembles de nœuds du réseau est effectué et une méthodologie d'identification des nœuds critiques pour le contrôle de la pression maximale basée sur les caractéristiques du trafic aux nœuds est développée, dans le but de réduire le coût de mise en œuvre de MP. Différents agencements de nœuds avec différents taux de pénétration du réseau sont évalués, à la fois en application indépendante du MP et dans le cadre combiné, en utilisant le même modèle de trafic qu'au chapitre 2. Les résultats de simulation fournissent des indications significatives en termes d'application indépendante et parallèle du contrôle PC et du contrôle MP efficace.

Enfin, une analyse détaillée de la distance totale restante à parcourir dans des réseaux multi-régionaux est effectuée par simulation microscopique, dans le but d'évaluer les avantages de l'utilisation du M-Model récemment proposé dans la littérature pour des applications de contrôle de réseau agrégé basées sur le diagramme fondamental macroscopique (MFD) et qui ne nécessite pas l'approximation en régime permanent du modèle PL conventionnel. Les résultats indiquent une amélioration potentielle significative en termes de précision de la prédiction, en particulier dans les cas de modèles d'évolution du trafic hautement dynamiques.

Mots clés: voies de bus réservées, store-and-forward, large neighborhood search, contrôle adaptatif des signaux, contrôle au cordon (PC), max pressure, contrôle hiérarchique, diagramme fondamental macroscopique (MFD), distance de voyage restante.

Contents

List of Figures xv

List of Abbreviations xxi

1 Introduction **1**

1.1 Motivation and Background . . . 1

1.1.1 State-of-the-art in Dedicated Bus Lanes network design . . . 8

1.1.2 State-of-the-art on network-wide traffic-responsive signal control 11

1.1.3 State-of-the-art on MFD-based modeling . . . 15

1.2 Thesis Objectives . . . 17

1.3 Thesis Contributions . . . 18

1.4 Thesis Structure . . . 21

2 Modeling and optimization of dedicated bus lanes space allocation in urban networks **23**

2.1 Introduction . . . 23

2.2 Problem description and modeling framework . . . 24

2.2.1 Problem description . . . 24

2.2.2 Network traffic model . . . 25

2.2.3 Decision variables . . . 29

2.2.4 Total travel time estimation . . . 29

2.2.5 Mode and route choices . . . 31

2.2.6 Optimization problem formulation . . . 33

2.3 Solution methods . . . 35

2.3.1 Algorithm 1: Link-by-link plan construction and link score calculation . . . 36

2.3.2 Large Neighborhood Search . . . 37

2.3.3 Summary of the solution scheme . . . 47

2.4 Numerical application . . . 48

2.4.1 Case study . . . 48

2.4.2 Problem variants . . . 50

2.4.3 Results . . . 51

2.4.4 Best found solutions 58
2.5 Summary 61

3 Two-layer hierarchical adaptive traffic signal control for congested urban networks 63
3.1 Introduction 63
3.2 Methodology 64
3.2.1 Max-Pressure Control 65
3.2.2 Perimeter Control 72
3.2.3 Overview of the two-layer controller 76
3.2.4 Traffic simulation details 77
3.3 Implementation to a large-scale network via simulation 82
3.3.1 Case study 82
3.3.2 Control scenarios 85
3.3.3 Experiment settings 86
3.4 Results 88
3.4.1 Single Max Pressure 88
3.4.2 Two-layer framework: Perimeter control combined with Max Pressure 94
3.5 Summary 102

4 Analysing the effects of remaining travel distance for MFD-based models 105
4.1 Introduction 105
4.2 Model description 106
4.2.1 A review of modeling with production-over-trip length approximation 106
4.2.2 Trip based models and approximations 108
4.3 Simulation and trajectory analysis 110
4.3.1 Network and simulation setup 110
4.3.2 Trajectory analysis 111
4.4 Results 112
4.5 Summary 117

5 Conclusions and future research 119
5.1 Main Findings 119
5.2 Future Research 123

References 127

Curriculum vitae 139

List of Figures

2.1 Evaluation process of any newly formed candidate solution **y** (DBL plan) with mode choice adjustment, for calculation of the respective cost $f(\mathbf{y})$ (PHT). 32
2.2 Sub-networks of different size in the neighborhood of the same reference link (in red). Distance refers to the number of consecutive and connected links included, upstream and downstream of the central link. 46
2.3 Overview of the proposed solution scheme and the relation between algorithms, inputs and outputs; dashed lines indicate alternative options. 47
2.4 The San Francisco road network; (a) Map of the studied area. (b) Model of the network in Aimsun. (c) Map of candidate roads for bus-only lane installation (in blue). 48
2.5 Case study input data about car and bus travel demand; (a) Profile of the initial total car travel demand over time; (b) Average passenger ridership of buses per line in peak-hour; (c) Bus routes distribution in the network; (d) Average bus passenger flow per link in peak-hour. 49
2.6 Algorithm 1: Gradual improvement of solution for problem variant 1 (PHT Only); (a) Step-by-step decrease of PHT (blue solid) and respective car share adjustment (red) as DBL plan is constructed. In dashed line the PHT of the initial state (no DBL) considering the adjusted car share (red line); (b) Step-by-step PHT decrease vs DBL solution length (red) 52
2.7 Link scores calculated by Algorithm 1: Mean ± standard deviation of all trials for candidate links in the sequence they are added to the solution (left to right). 53
2.8 Algorithm 1: Improvement of solution for problem variant 2 (PHT + Cost); (a) step-by-step decrease of the solution cost (blue solid) and the resulting car share (red) as DBL plan is constructed. In dashed line the cost of the initial state (no DBL) considering the car share of the red line; (b) step-by-step cost decrease and solution length (red) as DBL plan is constructed 54

2.9 Performance of LNS algorithms for every type of link scores used. The boxplots show the distribution of different attributes of the best found and the initial solutions for each case; (a) distribution of total cost (PHT + penalty cost); (b) distribution of total travel time (PHT); (c) distribution of total DBL length. 55
2.10 Evolution of the search process of LNS algorithm with different types of link scores; (a) Uniform; (b) based on bus frequency; (c) based on simulation trials (Alg. 1); (d) Uniform with update process; (e) based on bus frequency with update process. Figures show distribution of total cost of the best solution found so far for all performed replications in the form of boxplots. 56
2.11 Selection probability of all candidate links in the Initial and the best found set of solutions for every type of link scores used by LNS algorithm, for all performed replications per case. 57
2.12 Pareto frontier relating the % improvement in total PHT compared to the no DBL case, with the DBL length of the best solutions found by algorithm 1 and LNS algorithm 2 for the two problem variants. . 60
2.13 Spatial distribution of three of the best performing DBL networks identified by different runs of LNS algorithm (links with DBL shown in red) with the respective PHT improvement compared to the initial state of no DBL; (a) LNS using simulation-based scores for problem variant 2 (-12.52%, 2.13 km of DBL); (b) LNS using scores based on bus frequencies plus an update process for problem variant 2 (-12.57%, 4.13 km of DBL); (c) LNS using simulation-based link scores for problem variant 1 with a constraint of reserving maximum 25% of available road space (-12.56%, 2.76 km of DBL). 60

3.1 Schematic description of the two-layer controller. 65
3.2 Detailed structure of the multi-layer controller combining perimeter control and Max Pressure control. 77
3.3 (a) Map of the studied network of Barcelona city center; (b) Model of Barcelona network as a directed graph with annotation of nodes used for PC; (c) schematic representation of controlled approaches for perimeter and boundary flow control, with green time per approach as control variable. 83
3.4 Description of the two demand scenarios in peak period: medium (a)-(c) and high (d)-(f). (a) and (d): aggregated trip distribution between regions; (b) and (e): trip origin density; (c) and (f): trip destination density. 84

3.5 Production MFDs for the on-load of congestion in the case of FTC: (a) MFD of entire network; (b) MFDs per region after partitioning. 84

3.6 Visualization of the MP node selection process for the case of medium demand, according to the proposed method. Each row refers to different penetration rate (5%, 10% and 15%, from top to bottom). First and second column graphs show relations between selection variables m_2 - m_1 and m_2 - N_c, respectively, for all network nodes, for the benchmark case of FTC. Blue dots represent the selected nodes for MP control. Third column figures show the plan of the studied network, partitioned in 3 regions, with the spatial distribution of the selected MP intersections shown as red dots. 88

3.7 Comparison of Total Travel Time improvement, with respect to FTC scenario, of single Max Pressure application, for medium demand scenario. Boxplots refer to 10 randomly created MP node sets for every penetration rate, red triangles refer to node selection based on the proposed method and yellow dot represents the full-network implementation. 89

3.8 Visualization of the MP node selection process for the case of high demand, according to the proposed method. Each row refers to different penetration rate (5%, 10% and 15%, from top to bottom). First and second column graphs show relations between selection variables m_2 - m_1 and m_2 - N_c, respectively, for all network nodes, for the benchmark case of FTC. Blue dots represent the selected nodes for MP control. Third column figures show the plan of the studied network, partitioned in 3 regions, with the spatial distribution of the selected MP intersections. 91

3.9 Comparison of Total Travel Time improvement, with respect to FTC scenario, for single Max Pressure application for high demand scenario. Boxplots refer to 10 randomly created MP node sets for every penetration rate, red triangles refer to node selection based on the proposed method and yellow dot represents the full-network implementation. 92

3.10 Cumulative distribution functions for the changes in node selection criteria m_1, m_2, N_c after MP control implementation for medium demand scenario. Start superscript refers to the FTC case. Blue lines correspond to MP nodes and red lines to the rest. Each row represents a different MP node penetration rate (15%, 20% and 25%). 93

3.11 Cumulative distribution functions for the changes in node selection criteria m_1, m_2, N_c after MP control implementation for high demand scenario. Star superscript refers to the FTC case. Blue lines correspond to to MP nodes and red lines to the rest. Each row represents a different MP node penetration rate (15%, 20% and 25%). 94
3.12 Comparison of Total Travel Time improvement, with respect to FTC scenario, for single PC and combined PC plus MP implementation, for (a) medium, and (b) high demand. Graphs show the performance for different MP node penetration rates (0% is single PC) for the two-layer PC+MP framework. Boxplots refer to 10 random MP node selections per rate while yellow triangles refer to node selection based on the proposed method. 95
3.13 Simulation results for medium demand scenario. Comparison between FTC, MP to all network nodes (100%), MP to only 20% nodes selected by the proposed method, and combined PC with MP to 20% selected nodes. Figures refer to the entire network: (a) MFD of accumulation vs. production; (b) time-series of accumulation; (c) time-series of total virtual queue; (d) time-series of cumulative trip endings. . . . 97
3.14 Simulation results for high demand scenario. Comparison between FTC, single PC, single MP in 100% of nodes and combined PC with MP in 25% of nodes selected by the proposed method. Figures refer to the entire network: (a) MFD of accumulation vs. production; (b) time-series of accumulation; (c) time-series of total virtual queue; (d) time-series of cumulative trip endings 98
3.15 Difference of cumulative link throughput and cumulative time spent with respect to FTC, in medium demand scenario, at the end of simulation (6:00 h): (a)-(b) Case MP in 20% critical nodes; (c)-(d) Case MP in 20% critical nodes combined with PC. Dots and triangles represent MP and PC node locations, respectively. 99
3.16 Difference of cumulative link throughput and cumulative time spent with respect to FTC in high demand scenario at the end of the simulation (6:00 h): (a)-(b) Case single PC; (c)-(d) Case PC combined with MP in 25% of selected nodes. Dots and triangles represent MP and PC node locations, respectively. 101

3.17 Sensitivity analysis of the two best performing control schemes under demand fluctuations: (a) to (d) refer to the medium demand case, and FTC is compared to single MP in 25% of nodes selected by the proposed strategy; (e) to (h) refer to high demand case, and FTC is compared to combined scheme of PC and MP in 25% of selected intersections. Boxplots correspond to 20 different OD matrices where each OD pair is a random variable with normal distribution, with standard deviation equal to 5, 10, 15 and 20% of the mean, respectively. 104

4.1 (a) A well-defined outflow macroscopic fundamental diagram. (b) A four region network without route choice. 107
4.2 Microscopic (Aimsun) model of the network, with clustering results as links and controlled intersections as circles (intersections belonging to u_{12} in yellow, u_{21} in magenta, u_{23} in cyan, and u_{32} in black). . . . 110
4.3 (a) Trip length distribution f of all generating trips in the network; (b) Cumulative input-output diagram for the entire network and simulation time; the arrow indicates the time $t = 2$ hours of maximum accumulation, after which outflow is higher than inflow and the network starts unloading. 112
4.4 m/n plots: (a) Time series of average remaining distance m/n for the whole network. (b) m/n vs. accumulation for the whole network. Average remaining distance in steady state $l^* = 1.28$ km. 113
4.5 Mean value, 25th and 75th percentiles of (a) generating trip length and (b) average remaining distance to be traveled, for batches of 10k vehicles in order of departure time. Values are normalized over those of the first batch to highlight evolution over simulation time. 114
4.6 Cumulative distribution functions of (a) remaining distance of individual vehicles and (b) generating trip length distribution, for different times. 115
4.7 Time series of (a) average remaining distance, and (b) average speed per region ($l_1 = 1192$ m, $l_1^* = 818$ m, $l_2 = 1171$ m, $l_2^* = 888$ m, $l_3 = 852$ m, $l_3^* = 615$ m). 116
4.8 Comparison of m_i observed vs. steady state M model ($m_i = n_i \cdot l_i^*$) vs. PL approximation ($m_i = n_i \cdot l_i$) for each region i. The values of average l_i and l_i^* per region are: $l_1 = 1192$ m, $l_1^* = 818$ m, $l_2 = 1171$ m, $l_2^* = 888$ m, $l_3 = 852$ m, $l_3^* = 615$ m. 116

List of Abbreviations

BPR	Bureau of Public Roads
BRT	Bus Rapid Transit
DBL	Dedicated Bus Lane
GA	Genetic Algorithms
HOV	High Occupancy Vehicle
LNS	Large Neighborhood Search
LS	Local Search
MFD (or NFD)	Macroscopic (or Network) Fundamental Diagram
MP	Max Pressure
MPC	Model Predictive Control
NEF	Network Exit Function
PC	Perimeter Control
PI	Proportional-Integral (controller)
PTP	Public Transit Priority
TSP	Transit Signal Priority
SA	Simulated Annealing
SaF	Store-and-Forward

1

Introduction

This chapter is based on the articles:

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2021). “Modeling and optimization of dedicated bus lanes space allocation in large networks with dynamic congestion”. In: *Transportation Research Part C: Emerging Technologies* 127, p. 103082
- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2022c). “Two-layer adaptive signal control framework for large-scale dynamically-congested networks: Combining efficient Max-Pressure with Perimeter Control”. In: *Transportation Research Part C: Emerging Technologies* (under review)

1.1 Motivation and Background

Traffic congestion nowadays is more and more prevalent in densely populated cities, as a side effect of the contemporary, highly active, urban-centered lifestyle, affecting communities in multiple levels. Congestion repercussions expand from inconveniences such as long delays, low speed and idle in-vehicle time spent in congested roads, to significant environmental and societal impacts owed to the increased carbon emissions, energy waste and severe degradation of air quality, which can even consist a health hazard (Levy et al., 2010). Despite the drastic advancement and amplitude of technological resources that are available today and can enhance congestion-relieving strategies, such as information and communication systems, monitoring equipment, infrastructure intelligence, data-driven artificial intelligence applications and others, resolving the problem of urban traffic congestion still remains particularly challenging. The main reason is the increased complexity in understanding, predicting and controlling the conflicting objectives of the various components that form the basis of congestion: i) traffic network users, who are driven by their personal objectives that determine their mode, route and departure time preferences; ii) traffic managers and city authorities that decide on the transportation systems characteristics, namely the public transport offer, the availability and pricing of road space and parking, the control policies for accessing specific high-demand areas, the speed limit and available signal control strategies and infrastructure; and iii) private transportation companies offering mobility services, from bus, train or metro operators, to ride-hailing on-demand taxi services, that aim to maximize their profit.

Moreover, prevailing travel patterns, including locations of origins and destinations and frequency of desired trips, are highly related to the urban space use and zone separation. The location of activity hubs, which can change with time while the city expands, also affects the decisions and planning of all the aforementioned components, thus creating the need for regular monitoring and continuous planning of the transportation system. Moreover, similar to a game with competing players, each involved party's decision affects the decisions of the rest. For instance, any change of the available traffic infrastructure (e.g. introduction of dedicated bus lanes) can induce further behavioral changes both for commuters (e.g. mode shift to buses or rerouting of drivers to avoid congested bus paths) and for private companies (e.g. different bus frequencies, increased fares etc). The intertwined relations between the decisions of all the above involved parties as well as the dynamic nature of these interactions reveal the complexity of the traffic system which makes congestion a particularly challenging problem to resolve.

The basic mechanism of congestion forming is founded on the fact that demand overpasses system serving capacity, or in the specific case of traffic networks, this translates to many people traveling at the same time towards closely located destinations (e.g. city center) and thus, circulating vehicles tending to occupy the total available road space, creating queues that propagate upstream. On the basis of this simplistic interpretation of congestion, existing strategies aim at either altering demand patterns or increasing system supply. However, this often happens mainly at the local scale and network effects are not considered.

This thesis' principal objective is to contribute to the existing efforts in addressing urban traffic network congestion by utilizing and enhancing state-of-the-art modeling, optimization and control tools for the design of congestion-relieving strategies, with a special focus on considering the dynamics of congestion propagation and the practical applicability to large-scale networks. On this basis, different congestion-relieving strategies are revisited or developed from scratch in every chapter, and a structure of analysis, modeling, optimizing and testing of the proposed methods via simulation is followed. Additionally, users reactions to the interventions under consideration are taken into account in the design phase. Each chapter refers to a different approach with different congestion-relieving mechanisms but built on the same founding elements and objectives.

Demand-targeting strategies focus on reducing the number of vehicles that need to travel at peak hours as much as possible. A sustainable way to achieve vehicle reduction is by offering accessible, reliable, frequent and fast public transportation services, which can serve a large number of passengers while occupying significantly less road space than cars. In order to achieve this in large-scale, massive transportation needs to be a competitive alternative to private vehicle, and to this end, several strategies for prioritizing public transportation are developed, which are commonly known as Public Transit Priority (PTP). The long-term goal of such strategies is to motivate commuters to opt for public transport instead of private car in the city center, leading to reduced car traffic and more efficient road space use.

Various PTP measures have been proposed in the literature, ranging from infrastructure improvements, such as Bus Rapid Transit (BRT) systems (Deng and Nelson, 2011; Levinson et al., 2003; Wirasinghe et al., 2013), application of smart technologies, such as Transit Signal Priority (TSP), pre-signals, queue-jumper lanes or perimeter control (Ampountolas et al., 2017; Christofa et al., 2016, 2013; Farid et al., 2015; Guler and Cassidy, 2012; Guler et al., 2016), or planning for exclusive road space, e.g. bus-only lanes, bus-lanes with intermittent priority or dynamic priority lanes (see Viegas and Lu, 2001, Eichler and Daganzo, 2006, Anderson and Geroliminis, 2020).

Conceptualizing, designing and optimizing effective PTP strategies requires good understanding of the traffic dynamics and interactions of public transit vehicles with the non-public ones, while passenger behavior and travel demand patterns need to be carefully considered. A frequently implemented PTP measure is the introduction of dedicated bus lanes (DBL), which provide exclusive space to buses, allowing them to move easily through traffic even in congested states, and ensuring shorter and more reliable travel time for passengers. However, given that road space is limited in city centers, reserving one lane for public transportation can lead to increased congestion for private vehicles due to reduced available road space. Moreover, if not enough passengers are served by buses using the reserved lanes, road space might be inefficiently used, since the reserved lane might serve more commuters during the day if it was open to general traffic (Dahlgren, 1998). Also, commuter reactions to the newly installed DBLs and the disturbance of traffic equilibria should be taken into consideration during the planning phase, since the allocation of DBLs may significantly improve bus travel times, which might motivate commuters to opt for taking the bus instead of their private car. This can lead to increased bus ridership that would necessitate more frequent bus scheduling. Similarly, private car drivers might adjust their usual paths in order to adapt to the newly created traffic patterns that DBL setting may induce, e.g. in order to avoid narrow roads where DBL have been placed and reduce the available space, leading to lower speed. However, the decision process that is often followed in practice for DBL deployment usually does not consider all of the aforementioned elements, resulting to sub-optimal designs. More specifically, due to high problem complexity, most researchers have addressed the DBL optimal design problem by assuming static traffic conditions using the empirical Bureau of Public Roads functions to estimate travel time (Mesbah et al., 2011,Yu et al., 2015,Yao et al., 2012) and focused on the bi-level structure of the problem, in order to account for users reactions. However, congestion propagation due to queue spill-backs is not accurately considered in the DBL network design process when static traffic modeling and assignment is assumed, in spite of being closely interrelated to DBL relative locations. Moreover, most existing approaches are applied to small-scale networks with only a few roads and intersections, while they might be unsuitable for large networks due to high complexity of the involved formulations. Also, DBL location decisions are often made based on empirical rules or simple scenario evaluation and are not product of optimization.

Other demand-targeting strategies include offering incentives for departure time shifting (Li et al., 2021,Sun et al., 2020), imposing fees to vehicles that circulate in defined high-demand areas such as city center during peak times (cordon pricing, e.g. see Geroliminis and Levinson, 2009,Zheng et al., 2012, Zhang and Yang, 2004) or even banning some or all of them, offering free parking spaces outside the congested zones and connecting them to popular destinations with mass transit services for last-mile of trips, providing infrastructure and promoting the use of bicycle or other light modes of transport, occupying little road space and consuming less energy. Decentralizing activity hubs is also contributing to the same objective, which however can be effective mostly in the urban planning phase, and concern future projects development.

On the other hand, increasing the supply capacity of traffic systems is more cumbersome, as capacity is primarily associated with road space, which is scarce in city centers that usually attract the majority of travel demand. In fact, it has been shown that increasing road space, apart from being extremely costly and often impossible, it is not a sustainable solution to congestion, since it can induce new demand which will lead to similar congestion levels in the near future (Toth, 2007). Therefore, in order to increase system supply, more effort is directed towards effective flow control through traffic lights, which can have a significant impact on the spatio-temporal formation of queues and, consequently, improve mobility and increase serving capacity by reducing travel time. Conventional control of traffic lights on intersections aims at moderating circulation of competing streams in order to avoid accidents and provide safe and fair right-of-way to everyone. Conventional strategies for signal planning aim at discharging all queues after one control cycle, before the plan is repeated, and are based on historical data of flows, which the plan is designed to serve. The signal plan that results by this process is executed continuously without change (static plan) and is designed for undersaturated conditions (see Webster, 1958, Allsop, 1971, Allsop, 1976). Coordination of consecutive intersections is also found to increase serving rate, by implementing the logic of the "green wave" (e.g. Nagatani, 2007,Lämmer and Helbing, 2008, Kong et al., 2011), which aims at minimizing the number of times platoons of vehicles are forced to stop by a red light, when passing through a series of consecutive intersections. However, the effectiveness of conventional fixed-time static signal control as well as signal coordination, is limited under high levels of congestion, where adaptive control systems are more agile. This is why more advanced control systems have been proposed, as we will see below. A comprehensive review of network control strategies can be found in Papageorgiou et al., 2003.

Developing intelligent, adaptive flow control systems, able to improve mobility and reduce congestion in high demand cases of large-scale networks, requires in-depth understanding of the physical characteristics of congestion. Numerous research works have focused on understanding and describing urban congestion in an effort to develop descriptive models that can be used to simulate traffic evolution and consequently support the design and test of advanced control strategies. Several categories of traffic models exist, often classified based on the level of detail with which traffic is represented: microscopic models focus on individual spatio-temporal vehicle characteristics (e.g. acceleration, speed, position) and interactions between vehicles; macroscopic models describe traffic evolution in an aggregated way, using variables such as flow, density and space-mean speed, which refer to entire network links or neighborhoods, without monitoring individual vehicles; mesoscopic models represent traffic dynamics in an aggregate way but utilize probability distributions to consider individual user choices. The choice of model type is determined by the nature and requirements of the intended application.

Macroscopic traffic modeling, specifically, became highly popular mainly after the emergence and empirical validation of the concept of Macroscopic (or Network) Fundamental Diagram (MFD or NFD), which represents a demand-insensitive, unimodal relationship between the accumulation (the number of vehicles inside a region) or density, and the respective travel production or average speed, provided that traffic is homogeneously distributed within the region. Theoretical existence of MFD was firstly described by Godfrey, 1969, later revisited by Mahmassani et al., 1984 and Daganzo, 2007, while its existence with dynamic features was empirically validated with real data for the city of Yokohama by Geroliminis and Daganzo, 2008. This relationship enabled the development of aggregated modeling and control strategies for neighborhood-sized regions, that were much simpler and required significantly less input data to operate. Among MFD-based control strategies, perimeter control (PC) is established as a state-of-the-art strategy of traffic-responsive control in the network-scale, which attracted massive recent attention due to its effectiveness in preventing one or multiple, homogeneously congested, protected regions from reaching highly congested states. This is done by controlling flows crossing the perimeter of the protected regions via updating the respective traffic signal plans. Based on real-time traffic information, such as regional accumulation or density, which can be measured by a set of detectors, the controller is activated to increase or decrease the flow rate entering each region, with the aim of maintaining maximum possible travel production, according to the relationship dictated by the respective regional MFD. Given that low-scatter MFD can be identified only in regions with homogeneous traffic distribution (low variance between queues of links inside the region), clustering methods have been proposed for the partitioning of heterogeneous large networks to homogeneous neighborhoods. The same control logic, also known as gating, is proposed for application on the entrances of highways, known as "ramp metering" (e.g. see Papageorgiou et al., 1991; Papageorgiou and Kotsialos, 2002). In contrast to fixed-time static signal control, these strategies adjust and often optimize signal plans of the controlled intersections in real-time, depending on the type of implemented control scheme and based on the prevailing traffic conditions.

In fact, numerous different types of traffic-responsive signal control systems have been proposed and implemented in the field. According to architecture and communication requirements, they are classified to centralized and distributed systems. Centralized systems, such as PC, are designed for adaptive control of entire networks and update the timing of multiple traffic signals at the same time, based on aggregated traffic information coming from several points of the controlled network. Information is collected at a central computer and processed within a specific algorithm which generates new signal plans. Then the updated plans are communicated back to the controlled intersections and applied in the next cycle. Distributed systems on the contrary are applied to the local scale, often in isolated intersections, from where traffic data is collected and processed for updating the respective signal plan, without knowing or directly affecting signal planning of other intersections in the network. However, the impact of the control in neighboring nodes is indirectly taken into account through the real-time traffic data that is collected locally and can be affected by neighboring queues.

A well-known and frequently studied distributed controller applied to isolated intersections is Max Pressure (MP), which was formulated for traffic signal control by Wongpiromsarn et al., 2012 and Varaiya, 2013a,b, who theoretically proved its ability to stabilize queues and maximize throughput under any demand scenario that can be controlled, but based on some relatively restrictive assumptions (point-queues of unlimited capacity). The controller receives traffic information from links upstream (or also downstream) the intersection and updates the signal plan giving more green time to queues that have high probability of discharging vehicles during the entire green interval, thus maximizing green time utilization. Several variants of the initial MP version are proposed and promise increased efficiency and better performance in network-wide application. However, due to increased requirements in traffic monitoring equipment, practical implementation of MP becomes complicated and costly, thus remains limited. Nevertheless, distributed systems are easier and less costly to implement than centralized, due to limited communication infrastructure required, while being scalable, in the sense that they can be installed gradually in several intersections without affecting the previously installed ones. Both system types have shown promising results in decreasing travel delay in usual traffic conditions, each through its own mechanism.

Despite the significant contributions in the field of responsive signal control, both centralized and distributed systems face different challenges and can prove insufficient under heterogeneous traffic distribution or highly congested conditions. On one side, local controllers focus on limited areas and have no knowledge of the traffic conditions elsewhere in the network. While this can be seen as an asset in terms of simplicity in practical application, there is a limit on the impact they can have under highly congested conditions. This is justified by the fact that received traffic information only refers to the proximity of the controlled intersection, and it may arrive too late for local controllers to react and have a network-wide impact, since congestion often evolves rapidly and results in instability and fast degradation of network serving capacity. Then, in a state of completely gridlocked queues, local controllers such as MP, might function similarly to static ones, since all approaches would be in equal need of green time. Furthermore, the impact of partial implementation of local controllers in the performance of network-scale control is rarely studied and, thus, is still uncertain. On the other side, centralized, aggregated strategies such as PC, aim at holding vehicle flows outside the perimeter of a high-demand protected zone, by adapting traffic signal timing of the approaches located on the perimeter of the zone. However, MFD-based PC relies on the assumption of homogeneous traffic distribution within the controlled regions, which is not always realistic, especially when congestion starts building and local pockets of congestion unavoidably appear. Secondary sources of heterogeneity can even be endogenously generated by PC strategy, since the applied flow control often leads to queue creation on the borders between regions.

This thesis focuses on two principal types of congestion-relieving strategies. The first one refers to optimal DBL allocation in existing urban networks, with the focus being on modeling and optimization methods. The second one refers to traffic-responsive signal control scheme design, with the focus being on investigating parallel application of different types of controllers (centralized and distributed) with different and potentially conflicting objectives. Finally, a smaller-scale study on MFD-based modeling is included, where a recently proposed, novel modeling framework incorporating total remaining travel distance dynamics, which can potentially improve PC performance, is evaluated through microscopic simulation.

A detailed literature review about each distinct thematic circle of the thesis is presented in the following subsections. The current state-of-the-art is described and research gaps that motivate and determine the specified research objectives of this thesis are elaborated. Finally, objectives and contributions of the thesis are summarized and thesis structure is described at the end of this chapter.

1.1.1 State-of-the-art in Dedicated Bus Lanes network design

Among PTP strategies, Dedicated Bus Lane (DBL) installation has seen wide implementation due to its effectiveness in reducing bus delays caused by traffic interactions, while being inexpensive and relatively fast to implement. Space separation also results in increased punctuality and reliability for the bus service. However, space reserved for transit vehicles (simply speaking buses) is taken from non-transit traffic, which can potentially induce congestion due to local capacity reduction, e.g. in neighborhoods of infrequent bus service or inflexible non-transit demand (Dahlgren, 1998), indicating that spatial distribution plays an important role in the efficiency of DBL strategy. The great potential of DBL, however, is demonstrated in numerous studies performed in the last few decades, addressing the topic from several perspectives and in different scales.

Microscopic traffic simulation is often utilized for the evaluation of roads and road networks with existing DBL systems (e.g. Choi and Choi, 1995; Waterson et al., 2003; Wei and Chong, 2002), as well as for alternative scenario analysis for future development (e.g. Abdelghany et al., 2007; Arasan and Vedagiri, 2010; Chen et al., 2010; Shalaby, 1999). Abdelghany et al., 2007 developed a dynamic assignment and simulation framework for the evaluation and planning of BRT systems by integrating exclusive road space for buses. Stirzaker and Dia, 2007 performed microsimulation analysis for a real corridor in Brisbane, Australia, in order to assess the impact of setting a DBL or High Occupancy Vehicle (HOV) lane, aiming at road use maximization. Khoo et al., 2014 also utilized microscopic simulation for the DBL allocation and scheduling problem in a bi-objective formulation and applied Genetic Algorithms (GA) to solve it. Farid et al., 2018 developed analytical models based on the kinematic wave theory for person-based evaluation of the combined effect of DBL, Queue-Jumper Lanes and TSP strategies in signalized intersections using microsimulation. Numerical models have been applied by Basso et al., 2011 for mode choice and car-bus interactions modeling in a synthetic road, in order to compare different congestion management policies, including congestion pricing and transit subsidization, while optimizing bus operations and DBL provision for maximizing social welfare. Although microscopic simulation may provide an analysis of increased accuracy compared to macroscopic or empirical models, it is often too expensive computationally to be included in an optimization framework, especially for large networks. Most of these studies involve local scale DBL applications in specific arterial roads, freeways, or small corridors, and their outcomes result from scenario evaluation, rather than a well-established optimal design process.

Going a step further, strategic planning methodologies for DBL allocation have also been proposed, with a special focus on handling passenger response concerning mode and route choices. Bi-level programming is typically utilized, where a social optimum objective for the DBL location assignment is decided by traffic planners in the upper level, and the respective user reactions to these decisions are determined in the lower level, given that users seek to maximize their personal utilities by making appropriate mode, route, and departure time selection. Mesbah et al., 2011 have modeled the DBL location selection problem as a Stackelberg competition and applied decomposition and cutting planes techniques for solving the corresponding bi-level program. A similar formulation was later proposed by Yu et al., 2015, where column generation, branch-and-bound, and successive averages have been applied for deriving the solution. Miandoabchi et al., 2012 have formulated a discrete bi-modal Network Design Problem (NDP) as a bi-level program with equilibrium constraints and several decision variables, including new road construction, lane addition, lane direction assignment, and bus-only lane assignment, which they have addressed with hybrid metaheuristic algorithms. Yao et al., 2012 have formulated a bi-level programming model for DBL setting and bus frequency optimization, considering an integrated network equilibrium model in the lower level, and used GAs for optimization. Sun and Wu, 2017 and Zhao et al., 2019 also proposed bi-level programming structures while focusing on multiple objectives and operational intersection dynamics and applied GAs for the solution process.

While bi-level programming integrates modeling of passengers reactions in the strategic planning phase, it often leads to highly complex problem formulations requiring excessive computational resources, which, again, makes the application practically impossible to large-scale networks. Moreover, steady-state traffic conditions are typically assumed in such formulations, with link travel times being calculated by empirical relationships on the basis of “Bureau of Public Roads” (BPR) functions proposed in the Highway Capacity Manual NRC, 2010. However, static traffic assignment cannot capture the effects of queue formation and spill-backs related to backward propagation of congestion, which can be highly correlated to DBL presence, resulting to potentially unrealistic estimation of network performance.

Interesting studies on urban road space management based on macroscopic traffic modeling also exist, founded on the concept of MFD. Gonzales et al., 2010 utilize the MFD concept for urban space management and distribution between competing modes, and show that reserved lanes for transit or high occupancy vehicles can reduce traffic-related delays even for non-prioritized vehicles. Zheng and Geroliminis, 2013 propose the concept of a multi-modal MFD to capture the dynamic interactions between competing modes in multi-region cities, and develop a framework for road space allocation that optimizes passenger throughput. Geroliminis et al., 2014 and Chiabaut, 2015 study the concept of passenger MFD (p-MFD) as a useful tool to integrate in the development of optimal transit operation strategies. Zhang et al., 2018 propose an MFD-based framework for optimizing road capacity management together with transit operations, by integrating mode choice and dynamic user equilibrium in modeling dynamics. More recently, Anderson and Geroliminis, 2020 study the impact of a controlled bus lane, which allows regular vehicles to enter under certain conditions, according to a dynamic control framework based on MFD modeling. Nevertheless, although MFD-based approaches are able to capture dynamic interactions between competing modes, they do this in an aggregated way for a region or arterial road and cannot drive decision making processes on the link-level of an urban network.

In order to respond to the limitations of existing research, part of this thesis will address the problem of optimal DBL location selection by proposing a modeling framework on the basis of a dynamic urban traffic model with queuing characteristics, typically used in model-based control applications. This type of traffic representation can capture the topological variations of congestion propagation in the road level and evaluate the impact that candidate DBL location schemes will have in the resulting congestion patterns. At the same time, it is easily integrated in an optimization framework. To the best of our knowledge, there have been particularly few attempts in the same direction. Li and Ju, 2009 have proposed a modeling framework based on a point-Q model, where a Variational Inequality formulation is integrated to capture the mode, route, and departure time choices of passengers as a result of DBL presence, but only tested two alternatives in small network instances. More recently, Bayrak and Guler, 2018 have also used a dynamic traffic model with horizontal queues to identify the best DBL plan and used GA to solve the optimization problem for a small network instance.

With the aim of constructing a method simple enough to be applied to large networks, in chapter 2, we simplify the typical bi-level programming structure by removing the necessity of solving interconnected optimization problems in the lower lever to account for passenger mode and route responses. Instead, we utilize a simple Logit model to capture commuter mode preferences based only on the estimated travel time per mode, while we assume that route choice patterns remain unaffected by DBL introduction. Even though dynamic traffic assignment model could be integrated for a more realistic representation of route choices, it might add significant computational complexity in the problem without guarantee of significant increase in accuracy. Nevertheless, the developed optimization framework is quite general and could be utilized with enhanced modeling efforts in the future.

While GA is a typical method choice for this type of problems, in our case, dynamic traffic modeling significantly increases complexity, thus population-based metaheuristics become less efficient. Moreover, apart from the excessive computational cost for finding good-quality solutions, GAs require fine-tuning of several operational parameters, such as population size, mutation and crossover rate, while their searching process is characterized by low interpretability. This makes it difficult to introduce problem intuition to direct the search towards potentially better solutions and accelerate the optimization process, while hard constraints cannot be handled directly. Instead, local search (LS), often introduced in a Simulated Annealing (SA) framework, has shown promising results in tackling combinatorial optimization problems referring to location or resource allocation (e.g. Zockaie et al., 2018). More specifically, Large Neighborhood Search (LNS), firstly proposed by Shaw, 1998, has been successfully applied in vehicle routing problems (VRP), but also in public transport scheduling, facility location and logistics, due to its ability to efficiently explore large solution spaces by using heuristics. A comprehensive review on the various applications of LNS and its more recent variant A-LNS can be found in Pisinger and Ropke, 2019.

1.1.2 State-of-the-art on network-wide traffic-responsive signal control

Several types of traffic-responsive signal control systems have been proposed and applied in the field, incorporating different optimization methods, as well as modeling and control approaches. Some of the most commonly used are SCOOT (Hunt et al., 1981), OPAC (Gartner, 1983), PRODYN (Henry et al., 1984), SCATS (Lowrie, 1990), UTOPIA (Mauro and Di Taranto, 1990), and the more recent ones RHODES (Mirchandani and Head, 2001) and TUC (Diakaki et al., 2002). These systems employ a centralized control logic, in the sense that traffic information from the entire controlled region is required to be transferred to a central processing unit, where the respective control algorithm will process it and determine the control actions that need to be taken, which must be communicated back to every single intersection. This architecture, although it constitutes the state-of-the-art in the field of adaptive signal control due to higher number of degrees of freedom, is characterized by low applicability. This is due to its requirements in communication and computing infrastructure, which imposes a high installation, maintenance and operational cost. Also they provide relatively low scalability, in the sense that gradual installation or expansion of existing system to cover a larger region is often not possible or considerably difficult. Another implementation difficulty relates to the exponential complexity optimization algorithms that many of these systems employ, which prohibits network-wide, real-time central application due to computational cost. On the contrary, decentralized approaches, based on local control of isolated or coupled intersections, have significantly lower infrastructure requirements and algorithmic complexity, increased scalability, and yet high potential to improve network performance in dynamically-changing traffic conditions, even by starting from a local scale. An overview of the different aspects of centralized and decentralized control strategies is found in Chow et al., 2020, while simulation experiments on such strategies are performed by Manolis et al., 2018. A detailed review of existing research in decentralized adaptive signal control can be found in Noaeen et al., 2021.

Max Pressure (MP) is among the most popular distributed feedback-based controller (also known as Back Pressure) proposed for isolated traffic intersections. Its function is based on a simple algorithm that adapts the right-of-way assignment between competing approaches in real time, according to feedback information of queues forming around the intersection (upstream and downstream), in a periodic cyclic process. In the core of MP lies a pressure component, which quantifies the actual queue difference between upstream and downstream links of every phase, and determines phase activation or green time assignment during the following time slot. Initially proposed for packet scheduling in wireless communication networks by Tassiulas and Ephremides, 1990, MP was formulated as a signalized intersection controller through the works of Varaiya, 2013a, Varaiya, 2013b Wongpiromsarn et al., 2012, Zhang et al., 2012 and was theoretically proven by Varaiya, 2013a to stabilize queues and maximize network throughput for a controllable demand, based on specific assumptions, such as point-queues with unlimited capacity and separate queues for all approaches. The theoretical stability proof, despite the constraining assumptions on which is was based, together with the element of independence from any demand knowledge, and the distributed and scalable layout, render MP a promising control strategy, especially for signalized networks facing unstable, excessive or dynamically variable congestion. These traits have motivated a vast amount of research works focusing on MP controller, which generated numerous enhanced and/or improved versions, some of which are listed here. Gregoire et al., 2014b and Kouvelas et al., 2014 introduced normalized queues in the pressure calculation, thus implicitly taking into account link size and spill-back probability, and as a result, queue capacity, which was considered infinite in the initial MP version of Varaiya, 2013a. Gregoire et al., 2014a presented an MP version with unknown turn ratios but existing loop detectors for all directions at the exit line. Xiao et al., 2015b and Xiao et al., 2015a proposed extended MP versions able to address bounded queue length estimation errors and incorporate on-line turn ratio estimation, as well as dynamically update control settings over space according to demand. Le et al., 2015 applied a strict cyclic phase policy, in contrast to phase activation based on pressure which can induce long waiting time to some drivers, and provided stability proof, which applied also for non-biased turn ratios. Zaidi et al., 2015 integrated rerouting of vehicles in MP algorithms. Li and Jabari, 2019 proposed a position weighted back pressure control, on the basis of macroscopic traffic flow theory, and integrated spatial distribution of vehicle queues in pressure calculation. Wu et al., 2017 proposed a delay-based version of MP controller in order to increase equity of waiting time among drivers around the intersection. In Mercader et al., 2020, pressure is calculated by using travel time estimation instead of queue length, in an attempt of relaxing the need for expensive queue measuring equipment, and findings are supported by simulation and real field experiments. Levin et al., 2020 propose an alternative signal cycle structure with maximum cycle length. However, even though many of the above works provide stability and throughput maximization proofs, they usually refer to moderate and feasible demand sets, while MP performance in highly congested networks is questionable. Unstable behavior in such conditions can be attributed to the latency of local controllers, in general, in reacting to the rapid forming of congestion, given the lack of knowledge about traffic conditions upstream and out of the proximity of the controlled intersection. Also, in the case of MP, there is little area for improvement if most (or all) controlled queues are saturated, in which case, control tends to approximate the fixed-time plan. In addition, it should be noted that most of the relative theoretical proofs regarding the efficiency and robustness of MP controller are based on simplified assumptions of point queues and infinite link storage capacity. Nevertheless, at the network level, network capacity (as expressed by the MFD) cannot be reached for high vehicle densities due to long queues that spill-back from one link to another. Thus, it is interesting to evaluate the efficiency of max pressure controller at the network level under heavily congested conditions considering the spill-back effect of queues. This thesis tries to shed some light in this direction.

Perimeter control (PC), on the other hand, as briefly described above, is an aggregated, centralized approach that has been gaining momentum over the last decade, mainly due to significantly reducing complexity in network-wide congestion control. On top of the modeling simplicity, it has light requirements for system integration (only the aggregated regional accumulations are needed), while only a small number of traffic signals are required to function as actuators of the control decisions. The strategy, also known as gating, is based on the principle of regulating vehicle flows withing a defined, high-demand region (e.g. city center), in order to prevent high congestion, which can lead to gridlocks, delays and service rate decline. Exchange flow regulation happens via dynamical adjustment of the corresponding green times at the intersections located on the regional boundaries (perimeters). The method has intensively been studied on the basis of MFD modeling, and a large number of MFD-based PC schemes have been proposed, analyzed and evaluated in the recent years, utilizing different modeling and control methods and focusing on different control aspects. Proportional-Integral (PI) feedback regulator for single- and multi-regional networks is implemented in Keyvan-Ekbatani et al., 2012, Keyvan-Ekbatani et al., 2015a, Aboudolas and Geroliminis, 2013, Ingole et al., 2020, optimal model predictive control (MPC) is implemented in Geroliminis et al., 2012, Haddad, 2017a, Haddad, 2017b with boundary queue consideration, while route guidance is incorporated in PC schemes in Yildirimoglu et al., 2015, Sirmatel and Geroliminis, 2017. Stability analysis is done in Haddad and Geroliminis, 2012, Sirmatel and Geroliminis, 2021, integrated PC with freeway ramp metering is proposed in Haddad et al., 2013, robust control is implemented in Ampountolas et al., 2017, Mohajerpoor et al., 2020, adaptive control is implemented in Kouvelas et al., 2017, Haddad and Zheng, 2020, Haddad and Mirkin, 2020, demand boundary conditions are considered in Zhong et al., 2018, bi-modal MFD is proposed in Ampountolas et al., 2017; Geroliminis et al., 2014, cordon queues impact is assessed in Ni and Cassidy, 2020, remaining travel distance dynamics are integrated in MFD modeling and control in Sirmatel et al., 2021, while data-based model-free PC implementations are proposed in Ren et al., 2020, Chen et al., 2022, Zhou et al., 2016. However, despite the impressive amount of literature and the promising results in terms of network-wide traffic performance, there are still certain concerns regarding practical application of MFD-based PC. One is the requirement for homogeneous congestion distribution, which is not common in real congested networks and can even be endogenously compromised, by the PC gating strategy, that tends to create local queues on the boundaries between clusters, which are themselves source of heterogeneity and compromise PC effectiveness. A second one is that PC can have little to no effect in cases of low to medium demand patterns that yet lead to heterogeneous traffic distribution, where gridlocks appear in specific parts/paths of a seemingly uncongested network, creating spill-backs and local congestion pockets, and causing delays.

Driven by the above review of centralized and decentralized/distributed signal control, developing multi-layer control policies comprising of both local and aggregated strategies, seems an intuitive way of achieving network control with multiple objectives. Various multi-layer hierarchical control structures involving perimeter control in the upper layer have been proposed, often with a lower layer focusing on reducing heterogeneity in local scale. Ramezani et al., 2015 propose a two-layer feedback controller, where MPC is utilized to solve the optimal PC problem in the higher layer, while a feedback homogeneity controller is embedded in the lower layer, which acts as PC in a sub-regional level, aiming at decreasing heterogeneity within regions. However, this has only be tested in a simplified network structure. Similarly, Keyvan-Ekbatani et al., 2015b propose a multi-layer MFD-based gating strategy for concentric cities, where gating is applied to different perimeters at different times, depending on the spatio-temporal evolution of congestion. Zhou et al., 2016 propose a two-layer framework where demand-balance problem between homogeneous subnetworks is treated via MFD-based PC in the higher layer, and a more detailed traffic model is embedded in the lower layer for optimizing all signals within each subnetwork, with both layers being based on MPC optimal control formulations. Based on the same concept, Fu et al., 2017 propose a three-layer hierarchical control strategy, also including a stability analysis top layer. In all these works, hierarchical communication schemes between layers are required for exchange of information between controllers, which results in increased infrastructure requirements. Furthermore, MPC optimal control formulations increase complexity and computational cost, and performance is affected by the accuracy and fine-tuning of the utilized models. Yang et al., 2017 proposed an MPC-based controller incorporating two objectives of network-wide and intersection-scale delay minimization, by formulating and linearizing a complex multi-objective optimal control problem, applied in a connected vehicle environment. However, apart from the increased complexity, model-dependent accuracy and requirement of detailed demand information of MPC, local delay minimization was focused only on intersections on the perimeter of the protected region, while the coupled control scheme is centralized, meaning that communication infrastructure to central controller is required, in contrast to MP which does not have this requirement. Keyvan-Ekbatani et al., 2019 examine the effects of combining PC with two different local adaptive controllers, a volume-based strategy and a simplified SCATS strategy. Performance results showed significant improvement in the cases of combined PC and local traffic-responsive control strategies. However, due to the size of the test network, only a small number of intersections were available for the local controllers and no investigation of the spatial control layout was held.

