Modelling and Optimization of Demand Responsive Systems and Urban Congestion

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. 1234 2011
présenté le 07 Avril 2014
à la Faculté de L'Environnement Naturel, Architectural et Construit
Laboratoire de systèmes de transports urbains
programme doctoral en Génie Civil et Environnement
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Burak Boyacı

acceptée sur proposition du jury:

Prof François Golay, président du jury
Prof Nikolas Geroliminis, directeur de thèse
Prof António Pais Antunes, rapporteur
Prof Michel Bierlaire, rapporteur
Prof Kostantinos G. Zografos, rapporteur

Lausanne, EPFL, 2014

A person who never made a mistake
never tried anything new.
— Albert Einstein

To Belkıs . . .

Acknowledgements

Every single day in our life, we learn new things. PhD may be one of the periods this process boosts. I have learned a lot during my PhD and I am sure I will benefit from them in my whole life. Although sometimes it was challenging and tedious, I have never felt regret to do a PhD. It was most probably because of the people who helped and supported me through this process. It would not have been possible to write this thesis without these people around me, to only some of whom it is possible to give mention here.

I would like to express my sincere gratitude to my thesis supervisor Nikolas Geroliminis for his guidance, encouragement and help throughout this research. He is not only a great scientist and professor but also a good friend. I should thank him for his support not only for my thesis but also for his help in desperate situations I have encountered. I have learned a lot from him and his guidance will be always with me, in my whole life. I feel extremely lucky to know him and work with him.

I also would like to thank Konstantinos Zografos. Not only for taking part in my thesis committee but also for co-authoring on car-sharing works presented in Part II. He is a person that I have learned a lot and I will in the future. I think, to know him and the opportunity to work with him are the two of the greatest things I have gained with this PhD.

I am grateful to the rest of my thesis committee: Michel Bierlaire, António Pais Antunes and François Golay, for interesting and valuable discussion and, their honest and constructive comments on my thesis. Their suggestions helped to improve the manuscript a lot. They gave me ideas of new research directions for the future.

I think without the good mood at the working place, it would not have been possible to successfully conclude my PhD. I felt really lucky for taking part in the Urban Transport Systems Laboratory. I firstly want to thank to our lovely, caring and helpful secretary Christine. She helped me a lot to survive in Lausanne. I would like to thank all my colleagues, in my lab: Mehmet, Nan, Mohsen, Jack, Kostas, Yuxuan, Reza. It has been a pleasure to know you, work and spend time with you. I have benefited a lot from the friendly, relaxing but motivating environment. In particular, I want to thank Mehmet not only for helping me with printing and delivering the manuscript when I am away but also for his everlasting friendship. He is a friend that I want to have in my whole life. Knowing him is one of the great gains of this PhD. I also want to thank Martin for translating my abstract into French.

In addition to the friends in the lab, I had chance to meet great people in Lausanne. I specially thank to Barış not only for hosting me in his flat during my visits to Lausanne but also for interesting ski trips, BBQs and many other things. I feel like he is a friend that I have for a long

time. I also want to thank Çağla, Özcan, Suat, Okan, Bilge, Oğuz and the other members of Turkish community in Lausanne for great time we spent with and support they provided. I also in particular want to thank Stefan Minniberger and Daniel Rüffer for their friendship.
I am also grateful to my family, my parents and my brother, for their continuous support and endless love for me not only throughout my PhD but my whole life. Starting with my high school, I was usually away from them but they were always there for me to listen, help and support.
Last but not least, I want to thank my love, Belkıs for her endless support. During the PhD, although we were in separate continents, thousands of kilometers away from each other, I always felt she was next to me, supporting, encouraging and helping me. Thank you for proving me distance means so little when someone means so much. Thank you for everything we had together and we will have in the future.

Lausanne, 30 April 2014 B. B.

Abstract

This study is motivated by planning on-demand transportation systems in large scale urban networks. We specifically handle emergency response systems and car-sharing services in congested urban areas. Our ultimate aim is to improve the quality of service for both emergency response and car-sharing systems with the support of operational research tools.

In the first part of the thesis, we deal with a method that can evaluate the performance of spatial queueing systems with different server-customer service rates. More precisely, we propose two new spatial queueing models, which utilize service rates that are the functions of both the server providing the service and the customer receiving it. These two new models can be regarded as two extensions to the known hypercube queueing models. These models have a lot of application areas including improvement of emergency (e.g. ambulance, police, emergency repair) and transportation (taxicabs, on-demand transportation, para-transit) services in cities.

The first contribution of this part of the thesis to the literature is developing two hypercube models that apply different service rates according to the distance between server and the customer. In order to keep problems tractable, we assume that there are two different service rates for each server: The service rate when a server is serving its own region (i.e. intradistrict service) is different when the same server is serving outside its own region (i.e. interdistrict service). The second contribution is proposing an approximation algorithm for the problems that are intractable because of their size. We also test both methods inside some efficient heuristics to show the applicability of the two methods with algorithms for real case problems.

In the second part of the thesis, we work on improving the services in (non-floating) one-way (electric) car-sharing systems. We regard this problem in three different levels: Strategic, tactical and operational. Strategic decisions are regarded as the decisions related to the infrastructure (e.g. location and size of the stations) and tactical decisions are about the vehicles and the personnel (e.g. fleet