1.1.3 State-of-the-art on MFD-based modeling

As already mentioned previously, a framework for optimal PC was introduced by Daganzo, 2007, who described a theoretical model of adaptive control at an aggregate level, applicable to cities partitioned in neighborhood-size reservoirs. The work discussed the benefits of an inflow-withholding policy aimed at maintaining regional service rate close to maximum, showing how overall travel performance can increase, even for vehicles that are forced to wait before entering the protected region, similar to on-ramp metering for highways. This policy was based on a steady-state approximation of a demand-insensitive, uni-modal relationship between vehicle accumulation and travel completion rate, that became later known as the Macroscopic (or Network) Fundamental Diagram (MFD or NFD), and derived an optimal control policy for a single-region network. Nevertheless, while in a single-region PC, it is straightforward to show that the region should perform (if possible) in the maximum flow point of the MFD, multi-region PC might have very different features. Given the antagonistic decisions of boundary controllers, it is highly probable that not all regions can operate simultaneously at the critical flow, which requires more advanced strategies.

MFD concept, firstly introduced by Godfrey, 1969, was theoretically founded and described by Daganzo, 2007, and was empirically validated with real data for the city of Yokohama, Japan, by Geroliminis and Daganzo, 2008. Similar studies focusing on multi-modality and critical accumulations based on real traffic data are performed by

Paipuri et al., 2021 and Loder et al., 2019. Well-shaped low-scatter MFD is observed for homogeneously congested, compact regions (with low link-flow variance) with relatively low scatter, which constitutes an elegant, dynamic modeling and control tool, that can serve as a basis for simplified, aggregated, network-scale, adaptive flow control schemes. Numerous researchers focused on the concept of MFD producing significant contributions regarding its existence preconditions, characteristics, shape, hysteretic behavior (Buisson and Ladier, 2009, Ji et al., 2010, Mazloumian et al., 2010, Geroliminis and Sun, 2011b, Gayah and Daganzo, 2011, Mahmassani et al., 2013), and many more incorporated MFD theory in developing aggregated models for dynamic traffic-responsive urban network control for congestion mitigation (some recent works are Yildirimoglu et al., 2018, Mariotte et al., 2017, Mariotte and Leclercq, 2019, Laval et al., 2017, Batista et al., 2021). Bi-modal MFDs are utilized in Geroliminis et al., 2014, Loder et al., 2017, Haitao et al., 2019, Paipuri and Leclercq, 2020, Paipuri et al., 2021. MFD application to ride-sourcing systems in multi-modal networks is proposed by Wei et al., 2020 and Beojone and Geroliminis, 2021. Modeling and dynamic control of taxi operations based on MFD is described in Ramezani and Nourinejad, 2018.Network partitioning research for the purposes of MFD-based control was also intensified (Ji and Geroliminis, 2012, Saeedmanesh and Geroliminis, 2016, Ambühl et al., 2019). For an overview of MFD research with respect to traffic modeling, the reader could refer to Johari et al., 2021.

When Daganzo, 2007 re-introduced MFD, he associated them with the concept of network exit function (NEF), thereby highlighting their potential for regional traffic flow management. A NEF is a function expressing the outflow of the zone, i.e. trip completion rate. While outflow is difficult to be measured (unless numerous vehicles are equipped with tracking devices), network production (veh-km traveled per unit time) can be estimated using conventional methods (e.g., loop detectors data) if one assumes that outflow O and production P are linearly related with the average trip length l of a region, i.e., $O = P/l$. We will refer to this specific NEF as the outflow MFD or PL model, while the relations between accumulation and speed (resp. production) will be referred to as speed (resp. production) MFD.

Recently, several authors introduced an alternative description of congestion dynamics, based on the speed MFD but avoiding the steady state approximation (Arnott, 2013; Daganzo and Lehe, 2015; Fosgerau, 2015; Jin, 2020; Lamotte and Geroliminis, 2018), the so-called trip based model (i.e., TB model). This was mainly related to studying the departure time choice problem at the city scale. Trip-based model is computationally more demanding than PL model and may cause intractability; also, it cannot be written in a compact ODE form. Yet, it also provides a sounder treatment of propagation phenomena, avoiding some artifacts associated with PL model, such as the temporary reduction of experienced travel time that can follow a demand surge (see Lamotte and Geroliminis, 2016). A recent paper by Mariotte et al., 2017 analyses some of these issues, but focuses primarily on the consequences of a non-stationary inflow with homogeneous trip length. Nevertheless, such models require full knowledge of the trip length distributions, and, due to the complex dynamics, would potentially cause the numerical optimal control methods based on such models to suffer from problems associated with a high dimensional state.

Lamotte et al., 2018 and Murashkin, 2021 introduced a third type of NEF for single-region (i.e., the so-called M-model), which reproduces particularly well the behavior of the trip-based model at a much lower computational cost. This is achieved by keeping track of the average remaining distance to be traveled. By comparison, trip-based model keeps track of the distance remaining to be traveled by each individual user, while PL model does not keep any record of traveled distance. Such a model offers valuable intuition and represents an attractive trade-off for control applications. The same references showed that all three models (i.e., TB, PL and M) are equivalent in the steady state or when trip length follows an exponential distribution, but this might not be a realistic assumption; nevertheless, the M-model, even if it makes physical sense, has never been tested against real or simulated data.

1.2 Thesis Objectives

The primary and foremost objective of this thesis is to contribute to the existing research on strategies that aim at reducing congestion in large-scale urban space. Based on the preceding review of the existing literature on the main thematic areas of the thesis, there exist a set of research areas that can be further explored. Therefore, the following objectives are set, organized in chapters following the structure of the thesis:

- **Chapter 2: Modeling and optimization of DBL space allocation.** The aim of this chapter is to address the problem of optimal DBL assignment at the link level in existing urban networks by incorporating congestion dynamics and passengers mode choice. The framework for optimal DBL layout identification is to be designed on the basis of a link-level queue-based traffic replication model that can take into account backwards propagation of queues in congested conditions, while being able to properly model the impact of road space reduction for general traffic due to DBL installation. Given that implementation should be possible for large-scale networks while problem complexity can be high, proper optimization techniques are to be explored and customized, in order to provide good-quality solutions with acceptable computational cost.

- **Chapter 3: Two-layer hierarchical adaptive signal control.** The aim of this chapter is to design and evaluate the performance of a two-layer network control framework consisting of: (i) PC implemented through a Proportional-Integral (PI) feedback controller, which manages exchange flows between homogeneously congested regions by adapting traffic signal plans of intersections on their boundaries; and (ii) a set of independent MP controllers acting on signalized intersections in the interior of the regions, and balancing queues in their proximity, at a local scale. Additional objective is the investigation of partial implementation of MP controller in subsets of network nodes, and the development of a methodology to identify critical nodes, where MP control is potentially more beneficial in terms of global network performance. Node selection is based on node traffic characteristics such as variance and mean of normalized queues and spill-back occurrence of incoming node links. Each control layer is to be tested independently as well as in the form of a combined multi-layer framework, by using a modified version of link-based Store-and-Forward dynamic traffic model, which accounts for spill-back effects and capacity of queues. Potential rerouting of vehicles due to control effects is to be considered in the simulation experiments. To the best of the author's knowledge, no existing research up to date investigates partial MP application to subsets of network nodes.

- **Chapter 4: Analysing the effects of remaining travel distance dynamics for MFD models.** The aim of this chapter is to investigate the potentially improved accuracy of M-Model, a recently proposed macroscopic traffic model suitable for MFD-based control applications, which monitors the total remaining travel distance of all drivers. Microscopic simulation of a realistic demand scenario for large-scale network can provide detailed trajectory information for all generated trips. Analysis of the generated trajectories can reveal the evolution of average trip length distribution and remaining distance over time, which can provide valuable insights regarding the performance of the novel M-Model against conventional PL model, which assumes steady-state conditions. Results can be particularly illustrating with respect to the potential benefits of integrating M-Model to MPC-based optimal perimeter control.

1.3 Thesis Contributions

Driven by the stated objectives and based on the methods and results that will be elaborated in detail in the next chapters, the research conducted in the scope of this thesis leads to the following contributions, listed and elaborated per chapter as follows.

- **Chapter 2: Modeling and optimization of DBL space allocation.**

  A combinatorial optimization problem with binary decision variables is formulated for the problem of optimal deployment of DBLs in urban networks with existing bus system of known operational characteristics. The formulation builds on an enhanced version of link-level macroscopic SaF traffic model that is able to model congestion propagation due to queue spill-backs, thus realistically considering the effects of potential DBL-induced congestion due to reduction of road space for general traffic, which is influenced by the DBL relative positions. The expected adjustment of mode choice of commuters due to a potential DBL layout is also considered in an aggregated way by using a Logit model. A set of solution algorithms based on problem-specific local search heuristic and LNS metaheuristic are employed in the search for the optimal solution. A simulation-based learning process for assessing the potential importance of network links for receiving DBLs is developed and utilized for driving the searching process of LNS. A network decomposition strategy for reducing the computational cost of LNS for very large networks is also proposed. A Pareto frontier associating passenger hours traveled with lane-kilometers of DBL is constructed, that can support the decision making process in case of budget constraints. Testing of all solution algorithms is performed for a large network of more than 450 links and 260 nodes and near-optimal DBL layouts are identified, showing significant improvement in total passenger travel time compared to the benchmark case of no DBL.

- **Chapter 3: Two-layer hierarchical adaptive signal control.**

  The chapter develops a two-layer hierarchical adaptive control framework combining multi-region PC with distributed MP control, which was evaluated on the basis of a dynamic link-level macroscopic traffic model. With the aim of reducing MP implementation cost, partial deployment of MP controllers in fractions of network nodes is investigated for the first time. A method to classify and assess the level of node importance for MP control in terms of global network performance improvement based on measurable node traffic characteristics is proposed and supports the design of the two-layer controller. Both control strategies are implemented independently and as parts of the hierarchical framework, in a large network of more than 1500 links and 900 nodes, for two demand scenarios leading to moderately and highly congested states in the benchmark (fixed-time control) case, respectively. Finite capacity of queues and spill-back effects are properly modeled, while potential rerouting effect of drivers, in an effort to avoid queues induced by the control strategies, is also considered.

  Simulation results provide significant insights in terms of both independent and parallel application of the two control strategies, indicating significant performance gain in the parallel application for both demand scenarios tested. Partial MP implementation appears more efficient and in some cases achieve better performance gains than full-network implementation, both in independent application and in combination with PC. Therefore, significant reduction in the implementation cost of MP can be achieved without loss in performance. Especially in the highly congested scenario, the combined MP-PC scheme outperforms single application of both PC and MP, while single MP shows almost zero improvement. Spatial analysis of the impact of adaptive control schemes on cumulative throughput and total delay per link, with reference to MP and PC controlled node locations, reveals interesting correlation of behavior among queues in the proximity of controlled nodes. Finally, sensitivity analysis under demand fluctuations is performed for MP schemes installed only on critical nodes, showing satisfying performance of the selected node layouts for standard deviation of demand up to 20% of its mean.

- **Chapter 4: Analyzing the effects of remaining travel distance dynamics for MFD models.**

  The chapter investigates the potential benefits of integrating the recently proposed M-Model in aggregated MFD-based network control schemes, such as PC, through analysis of more than 200k detailed vehicle trajectories extracted from microscopic simulation experiment. Disengaged by the assumption of steady state or slowly changing conditions imposed by the conventional PL model, M-Model proposes an alternative relationship between regional outflow and speed MFD, by monitoring network state using total remaining travel distance, apart from regional accumulation, while it does not require the full trip length distribution that its counterpart trip-based model needs. Detailed analysis of simulated vehicle trajectories reveal that, even though average trip length may not vary significantly over time, the same does not apply to average remaining distance, which can increase with the building of congestion. The behavior of remaining travel distance over time is analyzed for a network partitioned in three homogeneously congested regions, showing significant deviation from the steady-state approximation, which indicates the potential improved accuracy of M-Model compared to PL in network control applications under highly dynamic demand. Being the first attempt to investigate the behavior of total remaining distance dynamics through microsimulation, the chapter reveals interesting insights regarding the potential increased accuracy of the novel M-Model, which is later verified by integrating M-Model in MPC-based PC schemes by Sirmatel et al., 2021.

1.4 Thesis Structure

The thesis is organized in 5 chapters, separated by thematic area. The interior structure of each chapter is described and the respective publications of parts of each chapter in conferences and in scientific journals are listed below. Chapters 2 and 3 are standalone articles published or under review in scientific journals, consisting of separate introduction, literature review and conclusions, while chapter 4 constitutes part of a broader research work, published as a whole. Introduction and literature review have been removed from the chapters and are presented for the entire thesis in chapter 1. Each chapter, being or belonging to a stand-alone publication, has its own notation and it might be the case that the same symbol is used to represent different quantities in different chapters.

Chapter 2 presents the detailed formulation of a combinatorial optimization problem for DBL allocation in networks with functional bus system and fixed-time signal control plans of known characteristics. The details of the modified SaF model used as base are presented and an algorithmic scheme based on local search and LNS is described and proposed for addressing the complex optimization problem. The performance of the proposed solution algorithms is discussed and the best performing solutions for a large-scale network are presented. Parts of this research is presented in:

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2019a). “An optimization framework for exclusive bus lane allocation in large networks with dynamic congestion”. In: *98th Annual Meeting of the Transportation Research Board (TRB 2019).* The National Academies of Sciences, Engineering, and Medicine, pp. 19–02738
- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2019b). “Modeling and optimization of dedicated bus lane network design under dynamic traffic congestion”. In: *8th Symposium of the European Association for Research in Transportation (hEART 2019).* ETH Zurich

Chapter 2 is a stand-alone article published in

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2021). “Modeling and optimization of dedicated bus lanes space allocation in large networks with dynamic congestion”. In: *Transportation Research Part C: Emerging Technologies* 127, p. 103082

Chapter 3 describes a two-layer hierarchical adaptive signal control framework combining aggregated PC with partial efficient distributed MP control. Detailed description of both control strategies is given, complemented by a proposed method for identifying critical intersections for MP control based on traffic characteristics. Additional elements regarding the simulation process are described, including an algorithm to dynamically adjust turn ratios in order to consider possible rerouting effects of drivers as a response to the actual control strategies. Simulation results of several control layouts are presented and discussed, leading to useful insights regarding parallel MP and PC application and partial MP application to critical node sets. Parts of this work are presented in:

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2022b). "Efficient Max-Pressure traffic management for large-scale congested urban networks". In: *101st Annual Meeting of the Transportation Research Board (TRB 2022).* The National Academies of Sciences, Engineering, and Medicine

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2022a). "Critical node selection method for efficient max-pressure traffic signal control in large-scale congested networks". In: *10th Symposium of the European Association for Research in Transportation (hEART 2022)*

Chapter 3 is a stand-alone article submitted for publication in *Transportation Research Part C: Emerging Technologies*, currently under review:

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2022c). "Two-layer adaptive signal control framework for large-scale dynamically-congested networks: Combining efficient Max-Pressure with Perimeter Control". In: *Transportation Research Part C: Emerging Technologies* (under review).

For the most recent version of this work, the reader can find the online version in the following URL: https://arxiv.org/abs/2210.10453.

Chapter 4 describes a detailed analysis on the evolution of total remaining travel distance, in the scope of evaluating the potential increased accuracy of integrating the recently proposed M-Model in MFD-based network control applications, replacing the conventional PL model. The analysis is done on more than 200k detailed vehicle trajectories extracted from microscopic simulation of a large scale network partitioned in three homogeneous regions. Various results are presented, mainly focusing on the relative relation between average remaining travel distance and trip length distribution, compared to the steady state approximation. The chapter describes part of a broader research work, which is published as

- I. I. Sirmatel, D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2021). "Modeling, estimation, and control in large-scale urban road networks with remaining travel distance dynamics". In: *Transportation Research Part C: Emerging Technologies* 128, p. 103157

and presented as a podium paper in the 24th International Symposium on Transportation and Traffic Theory (ISTTT24).

Finally, chapter 5 summarizes the findings and contributions of the thesis and discusses about possible future research directions.

2

Modeling and optimization of dedicated bus lanes space allocation in urban networks

This chapter is based on the article:

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2021). “Modeling and optimization of dedicated bus lanes space allocation in large networks with dynamic congestion”. In: *Transportation Research Part C: Emerging Technologies* 127, p. 103082

2.1 Introduction

Dedicated bus lanes (DBL) provide a low cost and easily implementable strategy to improve transit service by minimizing congestion-related delays. Identifying the best DBL spatial distribution in order to maximize traffic performance of urban networks while balancing the trade-off between bus priority and regular traffic disturbance is a challenging task. This chapter focuses on the problem of optimal DBL allocation and proposes a modeling framework based on a link-level dynamic traffic model, which is compatible with the dynamic characteristics of congestion propagation that can be correlated with DBL relative positions. The problem is formulated as a non-linear combinatorial optimization problem with binary variables. An algorithmic scheme based on a problem-specific heuristic and Large Neighborhood Search (LNS) metaheuristic, potentially combined with a network decomposition technique and a performance-based learning process for increased efficiency, is proposed for deriving good quality solutions for large-scale network instances.

Numerical application results for a real city center demonstrate the efficiency of the proposed framework in finding effective bus lane network configurations; when compared to the initial network state they exhibit the potential of DBL to improve travel time for car and bus users.

Following the motivation and detailed review of the literature describing the current modeling and decision-making approaches regarding DBL deployment, which is given in section 1.1.1 of chapter 1, this chapter is organized as follows: At first the problem is formally introduced and the proposed modeling and optimization framework, which integrates a dynamic link-based macroscopic traffic model with an iterative mode choice adjustment process, is described. The mathematical formulation of the optimization problem is also presented here. The proposed optimization algorithms, based on a problem-specific heuristic and LNS, are presented afterwards. Algorithm performance and generated solutions for a real case study of San Francisco network are then discussed and the best derived DBL plans are compared to the benchmark case (no DBL) in terms of simulated total travel time. Finally, the chapter ends by summarizing the main findings and providing some useful insights.

2.2 Problem description and modeling framework

2.2.1 Problem description

Consider an urban traffic network facing high levels of congestion during peak-hour. The geometrical, topological, and traffic control characteristics of the network are considered known. Two modes of transport are available in this network: buses and private cars. The operational characteristics of the bus system (routes, frequencies, and bus stop positions), as well as the average passenger occupancy for all bus lines in all roads over several time-periods of the day are also known. A deterministic time-dependent Origin-Destination demand matrix feeds the network with private car flow. The routing choices of vehicles are assumed known, in the form of time-dependent turning ratios, for all intersections approaches. Assuming that, for the purpose of improving mobility in the network by prioritizing public transport, a fraction of general purpose road space is to be given for DBL installation, we seek to decide upon the best possible spatial DBL distribution in the network, with the aim of achieving optimal system performance from a passenger perspective. The optimal subset of network roads (links) needs to be identified, where the right-most lane would be transformed to DBL, in order to achieve minimum total passenger travel time. The DBL temporal assignment is considered constant during the study period, representative of the morning or the evening peak, and no attribute of the bus system operational characteristics (e.g. headway) is modified. The mathematical modeling constructed to address this problem is described in the following sections.

2.2.2 Network traffic model

Traffic evolution in the network is simulated by utilizing a macroscopic, link-based traffic model inspired by queuing theory, and built on the basis of two existing models, the “S-Model” (see Lin et al., 2011, 2012) and the “Store-and-Forward” (SaF) model (see Aboudolas et al., 2009; Kouvelas et al., 2014), which are often used in model-based control applications and meet our requirements for efficiency and simplicity. While earlier versions of SaF may consider only vertical (point) queues, we utilize the adjusted version of SaF by Aboudolas et al., 2009, which considers horizontal (capacitated) queues for proper handling of spill-backs, and expand it by integrating the “moving” and “queuing” vehicle separation of “S-Model”, which provides increased accuracy in travel time calculation by properly integrating link lengths. Consisting of a mathematical formulation based on a time-discretized flow conservation equation, the model dynamically updates the number of vehicles per link, according to road space availability, traffic signals, saturated and unsaturated flows. The model is used for simulating only car flows in the network, while buses movement is not modeled. Bus travel time is calculated indirectly based on the observed car speeds and the existence or absence of DBL, as it will be explained below. The mathematical formulation of the simulation model is described below.

A traffic network is represented as a directed graph $G = (N, Z)$, consisting of a set of nodes (junctions) N and directed links (roads) Z. Every link $z \in Z$ is a unique, one-way connection between a pair of nodes (s_z, e_z), where $s_z, e_z \in N$ denote the upstream (start) and downstream (end) nodes of link z, respectively. Moreover, every link z is associated with a set of upstream and downstream links, U_z and D_z, respectively. Note that, every link $i \in U_z$, i.e. belonging to the set of upstream links of z, is directly connected to z and vehicles are allowed to move from i to z; the same applies for the relation between link z and any of its downstream links $j \in D_z$. The system state at every time step k is described by the number of cars per link $z \in Z$, denoted as $x_z(k)$. The dynamic equations are described below.

$$x_z(k) = m_z(k) + w_z(k) \tag{2.1}$$

$$m_z(k+1) = m_z(k) + T\left(u_{\mathrm{VQ}_z}(k) + (1 - t_{z0}(k)) \sum_{\forall i \in U_z} u_{iz}(k) - a_z(k)\right) \tag{2.2}$$

$$w_z(k+1) = w_z(k) + T\left(a_z(k) - \sum_{\forall i \in D_z} u_{zi}(k)\right) \tag{2.3}$$

$$\forall z \in Z, \;\; k = 1, 2, \ldots, K-1.$$

In the above equations, $z \in Z$ is the link index, k is the time-step index, T denotes the discrete time-step duration, and KT is the total simulation time. Equation (2.1) states that the number of vehicles inside link z at time step k, denoted as $x_z(k)$, is composed by the sum of "moving" vehicles $m_z(k)$, i.e. vehicles moving with free-flow speed in the non-occupied part of the road, and "queuing" vehicles $w_z(k)$, i.e. vehicles already queuing upstream the intersection at the end of the link. The dynamics of moving and queuing vehicles for every link z are described by equations (2.2) and (2.3), respectively, where u_{ij} denotes the transfer flow from upstream link i to downstream link j, where $i, j \in Z$ and $i \in U_j \equiv j \in D_i$; a_z refers to the flow arriving at the tail of the queue inside link z, i.e. flow leaving the "moving" and joining the "queuing" part of z; u_{VQ_z} denotes the outflow of the virtual queue (explained below) of link z, which relates to the newly generated demand at link z; finally, t_{z_0} denotes the time-dependent fraction of the incoming vehicle flow of z that end their trip in z. These vehicles are assumed to exit the network just upon entering their destination link.

Transfer flow $u_{zi}(k)$ between any pair z, i of consecutive links, with $z, i \in Z, z \in U_i \equiv i \in D_z$, is calculated by equation (2.4), by taking into account the current traffic signal state of the approach, indicated by binary variable $\eta_{zi}(k)$, storage capacity c_i and current state x_i of the receiving link i, and the saturated or unsaturated flow of the approach z–i (depending on current state of queue in sending link z). For signalized intersections, $\eta_{zi}(k)$ is equal to 1 if approach z–i takes green light at time-step k, and zero otherwise (equation (2.5)). For non-signalized intersections, $\eta_{zi}(k)$ can be constant $\forall k$, according to the priority of movements and geometry of merging links z and i. Storage capacity c_z, referring to the maximum number of vehicles that can be in z simultaneously, is given by equation (2.6), where l_z denotes the number of lanes in link z, L_z the link length, and l_{veh} the average vehicle length. Saturation flow of any link z is assumed to be equal to 1800 veh/h/lane (equation (2.8)). However, this number can be adjusted to the specific network or link, based on real flow data. Saturated flow of an approach z–i is the minimum between the saturated flows of consecutive links z and i, by considering the number of lanes that are available for this approach in each link. This is shown in equation (2.7), where $l_{zi} \leq l_z$ denotes the number of lanes in sending link z that allow moving to downstream link i, according to existing traffic regulations. We assume that in the receiving link, all lanes are available to all arriving vehicles, independently of their link of origin. Unsaturated flow, calculated by the right term of the minimum function in equation (2.4), refers to the flow allowing all vehicles currently queuing to move from z to i in one time-step. The time-dependent turn ratio that corresponds to approach z–i, denoted as $t_{zi}(k)$, refers to the fraction of the current queue in link z that will move to link i. Hence, according to equation (2.4), outflow of any approach z–i at any time-step k is zero when the traffic light is red or when unoccupied space in receiving link is less than the maximum flow that the link can receive in one time-step (queue almost occupies the whole link); otherwise, it is equal to approach' saturated or unsaturated flow, depending on the queue length in upstream link z.

$$u_{zi}(k) = \eta_{zi}(k) \times \begin{cases} 0 & \text{if } c_i - x_i(k) \leq S_i T \\ \min\left(S_{zi}, \frac{(w_z(k)+a_z(k))t_{zi}(k)}{T}\right) & \text{else} \end{cases} \tag{2.4}$$

$$\eta_{zi}(k) = \begin{cases} 1 & \text{if } z \to i \text{ has right-of-way at time-step } k \\ 0 & \text{else} \end{cases} \tag{2.5}$$

$$c_z = \frac{l_z L_z}{l_{\text{veh}}} \tag{2.6}$$

$$S_{zi} = 1800 \cdot \min(l_{zi}, l_i) \quad \text{(veh/h)} \tag{2.7}$$

$$S_z = 1800 \cdot l_z \quad \text{(veh/h)} \tag{2.8}$$

$$\text{(2.4), (2.5), and (2.7) } \forall z \in U_i,\ \forall z, i \in Z;\ \text{ (2.6) and (2.8) } \forall z \in Z.$$

In case of high congestion, links may reach their storage capacity and newly generated demand may not be able to be received due to space limitation. As the model forces incoming flows from upstream links to wait in queues in such cases, the same should happen to the newly generated inflow. Therefore, similarly to most traffic simulation models, a “virtual queue” is assumed upstream of every link that serves as entry to the network (can be origin of trips). At first, newly generated flow joins the virtual queue and remains there as long as the entrance link is full. This effect is included in the model by means of a virtual link with infinite storage capacity, assumed upstream of every origin link. As the time spent in virtual queues is important for system performance and should be considered when calculating the overall travel time of vehicles, dynamics of virtual queues, similarly to actual links, are updated based on the following equations:

$$x_{\text{VQ}z}(k+1) = x_{\text{VQ}z}(k) + T\left(d_z(k) - u_{\text{VQ}z}(k)\right) \tag{2.9}$$

$$u_{\text{VQ}z}(k) = \begin{cases} 0 & \text{if } c_z - x_z(k) \leq S_z T \\ \min\left(S_z, \frac{x_{\text{VQ}z}(k)}{T}\right) & \text{else} \end{cases} \tag{2.10}$$

$$\forall z \in Z.$$

In equation (2.9), $x_{\text{VQ}z}$ denotes the number of vehicles in the virtual queue of link $z \in Z$, d_z the newly generated demand of trips starting in z, and $u_{\text{VQ}z}$ the virtual queue's outflow towards link z. The latter is calculated similarly to any approach outflow u_{zi}, by equation (2.10).

Flow arriving at the end of a queue at time-step k, denoted as $a_z(k)$, is calculated based on the tail's current position, which depends on the number of queuing vehicles $w_z(k)$. At every time-step k, the number of discrete time-steps required for a vehicle to travel the distance between link start node and current queue end (with free-flow speed v_{ff}), is calculated by the term inside ceil$(\cdot)$ in equation (2.11). By subtracting this time from the current time-step k, we can determine the discrete time-step by which, the total link inflow since simulation started must have already reached the current queue end. This time point is denoted as $\rho_z(k)$ and is non-decreasing with k, as shown by equation (2.11). The current arriving flow $a_z(k)$ is then computed as the difference between the cumulative link inflow that is calculated up to time-step $\rho_z(k)$ minus the cumulative link inflow up to time-step $\rho_z(k-1)$, which is considered to have already arrived at the previous time-step. This process is described by equation (2.12).

$$\rho_z(k) = \max\left(\rho_z(k-1), k - \text{ceil}\left(\frac{(c_z - w_z(k))\, l_{\text{veh}}}{l_z v_{\text{ff}} T}\right)\right), \quad k \geq 1, \quad \rho_z(0) = 0 \tag{2.11}$$

$$a_z(k) = \sum_{j=1}^{\rho_z(k)} \sum_{\forall i \in U_z} u_{iz}(j) - \sum_{j=1}^{\rho_z(k-1)} \sum_{\forall i \in U_z} u_{iz}(j) \tag{2.12}$$

$$\forall z \in Z.$$

It should be noted that the paths of vehicles circulating in the network are not known to the model; however, the impact of route choice in propagation of congestion is expressed through turning ratios $t_{zw}(k)$ and exit ratios $t_{z_0}(k)$. These can either be constant for the entire simulation time or vary between time-steps in the same way as traffic patterns vary throughout the day.

At every time step $k = 1, 2, \ldots, K$, the states (number of moving, queuing, and total vehicles) of all links $z \in Z$ and virtual queues of origin links are updated according to equations (2.1)–(2.12). Necessary inputs include the detailed network representation (connectivity, length and number of lanes per link, right-of-ways), initial state of the system, i.e, $m_z(0)$, $q_z(0)$, $x_{\text{VQ}z}(0)$, $\forall z \in Z$, dynamic demand $d_z(k)$ at all origin links, traffic signal plans, time-dependent turn ratios $t_{zw}(k)$, and link exit rates $t_{z_0}(k)$. The above formulation, by considering link capacities, available road space and traffic signals for outflow calculation, can properly capture the spatio-temporal propagation of queues and spill-backs. It should be noted that within-node movements of vehicles are not explicitly considered in this model, and vehicle outflow is assumed to enter downstream links without any time lag after leaving upstream links. Also, the considered saturation flow values of links are assumed to be in accordance with those allowed by the respective nodes.

2.2.3 Decision variables

The effect of transforming existing general-purpose lanes to bus-only lanes, in terms of traffic flow modeling, is equivalent to reducing the number of available lanes for car traffic in the respective roads. Assuming that, in every link, one lane, at most, can be transformed to DBL, we define a binary variable y_z, for every link $z \in Z$, which is equal to 1 if a DBL is installed in link z, and 0 otherwise (equation (2.13)). The number of general purpose lanes l_z, used for determining storage capacity and saturation flow, is then replaced in all equations of the traffic model by the difference $l_z - y_z$.

$$y_z = \begin{cases} 1, & \text{if a DBL is assigned to link } z \\ 0, & \text{otherwise} \end{cases}, \qquad \forall z \in Z. \tag{2.13}$$

Variables y_z, $z \in Z$, constitute the decision variables of the optimization problem. A feasible solution to the problem is represented as a binary vector $\mathbf{y}$ of dimension equal to the number of links $z \in Z_f \subseteq Z$, where Z_f denotes the set of candidate links for DBL installation, i.e., all links that satisfy a number of case-specific predefined criteria, such as having at least two lanes and belonging to the path of at least one bus line. Obviously, $y_z = 0$, $\forall z \notin Z_f$.

2.2.4 Total travel time estimation

The objective of the problem is to maximize system performance through appropriate DBL allocation. This is translated into minimizing total travel time of all passengers, or Passenger Hours Traveled (PHT), either by car or bus, during the simulated time and for a specific demand scenario, which is given by the following set of equations:

$$\text{PHT}_c = \sum_{z \in Z} \sum_{k} \left(x_z(k) + x_{\text{VQ}z}(k)\right) \xi T \tag{2.14}$$

$$\text{PHT}_b = \sum_{z \in Z} \sum_{k} \sum_{l \in B} \left(\left((1 - y_z) \frac{L_z}{v_z^{\text{c}}(k)} + y_z \frac{L_z}{v_{\text{ff}}} \right) P_z^l(k) + D_z^l(k) \right) T \tag{2.15}$$

$$v_z^{\text{c}}(k) = \min \left(v_{\text{ff}}, \frac{\sum_{j=k-t_w+1}^{k} u_z(j) L_z}{\sum_{j=k-t_w+1}^{k} x_z(j)} \right), \quad k > t_w \tag{2.16}$$

$$D_z^l(k) = \delta_z^l \left(\beta_z^l P_z^l(k) s_p + s_f f_l(k) \right) \tag{2.17}$$

$$\text{PHT} = \text{PHT}_c + \text{PHT}_b \tag{2.18}$$

Total PHT for car users, denoted as PHT_c, is derived from the total time that cars spent inside the network and in virtual queues. It is calculated by summing the queues in all links over time, as shown by equation(2.14), where ξ denotes the average car occupancy (in passengers/hour). Regarding bus travel time, since the traffic model only monitors car flows, it has to be estimated indirectly. This is done by assuming that buses travel in free-flow conditions inside links with DBL ($y_z = 1$), while in links without DBL ($y_z = 0$) we assume that buses travel either with speed of cars, which can be estimated using link queues and outflows over a specific time window, or with free-flow speed v_{ff} if this is smaller, as shown in equation 2.16. Note that all bus stops are assumed to provide bus bays outside of traffic lanes with sufficient space, in order to minimize traffic disturbance during boarding and alighting of passengers. Total PHT for bus passengers, denoted as PHT_b, is calculated by equation(2.15), where notation is as follows: B is the set of bus lines running in the network; l is the bus line index; $v_z^{\text{c}}(k)$ denotes the estimated car speed in link z at time step k, which is calculated based on link queue and outflow, according to equation (2.16); $P_z^l(k)$ is the average flow of bus passengers (pax/h) traveling on-board line l buses through link z at time step k; and $D_z^l(k)$ is the dwell time of buses for line l at the bus stop in link z at time step k. Therefore, according to y_z value, travel time of buses inside link z is estimated by assuming either free-flow speed v_{ff} in case of DBL presence ($y_z = 1$) or an estimation of car speed v_z^{c} in case of no DBL presence ($y_z = 0$). Consequently, if a link gets congested and its car outflow drops, estimated bus speed in this link will be affected accordingly. The values $P_z^l(k)$ can be estimated as $P_z^l(k) = f_l(k)Pb_z^l(k)$, where $f_l(k)$ is the frequency of line l at time step k, and $Pb_z^l(k)$ the average number of passengers per bus of line l inside link z at time-step k. These values can be derived from measurements performed by the bus operator.

Car speed in link z, v_z^{c}, is estimated by equation (2.16), as the minimum between the assumed free-flow speed v_{ff} and the fraction of total distance traveled by cars inside link z during a time window of t_w time-steps (prior to time-step k), divided by the total time spent by all cars inside link z in the same time window. It should be noted that in the case where link z remains empty during the entire time window, speed is equal to free-flow speed. Service time at bus stops is calculated according to equation (2.17), where δ_z^l is a binary indicator about bus stop existence in link z for buses of line l; β_z^l is the average fraction of bus occupancy of line l inside link z, which corresponds to the sum of boarding and alighting passengers for the bus stop in link z; s_p is the time delay per boarding/alighting passenger; s_f is a fixed delay per bus stop. This empirical relation, formulated based on a survey reported by Meng and Qu, 2013, considers as $s_p = 1.5$ sec per boarding/alighting passenger plus a fixed delay of $s_f = 4$ sec per stop. The average bus passenger occupancy of bus lines over time and space $Pb_z^l(k)$, as well as bus frequencies $f_l(k)$, and approximate fractions β_z^l are considered as known inputs in the context of this work. Total travel time of all passengers, denoted as PHT is the sum of PHT_c and PHT_b (equation(2.18)).

2.2.5 Mode and route choices

DBL introduction is expected to change experienced travel times and congestion patterns for both cars and buses in several parts of the network. Consequently, user choices in terms of mode and routes may be affected by DBL location assignment and car/bus shares and turn ratios may change accordingly. Therefore, the evaluation process of every candidate solution should account for these changes. This is why, in most related works, the problem is formulated as a bi-level program, similar to a Stackelberg competition, where traffic managers act as leaders, deciding on the best DBL plan to optimize system performance, and users act as followers, deciding upon their mode and route preferences, given the decisions of the traffic managers, with the aim of optimizing their personal overall travel utility (often including in-vehicle travel time, out-of-pocket cost, access time and other attributes).

While a bi-level programming approach might be possible when static traffic assignment is used, in our case, since our dynamic macroscopic model uses turning ratios for traffic distribution in intersections, exact trip paths are not known. Updating turn ratios for every new DBL configuration through a suitable traffic assignment process might improve the accuracy of system performance estimation, but, at the same time would add an unrealistic computational burden in the optimization process. This is why a typical bi-level programming structure, where the optimal solutions of lower level optimization problems are included in the constraints of the optimization problem of the upper level, would drastically increase problem's complexity, making the method inapplicable to large networks, which is one of our primary objectives. Seeking to avoid this effect, we model mode choice in a simpler, aggregated way, by using a Logit model, while we assume, for simplicity, that route choices of cars remain unaffected by DBL installation. While we understand that this assumption might affect the accuracy of the evaluation process for candidate solutions during optimization, we believe that its impact would not significantly alter the optimal solution, especially in large networks, where highly complex models are not applicable. Nevertheless, this assumption could be removed in the future by integrating, in the same modeling framework, a turn ratio update process, resulting from suitable traffic assignment considering current DBL locations, which would be efficient and simple enough, in order to not drastically increase the overall computational cost. However, this addition exceeds the scope of the present work and is currently omitted.

In order to address the expected variations in mode shares related to the specific DBL plan under consideration, we utilize a Logit model embedded in an iterative scheme of traffic simulations for evaluation of every candidate DBL plan. The process is depicted in Figure 2.1. The selected DBL plan is provided as input to the urban traffic model, together with all other exogenous inputs, including current car demand and bus ridership for all lines. Based on queue evolution inside links and bus operational characteristics, we estimate the total PHT for car and bus passengers, from which, by dividing by the number of passengers and average trip length per mode, we obtain the average time per unit distance traveled for every mode (inverse of average speed). Then, the percentage of users opting for car or bus is calculated based on the utility of each mode, according to Logit model. The new mode shares are translated to new car demand and bus occupancy, which are compared to the previous ones. If their difference is above a specified threshold, a new traffic simulation is performed by considering the derived car and bus demand information. For simplicity, the utility function only considers travel time per kilometer of trip plus a constant term; however, more attributes can be included, such as out-of-pocket costs, accessibility, parking availability, and others. Utility per mode and percentage of users for each mode over all network users are calculated according to the following equations:

$$U_m = ASC_m + \beta_m \cdot UTT_m \tag{2.19}$$

$$p_m = \frac{e^{U_m}}{e^{U_c} + e^{U_b}} \tag{2.20}$$

In (2.19) and (2.20), we use m as the mode symbol, i.e. car (c) or bus (b), UTT_m denotes the inverse of average speed for mode m, i.e. travel time per unit distance, ASC_m and β_m are model parameters, and p_m denotes the percentage of total number of users opting for mode m, according to Logit model. Average car trip length and number of bus passengers can be obtained from a preliminary microsimulation analysis and are assumed constant. The adjustment of car demand and bus ridership

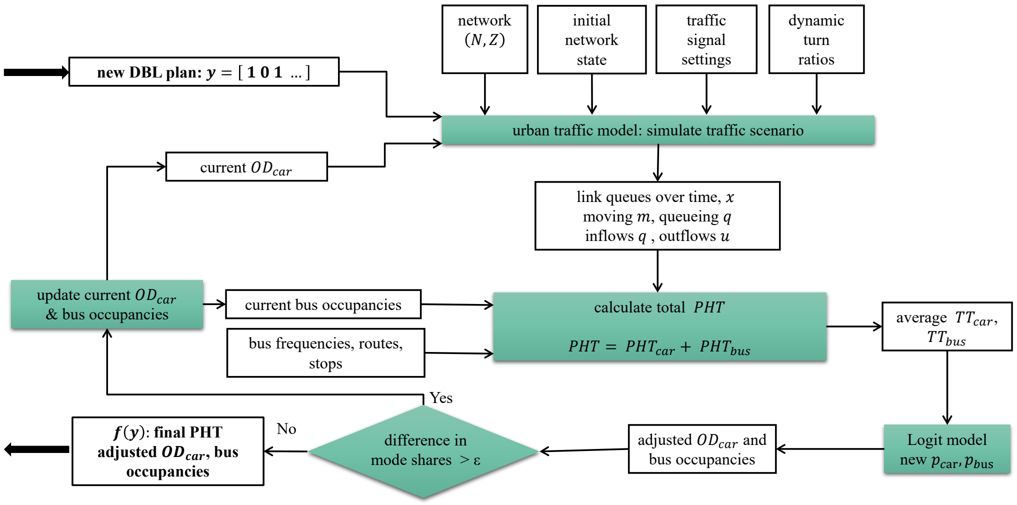

network (N, Z)
initial network state
traffic signal settings
dynamic turn ratios
new DBL plan: y = [1 0 1 ...]
urban traffic model: simulate traffic scenario
current OD_{car}
link queues over time, x moving m, queueing q inflows q , outflows u
update current OD_{car} & bus occupancies
current bus occupancies
bus frequencies, routes, stops
calculate total PHT
$PHT = PHT_{car} + PHT_{bus}$
average TT_{car}, TT_{bus}
Yes
No
$f(y)$: final PHT adjusted OD_{car}, bus occupancies
difference in mode shares > ε
adjusted OD_{car} and bus occupancies
Logit model new p_{car}, p_{bus}

Figure 2.1: Evaluation process of any newly formed candidate solution $\mathbf{y}$ (DBL plan) with mode choice adjustment, for calculation of the respective cost $f(\mathbf{y})$ (PHT).

per line in order to comply with a modified mode share, is performed in a uniform way across all OD pairs and bus lines, by maintaining the initial distribution of trips over OD pairs and bus lines. This is based on the assumption that population is homogeneous and all trips can be performed by both modes.

2.2.6 Optimization problem formulation

The objective of the problem is to identify the DBL layout that is expected to lead to minimum total passenger travel time. For every candidate road, a decision about whether a bus-only lane will be installed (by conversion of a general-purpose lane) or not, needs to be made. It should be noted that the total road space to be devoted to bus-only lanes may be predefined, especially in cases where road space is quite limited, or it may be indifferent/unlimited. To address different cases regarding total DBL length, different variants of the optimization problem can be constructed on the same framework: in one variant, the objective function may consist only of the PHT term (equation (2.18)) and a constraint may be considered, to ensure that feasible solutions have a total length close enough to (or simply less than) a predefined target length; in a second version, a penalty cost term, proportional to the total DBL solution length, can be included in the objective function to penalize excessive use of road space, and no additional constraint regarding total length is considered. This modification of the objective function aims at driving the search towards solutions of more efficient road space use than others, in the sense that they improve traffic performance while considering the amount of road space devoted to DBL. The value of the cost per unit length, denoted as γ, can be related to actual installation costs (e.g. horizontal signalization, mechanical infrastructure etc.), or it can simply reflect an approximate generalized cost associated with car traffic disturbance. Obviously, higher values of γ will lead to solutions with less DBL road space usage. The detailed mathematical formulation of the optimization problem is presented below:

$$\min_{y_z, \forall z \in Z_f} \quad \sum_{z \in Z} \sum_{k} \sum_{l \in B} \left(\left((1 - y_z) \frac{L_z}{v_z^{\mathrm{c}}(k)} + y_z \frac{L_z}{v_{\mathrm{ff}}} \right) P_z^l(k) + \delta_z^l \left(\beta_z^l P_z^l(k) s_p + s_f f_l(k) \right) \right) T + \sum_{z \in Z} \sum_{k} \left(x_z(k) + x_{\mathrm{VQ}z}(k) \right) \xi T + \gamma \sum_{z \in Z} y_z L_z \tag{2.21}$$

subject to:

Equations (2.1), (2.2), (2.3), (2.4), (2.5), (2.9), (2.10), (2.12)

$$S_{zi} = 1800 \cdot \min(l_{zi} - y_z\, rl_{zi}, l_i - y_i), \ \ \text{(veh/h)} \tag{2.22}$$

$$c_z = \frac{(l_z - y_z)\, L_z}{l_{\text{veh}}} \tag{2.23}$$

$$S_z = 1800 \cdot (l_z - y_z) \ \ \text{(veh/h)} \tag{2.24}$$

$$\rho_z(k) = \max\left(\rho_z(k-1), k - \text{ceil}\left(\frac{(c_z - w_z(k))\, l_{\text{veh}}}{(l_z - y_z) v_{\text{ff}} T}\right)\right), \ \ \rho_z(0) = 0 \tag{2.25}$$

$$\forall z \in Z, \ k = 1, \ldots, K, \ \forall\{(z,i) \in Z | z \in U_i\}$$

$$v_z^{\text{c}}(k) = \min\left(v_{\text{ff}}, \frac{\sum_{j=k-t_w+1}^{k} u_z(j) L_z}{\sum_{j=k-t_w+1}^{k} x_z(j)}\right), \ \ \forall z \in Z, k = t_w, \ldots, K \tag{2.26}$$

$$w_z(k), \ m_z(k), \ \ a_z(k) \geq 0, \ \ \ \forall z \in Z, k = 1, \ldots, K \tag{2.27}$$

$$y_z \in \{0, 1\}, \ \forall z \in Z_f \subseteq Z \tag{2.28}$$

$$y_z = 0, \ \forall z \notin Z_f \tag{2.29}$$

given:

$$m_z(0), w_z(0), x_{VQz}(0), v_z^{\text{c}}(0), d_z(k), t_{zi}(k), t_{z_0}(k), \eta_{zj}(k), \tag{2.30}$$

$$P_z^l(k), \beta_z^l, \delta_z^l, f_l(k) \tag{2.31}$$

$$\forall z \in Z, \forall\{(z,i) \in Z | z \in U_i\}, \forall l \in B, k = 1, .., K$$

Vector **y** contains the decision variables of the problem, i.e. all binary variables $y_z, \forall z \in Z_f \subseteq Z$, indicating the presence (or not) of bus-only lanes inside every link z belonging to the set of candidate links Z_f. The first term of the objective function (equation (2.21)) refers to total PHT of bus passengers (see equations (2.15)–(2.17)), the second term refers to PHT of car passengers (see equation (2.14)) and the third term represents the penalty cost associated with the total DBL operation and maintenance, which is proportional to total DBL length. In the cost term, γ denotes the cost per unit length of DBL added to the solution.The objective function has units of time, therefore cost γ is also expressed in units of time per DBL unit length.

Once a candidate solution is defined, the system performance is evaluated based on the evolution of queues in the network, as estimated by equations (2.1)–(2.12) of the urban traffic model. The dynamic equations of the traffic model become constraints for the optimization problem, with proper integration of the decision variables y_z, $\forall z \in Z_f \subseteq Z$, as shown in equations (2.22)-(2.25). Equation (2.22) refers to saturation flow of approach $z - i$, with $z \in U_i$, where rl_{zi} is a binary indicator of whether the right-most lane in upstream link z is included in the set of

l_{zi} lanes that allow movement to downstream link i. Equations (2.23)-(2.25) are simply equations (2.6), (2.8) and (2.11), where l_z is replaced by $l_z - y_z$. Constraint in Equation (2.27) ensures that all queues and arriving flows are non-negative, which is something that is also guaranteed by the traffic model itself. Equation (2.28) defines decision variables to be binary and equation (2.29) guarantees that all non-candidate links of set Z_f do not get DBL. All necessary inputs such as initial system state, dynamic car demand in all origins, time-dependent turn and exit ratios, and traffic signal plans for all signalized intersections are listed in (2.30) and considered known. Total PHT estimation, according to equation (2.21) requires knowledge of time-varying bus occupancy and operational characteristics of all bus lines (see (2.31)). In the case where a target or upper limit of total DBL length exists, instead of specifying a non-zero cost γ, constraint (2.32) can be considered

$$\left| \sum_{z \in Z} y_z L_z - LT \right| \leq \theta, \tag{2.32}$$

where LT denotes the total target length and θ the acceptable deviation from that target. Finally, for any candidate solution, an iterative process of mode share adjustment and traffic simulation is performed, until mode shares estimated by Logit model based on simulation-generated travel times are close enough to those that were taken as inputs for the traffic simulation (see figure 2.1).

2.3 Solution methods

The optimal DBL allocation problem that is presented in Section 2.2, is a non-linear combinatorial optimization problem with binary decision variables and a finite feasible set that grows exponentially with the network size. In fact, the solution space size, if no constraint is imposed for total solution length, is $2^{|Z_f|}$, where $|Z_f|$ is the cardinality of the DBL candidate links set Z_f. Hence, for real instances, complete enumeration is not possible with current computing technology. Due to increased complexity and size, heuristic and metaheuristic algorithms seem to be the most promising tools for finding good quality solutions in reasonable time.

In this work, we construct and test a set of algorithms based on local search and Large Neighborhood Search (LNS), while integrating performance-improving techniques to guide the search process towards promising areas of the solution space. Firstly, we propose an algorithm to construct a good-quality DBL plan by recursively adding DBL at one road per step, after trying all currently available roads one-by-one and choosing the one leading to highest performance improvement. During this process, the algorithm can also be used to calculate performance indicators, or “scores”, for all candidate links, according to the achieved system performance every time DBL placement is tested on each link. These scores can then be used in an LNS framework to drive the search towards potentially high performing solutions. Finally, a network decomposition technique is proposed as a complementary element of the repair process of LNS algorithm, with the aim of increasing search efficiency in cases of very large networks. The detailed description of the proposed algorithmic scheme is given in this section.

2.3.1 Algorithm 1: Link-by-link plan construction and link score calculation

A heuristic local search algorithm that constructs a good quality DBL plan, built on the principle of greedy, link-by-link addition, is formulated as shown in pseudo-code format, in Algorithm 1. The algorithm starts by setting as current solution the one corresponding to no DBL in the network, i.e. $y_z = 0, \forall z \in Z$ (Line 1); the objective function value for the current solution is calculated and stored (Line 2). Following a recursive process, one DBL link per step is added to the solution, after evaluating, one by one through distinct simulations, all candidate links currently available (Lines 6–10). The objective function values (i.e. cost) corresponding to all potential additions are stored in vector n. The best DBL addition of the current step is to link z^*, which results in the lowest overall cost compared to all other possible candidate link additions. If this cost is lower than the cost of the current solution (Line 13), the DBL is added to z^*, the current solution and the corresponding cost are updated (Lines 14–15), and the process is repeated until no further improvement is possible, i.e. none of the available candidate links can result in lower cost through DBL addition. Lines 17 and 19 relate to score calculation, are optional for Algorithm 1 and will be explained below.

Apart from constructing an improved solution from scratch, Algorithm 1 also serves the purpose of statistically assessing the potential of every link in improving system performance, if added to the DBL plan. In every iteration, the system performance resulting from all trials of DBL link addition is stored. At the end of every iteration, all links currently available for DBL addition are ranked in an increasing order of overall cost (Line 17). When no more improvement can be achieved by DBL addition, the algorithm stops and returns the current DBL plan $\mathbf{y}$ and a vector of rankings $\mathbf{r}$ per candidate link $z \in Z_f$, for every iteration of Algorithm 1. These scores represent the reported performance of the link, when added to the DBL plan, according to the solution construction process of Algorithm 1. These values can later be integrated in the search process of any general metaheuristic, in order to drive the search towards solutions with higher improvement potential. Score ω_z of every candidate link z ranges in $(0, 1)$ and is calculated based on the following formula:

$$\omega_z = \max\left(\omega_{\min},\ \frac{1}{|I^*|}\sum_{i\in I^*}\frac{N_{\mathrm{f}}(i) - r_z(i)}{N_{\mathrm{f}}(i)}\right) \tag{2.33}$$

Algorithm 1: Evaluation-based link-by-link plan construction and link score calculation

```
Data: initial solution $\mathbf{y}_0 = \mathbf{0}$   $(y_z = 0 \quad \forall z \in Z_f)$
Result: DBL plan $\mathbf{y}$, scores of candidate links $\omega$
$\mathbf{y} = \mathbf{y}_0$
$c = f(\mathbf{y}_0)$                                   $f$: objective function
$i^* = 0^+$
while $i^* > 0$ do
    initialize $\mathbf{n}$ : $n(z) = c, \forall z \in Z$
    for $z \in \{Z_f | y_z = 0\}$ do
        add trial DBL in $z$: $y_z = 1$
        calculate and store new obj. function value $n(z) = f(\mathbf{y})$
        remove trial DBL from $z$: $y_z = 0$
    end
    find best PHT improvement of step: $i^* = c - \min(n)$
    store the link of best improvement: $z^* = \mathrm{argmin}(n)$
    if $i^* > 0$ then
        place DBL in link $z^*$: $y_{z^*} = 1$
        c = f($\mathbf{y}$)
    end
    rank links according to increasing $n$ and store ranking $\mathbf{r}$
end
assign scores $\omega$ to links based on average overall ranking $\mathbf{r}$
return $\mathbf{y}$, $\omega$
```

In equation (2.33), ω_z denotes the score of link z, I^* the set of algorithm iterations during which link z is added in the trial solution, $N_\mathrm{f}(i)$ the number of available candidate links at iteration i, i.e. size of set $\{z | z \in Z_f \text{ and } y_z = 0\}$, $r_z(i)$ the ranking of link z among all available links $N_\mathrm{f}(i)$ at iteration i in increasing solution cost order ($r_z(i) = 1$ for link z, which leads to the lowest cost $f(y)$ if added to the DBL plan at iteration i), and $\omega_\mathrm{min} > 0$ the minimum score value that ensures a non-zero probability of selection. In summary, link scores ω_z, defined for every candidate link $z \in Z_f$, express the potential of the link to improve system performance by receiving a DBL, based on the assumption that links are included in the DBL plan in sequence of their ability to improve system performance (according to Algorithm 1). Link scores defined in this way are used, among other types of scores, to drive the search process of LNS algorithm described in the following section.

2.3.2 Large Neighborhood Search

The typical LNS algorithm, first proposed by Shaw, 1998 with application to Vehicle Routing Problems (VRP), explores a large solution space by recursively "destroying" and "repairing" an incumbent solution in an effort to reach solutions of improved

Algorithm 2: LNS for DBL allocation problem

```
   Data: feasible initial solution $\mathbf{y}_0$, link scores $\omega$
   Result: DBL plan $\mathbf{y}$
 1 $s = \mathbf{y}_0$
 2 $c = f(\mathbf{y}_0)$                                   $f$: objective function
 3 for $iter$ = 1 to $iterMax$ do
 4     $s' = \mathrm{destroy}(s, dd_{\max}, \omega)$
 5     $s'' = \mathrm{repair}(s', rd_{\max}, \omega)$
 6     $c' = f(s'')$
 7     if $c' < c$ then
 8         $s = s''$
 9         $c = c'$
10     end
11     * update link scores every $\kappa_{\mathrm{int}}$ iterations          * optional
12 end
13 return $s$
```

quality, i.e. lower objective function values (cost). The processes of destroying and repairing a current solution, initially defined for the VRP modeling structure, aim at resetting some of the decision variables of the solution and redefining them based on intuitive methods that can potentially lead to improved cost. The degree of destruction, referring to the number of decision variables that are reset in every iteration, is often large, thus allowing the algorithm to explore larger areas of the solution space and decreasing the chances of the algorithm getting trapped in local optima. If the newly constructed solution has lower cost with respect to the current one, it is accepted and replaces the current solution. Otherwise, it can either be immediately disregarded or, in order to diversify the search, it can be accepted with a probability depending on several criteria (as in Simulated Annealing, or SA). The stopping criterion can be based on a predefined number of iterations or execution time or it can depend on the current search performance. A detailed description of LNS metaheuristic as well as a review of its various applications in different fields and types of optimization problems can be found in Pisinger and Ropke, 2019.

Algorithm 2: Main LNS algorithm

The base structure of the proposed LNS framework is presented Algorithm 2. An initial feasible solution, given to the algorithm as input, is set as current solution s (Line 1). This solution can be either randomly constructed or generated according to intuitive traffic engineering principles. In case there is a target length constraint, i.e. equation (2.32) applies, the initial solution is constructed accordingly. Current solution s is evaluated and its cost is stored as current cost c (Line 2). The iterative destroy and repair process follows (Lines 4–6). The maximum possible destruction and repair degrees, $dd_{\max}$ and $rd_{\max}$ respectively, are set and given as inputs to the destruction and repair processes, together with a set of link scores ω, which take values in the interval $(0,1)$ and attempt to measure the efficiency, in terms of system performance, of a possible DBL installation in the link. These values can be generated by Algorithm 1, or defined differently, as it will be discussed later on. The destruction process modifies the current solution s by removing a set of DBL links. The result is the destructed solution s'. Afterwards, the repair process receives s' and reconstructs it by adding a set of DBL links, while respecting all feasibility constraints. The newly formed solution, s'', is evaluated (Line 6) and its cost c' is compared to the current solution cost c (Line 7). Solution s'' replaces s as the current solution only if it leads to a lower cost (Lines 8–9). While a different acceptance criterion can be used (e.g. as in SA), we choose a greedy approach, as we observed from a preliminary analysis, that in this case, the necessary flexibility in the search is sufficiently guaranteed and controlled by the destruction and repair degree of LNS mechanism and an SA acceptance criterion could decelerate the search (as more LNS steps are necessary), without achieving significant improvement in the final solution. Line 11 describes an optional process of link score updating, according to observed link performance in regular step intervals of LNS algorithm, that will be further discussed in a following section. LNS iterative process is repeated for a predefined number of times $iterMax$. After the last iteration, current solution s is the best found solution, which the algorithm returns as output to the user.

Destruction process

Typically, in LNS and A-LNS algorithms (see Ropke and Pisinger, 2006), the incumbent solution is destroyed and repaired through application of one or several specific methods that are expected to improve overall solution quality, according to the characteristics of the problem at hand. In the present approach, given the difficulty of identifying a deterministic way to describe correlations between DBL topology and the several elements influencing solution cost, such as queue spill-backs, resulting mode shares, or network traffic flows, we adopt a different approach. We utilize a system of link scores (weights) that describe the potential of every link to improve system performance by DBL addition. These scores are used as link selection probabilities, for DBL addition/removal during the destroy/repair processes for every iteration. The score values can be defined by simulation experiments, by prior execution of Algorithm 1, or based on a set of link characteristics that are intuitively connected to performance improvement in DBL presence (e.g. bus frequency per link), as we further discuss in a following section.

The destruction process of every LNS iteration (Line 4 of Algorithm 2), which consists of removing a number of DBLs from their current positions, is described by Algorithm 3. Given a maximum allowed destruction degree, $dd_{\max}$, set by the user, the actual destruction degree of every iteration is randomly drawn from an interval

$(0, dd_{\max}(iter))$ with a uniform probability, where $dd_{\max}(iter) \in (0,1)$, takes the user-defined maximum value $dd_{\max}$ at the first iteration and then decreases linearly with the number of iterations, i.e. $dd_{\max}(iter) = dd_{\max}\left(1 - \frac{iter-1}{iterMax}\right)$. This is done in order to allow larger possible search steps at the first iterations of LNS algorithm, which increase the chances of identifying promising areas of the solution space by escaping local optima, and smaller steps in the last iterations, in order to fine-tune the solution. The actual number of links to be removed at iteration $iter$, denoted as d_{iter}, is specified (Line 1) by the following formula:

$$d_{iter} = \texttt{ceil}\left(r \cdot dd_{\max} \cdot \left(1 - \frac{iter - 1}{iterMax}\right) \cdot \sum_{z \in Z_f} y_z\right) \tag{2.34}$$

In (2.34), r is a random number drawn from a uniform distribution $\mathcal{U}(0,1)$. Operator `ceil` ensures that an integer number of links greater than or equal to one will be removed from the solution in every iteration, provided that there is at least one link in the current DBL plan, i.e., $\sum_{z \in Z_f} y_z \geq 1$.

Links are then removed one-by-one, in an iterative process. A set of link scores is used to define selection probabilities for removal from the current DBL plan. Since link scores express the links' estimated potential in improving system performance by DBL addition, we simply use the difference of every link score ω_z from 1 to define link selection probability for removal. Therefore, in every step, one link z is selected for removal from the set of links currently having DBL, with probability

$$p_z^r = \frac{1 - \omega_z}{\sum_{z \in \{Z_f | y_z = 1\}} (1 - \omega_z)} \tag{2.35}$$

where, p_z^r denotes the probability of link z to be selected for removal and ω_z is the link score calculated for link z (e.g. provided by Algorithm 1 according to

Algorithm 3: Destroy current solution s (Module for Algorithm 2, Line 4)

```
   Data: current DBL plan $s$, $dd_{\max}, \omega_z$
   Result: destroyed solution $s'$
1  calculate no of links to remove $d_{iter}$
2  $s' = s$
3  $counter = 0$
4  while $counter \leq d_{iter}$ do
5      specify current set of links with DBL
6      calculate removal probabilities: $p_z^r \; \forall z$
7      draw one link $l$ from the distribution of $p_z^r$
8      update solution $s'$ by removing link $l$: $y_l = 0$
9      $counter = counter + 1$
10 end
11 return $s'$
```

formula (2.33)). Obviously, the probability of any link z to be selected for removal is higher if its score ω_z is lower compared to the scores of the rest of the links currently in the solution. After completing the draw from the distribution of p_z^r, with $z \in \{Z_f | y_z = 1\}$, the DBL is removed from the chosen link and the solution is updated. This process, described in lines 5–9 of Algorithm 3, is repeated until all d_{iter} links have been removed. Then the destructed solution s' is returned as output and the process of Algorithm 2 continues to the repair process of line 6.

Repair process

The repair process varies depending on the validity of constraint (2.32) about total DBL length. At first, we examine the case where no such constraint applies. The addition process is described in Algorithm 4, which receives as input the destroyed solution s' (output of Algorithm 3), the user-defined, maximum possible repair degree $rd_{\max}$, and link scores ω_z. The number r_{iter} of links to be added in the destroyed DBL plan s' during the repair process is defined by formula (2.36) below, similarly to (2.34).

$$r_{iter} = \texttt{ceil}\left(r \cdot rd_{\max} \cdot \left(1 - \frac{iter - 1}{iterMax}\right) \cdot \sum_{z \in Z_f} (1 - y_z)\right) \tag{2.36}$$

Links are added to the destroyed solution one-by-one, in an iterative process until the required number of additions, r_{iter}, is reached. In every iteration, the set of links currently available (i.e. $z \in Z_f$ with $y_z = 0$) is specified. One link z is drawn from this set with a probability

Algorithm 4: Repair destroyed solution s' (Module for Algorithm 2, Line 5)

```
Data: destroyed DBL plan $s'$, $rd_{\max}$, $\omega_z$
Result: repaired solution $s''$
*calculate no of links to add $r_{iter}$          * when no length constraint applies
$s'' = s'$
$counter = 0$
while $counter \leq r_{iter}$ do
    specify current set of links with no DBL
    calculate addition probabilities: $p_z^a$
    draw one link $l$ from the distribution of $p_z^a$
    update solution $s''$ by adding link $l$: $y_l = 1$
    $counter = counter + 1$
end
return $s''$
```

$$p_z^a = \frac{\omega_z}{\sum_{z \in \{Z_f | y_z = 0\}} \omega_z} \tag{2.37}$$

In (2.37), p_z^a denotes the probability of link z to be selected for addition in the current step. After the draw from the distribution of p_z^a is complete, the selected link is added to the destroyed solution s'. The same process (Lines 5–9) is repeated until all r_{iter} DBL links are added to the solution. The repaired solution, s'' which is the output that Algorithm 4 returns in the main LNS Algorithm 2 (Line 5), consists the newly formed solution of the current LNS iteration. It will be evaluated through the process depicted in figure 2.1 and it will replace the incumbent solution s only if it results in better system performance.

The probabilistic selection of links to be added and removed in every LNS iteration ensures the necessary diversification in the search process that helps the algorithm avoid getting trapped in local optima. In case where a length constraint such as (2.32) applies, the repair process is performed slightly differently. No repair degree is considered, since the number of links to be added back in the solution will depend on the current DBL solution length. Link addition is done again one-by-one in an iterative way but total solution length is calculated after every single link addition. The process terminates when the solution length reaches the desired value, so that length constraint (2.32) is satisfied. The selection process for link additions remains the same as shown in lines 5–8 of Algorithm 4.

Alternative link scores and score updates

Acquiring link score values ω_z for all candidate links, in order to be used in LNS processes can be done by running Algorithm 1, by integrating the score updating process of lines 17 and 19. Score values acquired likewise are expected to be highly effective in driving the search process of LNS algorithm towards promising solution space areas, because they are specified by trial-and-evaluation, which considers all possible effects of DBL setting on system characteristics, such as mode shift, queue spill-backs, bus passengers delay savings in the respective links, etc. Moreover, the solution building process of Algorithm 1 by definition leads to improved solutions in every step. However, this is a computationally expensive process, due to the large number of trials and simulation runs that Algorithm 1 needs to perform in order to identify the best link addition in every step and estimate the respective link scores.

Therefore, there might be cases where such a computationally expensive process cannot be afforded, for example when several different values of the initial inputs (e.g., dynamic travel demand profile, car trip routing patterns, bus operational characteristics, etc.) must be considered. For such cases, we propose the following alternative approaches:

- Utilize link scores according to link characteristics that are expected to lead to good quality solutions, e.g., bus frequencies, link bus passenger flow, road space availability, etc.
- Utilize uniform or other types of link scores (e.g., based on bus frequencies) for all candidate links and include a score update process in regular intervals in LNS algorithm (Line 11). Link scores that are used for removal and addition are updated, after a specific number of LNS iterations, according to the observed performance of the respective solutions. In this way, link scores are gradually adjusted to the specific case study during LNS execution, in a similar way as A-LNS metaheuristic evaluates and updates scores of different destroy and repair methods based on their performance (for details see Ropke and Pisinger, 2006).

It should also be noted that these options can be combined, e.g., if the update step of line 11 is included in LNS algorithm, the initial scores can either be uniform or based on current bus frequencies or even come from Algorithm 1 executed with different input data. However, LNS algorithm may require significantly more iterations in order to reach good quality solutions, as the learning process required for the proper tuning of scores is built through iterations. Score updates can be performed in regular step intervals according to the following equations:

$$\omega_z(\kappa+1) = \lambda\omega_z(\kappa+1-\kappa_{\text{int}}) + (1-\lambda)\left(\delta_z^a(\kappa) + (1-\delta_z^r(\kappa))\right) \tag{2.38}$$

$$\delta_z^r(\kappa) = \frac{\sum_{i=\kappa-\kappa_{\text{int}}+1}^{\kappa} \beta_z^r(i)\frac{c_i-c_i'}{c_i}}{\sum_{i=\kappa-\kappa_{\text{int}}+1}^{\kappa} \beta_z^r(i)} \tag{2.39}$$

$$\delta_z^a(\kappa) = \frac{\sum_{i=\kappa-\kappa_{\text{int}}+1}^{\kappa} \beta_z^a(i)\frac{c_i-c_i'}{c_i}}{\sum_{i=\kappa-\kappa_{\text{int}}+1}^{\kappa} \beta_z^a(i)}. \tag{2.40}$$

Equation (2.38) is used to update scores ω_z for all candidate links $z \in Z_f$, every κ_{int} iterations of LNS algorithm, according to the performance of the solutions tested in these κ_{int} steps and formed by adding or removing link z. In this equation, κ denotes the LNS iteration at the end of which scores are updated, and $\kappa-\kappa_{\text{int}}+1$ the iteration before which scores were last updated; λ is the decay factor, which is user-defined and dictates how much the new score values will be influenced by their previous value; $\delta_z^r(k)$ and $\delta_z^a(k)$ represent the score update terms, which are based on the average performance of all trial solutions that were formed by removing and adding link z, respectively, during iterations $\kappa-\kappa_{\text{int}}+1$ to κ. Their calculation is derived according to equations (2.39) and (2.40), where i is the index for LNS iterations; $\beta_z^r(i)$ is a binary indicator that is equal to 1 if link z is removed from the current solution at iteration i, and 0 otherwise; $\beta_z^a(i)$ is a similar binary indicator about link addition; c_i is the cost of current solution s at iteration i; c'_i is the cost of newly formed solution s'', after completion of destroy and repair processes of iteration i. In other words, terms $\delta^a_z(k)$ and $\delta^r_z(k)$ express the average relative cost change every time link z was added to or removed from the newly formed solution, respectively, during the last κ_{int} LNS iterations. Note that since link scores, by definition, express the potential of links to improve solution by DBL addition, δ^r_z is subtracted from 1 in equation (2.38), in order to translate the potential of removal to potential of addition. The relative change of cost, calculated by fraction $(c_i - c'_i)/c_i$, can also be normalized by the highest absolute value of relative cost change of this set of κ_{int} iterations.

Alternative repair process using sub-networks

A complementary element that can be integrated in LNS repair process is proposed and evaluated as part of this algorithmic scheme, intended especially for application in very large network instances. With the aim of increasing LNS algorithm's chances of finding good quality solutions, the repair process of destructed solutions is enhanced by including a trial and evaluation step including smaller networks, in the proximity of candidate links, which we call sub-networks. More specifically, the main LNS algorithm remains as is (Algorithm 2), and so does the destruction process (Algorithm 3), while the repair process is modified as shown in Algorithm 5 below. As described in lines 8–14, every DBL addition is made at the best performing link, out of a sample L of available links selected according to their scores, after evaluating the traffic performance of isolated sub-networks around them. By isolating a sub-network around a candidate link, a set of before-after DBL scenario evaluations that last for a fraction of total simulation time, can give a good estimation about the potential performance of DBL in this link, with much lower computational cost compared to a full-network and full-time simulation process. The difference of this enhanced repair module with respect to the process described in Algorithm 4 is that link scores are iteratively used to create sample sets of candidate links for DBL addition. Before every DBL addition, for every link of the sample set, the algorithm evaluates the performance of the respective sub-network with and without DBL (lines 10–11). Then, actual DBL addition is done to the best performing link of the set (lines 14–15). This process is repeated for every DBL addition with a new sample set of available links created each time. When all additions of the step are made, the repair process is completed and the newly formed solution s'' is returned to the main Algorithm 2 and evaluated with a full-network full-time simulation.

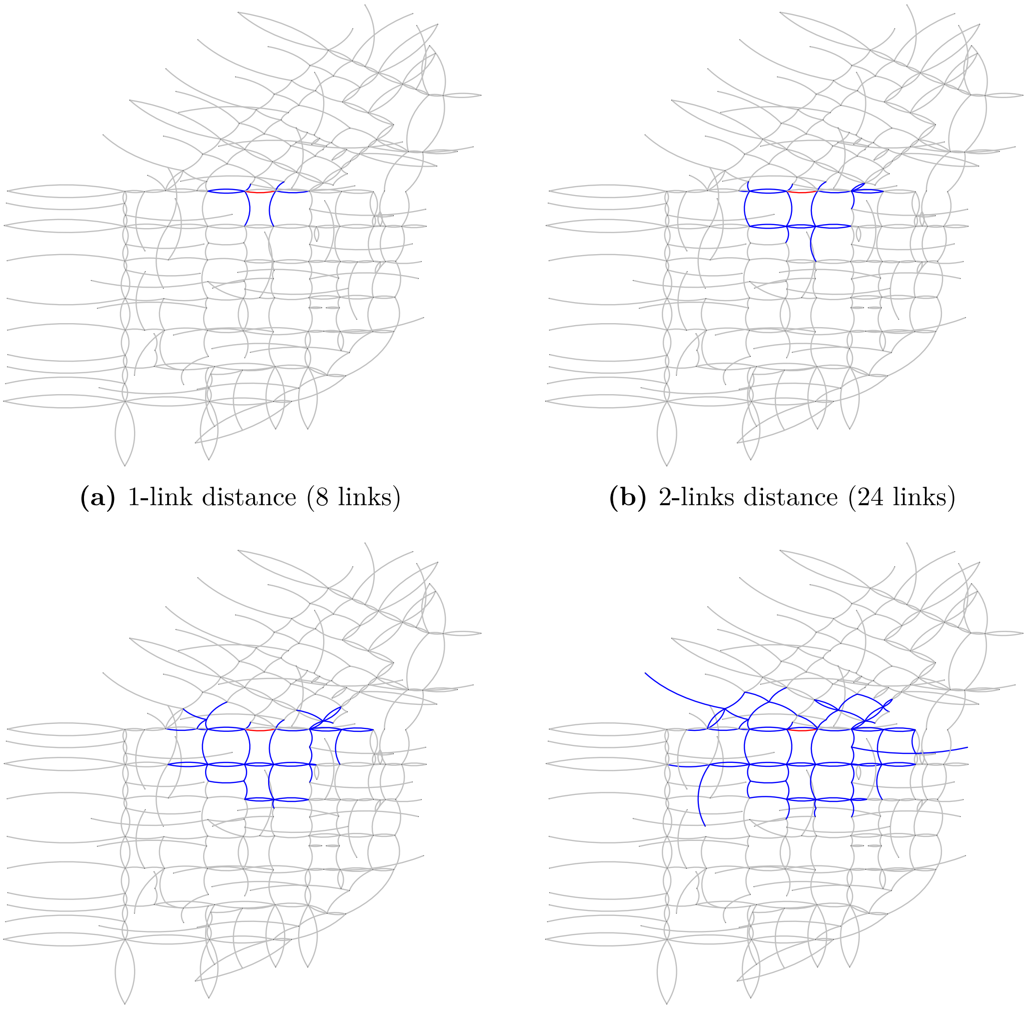
The sub-network surrounding any link of interest is composed by a set of interconnected links upstream and downstream of the central link of interest, in the form of tree branches, whose distance from the central link, measured in number of consecutive links, is less than a predefined number, which dictates the sub-network size. This set of links together with their start and end nodes form the basis of a

Algorithm 5: Repair the destroyed solution s' with sub-network evaluations (Alternative module for Algorithm 2, Line 5)

```
Data: destroyed DBL plan $s'$, $rd_{\max}$, $\omega_z$
Result: repaired solution $s''$
Calculate no of links to add, $r_{iter}$        (when no length constraint applies)
$s'' = s'$
$counter = 0$
while $counter \leq r_{iter}$ do
    specify current set of links with no DBL
    calculate addition probabilities: $p_z^a$
    draw a sample set of links $L$ from the distribution of $p_z^a$
    for $j \in L$ do
        create sub-network around link $j$
        evaluate performance of sub-network without DBL in $j$, $f_0$
        evaluate performance of sub-network with DBL in $j$, $f_1$
        store improvement $f_{sbn}(j) = f_0 - f_1$
    end
    find best link $l \in L$, so that $f_{sbn}(l) < f_{sbn}(j), \forall j \in L, j \neq l$
    update solution $s''$ by adding link $l$: $y_l = 1$
    $counter = counter + 1$
end
return $s''$
```

sub-network. To maintain local simulation accuracy, links upstream/downstream of the central link that are not directly connected to it but receive (or send) flow from (to) other sub-network links, are also included in the sub-network. An example of different size sub-networks around the same link of interest can be seen in Figure 2.2. Sub-network size, as well as the considered traffic simulation time window, affect the accuracy of the assessment of the potential impact of DBL in central link. A smaller size sub-network and a short selected simulation period might lead to misleading results, which would not be verified by a full-network full-time simulation test. However, smaller sub-network size and simulation time result in lower computational cost.

Traffic simulation in sub-networks is done in the same way as in full network, by using the same traffic model. Among sub-network links, those with no upstream link function as origin links, while links with no downstream link function as destination links. By using the time-series of observed link inflows of the sub-networks' origin links, which were calculated by the latest full-network simulation, as the input demand for the sub-network, as well as the turn and exit ratios of the full-network, we simulate traffic in every sub-network with and without considering DBL presence in the central link, for a short period of peak hour. The initial sub-network traffic state is the same as in the last full-network simulation, at the



(a) 1-link distance (8 links)

(b) 2-links distance (24 links)

(c) 3-links distance (47 links)

(d) 4-links distance (83 links)

Figure 2.2: Sub-networks of different size in the neighborhood of the same reference link (in red). Distance refers to the number of consecutive and connected links included, upstream and downstream of the central link.

time step corresponding to the start of the sub-network simulation. Also, DBLs placed in other sub-network links apart from the central, according to destroyed solution s', are always considered in sub-network simulations but their position does not change. Only central links are tested with and without DBL. PHT difference between the two scenarios (with and without DBL) is used as the evaluation criterion of the considered DBL addition. However, the accuracy of the sub-network evaluations is expected to be reduced in iterations with high destruction or/and repair degree, as DBL network configuration differs significantly at the beginning and end of the repair process, meaning that the first sub-network evaluations are done in significantly different surrounding conditions compared to the last ones. This is the reason that this method proves more efficient with smaller scale solution

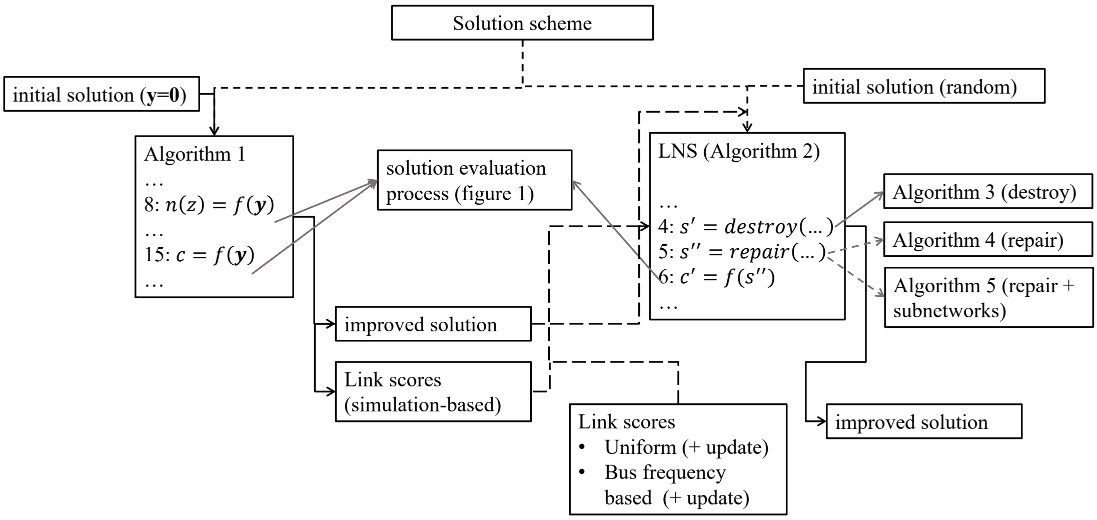
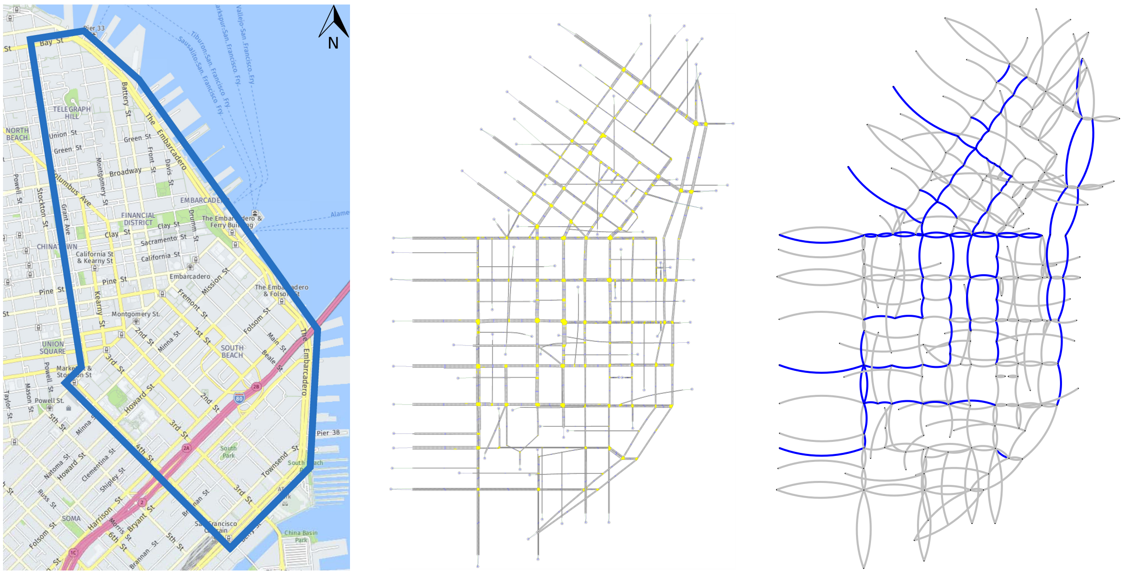

Solution scheme
initial solution (y=0)
initial solution (random)
Algorithm 1
…
8: $n(z) = f(\mathbf{y})$
…
15: $c = f(\mathbf{y})$
…
solution evaluation process (figure 1)
LNS (Algorithm 2)
…
4: $s' = destroy(\ldots)$
5: $s'' = repair(\ldots)$
6: $c' = f(s'')$
…
Algorithm 3 (destroy)
Algorithm 4 (repair)
Algorithm 5 (repair + subnetworks)
improved solution
Link scores (simulation-based)
Link scores
• Uniform (+ update)
• Bus frequency based (+ update)
improved solution

Figure 2.3: Overview of the proposed solution scheme and the relation between algorithms, inputs and outputs; dashed lines indicate alternative options.

modifications, similar to the effectiveness of LS techniques.

2.3.3 Summary of the solution scheme

The proposed solution scheme that is described in details in section 2.3, is graphically represented in Figure 2.3, where the connections between the algorithmic components discussed above, as well as the sequence of the proposed solution's procedure are depicted. Dashed lines indicate alternative options. For instance, improved (optimized) solutions can be produced either by Algorithm 1 or 2, executed separately, or by executing first 1 and then 2, where outputs of the first one can be used as inputs in the second one (e.g. a good-quality initial solution or the simulation-generated link scores can be used as inputs to LNS). As we will see in the following section, though, Algorithm 1 performs significantly higher number of solution evaluations that require longer execution time, for the generation of a single improved solution. However, at the same time, it can generate a set of simulation-based link scores, that are highly effective in indicating good candidate links for DBL introduction and can be used in LNS algorithm (Algorithm 2), to increase its efficiency. Nevertheless, the latter can be executed on its own without prior execution of Algorithm 1, by using link scores that are either uniform or generated based on bus operational characteristics, with the possibility of integrating an update process that adjusts link scores based on their performance in the process of LNS. The repair process of LNS can be performed by either Algorithm 4 or 5. The latter performs sub-network evaluations to increase the possibility of building improving solutions, especially in very large networks. Finally, the evaluation of any newly formed solution follows the process depicted in Figure 2.1.



(a) **(b)** **(c)**

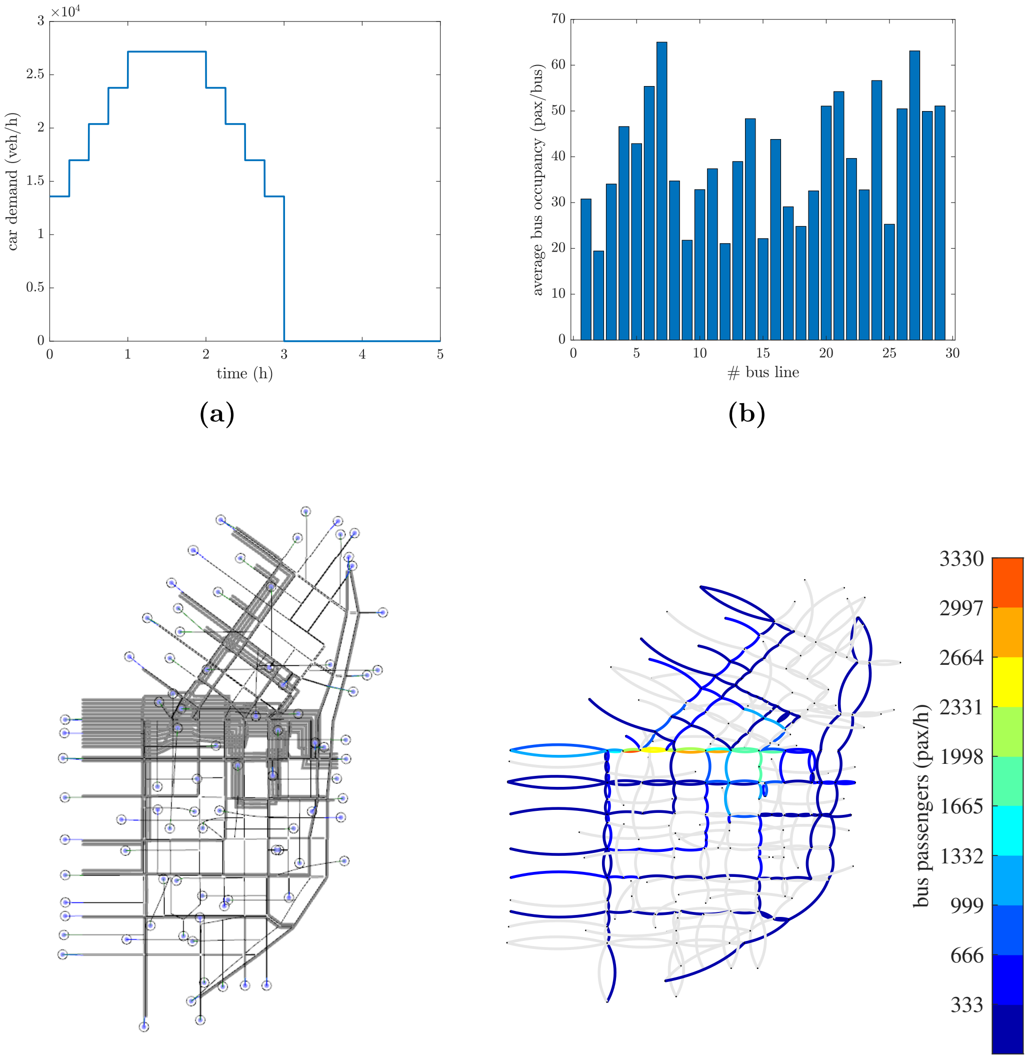
Figure 2.4: The San Francisco road network; (a) Map of the studied area. (b) Model of the network in Aimsun. (c) Map of candidate roads for bus-only lane installation (in blue).

2.4 Numerical application

2.4.1 Case study

The proposed method is applied to part of the traffic network of San-Francisco central area, in California, USA (see Figure 2.4(a)). The network is composed of 426 links and 267 nodes, out of which 156 represent signalized intersections on a pre-timed traffic signal control plan with cycle lengths up to 100 sec. Out of all links, 96 are labeled as candidate for bus-lane setting (see Figure 2.4(c)), corresponding to a total of 10.6 lane-kilometers. Candidate links are in principle all links that are included in the route of at least one bus line and have two or more lanes in total. In order to avoid excessive disturbance of car traffic, links with only two lanes were included in the candidate set only in case of considerably high bus frequencies.

The dynamic profile of car trip demand has a trapezoidal shape, as depicted in Figure 2.5(a), where the total generating car flow in the network is shown over time; this demand is distributed in 42 origin nodes. There exist 29 bus lines traveling in the studied region. The distribution of bus routes in the network can be seen in Figure 2.5(c). Bus frequencies and passenger ridership information are chosen so as to replicate realistic conditions: we assume 6 buses/hour (headway of 10 min) for all bus lines and bus occupancy that follow a trapezoidal profile over time, similar to the car demand profile. Average bus lines ridership in the central roads is assumed double than in peripheral roads. Bus average ridership is different for every bus line, ranging from 20 to 65 pax/bus during peak hour according



(a) **(b)**

(c) **(d)**

Figure 2.5: Case study input data about car and bus travel demand; (a) Profile of the initial total car travel demand over time; (b) Average passenger ridership of buses per line in peak-hour; (c) Bus routes distribution in the network; (d) Average bus passenger flow per link in peak-hour.

to Figure 2.5(b). Aggregated transit passenger flow during peak-hour is shown in Figure 2.5(d). Initial shares of total demand per mode are assumed equal to 74% for cars and 26% for buses, while total number of commuters is equal to 87K. Time-varying turn ratios, reflecting route choices of car users, are calculated by a preliminary microscopic simulation analysis performed with Aimsun software, for the case where no DBL is considered. Average turn ratios for every approach are calculated in periods of 15 mins. Without loss of generality, turn ratios are considered unaffected by DBL setting in the present study, based on the assumption that bus-only lanes introduction will not significantly affect the route choices of drivers; this can be changed in the future and use the same approach with turn ratios that can be adjusted to the specific DBL configuration tested.

Every potential DBL plan is evaluated by a sequence of traffic simulations followed by mode choice adjustments until convergence is achieved (as shown in Figure 2.1). Every traffic simulation is performed for a 10 hour period, in time steps of 5 sec, in order to guarantee that network is empty by the end of simulation time (in most cases network is empty after 4-5 hrs).It should be noted that in case where different DBL assignment for the evening peak is possible, the whole process needs to be repeated, by considering car/bus demand information of the evening period, in order to identify the best possible DBL assignment for the evening hours. In our case, we study only the morning peak period and derived solutions correspond to the morning DBL assignment. Average free-flow speed for cars and buses is assumed equal to 25 km/h. The same value is set for buses and cars for simplicity, since we observed in preliminary experiments that the results are not sensitive to small variations of bus free-flow speed. Average vehicle length is assumed equal to 5 m and average car occupancy 1 pax/car. For dwell time estimation, we assume that at every bus stop, the average fraction of boarding and alighting passengers, β_z^l, of Equation (2.17), is assumed to be roughly 30% of the current bus occupancy.

Logit model parameters are defined as follows: $ASC_c = 1.074$, $ASC_b = 0$, $\beta_c = -2.578$, $\beta_b = -9.294$. The iterative process of Figure 2.1 is terminated if the percentage of passengers changing mode with respect to previous state is less than 0.1% of the total number of passengers or after 50 iterations at most, in case no convergence can be achieved.

2.4.2 Problem variants

The proposed algorithmic scheme is applied for the case study described in Section 2.4.1. Algorithm 1 which generates a step-by-step constructed solution and link scores estimation, is executed first. Then, LNS Algorithm 2 is executed using link scores estimated by Algorithm 1, in addition to other approximations (see below), in order to efficiently explore the solution space. The two algorithms are applied for the following variants of the DBL optimization problem:

1. Minimize total passenger travel time without penalty related to total solution length ($\gamma = 0$ in equation (2.21)) and with an additional constraint for total DBL length (equation (2.32)).

2. Minimize sum of total passenger travel time plus penalty cost for road space occupied by DBL (equation (2.21)) without additional constraint for total DBL length. Penalty cost per unit length of DBL installed is set to $\gamma =$ 750 hours/(km day). We estimate this value by considering hourly total maintenance cost of DBL equal to 1715 USD per lane-kilometer, 10 hours of daily DBL operation and value of time equal to 22.90 USD/hour.

2.4.3 Results

Algorithm 1

Algorithm 1 is applied for both problem variants 1 and 2. However, link score calculation is only executed for variant 1, where only total PHT is included in the objective function. This is done so that scores reflect the potential of each link in improving system performance without considering their cost (per length unit). Algorithm 1 outputs include a good quality DBL plan and a set of link score values. Figure 2.6(a) shows the solution performance at the end of every iteration of the algorithm, for problem variant 1. As expected, total PHT is reduced in every step (with step 0 corresponding to no DBL scenario), after best link addition, following evaluation of all available options. The red line shows the car users percentage reduction, in response to improved bus travel time resulting from DBL setting, according to the utilized Logit model. For comparison reasons, the dashed line shows the system traffic performance for the initial network state (no DBL in the network), but considering the reduced car demand, according to the estimated mode shift of every step of Algorithm 1.

The difference between dashed and solid blue lines indicates that a significant part of the network performance improvement in presence of an efficient DBL plan, results exclusively from smart DBL distribution and is not related to the assumed mode shift from car to bus that is motivated by bus travel time improvement. In other words, even if we achieve mode shift from cars to buses by different means (dashed line), without DBL installation, the improvement of PHT is still smaller than in the presence of an efficient DBL plan (solid line). This can be explained by the fact that optimized DBL location selection can take advantage of unused road space surplus in specific roads and cause minimum car traffic disturbance will increasing bus speed. Another possible explanation for this is that for specific DBL configurations, the restricted car flow related to the reduced general-purpose lanes can produce an effect similar to perimeter control. Limiting vehicle access to congested parts of the network due to bottlenecks created by DBLs can sometimes have a beneficial effect on network performance, even if some vehicles are forced to wait longer in queues. This happens because the high-demand region remains below or close to capacity while, at the same time, bus delays are reduced due to DBL presence. It is useful to notice that the algorithm stops adding DBL links to the plan when no further improvement is possible, this is why the blue curve appears strictly monotonous. If any of the remaining links was added in the current plan, it would degrade the network performance and the PHT would increase, thus the blue curve would go up. While in this particular case the solution constructed by Algorithm 1 included almost all candidate links in the plan, this is not generally expected as it is related to the continuing mode shift that is indicated by the respective Logit model, whose characteristics differ from case to case.

Figure 2.6(b) shows PHT improvement in relation to the gradual increase of the total DBL lane-kilometers installed, as links are added in the DBL plan. The algorithm stops adding links when no further PHT improvement can be achieved. In the case of Figure 2.6 (problem variant 1), the best solution found includes DBL in almost all candidate locations (a total of 9.1 lane-kilometers) mainly because of the continuing mode shift, given that no constraint for total length is imposed in this case. Link scores ω, representing the average ranking-based scores of all trials, generated by Algorithm 1, are shown in Figure 2.7, with the whiskers' length representing the standard deviation over all link trials. Scores take values in the range (0,1), with larger scores indicating good performance of DBL if placed in the link. In the figure, links are sorted, from left to right, in the sequence they are added in the DBL plan. Links added earlier in the solution are tested fewer times and therefore have a smaller standard deviation. As expected, there is some correlation between average score values and sequence of addition, even though there are some low link scores with high standard deviation that are added in relatively early steps. However, it is expected that score values can increase the efficiency of LNS framework of Algorithm 2. Figure 2.8 shows results of Algorithm 1 applied for problem variant 2. As expected, the algorithm stops adding DBLs to the solution much earlier and the final solution is more efficient, meaning that the ratio of achieved PHT improvement over DBL-occupied road space is considerably higher. While Algorithm 1, in principle, results in good quality solutions (as it involves enumeration of a large number of possible DBL additions per step) and its execution is required if we need to extract simulation-based link scores, it is highly expensive computationally (4506 and 2313 DBL plans tested for problem variants 1 and 2, in 25.39 and 12.36 hours, respectively), while there is no guarantee

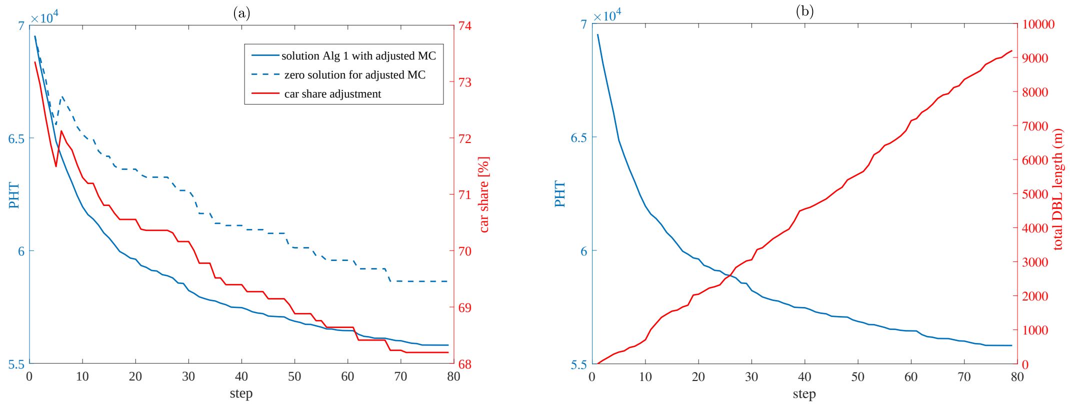


Figure 2.6: Algorithm 1: Gradual improvement of solution for problem variant 1 (PHT Only); (a) Step-by-step decrease of PHT (blue solid) and respective car share adjustment (red) as DBL plan is constructed. In dashed line the PHT of the initial state (no DBL) considering the adjusted car share (red line); (b) Step-by-step PHT decrease vs DBL solution length (red)

of optimal solution. This is why the optimization framework is enhanced by using LNS Algorithm 2, which can produce similar quality results with much fewer DBL plan evaluations (100 per replication in our case).

LNS algorithm

LNS, as described in Algorithm 2, is applied for problem variants 1 and 2. Since running Algorithm 1 to acquire link score values might not always be possible, due to its high computational cost, we test different performance indicators (scores) to estimate the potential of each candidate link in improving network performance if included in the DBL plan. In order to evaluate the role of link scores in algorithm performance, we execute LNS for problem variant 2 several times, using different sets of link score values for the destruction and repair of the solution, which we label as follows:

- *Uniform*: All link scores are equal. All links have the same probability to be chosen for removal or addition of DBL.
- *Bus Frequencies*: Every link score is equal to the ratio of total bus frequency of the link over the maximum bus frequency observed among candidate links.
- *Scores*: Link scores are found by Algorithm 1 based on simulation trials, according to equation (2.33).
- *Uniform + Update*: All link scores are initially equal but are being updated within LNS algorithm, according to the process described above
- *Bus Frequencies + Update*: Links are initially assigned scores according to their bus frequencies (see above) but are being updated within LNS algorithm, according to the process described above

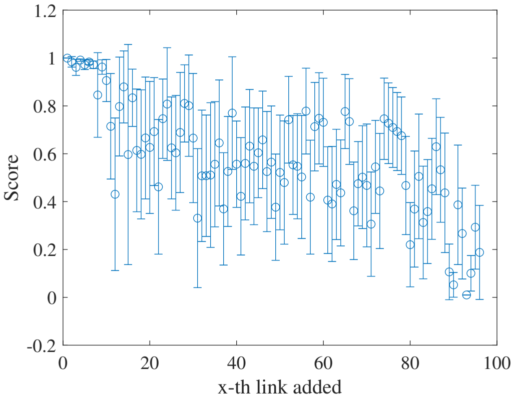


Figure 2.7: Link scores calculated by Algorithm 1: Mean ± standard deviation of all trials for candidate links in the sequence they are added to the solution (left to right).

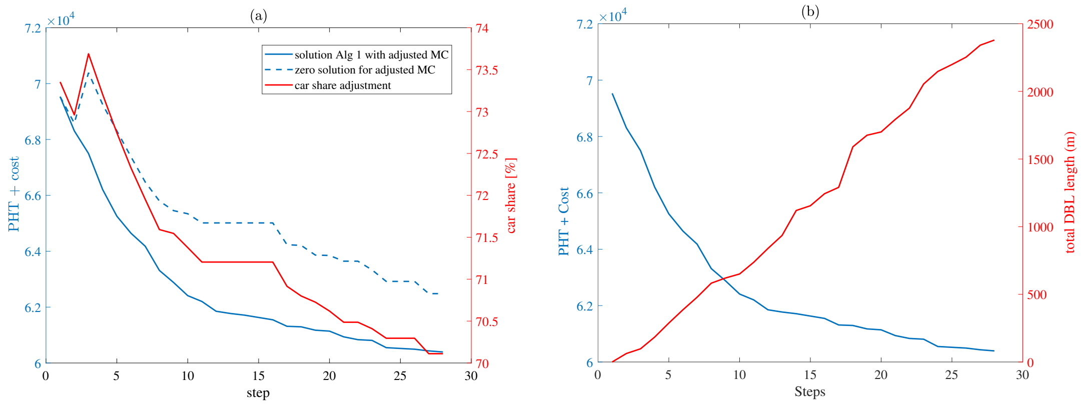


Figure 2.8: Algorithm 1: Improvement of solution for problem variant 2 (PHT + Cost); (a) step-by-step decrease of the solution cost (blue solid) and the resulting car share (red) as DBL plan is constructed. In dashed line the cost of the initial state (no DBL) considering the car share of the red line; (b) step-by-step cost decrease and solution length (red) as DBL plan is constructed

For all cases, LNS algorithm is executed 10 times, for 100 steps each, starting from a random initial solution with total DBL length corresponding to around 50% of the total candidate space. The algorithm takes as input the same set of initial solutions for all above cases. Maximum destroy/repair degree for all cases is set to $dd_{\max} = rd_{\max} = 30\%$ of number of links and re-adding the same links at the same step is not allowed, in order to avoid circling around the same solutions. For the last two cases of the above list, where link scores are being updated within LNS, the algorithm performs 250 steps, in order to include a warm-up part. The decay factor is set to $\lambda = 0.5$ and the update is done every $\kappa_{\text{int}} = 10$ steps.

In Figure 2.9 we show boxplots with distributions of total PHT with and without penalty cost and respective total DBL lane-kilometers of the initial and best found solution sets, after executing LNS algorithm for every score case listed above, for problem variant 2. In Figure 2.9(a) we observe that LNS is successful in improving a random initial solution of average quality, even by using uniform link scores for the destroy and repair processes, but the best performance is observed when link scores are computed by Algorithm 1. Small dispersion of the best found solutions for this case indicates higher probability of finding a good quality solution with fewer replications, compared to using other types of link scores. Link scores based on bus frequencies are also more efficient compared to uniform scores, showing, as expected, that higher bus frequency indicates more suitable link for DBL setting. Algorithm performance when less efficient link scores are used (such as uniform or bus-frequency-based), can be improved by applying the score update process described above. However, a larger number of iterations is necessary for the update process to be effective. We can see from Figure 2.9(a) that by updating link scores, best found solutions are improved with respect to cases without score update.

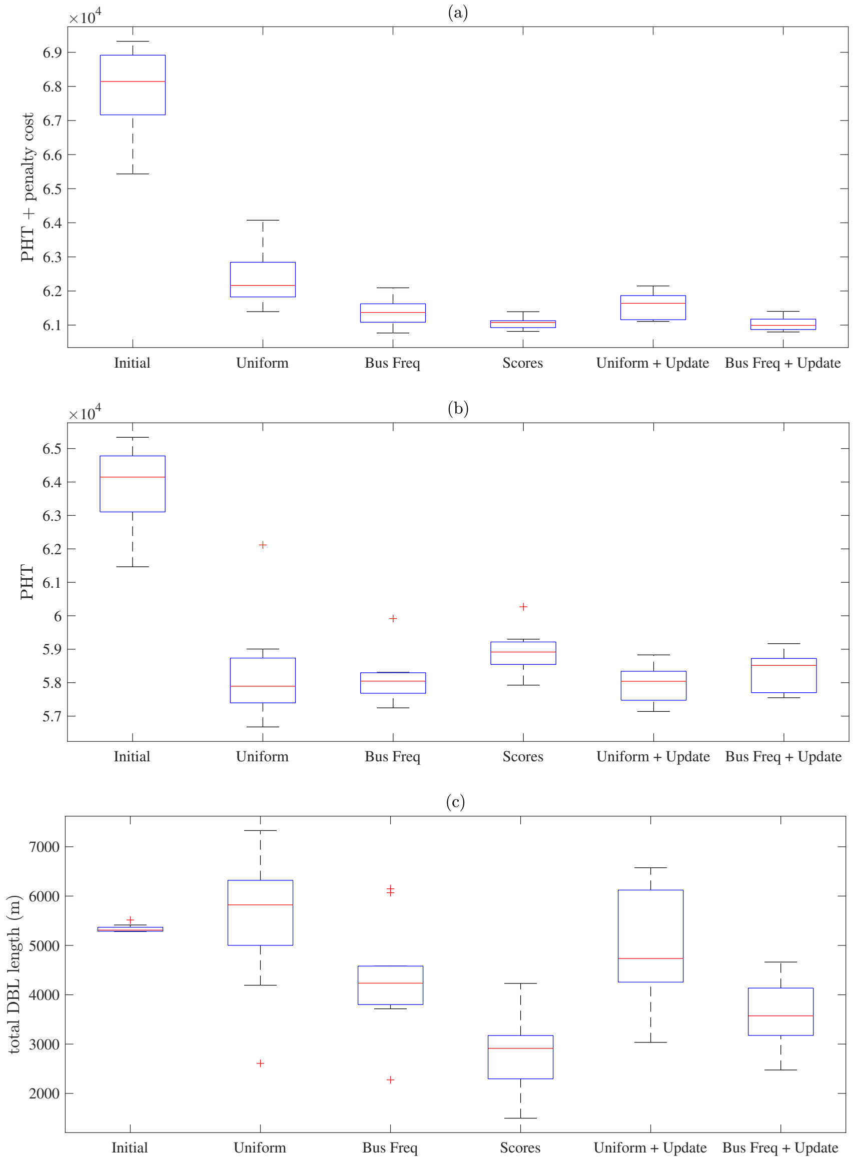
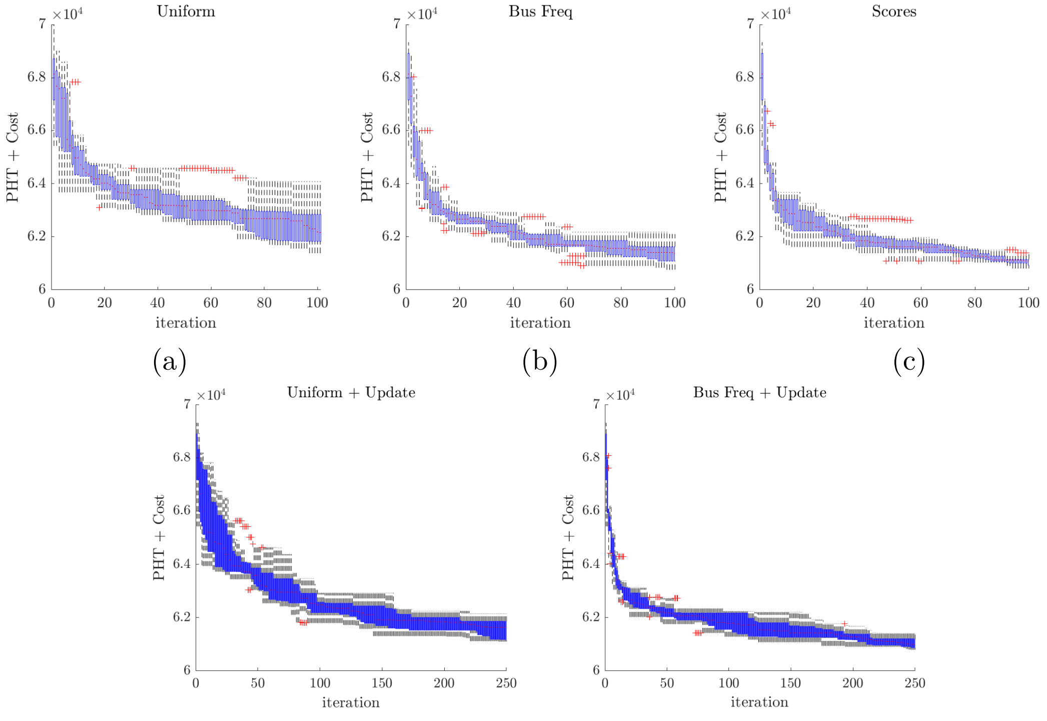


Figure 2.9: Performance of LNS algorithms for every type of link scores used. The boxplots show the distribution of different attributes of the best found and the initial solutions for each case; (a) distribution of total cost (PHT + penalty cost); (b) distribution of total travel time (PHT); (c) distribution of total DBL length.

This means that prior execution of Algorithm 1 is not necessary for efficient LNS application. The utility of an update process is more obvious in cases of different scenario evaluation, e.g., different demand or bus input data, where repeated execution of Algorithm 1 would be unrealistic. By comparison of Figures 2.9(a) and (b) we can see that solving the problem variant 2, i.e., with penalty cost included in the objective function and no solution length constraint, the algorithm seeks space-efficient solutions, where a compromise between PHT improvement and road space given to DBL is made. This is obvious if we compare the characteristics of the best found solutions of all tests in Figures 2.9(a)–(c), where we observe that even though the solutions found by “Scores” or “Bus Frequency + Update” tests have lower values of combined PHT plus penalty cost, PHT improvement is smaller compared to solutions of other cases; however, they occupy less road space for DBL. Therefore, the solutions that we consider best in variant 2 are those having the largest PHT improvement per lane-km of DBL. We also note that the total number of lane-kilometers of the initial random solutions is not binding, since LNS algorithm can increase or decrease total DBL plan length during the search process, due to the way destruction and repair are done and the greedy acceptance criterion for new solutions.

Effectiveness comparison between link score types used in LNS for problem variant 2 is complemented by Figure 2.10, where one can see the evolution of LNS search for every link score type used. The graphs show in boxplots the cost of the best solution found so far after every LNS iteration, for all replications of every score case. It is evident that using simulation-based link scores of figure (c), coming from Algorithm 1, leads to better solutions in fewer LNS iterations compared to cases (a) and (b), while final solution costs of 10 replications show small variance,



(a) (b) (c)

(d) (e)

Figure 2.10: Evolution of the search process of LNS algorithm with different types of link scores; (a) Uniform; (b) based on bus frequency; (c) based on simulation trials (Alg. 1); (d) Uniform with update process; (e) based on bus frequency with update process. Figures show distribution of total cost of the best solution found so far for all performed replications in the form of boxplots.

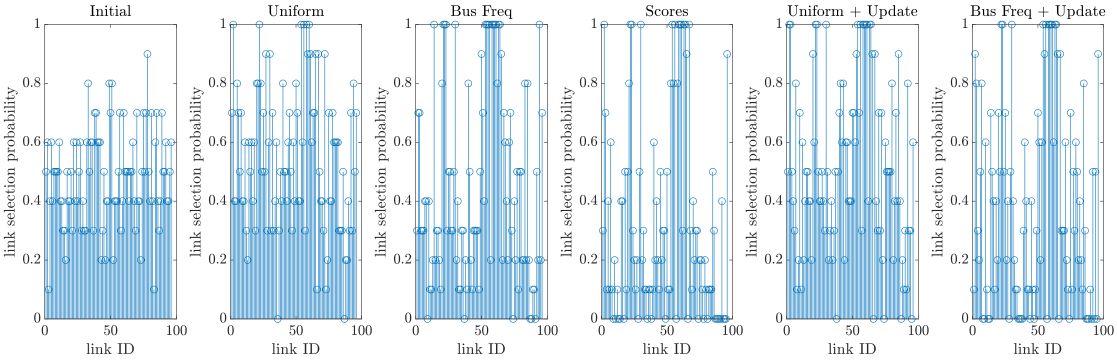


Figure 2.11: Selection probability of all candidate links in the Initial and the best found set of solutions for every type of link scores used by LNS algorithm, for all performed replications per case.

in contrast to the "Uniform" case, where costs of best found solutions after 100 steps vary significantly (larger IQR), indicating weaker optimization convergence (more steps required). The degree of similarity between best found DBL plans is shown in Figure 2.11, where the observed frequency of every link's appearance in the final DBL plan, after 10 LNS executions, is displayed. While the set of random initial solutions is composed of diverse solutions, the best solutions found by LNS by using any type of link scores show increased similarity. More specifically, several "good" DBL candidate links are identified and included in the final DBL plan, in most or even all replications, in more efficient score type cases, such as "Scores" and "Bus Frequency + Update". On the contrary, not promising links are rarely or never included in the final solution.

LNS using the alternative repair process described by Algorithm 5, which is mainly proposed for very large network instances, is tested for our case study with a smaller destroy and repair degree, resulting in small changes of the incumbent solution per LNS step. The smaller modification degree is considered important for the effectiveness of this repair strategy, because accuracy of sub-network evaluation, dictating the next best position for DBL addition out of a stochastically created subset of available links, depends on the surrounding network conditions. Therefore, in a large destruction case, where many of DBL links have been removed from the incumbent plan, sub-network evaluations performed in the first repair iterations assume surrounding conditions (neighboring DBLs) significantly different than those towards the last repair steps, when most DBL links have already been re-added. However, in very large networks, the process of sub-network evaluation, driving the repair process of the solution, is expected to increase search efficiency, even by using small destruction and repair degree. In our case, we apply this alternative repair process in LNS algorithm for 10 replications of 200 steps each, by starting from the same set of random initial solutions as in previous cases, for problem variant 2, with maximum destroy and repair degree $dd_{\max} = rd_{\max} = 5\%$. Link scores are derived from link bus frequencies and re-addition of the same links in the same step is not allowed. The sample size of link set L, that are tested in sub-network evaluations before every addition (Line 7 in Algorithm 5), is set to 10. The sub-networks used are composed of 4-links distance upstream and downstream of the central link (as in Figure 2.2(d)) and traffic simulation is done for a 2-hour peak period (from 0.5 to 4.5 hours). The initial state of the sub-networks is imported by the most recent full-network simulation. The characteristics of the best found solution set are listed in Table 2.1. While LNS with sub-networks did not provide significantly better results in this case study and performs worse than other LNS approaches without sub-networks, it can be of added value for even larger networks where a full simulation of the network might be prohibited; for example Chicago city has a network of 64000 links and close to 21000 buses (see Verbas et al., 2015), compared to only 426 links and 522 buses for the case study considered in this chapter.

2.4.4 Best found solutions

By considering the installation/maintenance cost in the objective function, the optimization process identifies the amount of required lane-kilometers of DBL indirectly, by balancing delay savings and costs. Setting a suitable value for cost factor γ can be done by accounting several maintenance cost components and the process depends on the specific case study and the objectives of the traffic authorities. In cases where cost is not important, the decision making process can be facilitated by applying LNS algorithm for problem variant 1, i.e. considering $\gamma = 0$ and introducing a constraint for a target total solution length. In fact, by solving the problem for a range of target length values, we can construct a Pareto frontier, as in figure 2.12, demonstrating the improvement in PHT as a function of the sum of lane-kilometers occupied by DBLs.

Adding to the cases tested for variant 2 that were discussed in the previous section, LNS algorithm is applied for problem variant 1 as well, by using link scores defined by Algorithm 1 and target length constraint corresponding to 25% and 50% of the total candidate lane-kilometers, as well as for a case where neither a penalty cost term nor a length constraint is included. The characteristics of the best solution found per case are listed in Table 2.1. The first six columns of the table list the percent change with respect to the initial network state (no DBL) in: PHT values with and without adding the penalty cost for the total lane-kilometers of DBL (as set in section 2.4.2, point 2); car and bus users preferences after DBL introduction; and average travel time per kilometer TT_c and TT_b for car and bus, respectively. The last two columns list the total number of DBL lane-kilometers of each plan and the median execution time per replication per case, respectively. The optimization algorithms, as well as simulation model, are coded in Matlab R2019b and run on an Intel-Core i7-7700 CPU at 3.6 GHz. We observe that all solutions found through the proposed framework lead to a significant PHT improvement, in the range of 15% to

Table 2.1: Performance of the best solutions found by LNS Algorithm 2 by using different link score types (upper part) and sub-network evaluations (middle part) for problem variant 1, and by using link scores defined by Algorithm 1 for problem variant 2 (lower part). The values represent the % change with respect to the values observed in initial scenario (no DBL), except for the last two columns, which show total DBL length (km) and median execution time (h) per replication per case, respectively.

Case	PHT+Cost	PHT	car	bus	TT_c/km	TT_b/km	DBL (km)	time (h)
Problem Variant 1								
Uniform	-11.36	-16.75	-5.73	+15.77	-11.79	-26.46	5.00	1.57
Bus Frequency	-12.50	-16.94	-5.24	+14.42	-12.46	-25.53	4.11	1.59
Scores (Alg. 1)	-12.52	-14.82	-4.21	+11.59	-11.17	-21.96	2.13	1.34
Uniform + update	-11.90	-15.17	-4.53	+12.48	-11.23	-22.98	3.03	3.45
Bus Freq + update	-12.57	-17.03	-5.57	+15.33	-12.24	-26.31	4.13	2.80
Scores + sub/works	-11.48	-18.27	-6.40	+17.62	-12.78	-28.82	6.30	2.48
Problem Variant 2								
Scores / const. 25 %	-12.56	-15.53	-4.34	+11.95	-11.79	-23.03	2.76	0.85
Scores / const. 50 %	-11.84	-17.93	-5.57	+15.34	-13.16	-27.46	5.64	1.04
Scores / no const.	-8.59	-19.45	-7.74	+21.31	-12.79	-32.10	10.07	1.92

19% with respect to the no DBL case. The solution leading to the highest possible improvement in total PHT is found by LNS algorithm when no length constraint or penalty cost is considered. This solution, however, is the least efficient as the algorithm tends to add a lot of lane-kilometers of DBL, mainly due to the continuing, assumed mode shift from cars to buses that is predicted by the mode choice model. Nevertheless, high accuracy of the prediction of the induced mode shift cannot be guaranteed, even with a well calibrated model, as commuters' behavior can be highly influenced by numerous factors that might not be considered by the model. This is why, the efficiency of a DBL plan in terms of lane-km requirement should be considered, meaning that DBL plan should aim at a high ratio of PHT improvement over road space usage. For instance, DBL plans found in case 'Scores' or 'Bus frequencies + update' are the most efficient choices for the present case study, as demonstrated by the respective values of PHT combined with penalty cost.

Best found solutions for all test cases are presented in Figure 2.12 where one can observe the relation between achieved PHT improvement compared to the no-DBL case, and total DBL plan length. The points of lower surrounding curve form the Pareto frontier for this case study, which can assist traffic authorities in the decision making process of DBL allocation. As expected, higher number of DBL lane-kilometers correspond to higher improvement in total travel time. The most efficient solutions seem to be those closer to total lengths of 2 to 3 km, after which, the required road space for DBL per unit of additional PHT improvement increases significantly. Algorithm 1 is proved very efficient in finding solutions of good quality but with a relatively high computational cost (25.39 h running time), which grows exponentially with the size of the network. LNS algorithm applied for problem variant 1 shows remarkable performance, as the best solutions found from

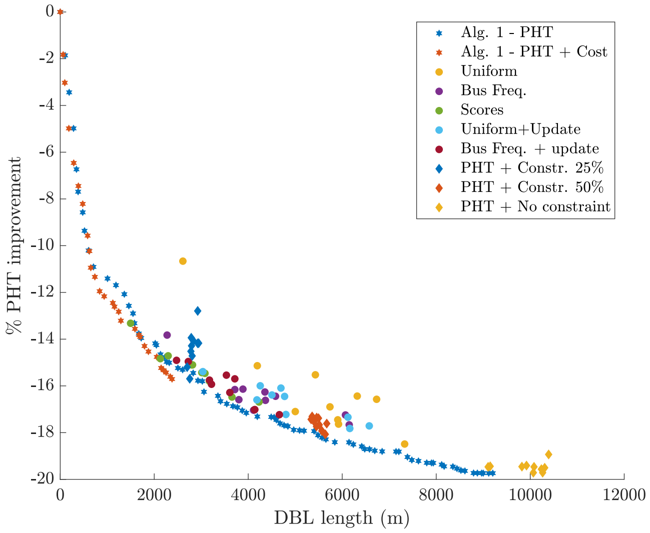


Figure 2.12: Pareto frontier relating the % improvement in total PHT compared to the no DBL case, with the DBL length of the best solutions found by algorithm 1 and LNS algorithm 2 for the two problem variants.

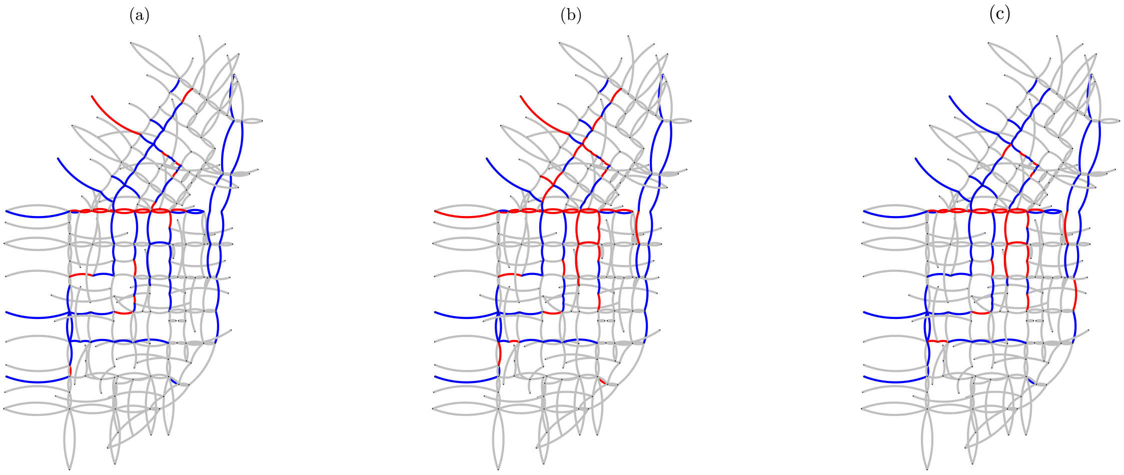


Figure 2.13: Spatial distribution of three of the best performing DBL networks identified by different runs of LNS algorithm (links with DBL shown in red) with the respective PHT improvement compared to the initial state of no DBL; (a) LNS using simulation-based scores for problem variant 2 (-12.52%, 2.13 km of DBL); (b) LNS using scores based on bus frequencies plus an update process for problem variant 2 (-12.57%, 4.13 km of DBL); (c) LNS using simulation-based link scores for problem variant 1 with a constraint of reserving maximum 25% of available road space (-12.56%, 2.76 km of DBL).

all tests are very close or even better that those produced by Algorithm 1, with much less computational cost, according to the values in the last column of Table 2.1. Regarding the observed execution time, the higher values reported in the cases that include a score update process and in the one using sub-networks, relate to the increased number of performed LNS steps (250 and 200, respectively), compared to the rest of the cases, which all performed 100 LNS steps. It is worth highlighting here the difference in computational costs between the proposed method and microscopic simulation. While one complete replication of LNS optimization, performing 100 evaluations of different solutions, lasts on average 1.5-2 hours, microscopic simulation might take 30-40 mins to evaluate just one candidate solution.

The DBL space distribution of the most promising solutions of Table 2.1 can be seen in Figure 2.13. As expected, links with high bus frequencies are almost always included in the DBL plan. However, this characteristic should not be seen as universal, as the amount of existing car traffic in the same links is a significant contributing factor, which can alter the result in a different case study. It should also be noted that, despite the lack of any type of connectivity constraint, DBLs appear mostly connected in most solutions found. This fact can support the idea that connected solutions are generally more efficient, although this is only implicitly considered by the optimization process, as no connectivity constraint was imposed.

2.5 Summary

This chapter focuses on the problem of optimal DBL allocation, in large-scale, bi-modal urban traffic networks, with the objective of minimizing total passenger travel time. We construct a modeling and optimization framework on the basis of a link-based traffic simulation model with queuing characteristics, which is consistent with the dynamic nature of congestion propagation. Mode choice of commuters is adjusted according to travel times that result from the corresponding DBL network configuration, through the use of a Logit model. A combinatorial optimization problem is formulated and a set of heuristic and metaheuristic algorithms based on LNS is executed, that can be combined with a learning process and a network decomposition technique, are proposed and tested in terms of effectiveness and efficiency in finding good quality solutions in reasonable time.

3

Two-layer hierarchical adaptive traffic signal control for congested urban networks

This chapter is based on the article:

- D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2022c). "Two-layer adaptive signal control framework for large-scale dynamically-congested networks: Combining efficient Max-Pressure with Perimeter Control". In: *Transportation Research Part C: Emerging Technologies* (under review)

3.1 Introduction

Traffic-responsive signal control has received considerable research attention, as a cost-effective and easy-to-implement network management strategy, bearing high potential to improve performance in heavily congested networks with time-dependent traffic characteristics. Max Pressure (MP) distributed controller gained significant popularity due to its theoretically proven ability of queue stabilization and throughput maximization under specific assumptions. However, its effectiveness is questionable under over-saturated conditions and queue spill-backs, while network-wide implementation is often practically limited due to high instrumentation cost that increases with the number of controlled intersections. Perimeter control (PC) based on the concept of Macroscopic Fundamental Diagram (MFD) is a state-of-the-art aggregated control strategy that regulates exchange flows between homogeneously congested regions, with the objective of maintaining maximum regional travel production and prevent over-saturation. However, homogeneity assumption is hardly realistic under congested conditions, thus compromising PC efficiency.

In this chapter, the effectiveness of network-wide parallel application of PC and MP strategies embedded in a two-layer control framework is assessed in a link-based macroscopic simulation environment. With the aim of reducing implementation cost of network-wide MP without significantly sacrificing performance gains, we evaluate partial MP deployment to subsets of nodes, indicated as critical by a node classification algorithm that we propose, based on node traffic characteristics. A modified version of Store-and-forward (SaF) dynamic traffic paradigm incorporating finite queue and spill-back consideration is used to test different configurations of the two-layer framework, as well as of each layer individually, for a real large-scale network in moderate and highly congested scenarios. Several scenarios of independent and combined application of PC and MP, with different MP node layouts and network penetration rates are tested for a real large-scale network. All control scenarios are tested for two trip demand settings, leading to moderate and high network congestion under fixed-time control settings. Results show that: (i) combined control of MP and PC outperforms separate MP and PC applications in both demand scenarios tested; (ii) MP control in reduced critical node sets selected by the proposed strategy leads to similar or even better performance compared to full-network implementation, thus allowing for significant cost reduction; iii) the proposed control schemes improve system performance even under demand fluctuations of up to 20% of mean.

Following a review of the most relevant PC and MP studies justifying the motivation and expected contributions of this work in section 1.1.2, the chapter is structured as follows. At first, the theoretical framework of MP and PC control is introduced, together with the required notation explanation. The relations transforming the controller outputs to adjusted traffic signal plans are described for both control strategies and characteristics of the traffic simulation model are briefly discussed. Afterwards, the experiment setup is described and the methodological framework of node rating and selection is discussed. Simulation results and analysis follow, together with an analysis of the system's behavior for the different scenarios, compared to the fixed-time control case. Finally, the chapter is concluded by summarizing the main findings.

This research work presented in this chapter is submitted for publication in Transportation Research Part C and is currently under review. For the most updated version of this work, the reader is kindly redirected to the online version of it (https://arxiv.org/abs/2210.10453).

3.2 Methodology

A schematic representation of the proposed two-layer controller is shown in figure 3.1. In the upper layer, perimeter control is applied in an aggregated scale between a set of homogeneously congested regions. At the end of every control cycle, the

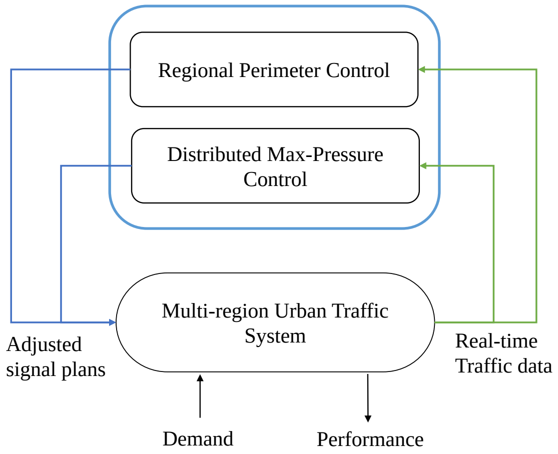

Regional Perimeter Control
Distributed Max-Pressure Control
Multi-region Urban Traffic System
Adjusted signal plans
Real-time Traffic data
Demand
Performance

Figure 3.1: Schematic description of the two-layer controller.

controller, based on inputs of aggregated regional vehicle accumulation, specifies the target inter-regional exchange flows for the next cycle, which are translated into the respective inter-regional green times between every pair of adjacent regions. The controller-specified inter-regional green times are then translated to exact green times per approach, for all PC controlled intersections, located on the boundaries between regions, by taking into account the actual boundary queues. In the lower layer, distributed control based on Max Pressure regulator is applied to a set of eligible intersections, in the interior of the regions. This set can contain all or a fraction of signalized intersections of the region, with the exception of those used for PC (if PC is applied in parallel). MP controllers do not communicate with each other or with any central control unit, but operate independently based on queue measurements directly upstream and downstream of the controlled intersections, by adjusting green times of the approaches accordingly, at the end of every control cycle. The control layers do not exchange information, however their combined effect is indirectly considered by both controllers through the real-time traffic measurements that they receive as inputs. The mathematical formulations of both controllers are described in the following subsections, followed by a brief description of the utilized traffic model and traffic simulation process.

3.2.1 Max-Pressure Control

Review of Max-Pressure feedback controller

Max-Pressure feedback-based control algorithm was initially formulated for traffic signal control independently by Wongpiromsarn et al., 2012 and Varaiya, 2013a,b, who theoretically proved its stability and throughput maximization properties, though under restrictive assumptions. It is a distributed and scalable control strategy, in the sense that its operation requires no communication with the rest of the network infrastructure and, therefore, it can be introduced to any signalized intersection at any time, without necessitating any readjustment of existing controllers elsewhere. In other words, gradual installation to network is theoretically possible. Moreover, it does not require any knowledge of the actual or expected traffic demand, in contrast to Model-Predictive Control (MPC) approaches, which simplifies practical implementation. The MP version applied in this work only requires real-time queue measurements of the links around controlled intersection and turning ratios of all alternative approaches that traverse the intersection. Both can be measured or estimated with proper instrumentation. Several modified versions of the original control algorithm have been introduced and tested by simulation experiments. The present version is similar to the one described in Kouvelas et al., 2014. The detailed description of MP algorithm together with the necessary notation is given below.

The traffic network is represented as a directed graph (N, Z) consisting of a set links $z \in Z$ and a set of nodes $n \in N$. At any signalized intersection n, I_n and O_n denote the set of incoming and outgoing links, respectively. The cycle time C_n and offset - which enables coordination with the neighboring intersections - are pre-defined (or calculated online by a different algorithm) and not modified by MP. Intersection n is controlled on the basis of a pre-timed signal plan (including the fixed total lost time L_n), which defines the sequence, configuration and initial timing of a fixed number of phases that belong to set F_n. During activation of each phase $j \in F_n$, a set of non-conflicting approaches v_j (i.e. connections between pairs of incoming-outgoing links of node n) get right-of-way (green light) simultaneously. The saturation flow of every link z, denoted as S_z, refers to the maximum possible flow that can be transferred to downstream links, depending on link and intersection geometry. The turning ratio of an approach between links $i-w$, where $i \in I_n, w \in O_n$ is denoted as $\beta_{i,w}$ and refers to the fraction of the outflow of upstream link i that will move to downstream link w. The present version of MP assumes that turning ratios are known to the controller. However, it has been shown that control effectiveness is not deteriorated if turning ratios are estimated (see Le et al., 2015). By definition, the following relation stands for every node n,

$$\sum_{j \in F_n} g_{n,j}(k_c) + L_n = (\text{or} \leq)\ C_n \tag{3.1}$$

where $k_c = 1, 2, \dots$ is the discrete-time control interval index, and $g_{n,j}(k_c)$ denotes the green time duration of phase j of node n at control interval k_c. The inequality may apply in cases where long all-red phases are imposed for any reason (e.g. gating).

The version of MP controller employed in this work assumes that phase sequencing is given as input (e.g. from pre-timed scheduling) and does not change during the control. Consequently, during every cycle, all phases will be activated for a minimum time in the same ordered sequence. As a result, the following constraint also applies for every node n,

$$g_{n,j}(k_c) \geq g_{n,j,\min}, j \in F_n \tag{3.2}$$

where, $g_{n,j,\min}$ is the minimum green time required for phase j of node n, which often matches the required amount of time for the respective pedestrian movements. The control variables of this problem, denoted as $g_{n,j}(k_c)$, represent the duration of the effective green of every stage $j \in F_n$ of all controlled intersections $n \in N$. Assuming that real-time measurements or estimates of the queue lengths (states) and turning ratios of all controlled intersections are available, the pressure $p_z(k_c)$ of every incoming link $z \in I_n$ of node n, at the end of control cycle k_c, is computed as

$$p_z(k_c) = \left[\frac{x_z(k_c)}{c_z} - \sum_{w \in O_n} \frac{\beta_{z,w} x_w(k_c)}{c_w} \right] S_z, \quad z \in I_n \tag{3.3}$$

In equation (3.3), $x_z(k_c)$ denotes the average number of vehicles that are present (moving or queuing) in link z during control cycle k_c and c_z denotes the storage capacity (maximum number of vehicles) of link z. Queue normalization by the link storage capacity aims at considering the link size, so that pressure of a smaller link is higher than that of a larger one with the same number of vehicles in it. In other words, pressure takes into account the likelihood of link queues - upstream and downstream - to spill-back in the following cycle. Pressures of all incoming links are calculated at the end of every cycle based on the latest queue measurements, which constitute the state feedback variables and are collected through proper instrumentation. Then, pressures are used by the controller for the signal settings update. It should be noted that the above formulation assumes that flow transfer is always possible between z and all $w \in O_n$. Otherwise, the second term of equation 3.3 refers only to downstream links for which flow transfer is allowed from z ($\beta_{z,w} > 0$). In different MP versions, x_z may refer to the instantaneous or the maximum observed queue length at the end of the control cycle. In this work, a preliminary analysis showed that mean queue length values generate better results and this way was adopted. By looking at the two terms, one can notice that, essentially, pressure depicts occupancy difference between upstream and downstream links. Therefore, higher pressure indicates higher potential in traffic production, i.e. significant volume waiting to be served and enough available space in downstream links to receive it. Low or close to zero pressure indicates lower need for right-of-way time, due either to small queue upstream, or to lack of space downstream (links close to capacity). We should note that negative pressures are meaningless, so constraint $p_z(k) \geq 0$ must always hold.

Based on equation 3.3 pressure is calculated for all incoming links $z \in I_n$ of node n. Then, the pressure corresponding to every stage j at control cycle k_c

is defined as the sum of the pressures of all incoming links that receive right-of-way in stage j, as follows.

$$P_{n,j}(k_c) = \max\left\{0, \sum_{z\in v_j} p_z(k_c)\right\}, \quad j \in F_n \tag{3.4}$$

This metric is then used as weight for the distribution of the total available green time between the competing stages of the intersection.

After pressure values $P_{n,j}$ are available for every phase $j \in F_n$ of intersection n, the total amount of effective green time G_n, calculated as

$$G_n = C_n - L_n = \sum_{j\in F_n} g^{\star}_{n,j}, \quad n \in N \tag{3.5}$$

is distributed to the phases of node n in proportion to pressure values. In equation (3.5), $g^{\star}_{n,j}$ denotes the green time assigned to phase j by static fixed-time analysis, using any of the standard algorithms. It holds that $g^{\star}_{n,j} \geq g_{n,j,\min}, \forall j \in F_n$.

There are several different approaches that have been proposed regarding green time calculation, some of which also include phase activation based on pressures. In this version, since phases are activated in a strictly defined and non-changing order with a guaranteed minimum green time, green duration, $\tilde{g}_{n,j}(k)$, is assigned to phases proportionately to the computed pressures, as follows

$$\tilde{g}_{n,j}(k_c) = \frac{P_j(k_c)}{\sum_{i\in F_n} P_i(k_c)} G_n, \quad j \in F_n \tag{3.6}$$

Eq. (3.6) provides the raw green times calculated according to MP controller. However, these values cannot be applied directly to the intersection signal plan, because it must first be guaranteed that they comply with a set of necessary constraints. Therefore, an additional step is added in the signal update process, whose objective is to translate MP outputs of eq. (3.6) to practically applicable green times $G_{i,j}$. This is done by solving online, for every control cycle k_c, the following optimization problem (similar to Diakaki et al., 2002 but in this case two additional constraints are added):

$$\begin{aligned}
\underset{G_{n,j}}{\text{minimize}} \quad & \sum_{j\in F_n} \left(\tilde{g}_{n,j} - G_{n,j}\right)^2 \\
\text{subject to} \quad & \sum_{j\in F_n} G_{n,j} + L_n = C_n \\
& G_{n,j} \geq g_{n,j,\min}, \; j \in F_n \\
& |G_{n,j} - G^{pr}_{n,j}| \leq g^{R}_{n,j} \\
& G_{n,j} \in \mathbb{Z}^+ \\
& \forall j \in F_n
\end{aligned} \tag{3.7}$$

According to the above formulation, the applicable green times for every phase $G_{n,j}$, $j \in F_n$, should be as close to the non-feasible regulator-defined greens $\tilde{g}_{n,j}$ as possible, while satisfying a set of constraints. The first constraint states that eq. (3.1) must always hold, therefore the sum of the updated feasible green times plus the total lost time L_n should be equal to cycle C_n. The second constraint ensures that all phases get a predefined minimum green duration $g_{n,j,\min}$. In order to avoid potential instability of the system due to large changes in the signal timing happening too fast, we impose a threshold to maximum absolute change of every phase duration between consecutive cycles. This is expressed in the third constraint, where $G^{pr}_{n,j}$ denotes the applied green times of the previous cycle and $g^{R}_{n,j}$ is the maximum allowed change of the duration of phase j between consecutive cycles. Finally, feasible green times must belong to the positive integers set. This type of integer quadratic-programming problem can easily be solved by any commercial solver fast enough to allow online solution for every control cycle. The solution of this optimization problem, i.e. variables $G_{n,j}, \forall j \in F_n$, are the new feasible phase duration for node n which will be applied in the next control cycle. The above process is repeated at the end of the cycle for every controlled intersection, regardless of what is happening to the rest of the network. The controller only requires real-time queue information of the adjacent intersections and respective turning ratios and the algorithm is executed once per cycle.

Developing a critical node selection framework for Max-Pressure control

While there has been significant research interest in finding more efficient or less infrastructure-dependent versions of MP controller as part of network-scale signal control systems, little attention has been directed towards defining the optimal number, relative location and traffic characteristics of the intersections that are included in the MP control scheme. In most case studies in literature, either all eligible intersections or only those across important arterial roads are controlled. However, given the high requirements in monitoring equipment that increase proportionately to the number of controlled intersections, it is interesting to investigate how the impact of the control scheme is affected in the following cases: if only a fraction of the eligible intersections are controlled; whether some nodes are more critical than others in the sense of MP control for the same fraction; and what are the characteristics that would allow us to identify them. In an effort to reply to these research questions, we develop a node selection methodology based on current network traffic characteristics, by using principles of traffic engineering combined with an optimization approach. We test several different schemes of MP control, where different fractions of eligible nodes are included in the MP control node set. The controlled nodes are chosen both by the proposed method and randomly, for comparison reasons, and are tested in independent MP schemes as well as combined with specific PC strategies. In this section, the proposed node selection process for MP nodes is described.

By analyzing the mechanism of the MP controller, we can infer that the process of green time re-assignment among competing phases/approaches is intuitively more beneficial in cases where queues of the competing approaches differ significantly from each other during the day, and thus, there are important pressure differences that can be balanced by the controller. This makes more sense if we think of an intersection where all approaches constantly have similar queues. In such case, pressures of phases would be relatively constant in time and therefore the controller would assign almost unchanged green time to all phases, similar to what the fixed-time control plan would do. In such cases, MP benefits are negligible. Thus, variance of queues of all approaches of an intersection, both incoming and outgoing, is one variable of interest in assessing node criticality. Furthermore, the mean normalized queue of the same approaches during peak period, or in other words the amount of traffic that the node serves with respect to its capacity, are also good indicators, since the impact of a poor or efficient signal plan is intensified if controlled node serves vehicles close to capacity. A node that gets congested in the FTC scenario, in the sense that some or all of its approaches reach jam densities and spill-backs occur upstream, might see higher improvement with MP control, which would affect higher number of trips, compared to a moderately congested node, where spill-backs do not occur. The charging level of each node can be estimated by the average queue over capacity ratio of all node approaches. Finally, the duration of the spill-back occurrence also seems important. For instance, a node might 'see' very high queues for only a short time, which can lead to a medium time-mean average approach queue and the same can happen for a node with medium to high queues that persist for a longer time period. In the former, spill-backs can occur for a short time while the node is relatively empty the rest of the time, while in the latter, spill-backs may not occur but short delays are affecting more drivers for longer time, therefore MP controller may create higher benefit for this node. To differentiate between the two node cases, we can count, for every node, the time during which at least one of the node approaches reaches jam density and spill-backs occur. Through this thinking process, we identify three quantities that should be taken into account in order to accurately assess the overall node significance for MP control.

Therefore, we define a set of three node assessment criteria, which we linearly combine into a kind of node (dis)utility function, whose coefficients can be determined through a suitable calibration process. In this work, a simple grid-search was performed, which is described in the following subsection. Given traffic information of the current network situation (e.g. FTC), a peak-period P is defined, based on the observed network state, as a set of time steps T_P. The selection process can be described, step by step, as follows:

- For every node n that is eligible to receive MP controller, the following three quantities are estimated: The first, denoted as m_1^n, represents the average node congestion level, as the mean over time of the mean occupancy (queue normalized over link capacity) of all incoming links $z \in I_n$ of node n, during peak period P. It is equal to

  $$m_1^n = \frac{1}{\|T_P\|}\frac{1}{\|I_n\|}\sum_{i\in T_P}\sum_{z\in I_n}\frac{x_z(i)}{c_z} \tag{3.8}$$

  where i is the simulation time-step index, T_P is the set of time-step indices corresponding to the peak period P and $\|T_P\|$ is the size of set T_P. The second, denoted as m_2^n, represents the mean over time of the variance of link occupancy of all incoming links $z \in I_n$ of node n during peak-period P, computed by

  $$m_2^n = \frac{1}{\|T_P\|}\sum_{i\in T_P}\mathrm{var}(X_z^n(i)) \tag{3.9}$$

  where $X_z^n(i) = \{x_z(i)/c_z | \forall z \in I_n\}$ is the set of normalized queues of all incoming links z of node n at time step i. The third quantity, denoted as N_c^n, represents the fraction of the peak period P, during which node n is considered 'congested'. In this analysis, we assume that a node is 'congested' during control cycle k if the average queue of at least one incoming link $z \in I_n$ of node n during k is higher than a preset threshold percentage p of its storage capacity, as shown by binary function $C_n(k)$ below.

  $$C_n(k) = \begin{cases} 1, & \text{if} \quad \frac{1}{t_c^n}\sum_{i=(k-1)t_c^n+1}^{kt_c^n} x_z(i) \geq p\, c_z, \quad \text{for any } z \in I_n \\ 0, & \text{else} \end{cases} \tag{3.10}$$

  $$N_c^n = \frac{t_c^n}{\|T_P\|}\sum_{\forall k\in P} C_n(k), \tag{3.11}$$

  In equation 3.10, t_c^n denotes the control cycle size in number of simulation time-steps, i.e. $t_c^n = C_n/t$, where C_n denotes the control cycle duration of node n and t the simulation time-step duration. In equation 3.11, the ratio $\|T_P\|/t_c^n$ is the number of control cycles that constitute peak period P. In other words, N_c^n represents (in the scale of 0 to 1) the fraction of the peak period P during which, at least one incoming link is congested and causes queue spill-back. In the current analysis, we set $p = 80\%$, since this is shown to significantly increase the probability for spill-back occurrence (see Geroliminis and Skabardonis, 2011).

- Then, the level of importance of each node n regarding MP control is estimated as a linear combination of the the above variables, denoted as R^n, as follows:

$$R^n = \alpha m_1^n + \beta m_2^n + \gamma N_c^n \tag{3.12}$$

  Quantity R^n is then used as a base to rank nodes and drive the selection of the most critical ones. The coefficients in equation 3.12 act as weights for the importance of every criterion and their values can be calibrated based on a trial and evaluation grid test, as described below.

- Finally, based on a target network penetration rate for MP control (i.e. percentage of eligible network nodes to receive MP controller), nodes are selected in sequence of increasing R^n, until the target number is reached.

An important step of the above process is finding proper values for the parameters α, β, γ of classification function R (equation 3.12), which serve as weights of node variables m_1, m_2 and N_c that indicate node importance. Since the relative importance of these variables is not straightforward, a trial-and-evaluation test is performed. More specifically, enumeration upon a grid of values of α, β, γ with subsequent MP simulation is performed. The combination of values leading to minimum total travel time for the same node penetration rate is then found and selected for all experiments.

3.2.2 Perimeter Control

Proportional-Integral regulator for MFD-based gating

The concept of gating in perimeter control strategies for single- or multi-region systems consists of controlling vehicle inflows in the perimeter or the boundaries of the protected regions, in order to prevent vehicle accumulation to rise excessively in their interior and lead to congestion phenomena, such as lower speeds, delays and gridlocks. Flow control can be applied by means of real-time adaptive traffic signals on the perimeter or the boundaries between regions, where green time of the respective approaches is periodically adjusted, based on a control law that takes into account the actual traffic state of the region. State information can be provided in real-time by loop detectors installed properly in the interior of the region, or by other types of traffic measuring equipment.

The concept of MFD enables the development of reliable feedback control strategies that assess the network state based only on measurements of vehicle accumulation in the system, which is associated with a specific travel production/service rate. Driven by the characteristics of the MFD curve, PC strategies can manipulate perimeter inflows during peak-hours so that vehicular accumulation in high-demand regions is maintained close to critical - for which travel production is maximum - and does not reach higher values belonging to the congested regime of the MFD. When accumulation tends to increase above this level, the allowed perimeter inflow is reduced by means of decreased green time for the approaches on the perimeter leading to the interior of the region. For large scale heterogeneously congested networks, proper clustering into homogeneous regions) that demonstrate a low-scatter MFD is required.

Several approaches for MFD-based perimeter control have been proposed and successfully tested, for single and multi-region systems, and different types of control laws are employed. In many studies, model-predictive control (MPC) schemes are employed for PC implementation, where dynamic aggregated traffic models are used to predict upcoming traffic states, based on which, an optimal control problem is solved for the prediction horizon and control variables are defined. However, in this work we consider no traffic states prediction, as we follow a simpler, less computationally expensive approach, where the system is controlled in real-time through a classical multi-variable Proportional-Integral (PI) feedback regulator (see Kouvelas et al., 2017), as follows:

$$\mathbf{u}(k_c) = \mathbf{u}(k_c - 1) - \mathbf{K}_P \left[\mathbf{n}(k_c) - \mathbf{n}(k_c - 1)\right] - \mathbf{K}_I \left[\mathbf{n}(k_c) - \hat{\mathbf{n}}\right] \tag{3.13}$$

In the above, $\mathbf{u}(k_c)$ denotes the vector of control variables u_{ij} for control interval k_c, which, in this work, represent the average green times corresponding to the controlled approaches between adjacent regions i and j (heading from i to j), as well as to the external perimeter approaches of every region (if external gates exist), denoted as u_{ii}; $\mathbf{n}$ is the state vector of aggregated regional accumulations n_i; $\hat{\mathbf{n}}$ is the vector of regional accumulation set-points $\hat{n_i}$; and $\mathbf{K}_P$, $\mathbf{K}_I$ are the proportional and integral gain matrices, respectively. If equation 3.13 is written in analytical form instead of matrix form, a system of equations will be produced. Every equation specifies the average green time, for the next control interval, for all nodes in specific direction between pairs of adjacent regions (e.g. u_{ij} and u_{ji} for adjacent regions $i, j \in \mathcal{N}$), while the last $\|\mathcal{N}\|$ equations refer to the average green time of all external approaches of each region. The control goal is to maintain accumulation close to critical in all controlled regions, in case of excessive demand, and impede reaching the congested regime of the MFD, where travel production drops significantly. Therefore, the set-points are decided based on the regional MFDs. The number of controlled regions and the respective MFD shapes depend on the network partitioning to a set of homogeneously congested regions $\mathcal{N}$, which can involve real or simulated traffic data. In order to be functional, the PI regulator requires as inputs the real-time regional accumulations, the set-points, as well as the proportional and integral gain matrices. Regional accumulations are supposed to be provided by loop detectors or other measuring equipment, properly distributed in the network. In this work we assume perfect knowledge of regional accumulations, that are averaged over the control interval.

The PI controller is activated at the end of every control interval and only when real-time regional accumulations are within specific intervals, in the proximity of the specified set-point, i.e. activated when $n_i \geq n_{i,\text{start}}$ and deactivated when $n_i \leq n_{i,\text{stop}}$, for $i \in \mathcal{N}$ and typically $n_{i,\text{stop}} < n_{i,\text{start}}$. This is important as early activation of the PI regulator (i.e. for low accumulation) can lead to signal settings aiming at increasing congestion in the controlled areas, so that production gets closer to critical, which is the target. However, such a policy can accelerate congestion and compromise the system performance. From equation 3.13, the average green time $u_{ij}(k_c)$ of all controlled approaches on the border between regions i and j with direction from i to j is calculated. Based on this average value, the exact green time for every specific intersection is calculated according to the process described in the following subsection. After deactivation of the controller, the FTC signal plan for all PC intersections is gradually restored.

Benefits of PC are more obvious in cases of high travel demand with highly directional flows towards the protected regions, that are usually areas of increased activity levels (e.g. city center). In such cases, the congestion-prone regions are kept close to capacity while queues are forming in the adjacent regions, due to PC. Usually, the adjacent regions are considered less probable to get highly congested and the PC-related queues do not cause significant performance degradation. However, multi-reservoir systems with multiple sources of congestion are more challenging to control with a multi-region PC scheme, as it may be the case that several neighboring regions can get congested at the same time and flow regulation between them is not straightforward. In such cases, the system optimal performance might come from a PC scheme where one region is ‘sacrificed’ and gets congested so that another more critical one is protected. In other words, finding the right values for the proportional and internal gain matrices is not trivial and depends on the actual traffic distribution patterns. In this work, since the focus is on the combination of PC and MP schemes, we calibrate $\mathbf{K}_P$ and $\mathbf{K}_I$ intuitively, based on a trial-and-error simulation process, in order to achieve some minimum performance improvement.

Green time calculation for PC intersections

After average green time u_{ij} for all PC controlled approaches between adjacent regions i and j (from i to j) is defined from equation 3.13, it is used as base to define the exact new green duration for the respective phases containing the approaches leading from region i to j. However, since not all intersections serve the same demand, green time assignment is more efficient if decision takes into account current queue lengths of the respective approaches. In other words, assignment should aim at providing longer green to approaches having larger queues, while ascertaining that the mean green of all approaches is as close as possible to the value defined by the PI controller, in an effort to balance queues in the boundaries between regions. Moreover, new greens are subject to a set of constraints, similar to the ones imposed in the case of MP signal update process, i.e. maximum allowed change between consecutive cycles, minimum and maximum green phase duration, constant cycle duration and integer green integrals.

Hence, for every u_{ij} that is specified by the PI controller, an optimization problem is solved for determining the exact green duration of the primary and secondary phases, denoted as p and s, respectively, of all controlled intersections with direction from i to j. Primary phase p includes approaches with direction from i to j and secondary phase(s) s include approaches of the same intersection but in the vertical (parallel to the boundary between regions) or the opposite direction (from j to i), depending on the traffic characteristics of the nodes. The sum of available green of primary and secondary phase remains constant, in other words, the green time that is removed from the primary phase goes to the secondary phase(s), while duration of any other phases of the node remains unchanged. Assuming that the set of controlled nodes in the direction from i to j is denoted as $\mathcal{M}_{ij}$, m is the node index, $G_{m,p}$ is the final green of the primary phase, $G_{m,s}$ is the final green of the secondary phase, $G_{m,t}$ is the sum of available green time for primary and secondary phases, $Q_{m,p}$ and $Q_{m,s}$ are the sum of the observed queues of all incoming links belonging to the primary and secondary phases during the last control cycle, respectively, and $S_{m,p}$ and $S_{m,s}$ denote the sum of saturation flows of all approaches of node m belonging to primary phase p and secondary phase s, respectively, the following optimization problem is formulated:

$$
\begin{aligned}
\underset{G_{m,p},G_{m,s}}{\text{minimize}} \quad & \theta_1 \left(\sum_{m\in\mathcal{M}_{ij}} G_{m,p} - u_{ij}\|\mathcal{M}_{ij}\| \right)^2 + \\
& \theta_2 \sum_{r\in\{p,s\}} \sum_{m\in\mathcal{M}_{ij}} Q_{m,r} \left(1 - \frac{G_{m,r}S_{m,r}}{Q_{m,r}+1}\right)^2 \\
\text{subject to} \quad & \\
& G_{m,p} + G_{m,s} = G_{m,t} \\
& G_{m,r} \geq g_{m,r,\min}, \\
& |G_{m,r} - G^{pr}_{m,r}| \leq g^{R}_{m,r} \\
& G_{m,r} \in \mathbb{Z}^+ \\
& \forall r \in \{p,s\}, \forall m \in \mathcal{M}_{ij}
\end{aligned}
\tag{3.14}
$$

In the above optimization problem, we seek to minimize the sum of two terms, the importance of which is weighted by parameters θ_1 and θ_2, respectively. The first term aims at minimizing the difference between the finally assigned total green of the primary phase of all PC intersections $m \in \mathcal{M}_{ij}$ of the approach $i - j$ and the total green indicated by the controller for the same approach. The second term aims at achieving green time distribution proportionally to the observed queues of the primary and secondary phases of the controlled intersections. This is done by minimizing the sum of the differences between the outflow that will be achieved in the primary and secondary phases of the PC nodes based on the finally assigned green time and the queue that was observed in the respective phases during the last control cycle. In case where queues are zero during last cycle, this term also becomes zero for the respective intersection and queue balancing effect is not taken into account. The denominators of the fractions in the second term are the recorded queues of the last cycle increased by 1, for avoiding the division by zero in the cases where queues become zero.

The constraints of the problem are the following: the first one is about maintaining cycle duration, i.e. ensuring that the sum of primary and secondary phases remains constant; the second one dictates that minimum green $g^R_{m,r}$ is assigned to all phases, where r is an index that indicates the type of phase (primary p or secondary s); the third one ensures that maximum absolute change between new green $G_{m,r}$ and green of the previous control interval $G^{pr}_{m,r}$ is below the preset threshold of $g^R_{m,r}$; and the forth one dictates that green time intervals are integer. The new control plans take effect, for every intersection, after the end of their ongoing cycle.

Regarding the gating of the approaches at the external perimeter of the regions, in the cases where no signal plan exists on the entry node, the control variable u_{ii} is translated to adjustment of the saturation flow of the entry links, i.e. the maximum allowed inflow from external virtual queues to entry links is adjusted accordingly. The constraints of minimum saturation flow and maximum absolute change of saturation flow between consecutive control intervals are also applied. However, in the case of external PC, no optimization problem is solved, but control variables u_{ii} are adjusted accordingly, in order to satisfy the constraints. The same settings are applied to all nodes of the external perimeter of every region.

3.2.3 Overview of the two-layer controller

A detailed schematic depiction of the proposed two-layer control framework is shown in figure 3.2. Based on the most recent collection of traffic information during the last control cycle, (aggregated regional accumulations $n_i, i \in \mathcal{N}$, where i is the region index), queues of primary and secondary phases of PC nodes ($Q_{m,p}, Q_{m,s}, \forall m \in \mathcal{M}_{ij}, \forall \{i,j\} \in \mathcal{N}$ with i adjacent to j), and queues of upstream and downstream links of MP controlled nodes ($x_z, \forall z \in I_n \cup O_n$, for all n belonging to the set of MP controlled nodes), the controllers decide on the updated signal plans of the respective intersections, according to the processes described in the previous sections. Perimeter control is activated/deactivated under specific conditions (when regional accumulations are above/below predefined thresholds). When PC is activated, first a PI controller defines the average new green of the controlled nodes for every approach $i-j$ between adjacent regions, as well as for the external perimeter of each region. Then this average green time is distributed to the specific controlled nodes proportionally to the recorded queues of the primary and secondary phases, in order

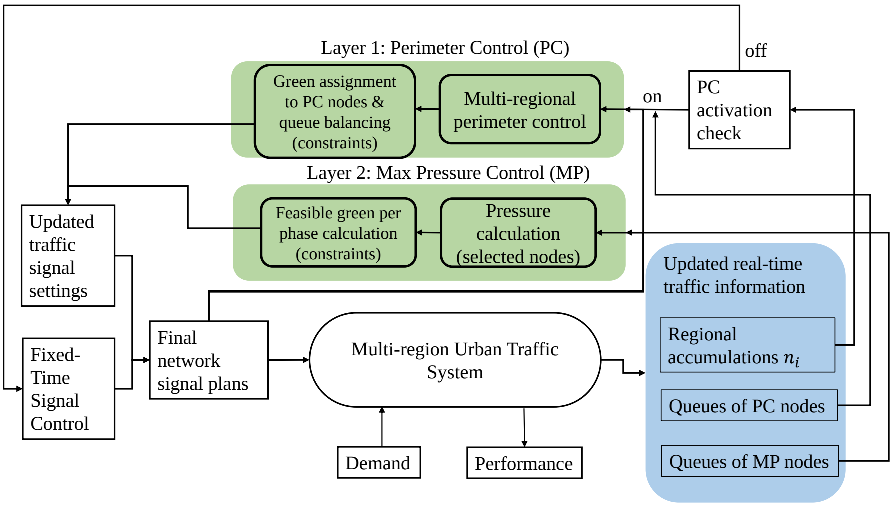

Layer 1: Perimeter Control (PC)
Green assignment to PC nodes & queue balancing (constraints)
Multi-regional perimeter control
off
on
PC activation check
Layer 2: Max Pressure Control (MP)
Feasible green per phase calculation (constraints)
Pressure calculation (selected nodes)
Updated traffic signal settings
Fixed-Time Signal Control
Final network signal plans
Multi-region Urban Traffic System
Demand
Performance
Updated real-time traffic information
Regional accumulations n_i
Queues of PC nodes
Queues of MP nodes

Figure 3.2: Detailed structure of the multi-layer controller combining perimeter control and Max Pressure control.

to prevent queues from growing excessively and impeding traffic upstream. In the second layer, Max Pressure calculates the pressure of each phase of every controlled node, based on the recorded queues upstream and downstream. Afterwards, the pressures are translated to new green time for every phase while satisfying a set of constraints. Finally, new signal settings are applied to all controlled intersections for the next cycle, before the process is repeated.

3.2.4 Traffic simulation details

Modified Store-and-Forward traffic model

The effectiveness of the various control scenarios is assessed through simulation experiments performed by a modified version of the mesoscopic, queue-based Store-and-Forward (SaF) paradigm (see Aboudolas et al., 2009), with enhanced structural properties derived by S-Model (see Lin et al., 2011). The exact version of the traffic model used in this work is described in details in Tsitsokas et al., 2021 and detailed presentation of its mathematical formulation is omitted here. However, in order to facilitate understanding of this work, a brief qualitative description of the model properties is provided in this section.

In accordance with the notation of MP mathematical description, the model monitors the state of the network, represented as a directed graph (N, Z), by updating the number of vehicles (or 'queue') inside every link $z \in Z$, denoted as $x_z(k)$, according to a time-discretized, flow conservation equation. Vehicle flow is generated in the network according to a dynamic Origin-Destination demand matrix while routing of the flow is performed via turn ratios received as inputs.

Queues are updated at every time-step. Link inflows include the sum of transit flows coming from all upstream links plus the newly generated demand of the link. Outflows consist of the sum of transit flows towards all downstream links plus trip endings inside the link. Backwards propagation of congestion is properly modeled, by replacing the initial assumption of point-queues in SaF by utilizing a more accurate link outflow representation. More specifically, the model assumes zero transfer flow for next time-step if the receiving link is already congested (receiving queue at the current time step is close to capacity). Therefore, in case of gridlock downstream, upstream queue is growing and spill-backs can occur, since no outflow is recorded. Therefore congestion propagation is properly captured by the traffic model. The regional accumulations n_i that appear as elements of vector $\mathbf{n}$ in equation 3.13, in real applications is measured through appropriate instrumentation of the network. However, in this work where perfect knowledge of real-time traffic is assumed, $n_i(k)$ is simply the sum of link accumulations $x_z(k)$ of all links assigned to region i at time step k, i.e. $n_i(k) = \sum_{z \in i} x_z(k), i \in \mathcal{N}$.

The inherent inaccuracy in estimating travel time and real queue length of links, owed to the dimensionless nature of the initial SaF version has been decreased by integrating the structure of S-Model (Lin et al., 2011), where every link queue x_z consists of two distinct queues, m_z and w_z, representing vehicle flows moving and waiting at the intersection at the end of each link, respectively. Transit flow vehicles that enter a new link firstly join the moving part of the link, where they are assumed to move with free flow speed. They transfer to the queuing part after a number of time steps, which is determined by the position of the queue tail at the moment when they entered the link. Moreover, traffic signal settings are properly taken into account. A binary function dictates whether an approach of an intersection gets green light at every time-step, according to the most recent signal plan. If an approach gets red light at any time step, this function will set the current outflow of this approach to zero for this time-step, regardless of the state of the queues upstream or downstream.

By iteratively updating all state variables according to the mathematical relations of the model for a number of time-steps, we get a complete traffic simulation for the specific demand scenario. By knowing the queues of every link of the network at every discrete time-step (state variables $x_z \forall z \in Z$), we can calculate the total travel time of all vehicles as

$$VHT = \sum_{k=1}^{K} \sum_{z \in Z} x_z(k)T + \sum_{k=1}^{K} \sum_{z \in Z} x_{VQ,z}(k)T \tag{3.15}$$

where the first term represents the total time spent inside the network and the second represents the time spent waiting in virtual queues, i.e. upstream network links that serve as origins. In equation 3.15, $x_{VQ,z}(k)$ denotes the virtual queue of link z, T denotes the time-step duration and KT is the total simulation time.

Dynamic turn ratio update

The modified version of SaF model that is used as traffic simulator in this work, requires knowledge of turn ratios between every pair of incoming-outgoing links for all network intersections. These values determine how traffic is distributed in the different parts of network, according to the considered origin-destination scenario, and drive the generation of congestion pockets, since they express aggregated driver decisions regarding route selection. It has been shown that adaptive routing decreases the scatter and the hysteresis of MFD as it avoids the development of local gridlocks (Daganzo et al., 2011, Mahmassani et al., 2013). In order to realistically model traffic distribution in the scope of assessing different control schemes, in this work, turn ratios are estimated and dynamically updated in regular intervals during simulation, on the basis of time-wise shortest path calculation. This is done in order to account for potential rerouting effects, since in reality drivers might react to the enforced control policies by adjusting their route in real time, based on the observed traffic conditions that can be associated with the applied control policy. This process assumes that drivers know or can accurately estimate the average speed of all roads that are included in their alternative paths, and thus select the fastest one among them. Once the final time-wise shortest path is determined for every origin-destination pair of the considered demand matrix, the amount of traffic transferring between all feasible connections of incoming-outgoing links of all network nodes, for a specific future time-horizon, can be counted and turn ratios can be estimated. The process is repeated in regular intervals and turn ratios are recalculated, while accounting for the evolving traffic conditions. Although an equilibrium analysis for path assignment would be a more appropriate approach, it would considerably increase modeling complexity and simulation computational cost, while exceeding the scope of this chapter, therefore the simpler shortest path approach was adopted.

The process of turn ratio recalculation is presented in Algorithm 6. Firstly, for a time window of calculation T_w, preceding the process, mean speed is estimated for every network link $z \in Z$ based on link outflow and queue values, according to the following relation:

$$v_z(t) = \min\left(v_{\text{ff}}, \frac{\sum_{j\in T_w(t-1)} u_z(j)L_z}{\sum_{j\in T_w(t-1)} x_z(j)}\right) \geq v_{\min} \tag{3.16}$$

where t is the time interval index for the turn ratio calculation, v_{ff} is the free flow speed, $T_w(t)$ is the set of simulation time step indices corresponding to time interval t, L_z is the length of link z, and $u_z(j)$ and $x_z(j)$ denote the outflow and accumulation of link z at simulation time step j, respectively. Since our objective here is to estimate the time to cross a link, a lower threshold $v_{\min}$ is imposed for the cases of gridlocks, for computational reasons, so that no link have infinite traversing time. When mean link speed is calculated for all links, algorithm 1 is called to estimate the new turn rates. Inputs include the origin-destination demand matrix, the simulated inflows to origin links during the previous calculation interval, the ongoing trips from previous calculation interval (in the format of origin-destination demand), as well as link length and mean speed vectors.

Firstly, if there are ongoing trips, they are added to the origin-destination demand matrix and ongoing trip stack is initialized. Traversing times for all links are calculated from speed and link lengths. Afterwards, for every origin-destination pair of the demand matrix, the shortest path is calculated by using traversing time as link cost function. If the path's estimated trip duration is longer than the calculation interval T_w, the trip is split and a new trip is added to the stack of the ongoing trips, where as origin link is set the link where path is split while destination is the original trip destination. The traffic volume of this path as well as the time already spent in the split link are stored in the stack as well. For the path that is estimated to be traversed entirely in the next interval, connection counters between consecutive links of the shortest path are updated with the respective vehicle volume corresponding to this trip. Trip volume is estimated by the simulated traffic volume that entered the network at the origin links (outflow of virtual queues) during the previous interval T_w, which is split to different destinations in proportion with the demand of the original origin-destination matrix. The use of simulated inflow instead of the nominal demand improves the accuracy of the turn ratio calculations by taking into consideration potential delays in trip starting, for example due to external perimeter control or spill-back of queues up to entry links. When all trips of the matrix have been counted, turn ratios for all approaches between upstream and downstream links are estimated by dividing the counted volume of each approach by the total volume of each upstream link. In cases where no vehicles are assigned to an approach, all approaches receive the same turn ratio. At the end, the algorithm returns to the simulator the updated turn rates and a stack of ongoing trips. Dynamic turn ratios determined in this way can reflect possible path adjustments due to congestion-related low speeds. Therefore, the impact of the control schemes under evaluation in terms of traffic distribution in the network is directly taken into account.

Congestion-based exit rate adjustment

The original version of SaF model assumes that vehicle flow exits the network at destination links, which are considered as 'sinks', i.e. having unlimited storage capacity and maximum saturation flow (they never spill-back). Although this might be realistic for exit links at the periphery of the network, where we consider that traffic is in general of lower importance, it is rarely accurate in internal links of high-demand regions, such as city centers, where trips usually end either in on-street parking spots or in specific parking facilities, both of which have limited capacity. In peak-time, trip-ending delays are expected when reaching popular destinations, due to reduced service rate of parking facilities or the cruising for

Algorithm 6: Turn ratio dynamic adjustment for modified SaF model

```
Data: origin-destination pairs $\mathbf{OD}$, inflows of origin links $\mathbf{u}$, ongoing trips
      o-d stack $\mathbf{OD_s}$, ongoing trips flows stack $\mathbf{u_s}$, time window $T_w$, speed
      vector $\mathbf{v}$, link length vector $\mathbf{L}$, previous turn ratios $\beta$
Result: updated turn ratios $\beta$, ongoing trip o-d pairs $\mathbf{OD_s}$ and flow $\mathbf{u_s}$
        stacks
Initialize connection counters: $c_n(z, w) = 0,\ \forall z \in I_n, w \in O_n, n \in N$
Merge ongoing trips stacks $\mathbf{OD_s}$,$\mathbf{u_s}$ with $\mathbf{OD}$,$\mathbf{u}$, respectively
Initialize stacks $\mathbf{OD_s} = 0, \mathbf{u_s} = 0$
Calculate link traversing time $\mathbf{t} = \mathbf{L} \div \mathbf{v}$
Create directed graph $G$ with link cost function $\mathbf{t}$
for all $(o, d) \in \mathbf{OD}$ do
    if $u(o, d) > 0$ then
        find shortest path $p$ from $o$ to $d$ in $G$
        if duration of $p > T_w$ then
            split $p$ up to link where duration is $< T_w$
            store remaining trip in o-d stack $\mathbf{OD_s}$ with the path link after
              the split as $o$ and same $d$ and volume stack $\mathbf{u_s}$
        end
        for every pair $z$ - $w$ of consecutive links in $p$ do
            update respective connection counter:
              $c_n(z, w) = c_n(z, w) + u(o, d)$
        end
    end
end
for all $n \in N$ do
    for all $\{(z, w) | z \in I_n, w \in O_n\}$ do
        if $\sum_{j \in O_n} c_n(z, j) > 0$ then
            Calculate turn ratios: $\beta(z, w) = \frac{c_n(z,w)}{\sum_{j \in O_n} c_n(z,j)}$
        else
            keep same turn ratios or set them all equal (if no previous value)
        end
    end
end
return $\beta, \mathbf{OD_s}, \mathbf{u_s}$
```

parking, which can deteriorate traffic conditions in the high-demand regions of the network. However, this effect is not taken into account by the original version of SaF model, and as a result, highly congested conditions are less accurately represented in simulation. In order to address this modeling weakness, for internal destination links located in the high-demand regions i (e.g. city center), saturation flow of the set of exit (destination) links E_i is considered a function of regional accumulation, according to the following relations:

$$S_z^r(k) = s_z(k) S_z = s_z(k) 1800 l_z \text{ (veh/h)}, \tag{3.17}$$

$$s_z(k) = \begin{cases} 1 & \text{if } n_i(k)/c_i \leq \theta \\ \max\left(s_{\min}, 1 - (\kappa_1 + \kappa_2 \ \frac{n_i(k)}{c_i})\right) & \text{else} \end{cases} \tag{3.18}$$

$$\forall z \in E_i \subseteq N$$

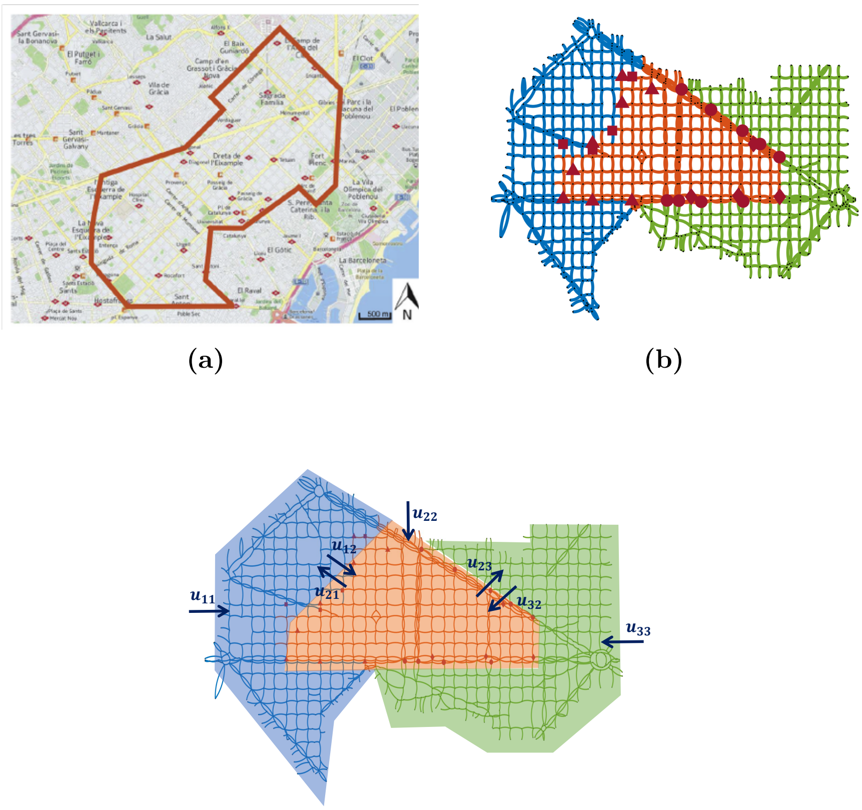
In equation 3.17, $S_z^r(k)$ is the congestion-adjusted maximum saturation flow of exit links $z \in E_i$ of region i at time step k, S_z is the nominal saturation flow of link z, which is proportional to the number of lanes l_z. In equation 3.18 the adjusting parameter $s_z(k)$ is equal to one, for lightly congested conditions where normalized regional accumulation is below threshold θ, while for higher congestion it is linearly reduced until it reaches a minimum threshold $s_{\min}$. The calibration of parameters $\theta, s_{\min}, \kappa_1, \kappa_2$ can be made based on real or microsimulation data, so that congestion is modeled in the most accurate way. In this work, we set $\theta = 0.25$, $s_{\min} = 0.1$, $\kappa_1 = 0.2$, $\kappa_2 = 1$.

3.3 Implementation to a large-scale network via simulation

The proposed adaptive signal control schemes are evaluated using the mesoscopic urban traffic model and settings discussed in subsection 3.2.4, which was coded from scratch and executed in Matlab R2020a, while optimization problems 3.7 and 3.14 are solved by Gurobi 9.1.2 solver called from Matlab via Yalmip toolbox Löfberg, 2004. Real-life large-scale signalized traffic network of Barcelona city center is used as case study and Fixed-Time Control (FTC) settings with no adaptive element, are used as benchmark case. The network model and FTC plans are provided by Aimsun. Both MP and PC schemes are applied separately, as well as in combination, for two different demand scenarios that create moderate and high levels of congestion in the FTC case, respectively. All MP cases are tested in full-network implementation and in node subsets for different penetration rates, selected by the proposed algorithm as well as randomly, for comparison. Detailed description of the network, the simulation settings and the performed experiments are provided in this section.

3.3.1 Case study

The traffic network utilized in this study is a replica of Barcelona city center, in Spain, as shown in Figure 3.3a. It consists of 1570 links and 933 nodes, out of which 565 represent signalized intersections with fixed-length signal control cycles ranging from 90 to 100 sec. Links have from 2 to 5 lanes. All control schemes are tested in two different demand scenarios, one leading to moderate and one to high congestion in FTC settings. For the purpose of PC implementation, the network is partitioned into three regions of similar traffic distribution, according to the clustering method described in Saeedmanesh and Geroliminis, 2016. The resulting regions are displayed in different colors in figure 3.3b, with region 1 shown in blue, region 2 in orange and region 3 in green. In the same figure, the controlled intersections used for gating through PC are displayed, with different annotation, based on the approach that they control: triangle for 1 to 2 approaches, circle for 3 to 2 approaches, square for 2 to 1 approaches and rhombus for 2 to 3 approaches. Figure 3.3c schematically represents the perimeter and boundary flow control variables u_{ij}, which denote the average green time of all approaches in the direction from i to j, while u_{ii} denotes the equivalent time for all external approaches of region i. All nodes on the external perimeter of all three regions are controlled. This partitioning leads to 4 state and 7 control variables, which are depicted in figure 3.3c. By running a simulation



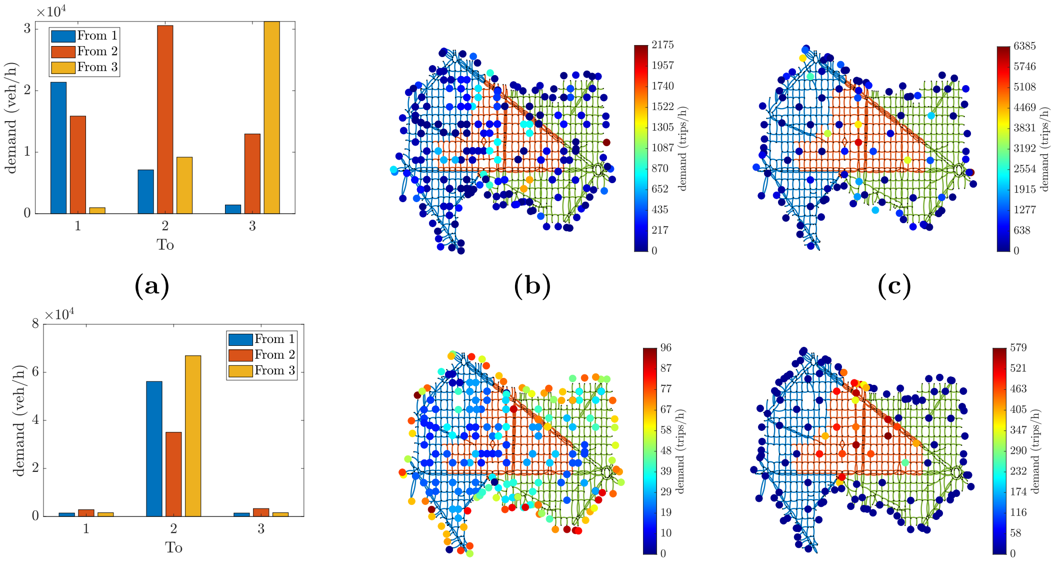
(a) **(b)**

(c)

Figure 3.3: (a) Map of the studied network of Barcelona city center; (b) Model of Barcelona network as a directed graph with annotation of nodes used for PC; (c) schematic representation of controlled approaches for perimeter and boundary flow control, with green time per approach as control variable.

experiment for the FTC case for both demands scenarios, we obtain MFD curves for each of the three regions, which are shown in figure 3.5.

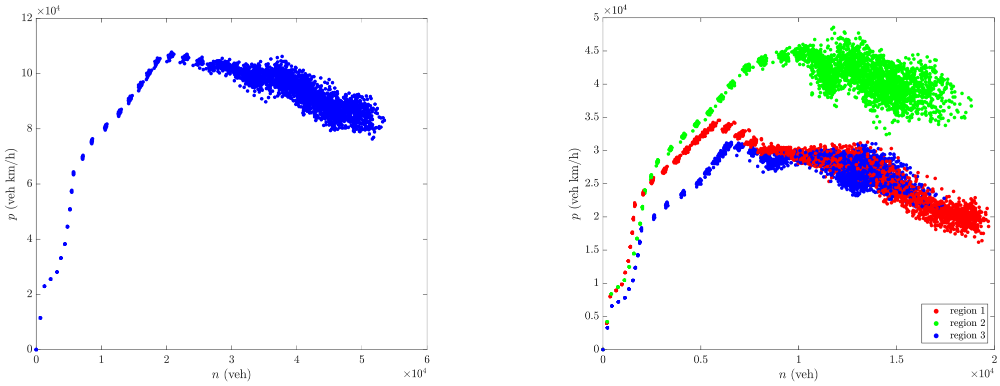
The dynamic profile of total generated demand, for both demand scenarios, consist of a 15-minute warm-up period followed by a 2-hour constant peak demand, for a total simulation period of 6 hours, which is representative of the morning-peak. Medium demand includes 251k trips generating at 88 origin links and heading towards 104 destination links, whereas high demand scenario includes 316k trips, from 123 origins to 130 destinations. Figure 3.4 graphically describes the spatial distribution of demand for both scenarios: (a) to (c) refer to medium demand



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 3.4: Description of the two demand scenarios in peak period: medium (a)-(c) and high (d)-(f). (a) and (d): aggregated trip distribution between regions; (b) and (e): trip origin density; (c) and (f): trip destination density.



(a) **(b)**

Figure 3.5: Production MFDs for the on-load of congestion in the case of FTC: (a) MFD of entire network; (b) MFDs per region after partitioning.

while (d) to (f) refer to high demand. The distribution of trips between regions, resulting from suitable clustering process as described below, is presented in the first graph of each row (a and d). Each bar corresponds to the total demand originating from each origin region, each represented by different color, while horizontal axis indicates destination region. Second graph of each row (b and c) presents the spatial distribution and density of origin points, represented by dots on the network map. Dot color indicates demand volume per origin point, as described by the respective colorbar. Similarly, third graph of each row (c and f) represents the same information for destination points. Apart from difference in total number of trips, the two demand scenarios lead to different traffic patterns, since high demand scenario has a clear directional profile, with trips mainly originating from the periphery (regions 1 and 3) and heading towards the center (region 2), whereas medium demand shows more diverse trip distribution between regions, with more intra-regional trips in all regions, plus a less intense directional pattern towards the city center (region 2). The objective of using different demand scenarios is to test the efficiency of the proposed control schemes under different traffic conditions. It should be noted, that no demand information is required by the controllers, which only receive real-time queue measurements and/or turn ratios as inputs. Accuracy and configuration of traffic measurements is not taken into account in this work and all necessary information to controllers is provided directly from the simulator (perfect knowledge assumed).

3.3.2 Control scenarios

The case of fixed-time, static signal control, labeled as 'FTC', is used as benchmark for all tested control schemes. The network performance improvement owed to traffic responsive control schemes based on single MP, single PC and on their combination in a hierarchical framework is calculated with reference to FTC, which is consider as the current network state. Firstly, MP is evaluated as single control strategy (no PC applied simultaneously). With the aim of investigating the performance of MP in relation to the number and location of the controlled nodes, we test the following scenarios:

- MP control of all eligible network nodes. All signalized nodes receive MP controller.
- MP control of fraction of network nodes, selected randomly. For each penetration rate of $5\%, 10\%, 15\%, 20\%, 25\%$, 10 randomly created MP node sets are evaluated through simulation.

- MP control of fraction of network nodes, selected by the proposed algorithm. For the same penetration rates as above, MP node sets are created according to decreasing values of R, after suitable parameter calibration. FTC simulation results are used for calculating variables m_1^n, m_2^n and N_c^n, and thus quantity R.

Afterwards, MFD-based PC based on the PI controller described in subsection 3.2.2, is applied first as a single control scheme and then in combination with distributed MP control, integrated in a two-layer framework. Similar to the case of single MP, the MP layer of the combined scheme is tested for several controlled node layouts, in various penetration rates, as well as in full network implementation (100 % eligible nodes). The following scenarios are evaluated:

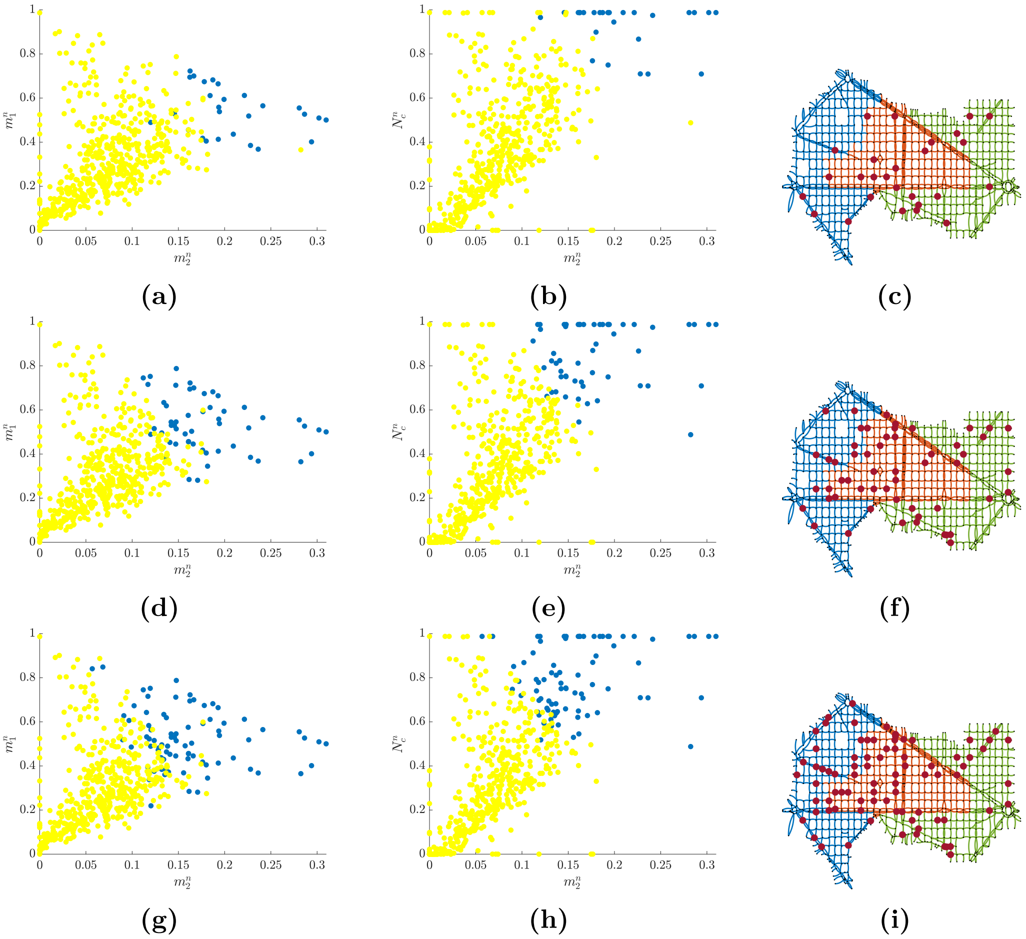
- Single PC for 3-region system
- PC for 3-region system combined with MP control to all eligible network nodes
- PC for 3-region system combined with MP control to fractions of eligible network nodes, selected randomly. For each penetration rate of 5%, 10%, 15%, 20%, 25%, 10 sets of randomly selected eligible nodes are formed (same as in single MP scenarios). MP control is evaluated in parallel with PC scheme.
- PC for 3-region system combined with MP control to fractions of eligible network nodes, selected by the proposed algorithm. For the same penetration rates as above, MP node sets are created according to decreasing values of R, after suitable parameter calibration. FTC simulation results are used for calculating variables m_1^n, m_2^n and N_c^n, and thus quantity R.

3.3.3 Experiment settings

The control scenarios under evaluation are simulated for a 6-hour time period for medium demand, and for a 8-hour period for high demand, representative of typical morning peak. High demand required longer simulation time to unload and completely empty the network, due to high number of trips. Simulation time step is set to 1 second. Free-flow speed of vehicles in the moving part of links, v_{ff} is set to 25 km/h and average vehicle length for capacity calculations is set to 5 meters. The time window for turn ratio update is 15 minutes.

MP regulator is active for all controlled intersections during the entire simulation time and signal plans are updated at the end of every cycle, based on traffic information collected during the last cycle. Only phases lasting longer than 7 seconds in the pre-timed scheduling are eligible for change by the regulator and minimum green time allowed per phase $g_{n,j,\min}$ is also 7 seconds. Maximum allowed fluctuation of green time between consecutive cycles, $g^R_{n,j}$ is set to 5 seconds, for all MP modified phases, to avoid instabilities (similar to Kouvelas et al., 2017). Inputs for MP regulator are link queues of incoming and outgoing links, as well as estimated turn ratios for all approaches of controlled nodes. For our experiments, this information is provided directed by the simulator, thus assuming that the controller receives perfect real-time traffic information.

Regarding PC implementation and based on the network partitioning in 3 regions described above, the PI controller of equation 3.13 regulates 7 control variables u_{ij}, which represent the average green time of all controlled intersections in the approaches between adjacent regions 1-2, 3-2, 2-1, 3-1, as well as those of the external perimeter of each regions 1, 2 and 3 (see figure 3.3c), using 3 state variables, i.e. the regional accumulations $n_i, i = 1, 2, 3$. Therefore, $\mathbf{u}$ is a 7x1 vector, $\mathbf{n}$ and $\hat{\mathbf{n}}$ are 3x1 vectors, while proportional and internal gain matrices $\mathbf{K}_P$ and $\mathbf{K}_I$ are of dimensions 7x3. Four first rows refer to boundary approaches in the order they are listed above and three last rows refer to external perimeter approaches of regions 1 to 3. Similar to MP, for inter-regional approaches, minimum green time per phase $g_{m,r,\min}$ is set to 7 sec and maximum allowed absolute change between consecutive cycles is $g^R_{m,r}$ is set to 5 seconds. For the external perimeter gating, where no signal plan is available and controller regulates saturation flow of entry links, the constraint of allowing at least 15% of the real saturation flow holds. Due to different directional patterns of the two demand scenarios, PI parameters differ slightly. For the medium demand scenario, setpoint accumulation for the three regions are $\hat{n}_1 = 10000, \hat{n}_2 = 12000, \hat{n}_3 = 6800$ veh, proportional gain matrix is $\mathbf{K}_P = [15, -10, 0; 0, -5, 10; -15, 10, 0; 0, 5, -10; -20, 0, 0; 0, -20, 0; 0, 0, -20]$ and integral gain matrix is $\mathbf{K}_I = \mathbf{K}_P \times 10$, $n_{i,\text{start}} = \hat{n}_i$, $n_{i,\text{stop}} = 0.85\,\hat{n}_i$, for every region $i = 1, 2, 3$. For the high demand scenario, accumulation setpoint is $\hat{n}_1 = 4500, \hat{n}_2 = 9200, \hat{n}_3 = 8000$ veh, proportional gain matrix is $\mathbf{K}_P = [18.5, -2.1, 0; 0, -3.3, 6.8; -13.3, 5.6, 0; 0, 4.6, -3.5; -16.4, 0, 0; 0, -9.8, 0; 0, 0, -10.5]$ and integral gain matrix is $\mathbf{K}_I = [18, -69, 0; 0, -69, 62; -44, 24, 0; 0, 1, -40; -54, 0, 0; 0, -30, 0; 0, 0, -51]$, $n_{i,start} = 0.99\,\hat{n}_i$ and $n_{i,stop} = 0.93\,\hat{n}_i$, for every region $i = 1, 2, 3$. In all cases, activation of the PI controller happens when at $n_i \geq n_{i,\text{start}}$ for at least 2 regions $i = 1, 2, 3$, while deactivation happens when $n_i < n_{i,\text{stop}}$ for all 3 regions. PC application requires inputs of aggregated regional accumulations n_i for all regions $i = 1, 2, 3$ for the PI controller, while for the phase of applicable green time calculation (optimization problem 3.14), queue measurements for all approaches of primary and secondary phases of PC intersections, $Q_{m,p}$ and $Q_{m,s}$ respectively, are required, together with the latest applied signal plan. After performing grid search optimization for parameters θ_1 and θ_2, we found that the best performing values are $\theta_1 = 0.4$ and $\theta_2 = 0.9$, which we used in all cases. Similar to MP case, all required real-time traffic information are considered given and accurate



(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) **(h)** **(i)**

Figure 3.6: Visualization of the MP node selection process for the case of medium demand, according to the proposed method. Each row refers to different penetration rate (5%, 10% and 15%, from top to bottom). First and second column graphs show relations between selection variables m_2 - m_1 and m_2 - N_c, respectively, for all network nodes, for the benchmark case of FTC. Blue dots represent the selected nodes for MP control. Third column figures show the plan of the studied network, partitioned in 3 regions, with the spatial distribution of the selected MP intersections shown as red dots.

and for the scope of this work are collected directly from the simulator. The process is repeated every 90 seconds.

3.4 Results

3.4.1 Single Max Pressure

Results of simulation experiments concerning single Max Pressure schemes are presented in this section. The case of full-network MP implementation, i.e. MP regulator assigned to all eligible nodes, is compared to the FTC case (no adaptive control) as well as to different scenarios of partial MP implementation, in different fractions of network nodes, selected both randomly and by the methodology

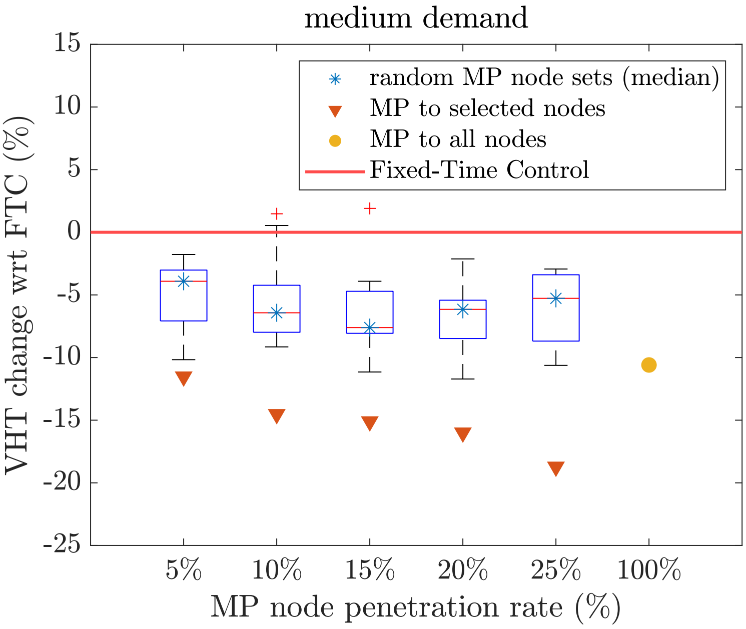


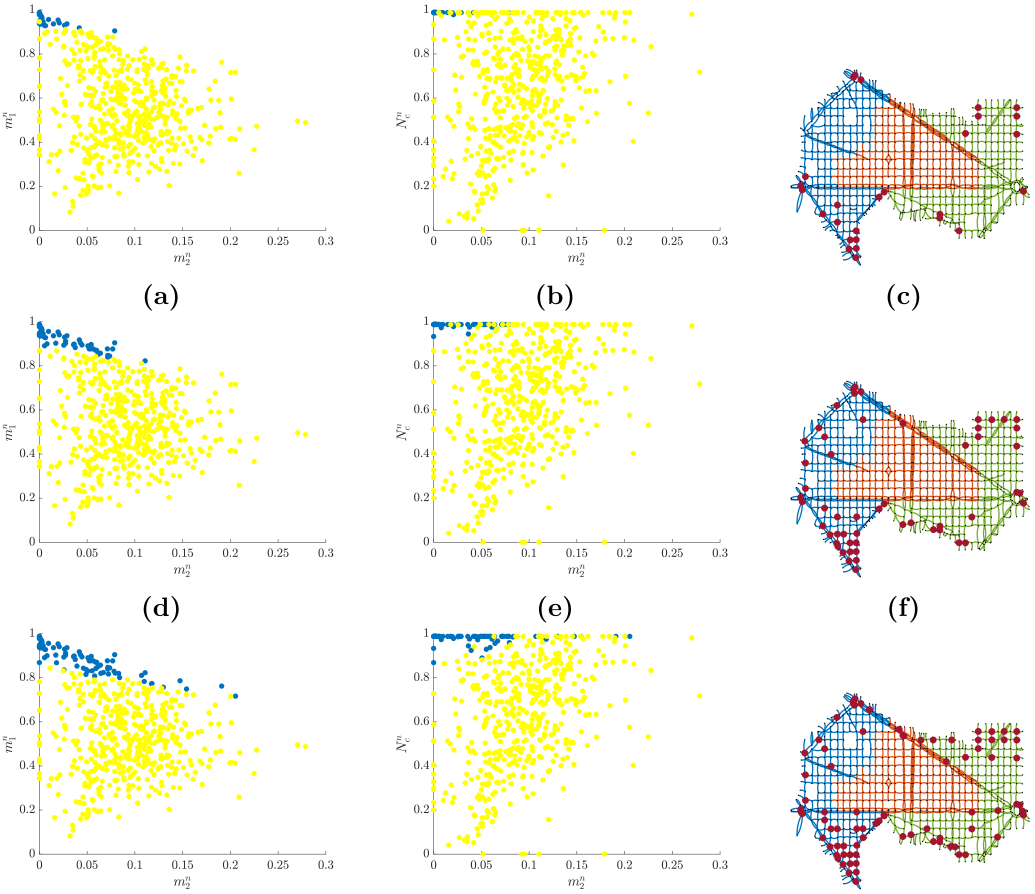
Figure 3.7: Comparison of Total Travel Time improvement, with respect to FTC scenario, of single Max Pressure application, for medium demand scenario. Boxplots refer to 10 randomly created MP node sets for every penetration rate, red triangles refer to node selection based on the proposed method and yellow dot represents the full-network implementation.

described in section 3.2.1. For the process of node selection, the considered peak-period is 2-hour long (starting from 0.5 and ending at 2.5 h) and consists of $T_P = 80$ control cycles of 90 seconds. FTC case is simulated and results are used for the calculation of m_1, m_2 and N_c. After performing a trial-evaluation test as described in section 3.2.1 for the medium demand scenario, the best performing values are $\alpha = 0.6$, $\beta = -1.8$ and $\gamma = -1$, and selection is done starting from nodes with lowest R. In this way, the algorithm prioritizes selection of nodes with relatively high queue length variance and spill-back occurrence during peak time but with moderate mean queue lengths. In figure 3.6 the node selection process for the case of medium demand is pictured for penetration rates 5%, 10% and 15%, where dots represent all network signalized nodes. First column graphs (a, d, and g) show the relation between m_1 and m_2, while second column graphs (b, e and h) show the relation between N_c and m_2, all calculated based on simulation results of FTC case. Blue dots represent nodes that are selected to receive MP controller according to the proposed algorithm. Third column graphs (c, f and i) visualize the spatial distribution of the selected MP nodes, depicted as red dots.

The performance of single MP network control for the medium demand scenario, for different fractions of MP controlled nodes, as well as for the standard full network implementation (penetration rate of 100%), is shown in figure 3.7. On the vertical axis the percentile improvement of total travel time (vehicle-hours traveled or VHT) with respect to FTC scenario is shown. Each boxplot refers to 10 cases of random selection of MP controlled intersections, corresponding to the respective penetration rate. The case of 100% rate refers to full MP network implementation. Red triangles represent scenarios where controlled nodes are selected according to the method described in section 3.2.1. Firstly, we observe that for the case of medium demand, almost all MP scenarios lead to improved total travel time, even those with randomly selected MP nodes. However, most cases of node selection made by the proposed method significantly outperform random assignment. In fact, we notice that the higher the number of controlled nodes, the larger the difference between random and targeted selection performance. These observations indicate that the proposed selection process is successful in identifying critical intersections for MP control. Interestingly, the case of installing MP regulator to all network nodes leads to smaller improvement than those including only a fraction of controlled nodes according to the proposed algorithm. More specifically, with only 10% of critical nodes the system travel time improves by 14.5% and with 25% of critical nodes, it improves by 18.8%, while in the case of controlling all nodes, it improves only by 10.6%. This remark indicates not only significant cost reduction can be achieved by reducing the number of controlled nodes through the proposed selection process, but also system performance can be increased. However, this behavior is observed for moderate demand, where the network does not reach highly congested states in the FTC case.

A different behavior is observed in the case of high travel demand, both in the node selection pattern and in the corresponding network performance. Firstly, we observed that using the same values for parameters α, β, γ as in moderate demand case and following the same selection pattern leads to poor system performance, very similar to the one of random selection. Therefore, parameter optimization is done again, by performing a new trial-evaluation test, specifically for the high demand scenario. The new values found are $\alpha = -0.72$, $\beta = -0.4$, $\gamma = -0.2$. In this case, we observe a change of sign in α, which directs the selection towards nodes with high mean queues (m_1) while high queue variance (m_2) and spill-back duration (N_v) are given lower weights. In other words, in this case, the best found selection pattern prioritizes nodes with high mean queues. However, as we can see in figure 3.8 which is made based on FTC results of the high demand scenario, there is correlation between m_1, m_2 and N_c. For instance, in the high demand scenario, a significant number of nodes experience complete gridlock (m_1 values close to 1), which corresponds to zero queue variance (m_2) and maximum gridlock duration (N_c equal or close to 1). As figure 3.8 shows, the selection process in this case prioritizes highly congested/gridlocked nodes for MP selection and, as we will see, leads to significant performance gains. Node selection is visualized for penetration rates of 5%, 10% and 15%, in the same format as in figure 3.6 for medium demand.

Regarding system performance of single MP in high demand scenario, for which the network reaches more congested states in the FTC case, results show relatively



(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) **(h)** **(i)**

Figure 3.8: Visualization of the MP node selection process for the case of high demand, according to the proposed method. Each row refers to different penetration rate (5%, 10% and 15%, from top to bottom). First and second column graphs show relations between selection variables m_2 - m_1 and m_2 - N_c, respectively, for all network nodes, for the benchmark case of FTC. Blue dots represent the selected nodes for MP control. Third column figures show the plan of the studied network, partitioned in 3 regions, with the spatial distribution of the selected MP intersections.

smaller improvement with respect to FTC than in the case of medium demand. This can be seen in figure 3.9, where percentile performance improvement with respect to FTC is shown on the vertical axis, for different MP node penetration rates. Again, boxplots refer to 10 cases of randomly selected node sets for MP control per penetration rate. Firstly, it is interesting that case of full-network MP control (yellow dot) leads to practically zero improvement compared to FTC. However, smaller MP penetration rates in best case result in improvement between 3% and 7%. The effectiveness of node selection method seems to drop, even with re-optimized parameters of function R, since we observe a few random sets performing better than the ones of the proposed method. Performance of targeted selection is always better than the median of the random set though, and in the case of

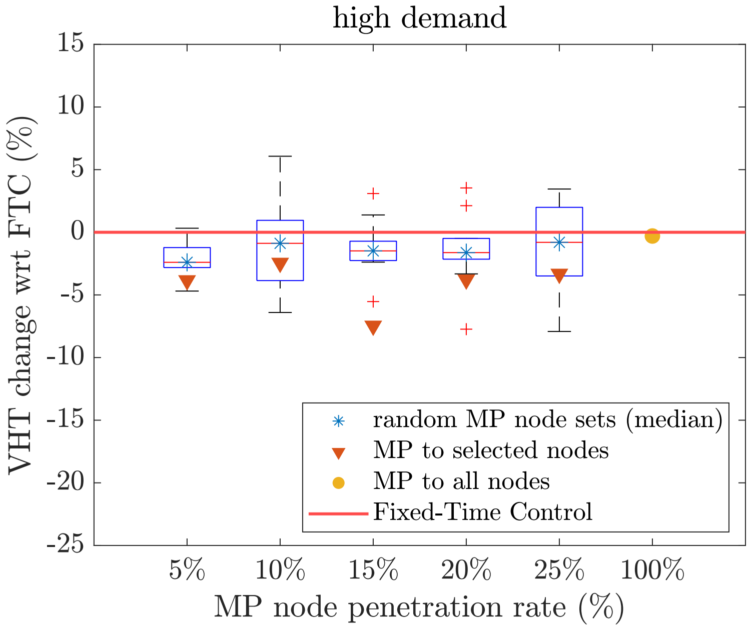


Figure 3.9: Comparison of Total Travel Time improvement, with respect to FTC scenario, for single Max Pressure application for high demand scenario. Boxplots refer to 10 randomly created MP node sets for every penetration rate, red triangles refer to node selection based on the proposed method and yellow dot represents the full-network implementation.

15%, for which parameter optimization was performed, performance is significantly better than random cases. Based on these remarks, we can infer that partial MP implementation, except for being less costly, can also lead to improved performance, especially in highly congested networks with single MP control, which implies that significant spatial correlation exist between performance of MP controlled nodes that can act detrimentally to system performance. In other words, even if performance generally improves in the proximity of a controlled intersection, adding MP controllers to all network intersections does not necessarily mean that system performance will globally improve as a result. Hence, optimizing spatial node distribution for MP control requires further investigation.

In the scope of investigating in detail the effectiveness of partial MP control in targeted node sets, for both demand scenarios, we plot cumulative distribution functions of selection variable m_1, m_2 and N_c changes, after implementing MP control, with respect to their values in FTC, for three MP node sets selected by the proposed method (15%, 20% and 25%), for medium demand case in figure 3.10 and for high demand case in figure 3.11. Blue lines correspond to MP node sets while red lines to the remaining nodes, where FTC applies. First, second and third column figures show changes in values of m_1, m_2 and N_c, respectively, where m_1^*, m_2^* and N_c^* refer to the FTC case, before MP control application. In medium demand, it is evident from second column graphs of figure 3.10 that queue variance m_2 was

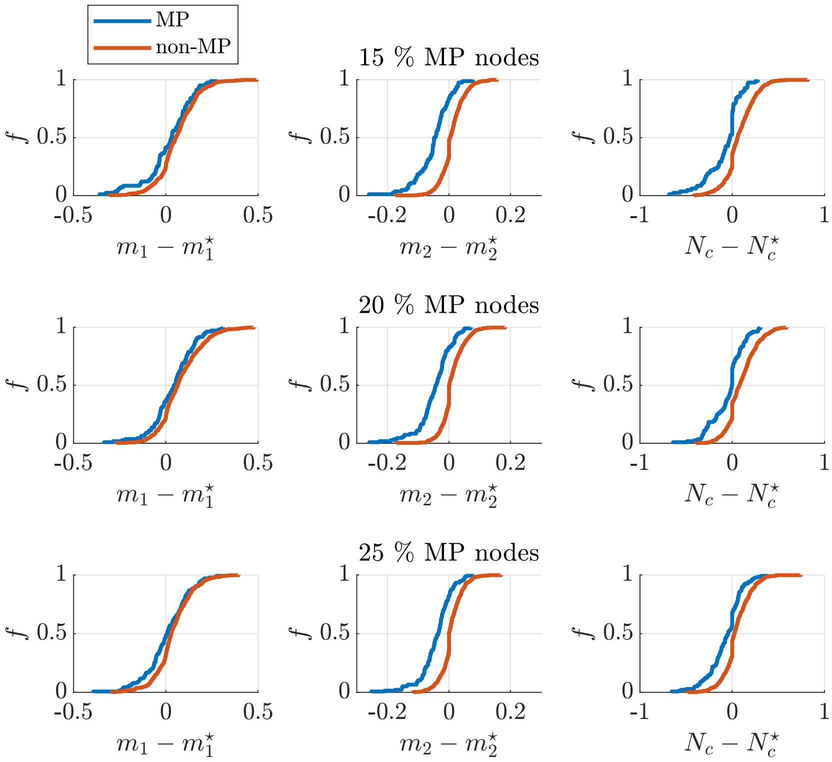


Figure 3.10: Cumulative distribution functions for the changes in node selection criteria m_1, m_2, N_c after MP control implementation for medium demand scenario. Start superscript refers to the FTC case. Blue lines correspond to MP nodes and red lines to the rest. Each row represents a different MP node penetration rate (15%, 20% and 25%).

decreased for 80% to 90% of nodes that received MP controller according to the proposed algorithm, for penetration rates of 15% to 25%, while for the remaining nodes, it was reduces for half and increased for the other half. Also, reduction to the non-MP nodes was smaller than the one to MP nodes. Moreover, spill-back duration during peak time n_c was also decreased in the majority of selected MP nodes in all three cases shown, though mean queue length was mostly increased. The increase in mean queue length in cases of both MP and non-MP nodes is justified, as a result of more efficient road space use that is achieved due to MP control, leading to more vehicles entering the network at the same time and increasing node occupancy. Different pattern of improvement is observed in the case of high demand as shown in figure 3.11, where we observe a significant decrease in spill-back occurrence N_c in almost all MP nodes (blue curve in third column graphs), and a decrease in mean queue length m_1, while no significant difference can be seen between MP and non-MP nodes in terms of queue variance m_2. This behavior can be explained with reference to the node selection pattern of the high demand case, where MP is assigned by priority to capacitated nodes with very high queues, where by definition variance is

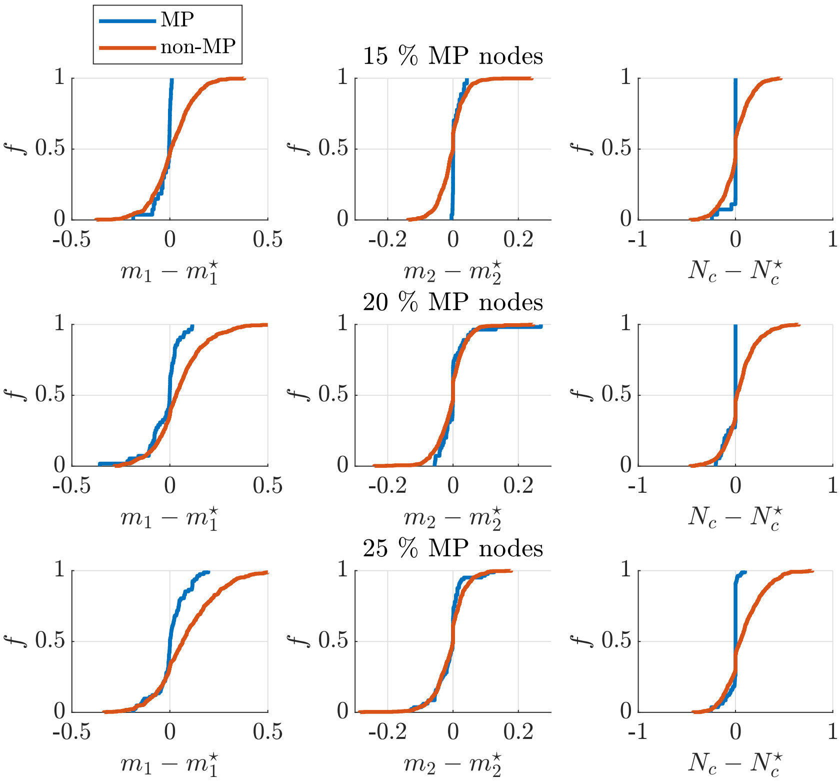
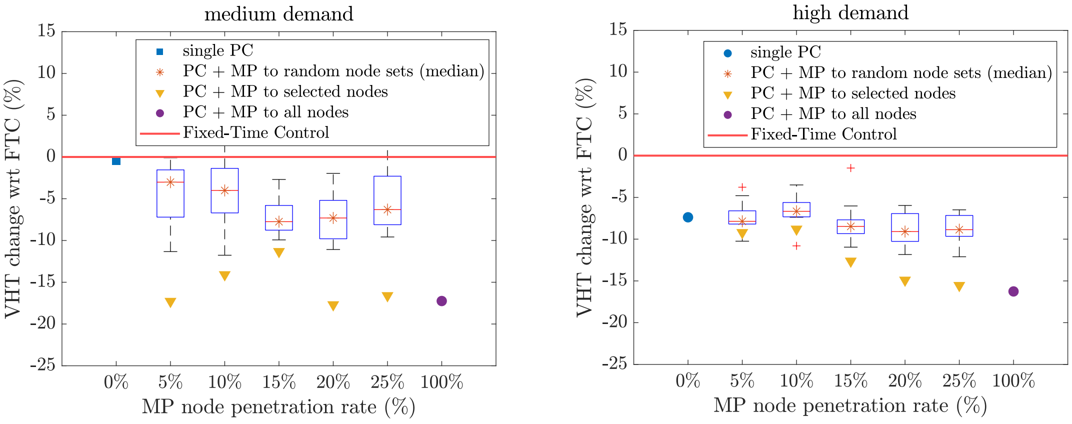


Figure 3.11: Cumulative distribution functions for the changes in node selection criteria m_1, m_2, N_c after MP control implementation for high demand scenario. Star superscript refers to the FTC case. Blue lines correspond to to MP nodes and red lines to the rest. Each row represents a different MP node penetration rate (15%, 20% and 25%).

very low and spill-back occurrence very high. Therefore, in both cases selected MP nodes are improved with respect to either spill-back occurrence and either queue variance or queue length, in moderate and high congestion scenarios, respectively.

3.4.2 Two-layer framework: Perimeter control combined with Max Pressure

Results of the two-layer framework combining PC with MP schemes are discussed in this section. Simulation results in terms of performance improvement with respect to FTC case are shown in figure 3.12, where (a) refers to medium and (b) to high demand. The case of 0% penetration rate (square) corresponds to typical PC application without any MP control (for comparison), while the remaining cases refer to combined control of PC and MP in different node penetration rates. Again, boxplots aggregate results of 10 scenarios of random MP node selection of the corresponding penetration rate, combined with the same PC scheme. Triangles refer to the combined scheme where MP nodes are selected according to the proposed methodology, while the case of 100% (dot) refers to the combined scheme with



(a) **(b)**

Figure 3.12: Comparison of Total Travel Time improvement, with respect to FTC scenario, for single PC and combined PC plus MP implementation, for (a) medium, and (b) high demand. Graphs show the performance for different MP node penetration rates (0% is single PC) for the two-layer PC+MP framework. Boxplots refer to 10 random MP node selections per rate while yellow triangles refer to node selection based on the proposed method.

MP installed in all nodes. Selection of MP nodes is based on FTC results and is done with the same parameter values α, β, γ as in single MP, for every demand scenario respectively, while PI controller parameters are the same for all cases of the same demand scenario.

For medium demand, we observe that single PC does not lead to considerable improvement with respect to FTC (only 0.5%). This is not surprising, since there is not enough demand to drive the network to heavily congested states, where PC would get activated for longer and would have a higher impact, and production MFD does not drop significantly for FTC case, as we will see also in figure 3.13 below. However, in all combined scheme cases, we observe significant travel time improvement with respect to the FTC case. Especially in the cases of 5% and 20% MP nodes selected by the proposed method, performance is slightly higher than the best performing single MP scheme of the respective penetration rate. Moreover, similar to the single MP cases, we observe that the proposed node selection algorithm leads to higher performance gains compared to random selection for most cases. The highest improvement for the two-layer controller, which is about 17.7% , is recorded for the case of 20% MP nodes, while both 100% and 5% penetration rates achieve about 17.2%. Therefore, in multiple cases, as in 5%, 20% and 100%, adding PC on top of MP increases network performance from 3% to 7% with respect to single MP.

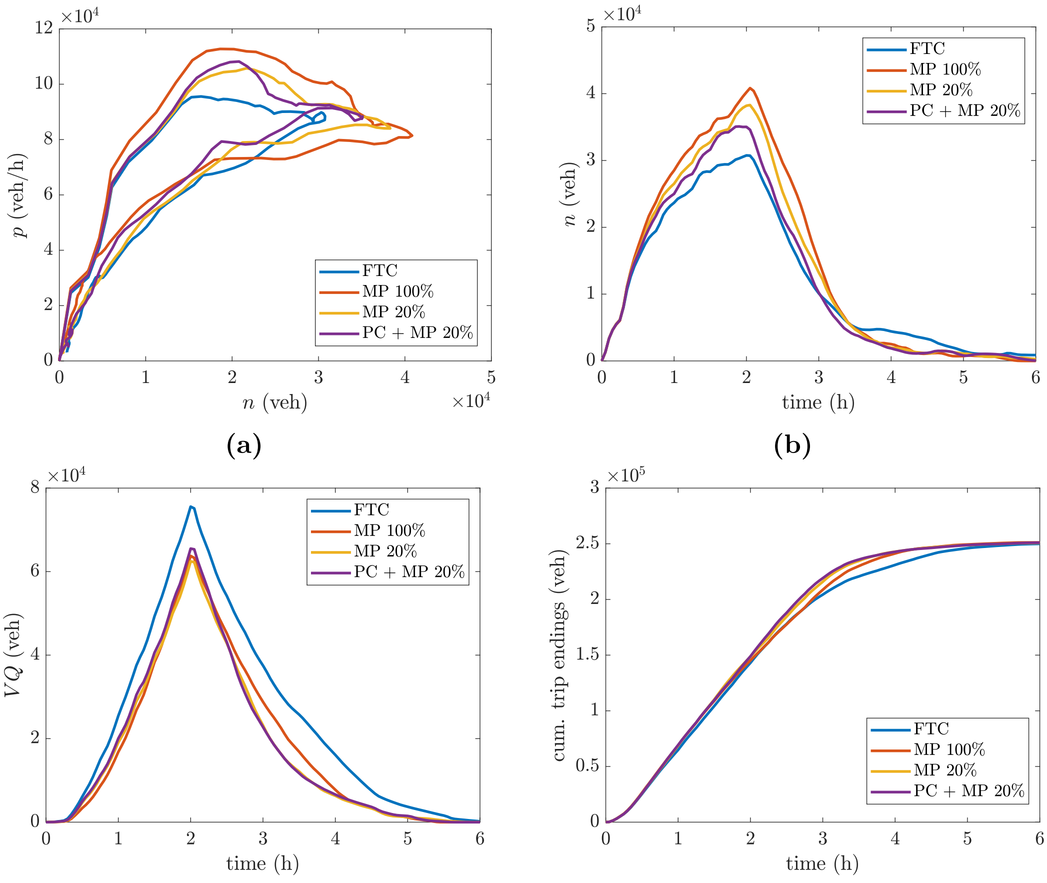
For high demand, results of the combined scheme are more promising than single MP, as we can see in figure figure 3.12(b). While single MP application, as well as single PC, only improve traffic performance by around 8% compared to FTC in the best case, the two-layer framework manages to improve up to 15%, in the case of

Table 3.1: Total travel time for all control scenarios, for medium and high demand, in vehicle hours traveled (VHT). Under ΔVHT, the percentile change of VHT with respect to FTC case is shown. For random MP node selection, the median VHT of 10 random node set replications is reported.

	Medium demand				High demand			
MP node selection	Targeted		Random		Targeted		Random	
Control scenario	VHT	ΔVHT(%)	VHT	ΔVHT(%)	VHT	ΔVHT(%)	VHT	ΔVHT(%)
FTC	221400	-	-	-	484000	-	-	-
MP 5%	195580	-11.7	212743	-3.9	464390	-4.1	472394	-2.4
MP 10%	189250	-14.5	207173	-6.4	471680	-2.5	479732	-0.9
MP 15%	187860	-15.1	204558	-7.6	447360	-7.6	476813	-1.5
MP 20%	185940	-16.0	207777	-6.2	465280	-3.9	476145	-1.6
MP 25%	179850	-18.8	209726	-5.3	467480	-3.4	480113	-0.8
MP 100%	197960	-10.6	-	-	483000	-0.2	-	-
PC	220380	-0.5	-	-	447000	-7.6	-	-
PC + MP 5%	183000	-17.3	214725	-3.0	438540	-9.4	445948	-7.9
PC + MP 10%	190070	-14.2	212544	-4.0	441060	-8.9	451741	-6.7
PC + MP 15%	196190	-11.4	204242	-7.8	422550	-12.7	442860	-8.5
PC + MP 20%	182230	-17.7	205216	-7.3	411510	-15.0	440004	-9.1
PC + MP 25%	184650	-16.6	207452	-6.3	408430	-15.6	441069	-8.9
PC + MP 100%	183220	-17.2	-	-	405300	-16.3	-	-

25% MP nodes selected by the proposed algorithm, and up to 17% in the case of full-network MP implementation. Therefore, in high demand scenario, PC strategy can be significantly enhanced by the additional distributed MP layer, even with only a fraction of properly selected, controlled MP nodes, which leads to performance very similar to the one of 100% controlled nodes, but with only 25% of the respective cost. Detailed performance of all evaluated controlled schemes can be found in Table 3.1.

Figure 3.13 depicts simulation results of the benchmark case of FTC, single MP case for all eligible nodes (‘MP 100%’), single MP controlling 20% of nodes selected according to the proposed method (‘MP 20%’), and the combined PC with MP to 20% of selected nodes (‘PC+MP 20%’), all for the case of medium demand. In 3.13(a) MFDs of accumulation versus production are shown for the four cases. We notice that all scenarios involving MP significantly increase the maximum production, compared to the FTC case, and therefore, increase both critical and maximum observed vehicle accumulation. This remark indicates that MP strategy can increase system serving capacity in conditions of moderate congestion, and by balancing queues around controlled nodes, it leads to better road space utilization. As a result it allows a higher number of vehicles to be in the system at the same time, which was not possible in FTC due to local gridlocks that were forcing excess demand to stay in virtual queues. This is evident in (b), where total network accumulation of all scenarios is higher than in FTC, as well as in (c), where total virtual queues are remarkably lower. Moreover, by comparing ‘MP 20%’ and ‘PC+MP 20%’ MFD curves in (a), we see that the latter leads to slightly lower maximum accumulation and, thus, smaller capacity drop and hysteresis loop in the unloading part. In this case, combined PC+MP performs slightly better than MP



(a) **(b)**

(c) **(d)**

Figure 3.13: Simulation results for medium demand scenario. Comparison between FTC, MP to all network nodes (100%), MP to only 20% nodes selected by the proposed method, and combined PC with MP to 20% selected nodes. Figures refer to the entire network: (a) MFD of accumulation vs. production; (b) time-series of accumulation; (c) time-series of total virtual queue; (d) time-series of cumulative trip endings.

by approximately 2%. However, the opposite is observed in the respective cases of 25% MP nodes, which is probably due to traffic correlation among additional MP controlled nodes. We should note here that some change of MFD curve in presence of MP is to be expected, especially with respect to critical accumulation, and this should be considered in the process of parameter tuning for PC. Another interesting remark is that, between the two single MP scenarios, ‘MP 100%’ results in higher increase in system serving capacity, but on the other side, it introduces more vehicles in the network and thus, reaches higher congestion levels and capacity drop in peak time than FTC. Or, production rises significantly, but also drops during peak, causing some delays and heterogeneity non-existent in FTC, as shown by the unloading part of the MFD curves. This effect can explain why MP installed in all nodes performs worse than partial installation to 20% of nodes, and is closely related to the shape of MFD and how fast production drops when network enters

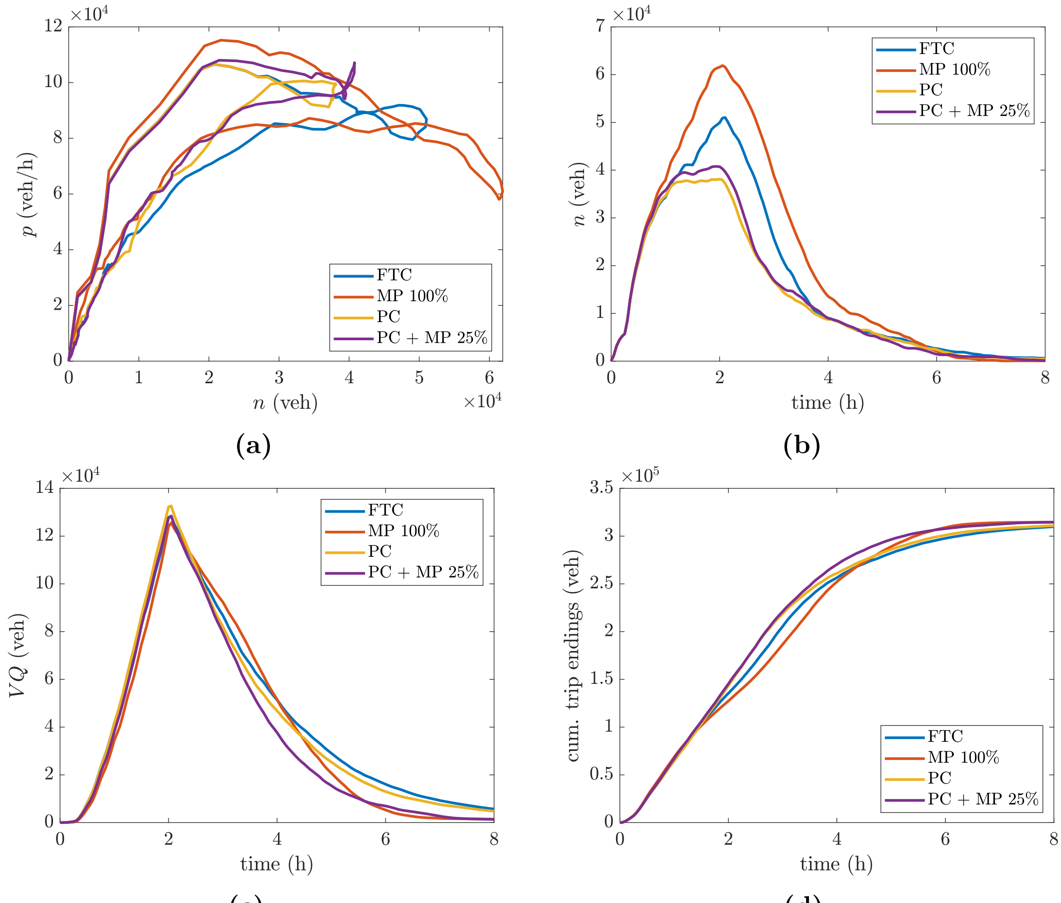
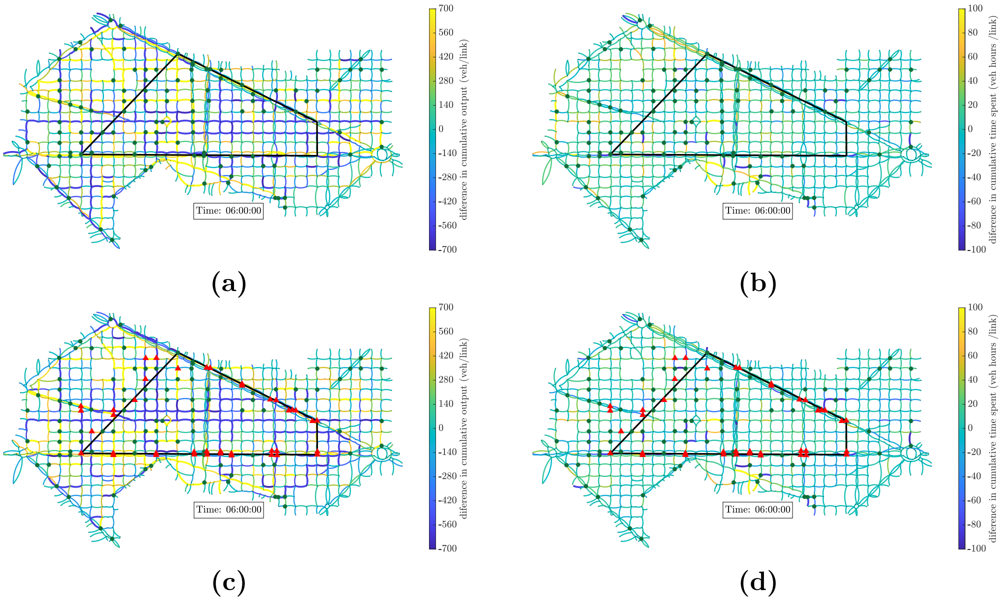


Figure 3.14: Simulation results for high demand scenario. Comparison between FTC, single PC, single MP in 100% of nodes and combined PC with MP in 25% of nodes selected by the proposed method. Figures refer to the entire network: (a) MFD of accumulation vs. production; (b) time-series of accumulation; (c) time-series of total virtual queue; (d) time-series of cumulative trip endings

the congested regime. However, in this case, minimum recorded production is not much lower than the one in FTC, while capacity increase is significant, thus, despite the importance of the production drop in MP 100% case, VHT savings compared to FTC are still significant. This effect is less intense in the 'MP 20%' and 'PC+MP 20%', where a slightly lower maximum production is recorded, but lower maximum congestion and capacity drop are observed as well. For medium demand, the highest delay savings are achieved for the case of PC + MP 25%.

Similarly, figure 3.14 shows simulation results of four, best-performing control scenarios, for the high demand scenario. FTC case is compared to the case of single MP with full-network control ('MP 100%'), the case of single PC, and the case of combined PC with distributed MP in subset of 25% of eligible nodes, selected according to the proposed method. Regarding single MP scheme, a behavior similar to medium demand case is also observed for the high demand, although in the


Time: 06:00:00
700
560
420
280
140
0
-140
-280
-420
-560
-700
difference in cumulative output (veh/link)
100
80
60
40
20
0
-20
-40
-60
-80
-100
diference in cumulative time spent (veh hours /link)

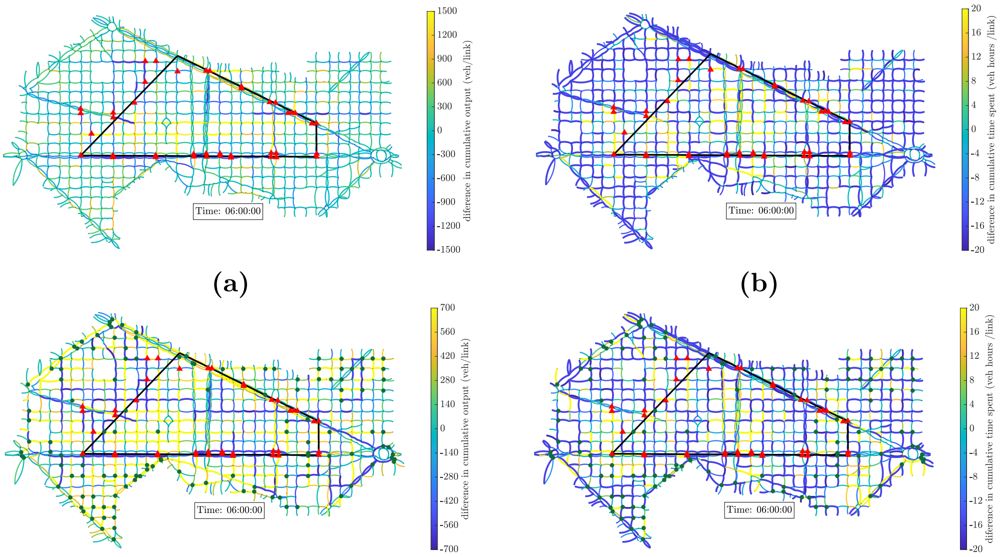
(a) **(b)** **(c)** **(d)**

Figure 3.15: Difference of cumulative link throughput and cumulative time spent with respect to FTC, in medium demand scenario, at the end of simulation (6:00 h): (a)-(b) Case MP in 20% critical nodes; (c)-(d) Case MP in 20% critical nodes combined with PC. Dots and triangles represent MP and PC node locations, respectively.

latter, capacity increase is relatively smaller compared to FTC (about 6.5%), while production drop in peak time is significantly higher (around 27%), which can be due to reaching more congested state by allowing more vehicles inside the network at the same time, as we can see in 3.14b. Overall, MP 100% case performs almost similar to FTC in terms of VHT but significant differences are observed between MFD curves. However, smaller hysteresis is recorder during network unloading in the case of MP 100%, thus reducing the damage made by the production drop. Interestingly this effect is eliminated in the case of the two-layer framework, where PC plays an important role in prohibiting the system from reaching highly congested states. Therefore, in the combined case of 'PC + MP 25%', the network reaches slightly higher production in peak period compared to single PC case, which drops with a smaller rate as accumulation increases above critical, due to MP control. Also, PC impedes the excessive increase of vehicle accumulation in the system and prevents highly hysteretic behavior owed to heterogeneity. Among the four cases shown, the combined framework leads to shorter total travel time, reduced by almost 15% with respect to FTC, when single PC achieves a decrease of around 7.5%. In short, adding a MP layer significantly improves single PC performance, while properly selected MP nodes allow for a smaller network penetration rate that leads to comparable performance as in full-network MP implementation.

Figure 3.15 depicts the impact of two well-performing control schemes, in terms of link throughput and link total time spent difference with respect to FTC scenario, for medium demand scenario, for the case of single MP in 20% of nodes selected by the proposed method (see a and b), and for the case of the combined control of PC and MP in 20% of nodes, selected in the same way (see c and d). All four graphs show the network map where each link's color corresponds to a range of values, as displayed on the bar on the right. Links appear slightly thicker when values approach or overpass bar limits. Graphs a and c depict the difference in link cumulative throughput with respect to FTC case, for each of the above control cases, respectively. Similarly, graphs b and d depict the difference in cumulative time spent per link, with respect to FTC case. Green dots denote MP controlled nodes while red triangles denote PC nodes. The black lines represent the approximate boundaries of central region 2. From graphs a and c, we observe a similar pattern of traffic redistribution, where intense blue links indicate significant cumulative throughput reduction in the case of responsive control compared to FTC, while yellow links indicate the opposite. We notice that cumulative throughput is reduced along several links that are also connected, but it is increased in the majority of network links, as warmer colors (green and yellow) indicate. However, around most MP nodes, we see at least one or several incoming or ongoing links, where throughput is overall higher, at least to one direction, which means that a higher number of vehicles left those links sooner than FTC case. This observation supports the idea that throughput is mostly increased around MP nodes. Nevertheless, a throughput decrease can also result from rerouting of vehicles due to change in congestion distribution around the network. Interestingly, we do not observe significant increase in cumulative total time spent in links where cumulative throughput is lower (by comparing left and right figures of each row), which indicates that throughput drop is related to rerouting rather than congestion increase. In fact, control scheme of single MP in 20% of critical nodes achieves 16% lower VHT than FTC, while combined PC + MP in 20% critical nodes achieves 17.7% lower VHT. This reduction is not particularly obvious in graphs b and d, since a large amount of this time gain comes from decreasing the time spent in virtual queues, which are not shown in these figures.

The same type of information is shown in figure 3.16, but for the cases of single PC (see a and b) and combined PC with MP in 25% of critical nodes selected by the proposed method, both for the high demand scenario. In a we observe significant increase in cumulative throughput along two main arterial roads in the interior of central region 2, which, for this demand scenario, attracts the majority of trips, while decrease is observed almost exclusively upstream or in the proximity of PC nodes (triangles), which is an expected effect of PC. However, throughput appears slightly higher in the majority of links also in peripheral regions, which is related to congestion drop, as shown in graph b, where most network links experience drop in cumulative time spent with respect to FTC case, with the exception of some links upstream PC nodes, which receive the queues that form due to gating. Performance appears even better for the combined scheme, as we see in c, where cumulative throughput is increased in most links (in green and yellow), and around



(a) **(b)**

(c) **(d)**

Figure 3.16: Difference of cumulative link throughput and cumulative time spent with respect to FTC in high demand scenario at the end of the simulation (6:00 h): (a)-(b) Case single PC; (c)-(d) Case PC combined with MP in 25% of selected nodes. Dots and triangles represent MP and PC node locations, respectively.

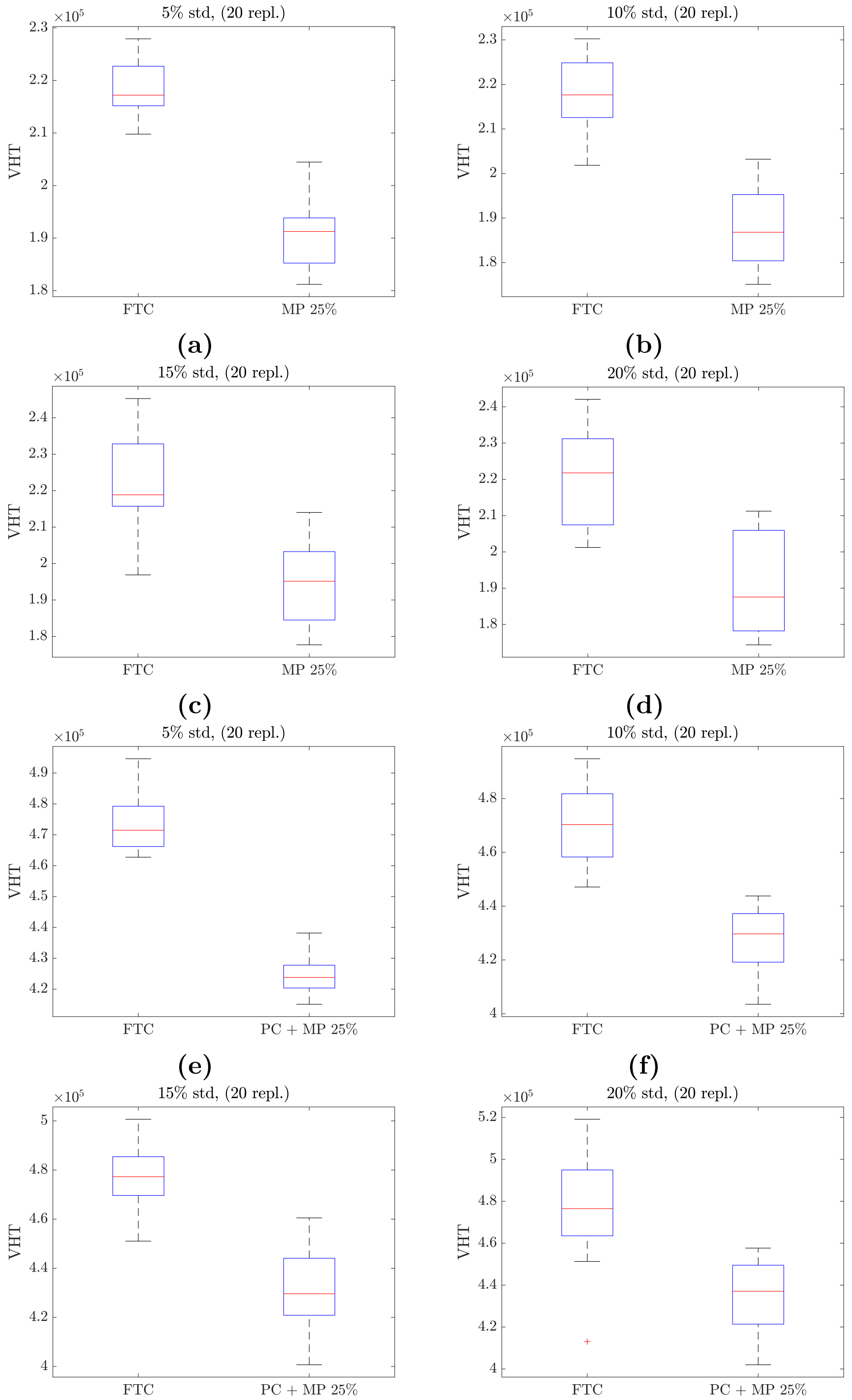
the majority of MP nodes (dots). Alternatively, the decrease in cumulative time spent is obvious in most network links, as we see from graph d, where increase is recorded mostly around the boundaries with region 2. The improvement in VHT compared to FTC case is 7.6% for single PC case (a and b), while for the combined case PC + MP 25% (c and d) reaches 15.6 %.

In order to examine the sensitivity of the proposed MP node selection process, in terms of performance, against fluctuations of spatial traffic distribution, a set of additional simulation experiments are performed. Specifically, for two well-performing cases of single MP and combined MP-PC schemes with MP nodes selected through the proposed method, simulation tests are performed, where instead of the mean demand value, for every origin-destination pair, a new value is considered generated by a stochastic process. More specifically, for the case of single MP with 25% controlled nodes selected by the proposed method, based on the results of FTC scenario of the medium demand, 20 new demand matrices are generated, by replacing mean demand of every origin-destination pair, by a random draw from normal distribution, with mean equal to the initial value and standard deviation that corresponds to 5, 10, 15 and 20% of the mean. The specific control scheme that was designed based on the mean demand, is tested for the 20 stochastically generated demand matrices for each of the aforementioned standard deviation cases, and performance is compared to the one of FTC case. Results are shown in figure 3.17 (a) to (d), where each graph refers to a different standard deviation level of demand. Boxplots represent the distribution of total recorder travel time of FTC and MP-25% scenarios for the same 20 modified demand matrices. We observe that MP-25% case on average outperforms FTC for all standard deviation levels (median of MP-25% is always less than median of FTC), although improvement decreases as deviation from mean increases. The same experiments are performed also for the combined MP-PC scheme with 25% MP nodes selected by the proposed method, for the high demand scenario and results are shown in figure 3.17 (e) to (h). Again, we observe that PC+MP 25% outperforms FTC for almost all standard deviation levels, but improvement decreases with the deviation from the mean. This result indicates that the node selection process that is based on mean OD values is not highly sensitive to demand fluctuations of up to 20% from the mean.

3.5 Summary

The chapter investigates the potential benefits of a two-layer signal control framework, combining centralized, aggregated perimeter control strategy, with partial and efficient distributed control, in critical intersections, based on the Max Pressure feedback controller. Cost-effective schemes of improved performance for network-wide MP control are produced by decreasing the number of controlled nodes, via selecting only nodes considered critical for MP implementation. A critical node identification method based on actual traffic characteristics is introduced and assessed. Simulation experiments performed with a modified version of Store-and-Forward model evaluate the effectiveness of PC and MP control schemes, both independently and combined in a two-layer framework. Spill-back effect and potential rerouting of drivers due to control-induced alteration of congestion patterns is taken into account. Regarding MP control, several controlled node layouts are tested, including subsets or all of network signalized nodes. Node subsets are selected both by the proposed method and randomly, for comparison. A sensitivity analysis of the best performing control schemes is done with respect to demand fluctuations. All scenarios are tested for two base origin-destination demand matrices, creating medium and high congestion, respectively. Results show promising performance of the hierarchical framework, in both demand scenarios, while several useful insights regarding the impact of each controller in the link service rates throughout the network and in relation to the location of MP and PC controlled intersections are gained.

Analysis of simulation results provide interesting insights about single and combined implementation of MP and PC, as well as regarding partial/selective MP schemes. Single MP control proves more effective in moderate congestion conditions rather than in over-saturated states, while partial MP application to subsets of critical nodes, properly selected by the proposed algorithm, can result to similar or even better performance than full-network implementation. Therefore, not only application cost can be drastically reduced, but control effectiveness can increase as well. Furthermore, partial MP implementation to critical nodes is shown to increase link throughput in the proximity of controlled nodes and balance surrounding queues, while spill-back occurrence of surrounding links is also reduced, for both demand scenarios. Also, significant production capacity increase is observed in single MP application, though it can lead to increased link density and thus increased congestion followed by production drop. This happens because MP achieves better road space use, leading to less local gridlocks that keep vehicles in virtual queues in the no control case, thus congestion is lower in the network. Yet, in the case of single MP, recovery of the network can be smoother with smaller hysteresis of the MFD curves.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3.17: Sensitivity analysis of the two best performing control schemes under demand fluctuations: (a) to (d) refer to the medium demand case, and FTC is compared to single MP in 25% of nodes selected by the proposed strategy; (e) to (h) refer to high demand case, and FTC is compared to combined scheme of PC and MP in 25% of selected intersections. Boxplots correspond to 20 different OD matrices where each OD pair is a random variable with normal distribution, with standard deviation equal to 5, 10, 15 and 20% of the mean, respectively.

4

Analysing the effects of remaining travel distance for MFD-based models

This chapter is based on the following article and includes only the parts where the author of this thesis was lead researcher:

- I. I. Sirmatel, D. Tsitsokas, A. Kouvelas, and N. Geroliminis (2021). “Modeling, estimation, and control in large-scale urban road networks with remaining travel distance dynamics”. In: *Transportation Research Part C: Emerging Technologies* 128, p. 103157

4.1 Introduction

This chapter aims at investigating the effects and behavior of total remaining travel distance dynamics, which have recently been incorporated in MFD-based macroscopic traffic modeling of single- and multi-region networks (see Lamotte et al., 2018, Murashkin, 2021) in an effort to provide more accurate and versatile representation of aggregated traffic modeling for large-scale control applications. More specifically, the recently proposed M-Model is based on a novel expression describing network outflow (trip ending) rate which does not depend on the steady state assumption that the frequently used PL (production over trip length) model requires, thus providing a more suitable ground to address rapidly changing demand scenarios that escape steady state or slow evolution assumptions. Moreover, in contrast to the alternative, more detailed trip-based (TB) model, M-Model does not require knowledge of the trip length distribution of all trips, but instead, it monitors the total remaining distance to be traveled by all vehicles as an additional state variable. In this chapter, a first attempt to investigate the evolution of total and average remaining distance to be traveled in a real large-scale network through microscopic simulation for a realistic demand scenario is described. Useful insights regarding the necessity and possible impact of utilizing M-Model instead of conventional PL approximation in aggregated, model-based control applications are provided based on the analysis of detailed trip trajectories. A short review of the model formulation is followed by a description of the simulation experiment setup and trajectory post-processing. Finally discussion on the results indicates that the newly-proposed M-Model can offer increased accuracy in describing aggregated network dynamics in specific highly-dynamic scenarios and therefore potentially improve control effectiveness of model predictive perimeter control applications.

4.2 Model description

4.2.1 A review of modeling with production-over-trip length approximation

Consider a city-scale road traffic network, consisting possibly of hundreds of links and intersections, with heterogeneous distribution of accumulation (i.e., number of vehicles) on its links. Using the macroscopic fundamental diagram (MFD) of urban traffic, it is possible to express the rate of vehicle-meters traveled in a region as a function of region accumulation. Clustering algorithms developed for such large-scale networks (see, e.g., Saeedmanesh and Geroliminis, 2016) can be used to partition the network into regions (i.e., a set of links) to obtain low intra-regional accumulation heterogeneity. Empirical results indicate that MFD can be approximated by an asymmetric unimodal curve skewed to the right (see Geroliminis and Daganzo, 2008). MFD-based models require the outflow MFD to express inter-regional vehicle transfer and exit flows, which is usually obtained by approximating the outflow MFD by the production-over-trip length approach formulated as

$$o_i(n_i(t)) = \frac{p_i(n_i(t))}{l_i} = \frac{n_i(t) \cdot v_i(n_i(t))}{l_i}, \tag{4.1}$$

where $t \in \mathbb{R}$ is the real time, $o_i(n_i(t))$ (veh/s) is the trip completion flow of the region as expressed by the outflow MFD (i.e., rate of vehicles exiting traffic), n_i (veh) is the accumulation in region i, $p_i(n_i(t))$ (veh.m/s) is the production MFD (i.e., rate of vehicle-meters traveled), l_i (m) is the average trip length for vehicles in region i (until exiting their current region i), whereas $v_i(n_i(t))$ (m/s) is the speed MFD expressing space-mean speed of region i.

Given a network $\mathcal{R}$ consisting of a set of R regions ($\mathcal{R} = \{1, 2, \ldots, R\}$) (see, e.g., fig. 4.1(b)), each with a well-defined MFD, aggregated dynamical models of

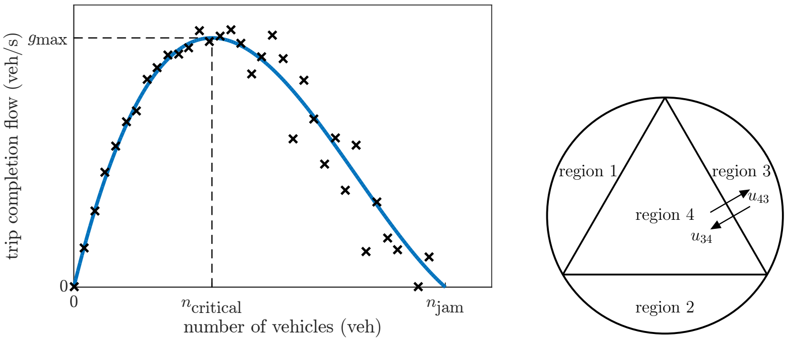


Figure 4.1: (a) A well-defined outflow macroscopic fundamental diagram. (b) A four region network without route choice.

large-scale road traffic networks can be developed based on inter-regional traffic flows as the following vehicle conservation equations (see Ramezani et al., 2015):

$$\dot{n}_{ii}(t) = q_{ii}(t) - o_{ii}(t) + \sum_{h\in\mathcal{N}_i} u_{hi}(t) o_{hii}(t) \tag{4.2a}$$

$$\dot{n}_{ij}(t) = q_{ij}(t) - \sum_{h\in\mathcal{N}_i} u_{ih}(t) o_{ihj}(t) + \sum_{h\in\mathcal{N}_i; h\neq j} u_{hi}(t) o_{hij}(t), \tag{4.2b}$$

where $n_{ii}(t)$ (veh) and $n_{ij}(t)$ (veh) are state variables expressing the accumulation in region i with destination region i and j, respectively, with $n_i(t) = \sum_{j=1}^{R} n_{ij}(t)$, $q_{ii}(t)$ (veh/s) and $q_{ij}(t)$ (veh/s) are disturbances expressing the rate of vehicles appearing in region i demanding trips to destination region i and j, respectively, $u_{ih}(t) \in [\underline{u}, \bar{u}]$ (with $0 \leq \underline{u} < \bar{u} < 1$) are perimeter control inputs between each pair of adjacent regions i and h, expressing actions of perimeter control actuators (with $h \in \mathcal{N}_i$; where $\mathcal{N}_i$ is the set of regions adjacent to i) that can adjust vehicle flows transferring between the regions, $o_{ihj}(t)$ (veh/s) is the vehicle flow transferring from i to h with destination j:

$$o_{ihj}(t) \triangleq \theta_{ihj}(t) \frac{n_{ij}(t)}{n_i(t)} o_i(n_i(t)), \tag{4.3}$$

where $\theta_{ihj}(t) \in [0,1]$ is the route choice term expressing, for the vehicles exiting region i with destination j, the ratio that is transferring to region h (with $o_{hii}(t)$ and $o_{hij}(t)$ defined similarly), whereas $o_{ii}(t)$ (veh/s) is the exit (i.e., internal trip completion) flow of region i:

$$o_{ii}(t) \triangleq \frac{n_{ii}(t)}{n_i(t)} o_i(n_i(t)). \tag{4.4}$$

Note that route choice effect can be omitted in modeling if the network topology leads to a single obvious route choice, in which case $\theta_{ihj}(t) = 1$ for all time for only one region $h \in \mathcal{N}_i$ for each i-j pair (with $j \neq i$). For example, for the network depicted in fig. 4.1(b), $\theta_{4hj}(t) = 1$ if $h = j$, and $\theta_{4hj}(t) = 0$ otherwise. The focus of this paper is on those networks where route choice can be omitted (see Sirmatel and Geroliminis, 2018 for a study where it is included).

4.2.2 Trip based models and approximations

Having revisited production-over-trip length approximation based dynamical models commonly encountered in the MFD literature in the previous section, in this section we derive the foundations of the M model (based on the work in Lamotte et al., 2018), and extend it to develop a multi-region dynamical M model.

All MFD type dynamic models rely on the conservation equation

$$\dot{n}(t) = I(t) - O(t), \tag{4.5}$$

where $n(t)$ denotes the accumulation of vehicles inside the zone, $\dot{n}$ its time derivative, $O(t)$ the outflow rate and $I(t)$ the inflow rate (note that trips may start either inside the zone, or by crossing the perimeter). While the inflow rate is often exogenous, outflow rate is estimated via a network exit function (NEF). Denoting the average trip length as l, outflow MFD can be derived as a simple expression for this outflow:

$$O(t) = \frac{n(t) \cdot v(t)}{l} = \frac{P(t)}{l}. \tag{4.6}$$

Daganzo, 2007 postulated that this result still holds approximately as long as the production MFD exists and inflow varies slowly enough, so that $O(t)$ can be approximated as $O_{\mathrm{PL}}(t) = O(n(t)) = P(n(t))/l$ (hence the appellation "PL model"). Validity of this assumption was empirically observed by Geroliminis and Daganzo, 2008, who observed using both loop detector and taxi data that the ratio of production over trip completion rate remained approximately constant over time for the city of Yokohama, Japan.

Trip-based model (TB) builds on the existence of a speed MFD $v(n(t))$, and derives outflow without requiring the PL model steady state assumption. An intuitive way to introduce the trip based model analytically for a single region is to consider that a user with trip length l_0 that enters the system at time t_0 should exit after traveling distance l_0 with travel time τ_0, satisfying

$$\int_{t_0}^{t_0+\tau_0} v(n(\tau))d\tau = l_0. \tag{4.7}$$

Among the users that entered the network at time s, the proportion that is still in the network at time $t > s$ is given by $1 - F_{\mathrm{cdf}}(\int_s^t v(n(\tau))d\tau)$, where $F_{\mathrm{cdf}}(\cdot)$ is the cumulative distribution function (cdf) of trip length, corresponding to the trip-generating process, with the corresponding probability density function (pdf) denoted by f. While in the general case $F_{\mathrm{cdf}}(\cdot)$ might be time dependent (e.g., due to rerouting and longer paths under congestion), it can be considered constant (as for example empirical data from Yokohama have shown in Geroliminis and Daganzo, 2008). Assuming that flow $I(s)$ entering the zone is known for all times $s < t$, and that $I(s) = 0$ for all times $s < 0$, the accumulation at time t is

$$n(t) = \int_0^t I(s) \left(1 - F_{\text{cdf}} \left(\int_s^t v(n(\tau)\Big)\, d\tau \right) \right) ds. \tag{4.8}$$

By differentiating the above equation, we obtain an expression that has the same form as eq. (4.5), but where the outflow is described by:

$$O_{TB}(t) = v(n(t)) \int_0^t I(s) f \left(\int_s^t v(n(\tau)) d\tau \right) ds. \tag{4.9}$$

While this expression is difficult to solve analytically, it can be easily implemented in an event- and agent-based simulation (see Section 4.2 of Lamotte et al., 2018). Lamotte et al., 2018 also define an approximation of the trip-based model, named M model with an additional state variable $m(t)$ representing the total remaining distance to be traveled by all vehicles currently in the network. The dynamics of m state can be expressed as:

$$\dot{m}(t) = I(t)l - n(t)v(n(t)), \tag{4.10}$$

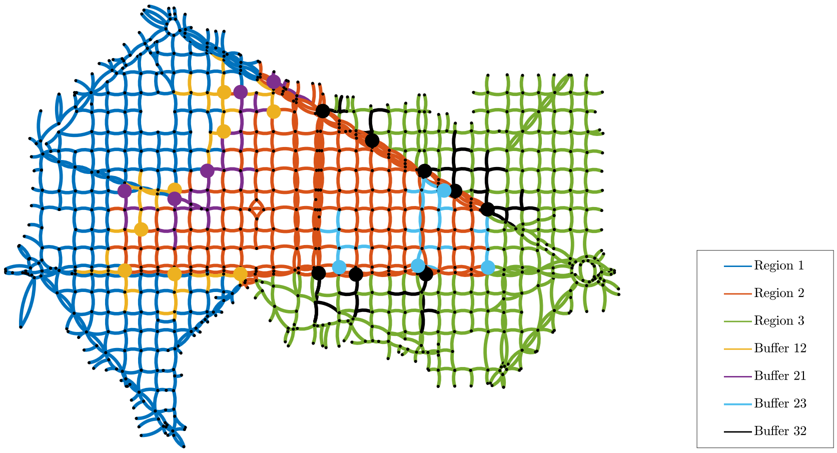
simply meaning that the rate of the total remaining distance increases by the generating demand multiplied by the average trip length and decreases by the current production rate $n(t)v(n(t))$.

By considering now multiple regions with a speed MFD $v_i(n_i(t))$ for each region i (with $i \in \mathcal{R}$), i.e., $v_i(n_i(t)) = p_i(n_i(t))/n_i(t)$, the outflow MFD $o_i(n_i(t))$ can be reformulated to account for variations in the remaining distance to be traveled as follows, i.e., with the M-model approach (building on the single region model of Lamotte et al., 2018):

$$o_i(n_i(t), m_i(t)) = \frac{n_i(t) \cdot v_i(n_i(t))}{l_i} \left(1 - \alpha_i \cdot \left(\frac{m_i(t)}{n_i(t) \cdot l_i^*} - 1 \right) \right), \tag{4.11}$$

where $\alpha_i \geq 0$ is a model parameter expressing the sensitivity of outflow $o_i(n_i(t), m_i(t))$ in relation to variations in the remaining distance to be traveled, $m_i(t)$ (veh.m) is the total remaining distance to be traveled by all vehicles in region i, whereas l_i^* is the average remaining distance to be traveled in steady state for vehicles in region i (until exiting their current region i) and l_i is the average trip length of vehicles in region i (according to the definition of the f distribution defined above). The ratio $m_i(t)/n_i(t)$ expresses the average remaining distance to be traveled by vehicles currently in region i, which might be different than the steady state value l_i^*. Thus, the dimensionless quantity inside the parenthesis multiplied by α_i can be positive or negative.

In steady state, the average distance remaining to be traveled is simply given by $l^* = \int_0^{+\infty} g(\delta) \frac{\delta}{2} d\delta$, where g is the pdf of the trip length distribution among all



(a) **(b)**

Figure 4.2: Microscopic (Aimsun) model of the network, with clustering results as links and controlled intersections as circles (intersections belonging to u_{12} in yellow, u_{21} in magenta, u_{23} in cyan, and u_{32} in black).

users present in a snapshot (note that g is different than f). Since users remain in the network for a duration proportional to their trip length, $g(\delta)$ is proportional to $f(\delta)\delta$. Imposing that $\int_0^{+\infty} g(\delta)d\delta = 1$ implies that $g(\delta) = f(\delta)\frac{\delta}{l}$. Thus:

$$l^* = \int_0^{+\infty} g(\delta)\frac{\delta}{2}d\delta = \int_0^{+\infty} f(\delta)\frac{\delta^2}{2l}d\delta = \frac{l^2+\sigma^2}{2l}. \tag{4.12}$$

Following the above description of aggregated traffic modeling with remaining travel distance dynamics according to recently proposed M-Model, we are interested to investigate the potential accuracy compared to conventional PL model, which relies on steady state conditions. More specifically, we examine how total remaining distance to be traveled evolves over time and in relation to trip length distribution changes, in detailed microscopic simulation, for a real large-scale network on realistic travel demand scenario, creating increased congestion states. Findings are expected to answer the question of whether M-Model formulation have the potential of describing more accurately aggregated traffic dynamics and thus improving the performance of model predictive perimeter control schemes, compared to conventional PL model.

4.3 Simulation and trajectory analysis

4.3.1 Network and simulation setup

An urban road network consisting of roughly 1500 links and 600 intersections is replicated as a computer model using the microscopic simulation package Aimsun (see fig. 4.2). The model represents a part of the urban network of the city of

Barcelona in Spain, covering an area of 12 km^2. The network is partitioned into three regions using the optimization-based clustering method of Saeedmanesh and Geroliminis, 2016. The direction of the road is represented by the curvature of the links in the graph (fig. 4.2), considering counter-clockwise movement of vehicles. Experiments are performed based on a realistic origin-destination demand matrix, generating approximately 202k trips. Simulation covers a 5 hours period, for which demand generation happens during the first 1.5 hours, with low rate for the first 15 mins (warm-up period) and high constant rate for the rest 1.25 hours. Trips start from 123 origin centroids and head towards 132 destination centroids, creating realistic traffic congestion patterns. A static fixed-time signal control plan is active for signalized intersections. Drivers adapt their selected paths during simulation and routes are updated dynamically based on real-time traffic information, according to a C-Logit route choice model (integrated element of Aimsun simulator) where travel time is used to define links disutility. The route update process is repeated every 3 minutes.

4.3.2 Trajectory analysis

In order to study the evolution of remaining distance to be traveled in microsimulation settings, detailed traffic information is produced and recorded through the Aimsun API tool. More specifically, a detailed record of the trajectories of all generated vehicles during simulation time is extracted. Each trajectory record consists of unique vehicle ID, the section (link) ID and region number of network entry and exit, the exact entry and exit simulation times, the total experienced trip length, and the vehicle location (section ID) every 1 second of simulation from its entry until its exit time. With suitable post processing of the trajectory files, a time-series of distance traveled from the beginning of the trip for the entire network and for every different region where the vehicle travels (if path expands to multiple regions) is created, providing useful information such as total distance traveled inside every region and time of border crossing, thus exiting one region and entering the neighboring one. This level of detail is useful for calculating different types of metrics such as remaining distance to be traveled per region, in reference to origin and destination of the trip. These metrics are useful for multi-regional MFD-based modeling and control using M-Model formulations, as they are proposed in the respective research article (see top of this chapter), of which this analysis makes part. The described dataset is then used to produce several figures regarding the evolution of trip length distributions and remaining travel distance, which are presented and discussed in the following section. Note that the first 30 minutes of the simulation are excluded from the analysis of remaining travel distance, since network starts loading from zero. Also, due to some atypical situation that is observed in the last simulation hour, where a group of few links demonstrate particularly high density

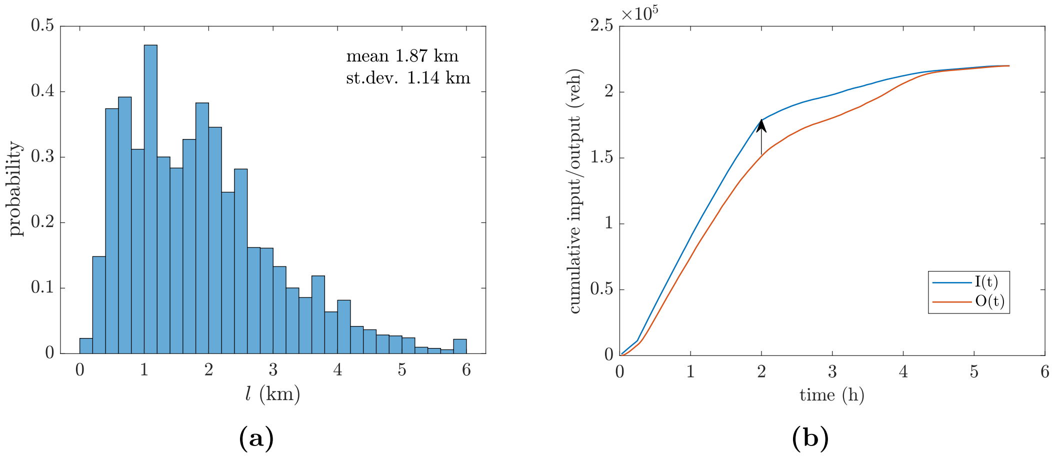


Figure 4.3: (a) Trip length distribution f of all generating trips in the network; (b) Cumulative input-output diagram for the entire network and simulation time; the arrow indicates the time $t = 2$ hours of maximum accumulation, after which outflow is higher than inflow and the network starts unloading.

(local gridlocks) and strong rerouting (artifact of the simulation software), this period is also excluded from the analysis of trip length distributions.

4.4 Results

The results of the analysis on the trajectories of the simulated trips are presented here. Remaining distance to be traveled to reach each destination is calculated every 90 seconds for every vehicle circulating in the network at this time (snapshot). The trip length for each vehicle during its whole trip is also extracted from the dataset. Consequently, we perform the same calculations for a network partitioned as per fig. 4.2 with 3 regions. Finally, the aforementioned two variables are calculated per region until a vehicle crosses in another region or finishes its trip within this region.

Figure 4.3(a) represents the trip length distribution of all generating trips while figure 4.3(b) represents the cumulative input-output of the entire network including all trips of the simulation. Note that before $t = 2$ hours, generating input is higher than output (trip endings) while the opposite happens after this time. The average remaining distance evolution during the simulation horizon is presented in fig. 4.4. For a given time, this is calculated as the sum of remaining distances of all vehicles currently in the network m divided by the total network accumulation n; fig. 4.4(a) shows the evolution of m/n over time. Note here that m/n is the actual average remaining distance, while l^* is the one in steady state. The difference between m/n and l^* is the main reason that the PL model may not be sufficiently accurate. Note that if $m/n = l^*$ then eq. 4.11 simplifies to the outflow of a PL model (i.e., eq. 4.1). Interestingly, we observe in fig. 4.4(a) that there is strong variation across time, due to the fact that congestion builds. As inflow is higher than outflow and

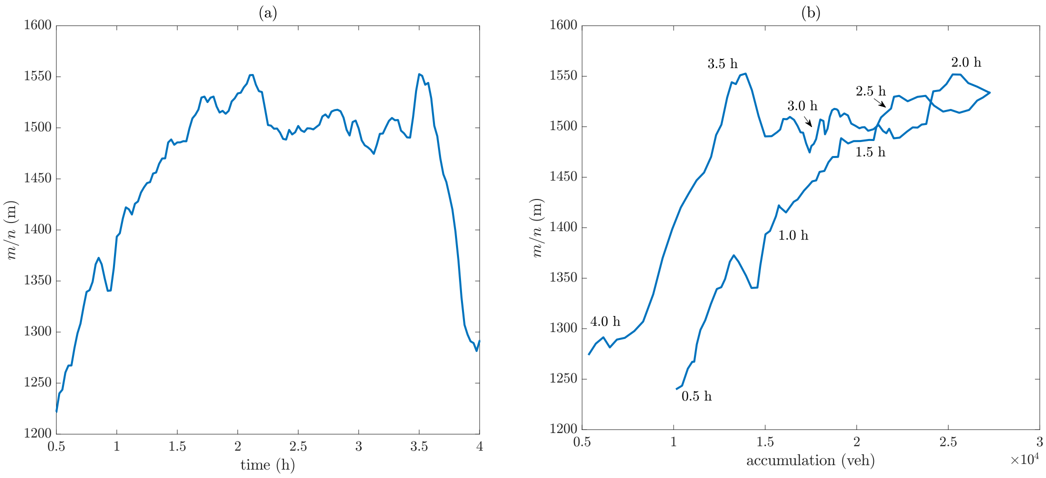


Figure 4.4: m/n plots: (a) Time series of average remaining distance m/n for the whole network. (b) m/n vs. accumulation for the whole network. Average remaining distance in steady state $l^* = 1.28$ km.

accumulation increases, the average remaining distance increases since the number of vehicles with longer trips remain in the network for longer and are influenced more by congestion. Note that for PL model that has memoryless properties, average remaining distance has to remain constant if trip length distribution is constant. The reason that m/n increases in the onset of congestion is the following: when inflow $I(t)$ is considerably higher than outflow $O(t)$, newly generated trips are entering in the network sampled from the distribution f as defined in section 4.2.1. As time passes, more and more large trips, which according to trip length distribution f have lower probability, are generated while circulating flow has already reached maximum values, and congestion starts forming. Since outflow is smaller due to network limitations or congestion propagation, the distribution of remaining distance shifts to the right and its average m/n is expected to increase. The opposite is expected to happen in the offset of congestion, when $O(t) > I(t)$, provided that the snapshot distribution g of trip length remains relatively constant. In this case, however, one can notice that while accumulation starts decreasing after about 2 hours and outflow is higher than inflow, m/n does not change significantly between 2 and 3.5 hours. One plausible explanation is that this happens due to the increased average trip length of snapshot distribution g, which is about 25% to 30% higher in this period compared to the onset of congestion. Therefore, while fewer trips are generated than completed, the increased average trip length of vehicles present in the network during this time, probably related to rerouting effect in congested conditions, keeps m/n from decreasing. Moreover, fig. 4.4(b) shows the relation between m/n and accumulation over the simulation time. Interestingly, a counter-clockwise hysteresis loop is observed. For the same accumulation, average remaining distance is higher in the offset of congestion. The explanation is the

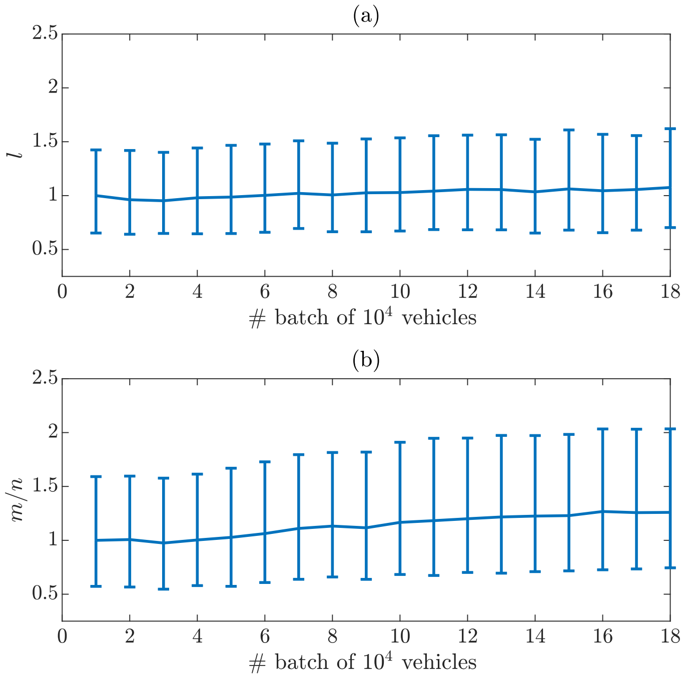


Figure 4.5: Mean value, 25th and 75th percentiles of (a) generating trip length and (b) average remaining distance to be traveled, for batches of 10k vehicles in order of departure time. Values are normalized over those of the first batch to highlight evolution over simulation time.

same as in fig. 4.4(a): This hysteresis is an inherent property of trip based model when congestion changes over time. Note that this hysteresis loop is not related to the spatial distribution of congestion, where more heterogeneous distribution can create lower production or speed for the same accumulation (see for example Geroliminis and Sun, 2011a or Saberi and Mahmassani, 2012). Interestingly, the hysteresis of fig. 4.4(b) is quite small for values of accumulation larger than the critical accumulation, which is around $1.4 \cdot 10^4$ vehicles.

While these results are consistent with the numerical simulations of an ideal trip based model as presented in Lamotte et al., 2018, assuming a *perfect* speed MFD (i.e., no errors or heterogeneity) and a constant distribution of generating trip length over time, vehicle rerouting and dynamic origin-destination tables might influence the distribution of trip length f, and as a result the remaining distance. Thus, we further investigate this in fig. 4.5, which shows the mean, 25 and 75 percentiles for average trip length l (of the distribution f), and the average remaining distance over time. The x-axis does not represent equal time intervals, but rather analyzes vehicles in groups of the same size (in batches of $10,000$ vehicles – approximately 5% of the total demand) as they appear in the simulation. While trip length value is unique for each vehicle, remaining distance changes with time. To estimate the distribution of remaining distance we need to choose a specific time and take a

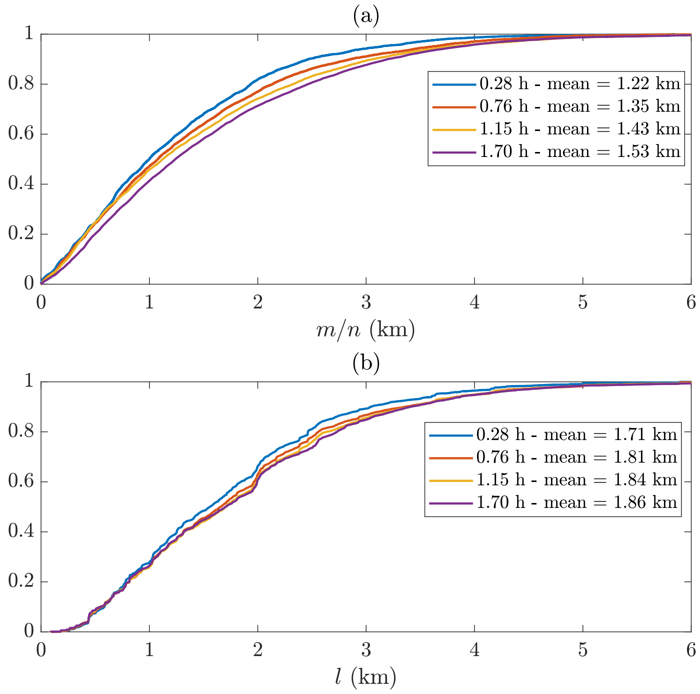


Figure 4.6: Cumulative distribution functions of (a) remaining distance of individual vehicles and (b) generating trip length distribution, for different times.

snapshot of the network. For each batch of $10,000$ vehicles we take a snapshot at the time that the middle vehicle in the batch starts its trip. To perform proper comparisons fig. 4.5(a) and (b) show dimensionless quantities by dividing all values with the average value of the first batch of 10,000 vehicles, at the beginning of the simulation (both for average trip length and average remaining distance). While l increases slightly during the simulation (not more than 7%), m/n increases following a different trend, reaching even magnitudes of 20%. Note also the higher deviation of the percentiles from the mean for m/n. Both l and m/n distributions are skewed to the right during all times and even more during the congested period.

Furthermore, fig. 4.6 shows the cumulative distribution functions for 4 different time intervals, where f is moderately similar while the remaining distance has a higher degree of variation (mean value l varies around 8% between $t = 0.28$ h and 1.7 h while mean m/n varies about 20% for the same period). These observations highlight that outflow estimates based on a PL model with memoryless characteristics are expected to be more inaccurate compared to an M-model that keeps track of remaining distance to be traveled and would possibly result in better control performance.

Moreover, we are interested in investigating the evolution of m/n over time for individual regions after partitioning. This is shown in fig. 4.7, where fig. 4.7(a) shows m/n and fig. 4.7(b) speed over time, respectively. While the 3 regions have

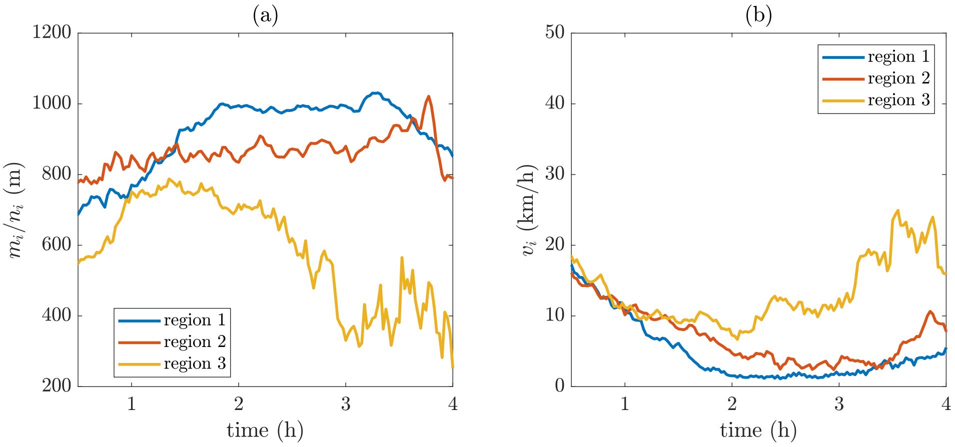


Figure 4.7: Time series of (a) average remaining distance, and (b) average speed per region ($l_1 = 1192$ m, $l_1^* = 818$ m, $l_2 = 1171$ m, $l_2^* = 888$ m, $l_3 = 852$ m, $l_3^* = 615$ m).

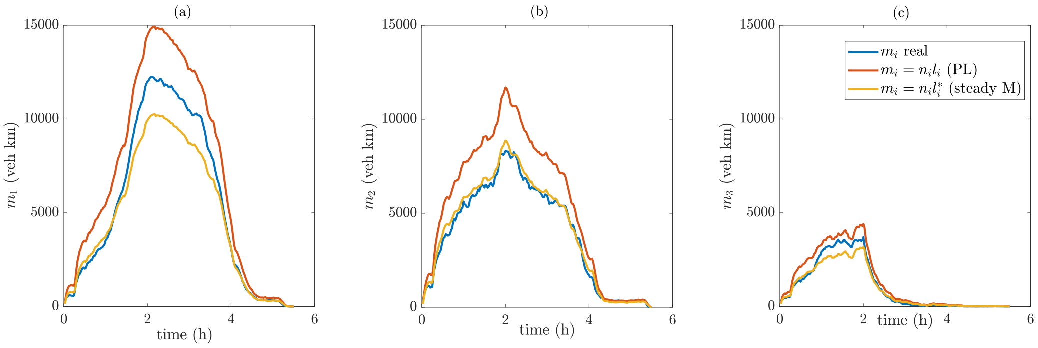


Figure 4.8: Comparison of m_i observed vs. steady state M model ($m_i = n_i \cdot l_i^*$) vs. PL approximation ($m_i = n_i \cdot l_i$) for each region i. The values of average l_i and l_i^* per region are: $l_1 = 1192$ m, $l_1^* = 818$ m, $l_2 = 1171$ m, $l_2^* = 888$ m, $l_3 = 852$ m, $l_3^* = 615$ m.

similar speeds in the onset of congestion, as demand increases further, regions 1 and 2 experience lower speeds, while region 3 recovers earlier. Nevertheless, average remaining distance follows quite different trends even when speed is equal for the 3 regions (during the first hour of the simulation). These trends are expected to influence the dynamics of accumulations as estimated by the PL and M models. Note for example that the time interval for which $m_i/n_i > l_i^*$ is different among regions.

Finally, we estimate the accuracy of the steady state representation of remaining trip distance based on M-model, $m_i = n_i \cdot l_i^*$, and the one if we assume the memoryless conditions of PL model, and compare with the exact observed value from the simulation. Given that PL, M and trip based models are identical for exponential trip length distribution (for a proof refer to Lamotte et al., 2018), remaining distance according to a PL model can be estimated for l^* of an exponential distribution (standard deviation equal to the mean), thus $m_i = n_i \cdot l_i$. Interestingly, fig. 4.8 shows that the real value of m_i is between the two aforementioned representations, as PL always overestimates m while steady-state M underestimates for region 1 and 3, while it is closer to the real value for region 2. Values of l_i and l^* are given in the caption of the figure. Note that these results are for the whole simulation duration (5 hours).

4.5 Summary

In this chapter, an investigation of the evolution of aggregated remaining travel distance for a realistic case study of a large-scale traffic network is performed based on microscopic simulation results. Research motivation derives by the recently proposed formulation for aggregated modeling and control based on total remaining distance to be traveled (called M Model), which is disconnected from the steady state approximation of outflow that is assumed by conventional PL model, while it requires less input information than its more general equivalent, the trip-based model. Analysis of the simulated vehicle trajectories reveal that recorded average remaining distance differs from the one corresponding to steady state for the realistic scenario that we examine, leading to the conclusion that conventional PL model may not be sufficiently accurate in predicting how traffic dynamics will evolve, especially in cases of highly dynamic congestion.

5

Conclusions and future research

This thesis develops modeling, optimization and control formulations for congestion-relieving strategies that are designed to be implemented in large-scale urban traffic networks facing high levels of congestion with dynamic characteristics. The focus is on dedicated bus lanes allocation in chapter 2, hierarchical traffic-responsive signal control in chapter 3, while in chapter 4 a recently-proposed modeling approach intended for aggregated MFD-based network control strategies is analyzed by microscopic simulation. In this final chapter of the thesis, a summary of the conducted research is made, highlighting the main findings and contributions, and proposing promising areas of related future research.

5.1 Main Findings

Chapter 2 focuses on the problem of optimal Dedicated Bus Lanes allocation in urban networks with existing bus systems, with the aim of balancing the trade-off between prioritizing buses and disturbing general traffic. A combinatorial optimization problem is formulated on the basis of a link-level dynamic traffic model and a set of algorithms are developed and tested for solving the high-complexity non-linear problem. The main contributions are the following:

- The DBL optimal allocation problem is formulated on the basis of an enhanced version of SaF macroscopic traffic model, which is able to capture the dynamic characteristics of congestion propagation and the spill-back effect of queues. Therefore, the impact of DBL location on general traffic is considered.
- Possible shift in mode preferences of travelers due to effects of DBL presence is considered in an aggregated way by a Logit model.

- An algorithmic scheme based on problem-specific local search heuristics and LNS metaheuristic is developed for addressing the problem. Several destroy and repair operators are proposed for the specific problem.
- A simulation-based learning process is proposed for estimating the importance of candidate links for receiving DBL, which can also be integrated in LNS.
- A network decomposition technique is proposed for accelerating the solution process in cases of very large networks.
- The proposed formulation and solution methodology is applied for a real network of more than 450 links and 260 nodes and good-quality DBL plans are found.
- A Pareto frontier associating the travel time gains with the lane-km of DBLs placed is constructed with the proposed method, which can assist authorities in the decision-making process regarding DBL installation.

More specifically, numerical results of the application of the proposed scheme in our case study show significant potential improvement in the system performance compared to the case of no existing DBL, proving the effectiveness of the proposed methodology in addressing the DBL allocation problem in large-scale networks considering dynamic congestion. The proposed heuristic for link-by-link construction of DBL plan (Algorithm 1) is shown particularly efficient for the construction of a good quality solution and estimation of link scores, even though its high computational cost might hinder its use in large scale networks. The formulated LNS algorithm is also shown effective in identifying good quality solutions in short time and its efficiency can be increased by proper setting of link selection probabilities (scores) or by including an update process during LNS execution. The proposed update process, similar in concept to A-LNS (see Ropke and Pisinger, 2006), proves efficient even when the initial link selection probabilities are not well tuned, which facilitates the application of LNS. Moreover, the optimized DBL plans are also efficient with respect to the relation between the amount of reserved road space and system performance improvement. Efficient road space allocation schemes are identified, which can lead to improvement of the system traffic performance even without considering the resulting mode shift from car to bus. This is achieved by identifying candidate DBL locations that improve bus travel time while causing the least possible disturbance to regular traffic, e.g. in wide roads with low traffic flow. A Pareto frontier describing the decrease of the experienced passenger travel time in relation to the road space occupied by DBLs is created through application of the proposed algorithmic scheme, which can actively support the decision making process regarding DBL network design, depending on the specific objectives and requirements of each case study.

Chapter 3 refers to traffic-responsive signal control for urban networks and proposes a two-layer hierarchical control framework, which combines perimeter control, implemented after partitioning of the network in homogeneously congested regions, and distributed Max Pressure control, implemented to isolated network intersections. The main contributions of the chapter are the following:

- Combined implementation of multi-region PC with distributed MP in a two-layer hierarchical control framework is proposed for large-scale network control.
- Partial implementation of MP control in subsets of network nodes is tested.
- An algorithm to help identify critical nodes for MP control according to queue-related metrics around the node (mean, variance and spill-backs of queues during peak hour) is developed and tested.
- Several control layouts for MP involving different penetration rates of controlled nodes, the selection of which was both targeted and random (for comparison) are tested by an enhanced version of link-based dynamic macroscopic SaF model, which integrates traffic signals, queue capacities and spill-backs.
- The impact of on-road vehicle re-routing due to control-induced changes in travel times on specific network links is taken into account. An algorithm to update turn ratios of SaF model in regular intervals during simulation, based on time-wise shortest paths and origin-destination travel demand is developed.
- Spatial analysis of the impact of adaptive control on cumulative throughput and total delay experienced on links in the proximity of MP and PC controlled node locations is performed.
- A sensitivity analysis on the performance of partial MP schemes on critical nodes under demand fluctuations is performed showing promising results for the tested schemes created by the proposed method.

More specifically, the proposed node selection method proves effective in identifying critical node sets for MP control, since it outperforms random selection for all network penetration rates in the medium demand scenario. Even though its effectiveness seems to drop in the high demand scenario for the single MP scheme, it remains effective in all combined schemes of PC and MP. However, it should be noted that the proposed selection method involves a parameter optimization step that resulted in different values for the medium and high demand, which indicates that the relative importance of the selection criteria can vary between moderate and highly congested conditions. Nevertheless, the proposed selection variables

(m_1, m_2 and N_c) seem to play an important role as indicators of node importance with respect to MP, while further research can help unravel the mechanism that relates selection variable importance to demand patterns, and thus determine optimal parameter values in a universal way by dropping parameter optimization requirement. Overall, it seems that significant correlation exists between controlled nodes, which affect each other in a way not necessarily beneficial for the system, since performance gains can decrease for penetration rates above 25% and can even drop to zero in 100% in highly congested scenarios. This phenomenon highlights the importance of partial MP implementation, especially in increased congestion, and the role of spatial distribution of the controlled nodes, which the proposed selection method tries to unravel.

Regarding the two-layer combined scheme of PC and partial MP, results are promising in most tested cases, especially in the high demand scenario, where in our case study, adding MP in only 25% of properly selected network nodes leads to doubling the performance gains of single PC compared to FTC case, from 7.5% to more than 15%. Moreover, almost the same performance gain is achieved in the case of full-network MP implementation, proving the proposed selection method effective and, as a result, reducing implementation cost to one fourth, compared to full-network scheme. Furthermore, PC protects high-demand regions from reaching saturated states, and therefore from capacity drop, which seems to also increase MP efficiency, given that single MP shows zero improvement for full network implementation in high demand scenario, while combined MP with PC achieves twice the gain of single PC. Moreover, the node schemes generated by the proposed selection method do not appear sensitive to small demand fluctuations that alter traffic distribution in the network, and can still lead to performance improvement even if demand deviates up to 20% from its mean value, which was used to generate the node scheme. However, as expected, improvement gains decrease as fluctuation increases.

Chapter 4 performs an analysis on the total remaining travel distance for a multi-region large-scale network based on a set of detailed trip trajectories extracted from microscopic simulation, with the objective of assessing the potential benefits of the recently proposed M-Model, which monitors remaining travel distance within homogeneous regions, in aggregated MFD-based network control applications. The main contributions of the chapter are the following:

- Evolution of aggregated remaining travel distance in a realistic demand scenario for a multi-region large scale network is analyzed through detailed microscopic simulation. Dynamic route choice update as a response to real-time traffic conditions is considered in the simulation for a more realistic representation of drivers decisions.

- Microsimulation results show that even though mean trip length does not change significantly during simulation, average remaining distance increases significantly when congestion forms, indicating that remaining travel distance can capture and carry along information related to congested states, in contrast to memoryless PL model.
- Even when speed time profiles are similar among different regions, it is shown that remaining distance may follow a different pattern over time, which diverts significantly from steady state approximation, indicating that M-Model can result in increased accuracy in MFD-based control applications.

Elaborating more on the findings, we observed that dynamic phenomena, such as drivers rerouting to avoid congested arterial roads and increasing their traveled distance, cause deviations from steady state, which is demonstrated by a gradual increase of average remaining distance as congestion builds that is much higher than the equivalent average trip length increase. As a result, counter-clockwise hysteresis phenomena can appear in the relation associating accumulation and average remaining distance to be traveled, where the latter is higher in congested than in uncongested conditions. Furthermore, even with similar speed profile, average remaining distance might behave differently among homogeneously congested regions. Specifically, it is shown that if steady state is assumed, as in the case of PL model, remaining travel distance is overestimated, causing concerns about PL accuracy, while creating expectations of a more accurate representation by M-Model, which monitors total remaining travel distance. Further investigation is necessary, both in simulated environment and ideally in the field, under different congestion scenarios, in order to definitively assess M-Model effectiveness compared to conventional PL model in terms of network-wide MPC-based gating. Results of a first attempt towards this direction are promising for M-Model use, for which the interested reader can refer to the respective published article, of which this chapter makes part (i.e. Sirmatel et al., 2021).

5.2 Future Research

On the basis of the developments and findings of the research included in this thesis, interesting future research directions emerge, related to the three main research fields explored.

Regarding optimal DBL allocation problem of chapter 2, in the presented approach no re-routing of drivers due to DBL presence was considered in the simulation scheme. However, this component was developed at a later stage for chapter 3, based on the turn ratio update algorithm 6 presented in section 3.2.4, and could be easily integrated as is, in the simulations that evaluate DBL candidate solutions. A useful but potentially computationally costly addition to the current formulation would be to include more decision variables in the optimization problem related to bus operations, such as bus frequencies, or even size of bus fleet and bus capacities depending on the expected ridership.

Moreover, introducing dynamic activation of DBL, in the sense of indicating with variable message signs or special lane signals whether a lane is open or closed for vehicles other than buses, would allow higher flexibility in space allocation, as all candidate lanes can potentially be reserved for buses, according to a central controller who would do the assignment dynamically, based on real-time traffic information and mainly the location of buses. In this way, a lane can become DBL when more buses carrying more passengers are expected to use it, while it can be open to all drivers in periods of lower bus frequencies. The concept of bus-exclusive lanes can also be replaced by a mixed-use dynamically controlled lane, similar to dynamic lanes introduced in Anderson and Geroliminis, 2020 for highways, the use of which can be shared among buses and some private cars, in an analogy defined by a feedback controller, based on the current traffic situation in the proximity of the lane. Several research questions arise there, such as what is the relationship between congestion in the general purpose lanes and the importance of reserved space for buses, what is the optimal number of vehicles that can use the priority lane given the number of buses, or which control practice, reactive or proactive, leads to higher system performance. The desired fraction of private vehicles that are allowed in the priority lanes can also be implemented through a pricing scheme, similar to the concept of High Occupancy Toll (HOT) lanes, where the price of using the priority lanes can vary dynamically, based on the time of the day and the level of congestion. An enhanced version of the presented modeling approach would be necessary in order to consider and monitor bus trips, as well as bus interactions with general traffic and impact of congestion in bus trips, given that in the current formulation, bus travel time is estimated exogenously to the utilized traffic model. Also, TSP based on strategies such as phase re-activation or green time extension, could be considered for DBLs in presence of buses in the lane, thus reducing bus delays even more.

Regarding adaptive signal control for large-scale networks, future research can focus on considering PTP polices on top of MP and PC frameworks in the two-layer controller. For instance, DBL dynamic activation can be introduced as a third layer in the proposed control framework, while incorporating bus system dynamics in the problem. This addition would also require revisiting the traffic model in order to simulate bus trips separately from general traffic that is monitored by the utilized version of SaF model. Furthermore, dynamic activation of subset of MP controlled nodes in real-time from a centralized approach would be an interesting future research direction, especially with the arrival of connected and automated vehicles, where required traffic information (such as presence in links and next link in the path, which can be translated to link accumulation and turn ratios) can be provided to MP controller through vehicle-to-infrastructure communication and MP implementation cost related to node instrumentation will be zero. This would mean that any node could function as a MP node for some time without any implementation cost. In this control framework, the objective would be to identify the subset of nodes that should be activated in the next control interval, given the actual traffic state of the network, in order to maximize performance. However, this would mean that the approach would be centralized and communication infrastructure would be necessary. Yet, the scalability of MP strategy would remain, as the system would optimize the assignment based on the currently available intersections, without necessitating all intersections to be available for MP control immediately, allowing gradual installation,

Additionally, on-line adjustment of PC parameters, in order to address possible MFD alterations, due to either MP effects or unusual spatio-temporal traffic distribution among regions, while monitoring current traffic conditions, would be an important complementary element, especially for multi-region PC. Machine learning techniques could be employed for this process, as proposed in Kouvelas et al., 2017. Particularly in the case of dynamic activation of MP in different node sets depending on the actual traffic patterns, it would be interesting to investigate how the shape of MFD would be affected, and therefore adapt the PC parameters. Adjusting MFD curve due to MP control would be paramount in MPC-based approaches, which utilize the entire MFD curve to predict system evolution and optimize the control over it. Besides, PC implementation through optimal MPC instead of PI regulator could prove even more beneficial for network performance. Robust estimation and control schemes, able to address demand and input data noise in saturated conditions is another interesting topic to explore. Yet, investigating system reaction under sudden incidents causing link or intersection closure, in presence of adaptive control based on PC and MP, can provide significant insights on the ability of these adaptive schemes to help the system recover, compared to fixed-time control. Path-assignment in the link-level, combined with adaptive signal control is also an interesting research topic to explore, whose importance is expected to increase with the advent of autonomous vehicles, similar to the work of Yildirimoglu et al., 2018 which was done in the regional level.

Finally, there is significant space in exploring the potential of integrating M-Model in MFD-based modeling and control strategies, especially under rapidly changing traffic conditions that deviate from steady state. A multi-region extension of M-Model with boundary queues is already developed and tested in Sirmatel et al., 2021, where suitable system identification framework is formulated and a nonlinear moving horizon observer is proposed for constructing remaining distance states only from accumulation data. Evaluating the performance of M-Model in describing traffic dynamics of single- and multi-region systems by comparison with real data would be an interesting next step. Comparing the performance of M-Model to the one of PL with buffer zones (similar to Ni and Cassidy, 2020) in a microsimulation environment for realistic, dynamic demand scenarios, would give more insights about the potential of the two models regarding network-wide control. Finally, integrating dynamic trip lengths in the modeling structures could potentially lead to improved prediction performance, on condition that efficient methods of measuring or observing individual trips can be developed.

References

Abdelghany, K. F., Mahmassani, H. S., and Abdelghany, A. F. (2007). “A modeling framework for bus rapid transit operations evaluation and service planning”. In: *Transportation Planning and Technology* 30.6, pp. 571–591.

Aboudolas, K. and Geroliminis, N. (2013). “Perimeter and boundary flow control in multi-reservoir heterogeneous networks”. In: *Transportation Research Part B: Methodological* 55, pp. 265–281.

Aboudolas, K., Papageorgiou, M., and Kosmatopoulos, E (2009). “Store-and-forward based methods for the signal control problem in large-scale congested urban road networks”. In: *Transportation Research Part C: Emerging Technologies* 17.2, pp. 163–174.

Allsop, R. E. (1971). “SIGSET: A computer program for calculating traffic signal settings”. In: *Traffic Engineering & Control*.

Allsop, R. E. (1976). “SIGCAP: A computer program for assessing the traffic capacity of signal-controlled road junctions”. In: *Traffic Engineering & Control* 17.Analytic.

Ambühl, L. et al. (2019). “Approximative network partitioning for MFDs from stationary sensor data”. In: *Transportation Research Record* 2673.6, pp. 94–103.

Ampountolas, K., Zheng, N., and Geroliminis, N. (2017). “Macroscopic modelling and robust control of bi-modal multi-region urban road networks”. In: *Transportation Research Part B: Methodological* 104, pp. 616–637.

Anderson, P. and Geroliminis, N. (2020). “Dynamic lane restrictions on congested arterials”. In: *Transportation Research Part A: Policy and Practice* 135, pp. 224–243.

Arasan, V. T. and Vedagiri, P (2010). “Microsimulation study of the effect of exclusive bus lanes on heterogeneous traffic flow”. In: *Journal of Urban Planning and Development* 136.1, pp. 50–58.

Arnott, R. (2013). “A bathtub model of downtown traffic congestion”. In: *Journal of Urban Economics* 76, pp. 110–121.

Basso, L. J. et al. (2011). “Congestion pricing, transit subsidies and dedicated bus lanes: Efficient and practical solutions to congestion”. In: *Transport Policy* 18.5, pp. 676–684.

Batista, S., Seppecher, M., and Leclercq, L. (2021). “Identification and characterizing of the prevailing paths on a urban network for MFD-based applications”. In: *Transportation Research Part C: Emerging Technologies* 127, p. 102953.

Bayrak, M. and Guler, S. I. (2018). “Optimizing Bus Lane Placement on Networks while Accounting for Queue Spillbacks”. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 920–925.

Beojone, C. V. and Geroliminis, N. (2021). “On the inefficiency of ride-sourcing services towards urban congestion”. In: *Transportation research part C: emerging technologies* 124, p. 102890.

Buisson, C. and Ladier, C. (2009). “Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams”. In: *Transportation Research Record* 2124.1, pp. 127–136.

Chen, C et al. (2022). “Data efficient reinforcement learning and adaptive optimal perimeter control of network traffic dynamics”. In: *Transportation Research Part C: Emerging Technologies* 142, p. 103759.

Chen, X. et al. (2010). “Microscopic traffic simulation approach to the capacity impact analysis of weaving sections for the exclusive bus lanes on an urban expressway”. In: *Journal of Transportation Engineering* 136.10, pp. 895–902.

Chiabaut, N. (2015). “Evaluation of a multimodal urban arterial: The passenger macroscopic fundamental diagram”. In: *Transportation Research Part B: Methodological* 81, pp. 410–420.

Choi, D and Choi, W (1995). “Effects of an exclusive bus lane for the oversaturated freeway in Korea”. In: *1995 Compendium of Technical Papers. Institute of Transportation Engineers 65th Annual Meeting. Institute of Transportation Engineers (ITE).*

Chow, A. H., Sha, R., and Li, S. (2020). “Centralised and decentralised signal timing optimisation approaches for network traffic control”. In: *Transportation Research Part C: Emerging Technologies* 113, pp. 108–123.

Christofa, E., Ampountolas, K., and Skabardonis, A. (2016). “Arterial traffic signal optimization: A person-based approach”. In: *Transportation Research Part C: Emerging Technologies* 66, pp. 27–47.

Christofa, E., Papamichail, I., and Skabardonis, A. (2013). “Person-based traffic responsive signal control optimization”. In: *IEEE Transactions on Intelligent Transportation Systems* 14.3, pp. 1278–1289.

Daganzo, C. F. (2007). “Urban gridlock: Macroscopic modeling and mitigation approaches”. In: *Transportation Research Part B: Methodological* 41.1, pp. 49–62.

Daganzo, C. F., Gayah, V. V., and Gonzales, E. J. (2011). “Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability”. In: *Transportation Research Part B: Methodological* 45.1, pp. 278–288.

Daganzo, C. F. and Lehe, L. J. (2015). “Distance-dependent congestion pricing for downtown zones”. In: *Transportation Research Part B: Methodological* 75, pp. 89–99.

Dahlgren, J. (1998). “High occupancy vehicle lanes: Not always more effective than general purpose lanes”. In: *Transportation Research Part A: Policy and Practice* 32.2, pp. 99–114.

Deng, T. and Nelson, J. D. (2011). “Recent developments in bus rapid transit: a review of the literature”. In: *Transport Reviews* 31.1, pp. 69–96.

Diakaki, C., Papageorgiou, M., and Aboudolas, K. (2002). “A multivariable regulator approach to traffic-responsive network-wide signal control”. In: *Control Engineering Practice* 10.2, pp. 183–195.

Eichler, M. and Daganzo, C. F. (2006). “Bus lanes with intermittent priority: Strategy formulae and an evaluation”. In: *Transportation Research Part B: Methodological* 40.9, pp. 731–744.

Farid, Y. Z., Christofa, E., and Collura, J. (2018). “An analytical model to conduct a person-based evaluation of transit preferential treatments on signalized arterials”. In: *Transportation Research Part C: Emerging Technologies* 90, pp. 411–432.

Farid, Y. Z., Christofa, E., and Collura, J. (2015). “Dedicated bus and queue jumper lanes at signalized intersections with nearside bus stops: Person-based evaluation”. In: *Transportation Research Record* 2484.1, pp. 182–192.

Fosgerau, M. (2015). “Congestion in the bathtub”. In: *Economics of Transportation* 4.4, pp. 241–255.

Fu, H., Liu, N., and Hu, G. (2017). “Hierarchical perimeter control with guaranteed stability for dynamically coupled heterogeneous urban traffic”. In: *Transportation Research Part C: Emerging Technologies* 83, pp. 18–38.

Gartner, N. H. (1983). *OPAC: A demand-responsive strategy for traffic signal control.* 906.

Gayah, V. V. and Daganzo, C. F. (2011). “Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability”. In: *Transportation Research Part B: Methodological* 45.4, pp. 643–655.

Geroliminis, N. and Daganzo, C. F. (2008). “Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings”. In: *Transportation Research Part B: Methodological* 42.9, pp. 759–770.

Geroliminis, N., Haddad, J., and Ramezani, M. (2012). “Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach”. In: *IEEE Transactions on Intelligent Transportation Systems* 14.1, pp. 348–359.

Geroliminis, N. and Levinson, D. M. (2009). “Cordon pricing consistent with the physics of overcrowding”. In: *Transportation and Traffic Theory 2009: Golden Jubilee.* Springer, pp. 219–240.

Geroliminis, N. and Skabardonis, A. (2011). “Identification and analysis of queue spillovers in city street networks”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.4, pp. 1107–1115.

Geroliminis, N. and Sun, J. (2011a). “Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks”. In: *Procedia-Social and Behavioral Sciences* 17, pp. 213–228.

Geroliminis, N. and Sun, J. (2011b). “Properties of a well-defined macroscopic fundamental diagram for urban traffic”. In: *Transportation Research Part B: Methodological* 45.3, pp. 605–617.

Geroliminis, N., Zheng, N., and Ampountolas, K. (2014). “A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks”. In: *Transportation Research Part C: Emerging Technologies* 42, pp. 168–181.

Godfrey, J. (1969). “The mechanism of a road network”. In: *Traffic Engineering & Control* 8.8.

Gonzales, E. J. et al. (2010). “On the allocation of city space to multiple transport modes”. In: *Transportation planning and technology* 33.8, pp. 643–656.

Gregoire, J. et al. (2014a). “Back-pressure traffic signal control with unknown routing rates”. In: *IFAC Proceedings Volumes* 47.3, pp. 11332–11337.

Gregoire, J. et al. (2014b). “Capacity-aware backpressure traffic signal control”. In: *IEEE Transactions on Control of Network Systems* 2.2, pp. 164–173.

Guler, S. I. and Cassidy, M. J. (2012). “Strategies for sharing bottleneck capacity among buses and cars”. In: *Transportation research part B: methodological* 46.10, pp. 1334–1345.

Guler, S. I., Gayah, V. V., and Menendez, M. (2016). “Bus priority at signalized intersections with single-lane approaches: A novel pre-signal strategy”. In: *Transportation Research Part C: Emerging Technologies* 63, pp. 51–70.

Haddad, J. (2017a). “Optimal coupled and decoupled perimeter control in one-region cities”. In: *Control Engineering Practice* 61, pp. 134–148.

Haddad, J. (2017b). “Optimal perimeter control synthesis for two urban regions with aggregate boundary queue dynamics”. In: *Transportation Research Part B: Methodological* 96, pp. 1–25.

Haddad, J. and Geroliminis, N. (2012). “On the stability of traffic perimeter control in two-region urban cities”. In: *Transportation Research Part B: Methodological* 46.9, pp. 1159–1176.

Haddad, J. and Mirkin, B. (2020). “Resilient perimeter control of macroscopic fundamental diagram networks under cyberattacks”. In: *Transportation research part B: methodological* 132, pp. 44–59.

Haddad, J., Ramezani, M., and Geroliminis, N. (2013). “Cooperative traffic control of a mixed network with two urban regions and a freeway”. In: *Transportation Research Part B: Methodological* 54, pp. 17–36.

Haddad, J. and Zheng, Z. (2020). “Adaptive perimeter control for multi-region accumulation-based models with state delays”. In: *Transportation Research Part B: Methodological* 137, pp. 133–153.

Haitao, H. et al. (2019). “Providing public transport priority in the perimeter of urban networks: A bimodal strategy”. In: *Transportation Research Part C: Emerging Technologies* 107, pp. 171–192.

Henry, J.-J., Farges, J. L., and Tuffal, J. (1984). “The PRODYN real time traffic algorithm”. In: *Control in Transportation Systems.* Elsevier, pp. 305–310.

Hunt, P. et al. (1981). *SCOOT-a traffic responsive method of coordinating signals.* Tech. rep.

Ingole, D., Mariotte, G., and Leclercq, L. (2020). “Perimeter gating control and citywide dynamic user equilibrium: A macroscopic modeling framework”. In: *Transportation research part C: emerging technologies* 111, pp. 22–49.

Ji, Y et al. (2010). “Macroscopic fundamental diagram: investigating its shape using simulation data”. In: *89th Annual Meeting Transportation Research Board, Washington DC.* Mira Digital Publishing, pp. 1–12.

Ji, Y. and Geroliminis, N. (2012). “On the spatial partitioning of urban transportation networks”. In: *Transportation Research Part B: Methodological* 46.10, pp. 1639–1656.

Jin, W.-L. (2020). “Generalized bathtub model of network trip flows”. In: *Transportation Research Part B: Methodological* 136, pp. 138–157.

Johari, M. et al. (2021). “Macroscopic network-level traffic models: Bridging fifty years of development toward the next era”. In: *Transportation Research Part C: Emerging Technologies* 131, p. 103334.

Keyvan-Ekbatani, M., Papageorgiou, M., and Knoop, V. L. (2015a). “Controller design for gating traffic control in presence of time-delay in urban road networks”. In: *Transportation Research Procedia* 7, pp. 651–668.

Keyvan-Ekbatani, M. et al. (2012). “Exploiting the fundamental diagram of urban networks for feedback-based gating”. In: *Transportation Research Part B: Methodological* 46.10, pp. 1393–1403.

Keyvan-Ekbatani, M. et al. (2015b). “Multiple concentric gating traffic control in large-scale urban networks”. In: *IEEE Transactions on Intelligent Transportation Systems* 16.4, pp. 2141–2154.

Keyvan-Ekbatani, M. et al. (2019). “Traffic-responsive signals combined with perimeter control: investigating the benefits”. In: *Transportmetrica B: Transport Dynamics* 7.1, pp. 1402–1425.

Khoo, H. L., Teoh, L. E., and Meng, Q. (2014). “A bi-objective optimization approach for exclusive bus lane selection and scheduling design”. In: *Engineering Optimization* 46.7, pp. 987–1007.

Kong, X. et al. (2011). “Urban arterial traffic two-direction green wave intelligent coordination control technique and its application”. In: *International Journal of Control, Automation and Systems* 9.1, pp. 60–68.

Kouvelas, A., Saeedmanesh, M., and Geroliminis, N. (2017). “Enhancing model-based feedback perimeter control with data-driven online adaptive optimization”. In: *Transportation Research Part B: Methodological* 96, pp. 26–45.

Kouvelas, A. et al. (2014). “Maximum pressure controller for stabilizing queues in signalized arterial networks”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2421, pp. 133–141.

Lämmer, S. and Helbing, D. (2008). “Self-control of traffic lights and vehicle flows in urban road networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.04, P04019.

Lamotte, R. and Geroliminis, N. (2018). “The morning commute in urban areas with heterogeneous trip lengths”. In: *Transportation Research Part B: Methodological* 117, pp. 794 –810.

Lamotte, R. and Geroliminis, N. (2016). “The Morning Commute in Urban Areas: Insights from Theory and Simulation”. In: *Transportation Research Board 95th Annual Meeting.* URL: `https://trid.trb.org/view/1392730`.

Lamotte, R. et al. (2018). “Dynamic modeling of trip completion rate in urban areas with MFD representations”. In: *2018 TRB Annual Meeting Online.* Transportation Research Board, pp. 18–06192.

Laval, J. A., Leclercq, L., and Chiabaut, N. (2017). “Minimal parameter formulations of the dynamic user equilibrium using macroscopic urban models: Freeway vs city streets revisited”. In: *Transportation research procedia* 23, pp. 517–530.

Le, T. et al. (2015). “Decentralized signal control for urban road networks”. In: *Transportation Research Part C: Emerging Technologies* 58, pp. 431–450.

Levin, M. W., Hu, J., and Odell, M. (2020). “Max-pressure signal control with cyclical phase structure”. In: *Transportation Research Part C: Emerging Technologies* 120, p. 102828.

Levinson, H. S. et al. (2003). “Bus rapid transit: Synthesis of case studies”. In: *Transportation Research Record* 1841.1, pp. 1–11.

Levy, J. I., Buonocore, J. J., and Von Stackelberg, K. (2010). “Evaluation of the public health impacts of traffic congestion: a health risk assessment”. In: *Environmental health* 9.1, pp. 1–12.

Li, L. and Jabari, S. E. (2019). “Position weighted backpressure intersection control for urban networks”. In: *Transportation Research Part B: Methodological* 128, pp. 435–461.

Li, S. and Ju, Y. (2009). “Evaluation of bus-exclusive lanes”. In: *IEEE Transactions on Intelligent Transportation Systems* 10.2, pp. 236–245.

Li, T., Chen, P., and Tian, Y. (2021). “Personalized incentive-based peak avoidance and drivers’ travel time-savings”. In: *Transport Policy* 100, pp. 68–80.

Lin, S. et al. (2011). “Fast model predictive control for urban road networks via MILP”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.3, pp. 846–856.

Lin, S. et al. (2012). “Efficient network-wide model-based predictive control for urban traffic networks”. In: *Transportation Research Part C: Emerging Technologies* 24, pp. 122–140.

Loder, A. et al. (2017). “Empirics of multi-modal traffic networks–Using the 3D macroscopic fundamental diagram”. In: *Transportation Research Part C: Emerging Technologies* 82, pp. 88–101.

Loder, A. et al. (2019). “Understanding traffic capacity of urban networks”. In: *Scientific reports* 9.1, pp. 1–10.

Löfberg, J. (2004). “YALMIP : A Toolbox for Modeling and Optimization in MATLAB”. In: *In Proceedings of the CACSD Conference.* Taipei, Taiwan.

Lowrie, P. (1990). “Scats, sydney co-ordinated adaptive traffic system: A traffic responsive method of controlling urban traffic”. In.

Mahmassani, H. S., Saberi, M., and Zockaie, A. (2013). “Urban network gridlock: Theory, characteristics, and dynamics”. In: *Procedia-Social and Behavioral Sciences* 80, pp. 79–98.

Mahmassani, H. S., Williams, J. C., and Herman, R. (1984). “Investigation of network-level traffic flow relationships: some simulation results”. In: *Transportation Research Record* 971, pp. 121–130.

Manolis, D. et al. (2018). “Centralised versus decentralised signal control of large-scale urban road networks in real time: a simulation study”. In: *IET Intelligent Transport Systems* 12.8, pp. 891–900.

Mariotte, G. and Leclercq, L. (2019). “Flow exchanges in multi-reservoir systems with spillbacks”. In: *Transportation Research Part B: Methodological* 122, pp. 327–349.

Mariotte, G., Leclercq, L., and Laval, J. A. (2017). “Macroscopic urban dynamics: Analytical and numerical comparisons of existing models”. In: *Transportation Research Part B: Methodological* 101, pp. 245–267.

Mauro, V. and Di Taranto, C (1990). “Utopia”. In: *IFAC Proceedings Volumes* 23.2, pp. 245–252.

Mazloumian, A., Geroliminis, N., and Helbing, D. (2010). “The spatial variability of vehicle densities as determinant of urban network capacity”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368.1928, pp. 4627–4647.

Meng, Q. and Qu, X. (2013). “Bus dwell time estimation at bus bays: A probabilistic approach”. In: *Transportation Research Part C: Emerging Technologies* 36, pp. 61–71.

Mercader, P., Uwayid, W., and Haddad, J. (2020). “Max-pressure traffic controller based on travel times: An experimental analysis”. In: *Transportation Research Part C: Emerging Technologies* 110, pp. 275–290.

Mesbah, M., Sarvi, M., and Currie, G. (2011). “Optimization of transit priority in the transportation network using a genetic algorithm”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.3, pp. 908–919.

Miandoabchi, E., Farahani, R. Z., and Szeto, W. Y. (2012). “Bi-objective bimodal urban road network design using hybrid metaheuristics”. In: *Central European Journal of Operations Research* 20.4, pp. 583–621.

Mirchandani, P. and Head, L. (2001). “A real-time traffic signal control system: architecture, algorithms, and analysis”. In: *Transportation Research Part C: Emerging Technologies* 9.6, pp. 415–432.

Mohajerpoor, R. et al. (2020). “H∞ robust perimeter flow control in urban networks with partial information feedback”. In: *Transportation Research Part B: Methodological* 137, pp. 47–73.

Murashkin, M. (2021). *The influence of trip length distribution on urban traffic in network-level models.* Tech. rep. EPFL.

Nagatani, T. (2007). “Vehicular traffic through a sequence of green-wave lights”. In: *Physica A: Statistical Mechanics and its Applications* 380, pp. 503–511.

Ni, W. and Cassidy, M. (2020). “City-wide traffic control: modeling impacts of cordon queues”. In: *Transportation research part C: emerging technologies* 113, pp. 164–175.

Noaeen, M. et al. (2021). “Real-time decentralized traffic signal control for congested urban networks considering queue spillbacks”. In: *Transportation research part C: emerging technologies* 133, p. 103407.

NRC (2010). *Highway capacity manual.* Transportation Research Board, The National Academy of Sciences.

Paipuri, M. and Leclercq, L. (2020). “Bi-modal macroscopic traffic dynamics in a single region”. In: *Transportation research part B: methodological* 133, pp. 257–290.

Paipuri, M. et al. (2021). “Empirical observations of multi-modal network-level models: Insights from the pNEUMA experiment”. In: *Transportation Research Part C: Emerging Technologies* 131, p. 103300.

Papageorgiou, M., Hadj-Salem, H., Blosseville, J.-M., et al. (1991). “ALINEA: A local feedback control law for on-ramp metering”. In: *Transportation research record* 1320.1, pp. 58–67.

Papageorgiou, M. and Kotsialos, A. (2002). “Freeway ramp metering: An overview”. In: *IEEE transactions on intelligent transportation systems* 3.4, pp. 271–281.

Papageorgiou, M. et al. (2003). “Review of road traffic control strategies”. In: *Proceedings of the IEEE* 91.12, pp. 2043–2067.

Pisinger, D. and Ropke, S. (2019). “Large Neighborhood Search”. In: *Handbook of Metaheuristics.* Springer, pp. 99–127.

Ramezani, M., Haddad, J., and Geroliminis, N. (2015). “Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control”. In: *Transportation Research Part B: Methodological* 74, pp. 1–19.

Ramezani, M. and Nourinejad, M. (2018). “Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach”. In: *Transportation Research Part C: Emerging Technologies* 94, pp. 203–219.

Ren, Y. et al. (2020). “Data driven model free adaptive iterative learning perimeter control for large-scale urban road networks”. In: *Transportation Research Part C: Emerging Technologies* 115, p. 102618.

Ropke, S. and Pisinger, D. (2006). “An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows”. In: *Transportation science* 40.4, pp. 455–472.

Saberi, M. and Mahmassani, H. S. (2012). “Exploring properties of networkwide flow–density relations in a freeway network”. In: *Transportation research record* 2315.1, pp. 153–163.

Saeedmanesh, M. and Geroliminis, N. (2016). “Clustering of heterogeneous networks with directional flows based on “Snake” similarities”. In: *Transportation Research Part B: Methodological* 91, pp. 250–269.

Shalaby, A. S. (1999). “Simulating performance impacts of bus lanes and supporting measures”. In: *Journal of transportation engineering* 125.5, pp. 390–397.

Shaw, P. (1998). “Using constraint programming and local search methods to solve vehicle routing problems”. In: *International conference on principles and practice of constraint programming.* Springer, pp. 417–431.

Sirmatel, I. I. and Geroliminis, N. (2017). “Economic model predictive control of large-scale urban road networks via perimeter control and regional route guidance”. In: *IEEE Transactions on Intelligent Transportation Systems* 19.4, pp. 1112–1121.

Sirmatel, I. I. and Geroliminis, N. (2018). “Economic model predictive control of large-scale urban road networks via perimeter control and regional route guidance”. In: *IEEE Transactions on Intelligent Transportation Systems* 19.4, pp. 1112–1121.

Sirmatel, I. I. and Geroliminis, N. (2021). “Stabilization of city-scale road traffic networks via macroscopic fundamental diagram-based model predictive perimeter control”. In: *Control Engineering Practice* 109, p. 104750.

Sirmatel, I. I. et al. (2021). “Modeling, estimation, and control in large-scale urban road networks with remaining travel distance dynamics”. In: *Transportation Research Part C: Emerging Technologies* 128, p. 103157.

Stirzaker, C. and Dia, H. (2007). “Evaluation of transportation infrastructure management strategies using microscopic traffic simulation”. In: *Journal of Infrastructure Systems* 13.2, pp. 168–174.

Sun, J. et al. (2020). “Managing bottleneck congestion with incentives”. In: *Transportation research part B: methodological* 134, pp. 143–166.

Sun, X. and Wu, J. (2017). “Combinatorial optimization of bus lane infrastructure layout and bus operation management”. In: *Advances in Mechanical Engineering* 9.9, p. 1687814017703341.

Tassiulas, L. and Ephremides, A. (1990). “Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks”. In: *29th IEEE Conference on Decision and Control.* IEEE, pp. 2130–2132.

Toth, G. (2007). “Reducing Growth in Vehicle Miles Traveled: Can We Really Pull It Off?” In: *Driving Climate Change.* Elsevier, pp. 129–142.

Tsitsokas, D., Kouvelas, A., and Geroliminis, N. (2019a). “An optimization framework for exclusive bus lane allocation in large networks with dynamic congestion”. In: *98th Annual Meeting of the Transportation Research Board (TRB 2019).* The National Academies of Sciences, Engineering, and Medicine, pp. 19–02738.

Tsitsokas, D., Kouvelas, A., and Geroliminis, N. (2019b). “Modeling and optimization of dedicated bus lane network design under dynamic traffic congestion”. In: *8th Symposium of the European Association for Research in Transportation (hEART 2019).* ETH Zurich.

Tsitsokas, D., Kouvelas, A., and Geroliminis, N. (2021). “Modeling and optimization of dedicated bus lanes space allocation in large networks with dynamic congestion”. In: *Transportation Research Part C: Emerging Technologies* 127, p. 103082.

Tsitsokas, D., Kouvelas, A., and Geroliminis, N. (2022a). “Critical node selection method for efficient max-pressure traffic signal control in large-scale congested networks”. In: *10th Symposium of the European Association for Research in Transportation (hEART 2022)*.

Tsitsokas, D., Kouvelas, A., and Geroliminis, N. (2022b). “Efficient Max-Pressure traffic management for large-scale congested urban networks”. In: *101st Annual Meeting of the Transportation Research Board (TRB 2022)*. The National Academies of Sciences, Engineering, and Medicine.

Tsitsokas, D., Kouvelas, A., and Geroliminis, N. (2022c). “Two-layer adaptive signal control framework for large-scale dynamically-congested networks: Combining efficient Max-Pressure with Perimeter Control”. In: *Transportation Research Part C: Emerging Technologies*.

Varaiya, P. (2013a). “Max pressure control of a network of signalized intersections”. In: *Transportation Research Part C: Emerging Technologies* 36, pp. 177–195.

Varaiya, P. (2013b). “The max-pressure controller for arbitrary networks of signalized intersections”. In: *Advances in Dynamic Network Modeling in Complex Transportation Systems*. Springer, pp. 27–66.

Verbas, İ. Ö., Mahmassani, H. S., and Hyland, M. F. (2015). “Dynamic assignment-simulation methodology for multimodal urban transit networks”. In: *Transportation Research Record* 2498.1, pp. 64–74.

Viegas, J. and Lu, B. (2001). “Widening the scope for bus priority with intermittent bus lanes”. In: *Transportation Planning and Technology* 24.2, pp. 87–110.

Waterson, B., Rajbhandari, B, and Hounsell, N. (2003). “Simulating the impacts of strong bus priority measures”. In: *Journal of Transportation Engineering* 129.6, pp. 642–647.

Webster, F. V. (1958). *Traffic signal settings*. Tech. rep.

Wei, B. et al. (2020). “Modeling and managing ridesharing in a multi-modal network with an aggregate traffic representation: A doubly dynamical approach”. In: *Transportation Research Part C: Emerging Technologies* 117, p. 102670.

Wei, L. and Chong, T. (2002). “Theory and practice of bus lane operation in Kunming”. In: *DISP-The Planning Review* 38.151, pp. 68–72.

Wirasinghe, S. et al. (2013). “Bus rapid transit–a review”. In: *International Journal of Urban Sciences* 17.1, pp. 1–31.

Wongpiromsarn, T. et al. (2012). “Distributed traffic signal control for maximum network throughput”. In: *2012 15th international IEEE conference on intelligent transportation systems*. IEEE, pp. 588–595.

Wu, J. et al. (2017). “Delay-based traffic signal control for throughput optimality and fairness at an isolated intersection”. In: *IEEE Transactions on Vehicular Technology* 67.2, pp. 896–909.

Xiao, N. et al. (2015a). “Further study on extended back-pressure traffic signal control algorithm”. In: *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, pp. 2169–2174.

Xiao, N. et al. (2015b). “Throughput optimality of extended back-pressure traffic signal control algorithm”. In: *2015 23rd Mediterranean Conference on Control and Automation (MED)*. IEEE, pp. 1059–1064.

Yang, K., Zheng, N., and Menendez, M. (2017). “Multi-scale perimeter control approach in a connected-vehicle environment”. In: *Transportation research procedia* 23, pp. 101–120.

Yao, J. et al. (2012). “Combinatorial optimization of exclusive bus lanes and bus frequencies in multi-modal transportation network”. In: *Journal of Transportation Engineering* 138.12, pp. 1422–1429.

Yildirimoglu, M., Ramezani, M., and Geroliminis, N. (2015). “Equilibrium analysis and route guidance in large-scale networks with MFD dynamics”. In: *Transportation Research Procedia* 9, pp. 185–204.

Yildirimoglu, M., Sirmatel, I. I., and Geroliminis, N. (2018). “Hierarchical control of heterogeneous large-scale urban road networks via path assignment and regional route guidance”. In: *Transportation Research Part B: Methodological* 118, pp. 106–123.

Yu, B. et al. (2015). “A bi-level programming for bus lane network design”. In: *Transportation Research Part C: Emerging Technologies* 55, pp. 310–327.

Zaidi, A. A., Kulcsár, B., and Wymeersch, H. (2015). “Traffic-adaptive signal control and vehicle routing using a decentralized back-pressure method”. In: *2015 European Control Conference (ECC)*. IEEE, pp. 3029–3034.

Zhang, F. et al. (2018). “A systematic analysis of multimodal transport systems with road space distribution and responsive bus service”. In: *Transportation Research Part C: Emerging Technologies* 96, pp. 208–230.

Zhang, R. et al. (2012). “Traffic routing guidance algorithm based on backpressure with a trade-off between user satisfaction and traffic load”. In: *2012 IEEE Vehicular Technology Conference (VTC Fall)*. IEEE, pp. 1–5.

Zhang, X. and Yang, H. (2004). “The optimal cordon-based network congestion pricing problem”. In: *Transportation Research Part B: Methodological* 38.6, pp. 517–537.

Zhao, J. et al. (2019). “Exclusive bus lane network design: A perspective from intersection operational dynamics”. In: *Networks and Spatial Economics* 19.4, pp. 1143–1171.

Zheng, N. and Geroliminis, N. (2013). “On the distribution of urban road space for multimodal congested networks”. In: *Transportation Research Part B: Methodological* 57, pp. 326–341.

Zheng, N. et al. (2012). “A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model”. In: *Transportation Research Part A: Policy and Practice* 46.8, pp. 1291–1303.

Zhong, R. et al. (2018). “Boundary conditions and behavior of the macroscopic fundamental diagram based network traffic dynamics: A control systems perspective”. In: *Transportation Research Part B: Methodological* 111, pp. 327–355.

Zhou, Z. et al. (2016). “Two-level hierarchical model-based predictive control for large-scale urban traffic networks”. In: *IEEE Transactions on Control Systems Technology* 25.2, pp. 496–508.

Zockaie, A., Saberi, M., and Saedi, R. (2018). “A resource allocation problem to estimate network fundamental diagram in heterogeneous networks: Optimal locating of fixed measurement points and sampling of probe trajectories”. In: *Transportation Research Part C: Emerging Technologies* 86, pp. 245–262.

Curriculum vitae

Dimitrios Tsitsokas

📫 Chemin du Croset 1, 1024 Ecublens, ☎ +41 21 693 2431,
✉ dtsitsokas@gmail.com, in linkedin.com/in/dimitrios-tsitsokas/, Dimitrios Tsitsokas

RESEARCH INTERESTS

Network traffic modeling, simulation and control; traffic signal control; intelligent transportation systems; mathematical programming; combinatorial optimization; metaheuristic algorithms

EDUCATION

11.2017 – present **Ph.D. Civil & Environmental Engineering**

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

- Thesis: Large-scale traffic signal control and multimodal network design
- Supervisor: Prof. Nikolas Geroliminis, Co-supervisor: Dr. Anastasios Kouvelas

09.2014 – 06.2016 **M.Sc. in Civil Engineering: Intelligent Transportation and Construction Management Systems**

University of Patras, Greece

- GPA: 9.79 / 10 – 'Excellent'
- Thesis: Multi-objective optimization for the resource-constraint project scheduling problem using evolutionary algorithms (Grade: 10/10)

09.2008 – 04.2014 **Diploma in Civil Engineering (M.Eng., 5-year joint degree)**

University of Patras, Greece

- GPA: 8.65 / 10 - 'Excellent' (2nd of class, top 2% of all graduates of past 10 years)
- Major: Structural Engineering
- Thesis: Experimental study of a three-story masonry-infilled Reinforced Concrete frame seismically retrofitted with Textile-Reinforced Mortar (TRM) (Grade: 10/10)

EXPERIENCE

11.2017 – present **Doctoral Assistant, EPFL**

- Enhancement of network traffic model incorporating traffic signal settings and link length in calculation of travel time
- Development of traffic simulator for research applications (coded from scratch)
- Formulation of optimization problem for dedicated bus lanes location assignment considering dynamic traffic conditions and queue spill-backs
- Testing of various metaheuristic optimization strategies (Large Neighborhood Search, Adaptive-LNS, Variable Neighborhood Search, Local Search)
- Development of hierarchical traffic-responsive signal control framework for congested networks combining centralized perimeter control and distributed control
- Microscopic simulation experiments for testing of different network control applications using Aimsun software
- Analysis of 200k detailed vehicle trajectories for model validation purposes
- Contributed to the development of MOOC "Intro to traffic flow modeling and intelligent transportation systems" (online setup, development of exercises and automatically-graded assessments)

- Involved in teaching activities of BSc course “Transportation Systems Engineering I”
- Supervision of 7 BSc and MSc semester projects and 1 MSc thesis
- Part of the organization team of international conference ISTTT in Lausanne (2019)

06.2021 – 07.2021 **Visiting researcher, ETH Zurich**

- Hosted by IVT – SVT (Dr. A. Kouvelas)
- Development of hierarchical adaptive traffic signal control (perimeter control part)

11.2016-08.2017 **Corporal of Engineering Arm (Compulsory military service), Greek Army**

- Overseeing construction works of new military facilities,
- Making topographical plans for military purposes.

TEACHING EXPERIENCE

Fall 2018 -Fall 2021 **Teaching Assistant for BSc course (EPFL): CIVIL351 - *Transportation Systems Engineering I***

- Instructor: Prof. Nikolas Geroliminis

Fall 2018 -today **Supervision of students (EPFL)**

- Kanae Iizuka (2022) - Perimeter control and Dynamic lane control of transit priority lanes in urban networks: A microsimulation analysis (semester project)
- Kanae Iizuka (2022) - Dynamic Control of transit priority lanes in urban networks: A microsimulation analysis (semester project)
- Antille Yoann, Siwar Othmane (2020) - Evaluation of the impact of dynamic transit priority lanes in urban networks through microsimulation (semester project)
- Jonas Jaeggi (2020) - Dynamic Transit Priority Lanes in Urban Networks (master thesis)
- Jonas Jaeggi (2019) - Exploiting the impact of dynamic transit priority lanes in urban networks via microsimulation (semester project)
- Maximo Jara (2019) - Dynamic Bus Lanes (semester project)
- Younes Bensaid (2019) - Optimization of dedicated bus lane space allocation: An application of Variable Neighborhood Search (semester project)
- Adrien Nicolet, Fabien Jacot-Descombes (2018) - An optimization framework for exclusive bus lanes allocation in large networks with dynamic congestion (semester project)

Fall 2018 **Contributed to the development of online course (EPFL): *Into to Traffic Flow Management and Intelligent Transportation Systems***

- Preparation of exercises related to fundamentals of traffic flow and traffic signal scheduling
- Digitalization of course content and uploading/structuring of all the material on edx.org platform
- Following forum and replying to students’ questions

Spring 2016 – Spring 2017 **Teaching Assistant for BSc course (University of Patras): *Transportation Infrastructure Management***

- Instructor: Prof. Athanasios Chassiakos
- Supervising MSc students in semester projects involving evolutionary optimization algorithms (Genetic Algorithms, Ant Colony Optimization, Swarm Intelligence)

PUBLICATIONS

Peer-reviewed journal articles

- **Tsitsokas, D**., Kouvelas, A., & Geroliminis, N. (2021). Modeling and optimization of dedicated bus lanes space allocation in large networks with dynamic congestion. *Transportation Research Part C: Emerging Technologies*, 127, 103082.
- Sirmatel, I. I., **Tsitsokas, D**., Kouvelas, A., & Geroliminis, N. (2021). Modeling, estimation, and control in large-scale urban road networks with remaining travel distance dynamics. Transportation Research Part C: Emerging Technologies, 128, 103157.
- **Tsitsokas, D**., Kouvelas, A., Geroliminis, N. (2022). Two-layer adaptive signal control framework for large-scale dynamically congested networks: Combining efficient Max-Pressure with Perimeter Control. *Transportation Research Part C: Emerging Technologies* (under review)

Conference papers

- **Tsitsokas, D**., Kouvelas, A., & Geroliminis, N. (2019). An optimization framework for exclusive bus lane allocation in large networks with dynamic congestion.
  Parts of this work presented in:
  - Swiss Transport Research Conference (STRC 2018), Ascona, Switzerland
  - Swiss Transport Research Conference (STRC 2019), Ascona, Switzerland
  - *98th Annual Meeting of the Transportation Research Board (TRB 2019)* (pp. 19-02738). The National Academies of Sciences, Engineering, and Medicine., Washington DC, USA
- **Tsitsokas, D**., Kouvelas, A., & Geroliminis, N. (2019). Modeling and optimization of dedicated bus lane network design under dynamic traffic congestion.
  Parts of this work presented in:
  - 8th Symposium of the European Association for Research in Transportation (hEART 2019), Budapest, Hungary
  - Swiss Transport Research Conference (STRC 2020), Ascona, Switzerland
  - 9th Symposium of the European Association for Research in Transportation (hEART 2020), Lyon, France
- **Tsitsokas, D.,** Kouvelas, A., and Geroliminis, N. "Efficient Max-Pressure traffic management for large-scale congested urban networks".
  Parts of this work presented in:
  - Swiss Transport Research Conference (STRC 2021), Ascona, Switzerland
  - *101st Annual Meeting of the Transportation Research Board* (TRB 2022). The National Academies of Sciences, Engineering, and Medicine, Washington DC, USA
- **Tsitsokas, D.,** Kouvelas, A., & Geroliminis, N. Critical node selection method for efficient max-pressure traffic signal control in large-scale congested networks. In *10th Symposium of the European Association for Research in Transportation (hEART 2022)*, Leuven, Belgium
- **Tsitsokas D.,** Kouvelas A. and Geroliminis N. "Two-layer adaptive signal control framework for large-scale networks combining efficient Max-Pressure and Perimeter Control".
  Parts of this work presented in:
  - Swiss Transport Research Conference (STRC 2022), Ascona, Switzerland
  - ISTRC22 - 2nd Israeli Smart Transportation Research Center (ISTRC) Annual Conference (2022), Haifa, Israel
  - *102nd Annual Meeting of the Transportation Research Board* (TRB 2023). The National Academies of Sciences, Engineering, and Medicine (2023) **,** Washington DC, USA

REVIEWING

Conferences	TRB Annual Meeting, hEART, STRC, IEEE ITS
Journals	Transportation Research Part C: Emerging Technologies

PERSONAL SKILLS

Technical Skills:

Software	Matlab, Aimsun, Mathematica, Minitab, Origin, MS Project 2013, Palisade - The Decision Tools Suite , SAP2000, AutoCAD, AutoCAD Architecture, Microsoft Office Suite
Programming	Matlab, Python, C++, C, Fortran, Gurobi, Yalmip Github, LaTeX

Spoken Languages:

Greek:	Native speaker
English:	Proficient - Level C2 (CPE, *University of Cambridge, 2014)*
French:	Proficient – Level C2 (*DELF – 2eme Degré, 2005)*
German:	Basic – Level A2 (currently taking courses)

AWARDS AND SCHOLARSHIPS

- **Limmat Stiftung (2014) "Award of Academic Excellence"** - Award for achieving the 2nd highest GPA in graduating year of 2014 in the Department of Civil Engineering, University of Patras, Greece
- **State Scholarships Foundation of Greece (I.K.Y.)** - Award for the 2nd highest GPA in class during the 2nd year of undergraduate studies (Academic year 2010-2011)
- **State Scholarships Foundation of Greece (I.K.Y.)** - Award for achieving the 1st place in entering the Department of Civil Engineering, University of Patras (1st out of 150 successful candidates), during National (Panhellenic) Exams (Academic Year 2008-2009)
- **Eurobank EFG (2008)** - Award for graduating from Senior High School (1st General Lykeio of Aigio) with the highest overall grade in the class of 2008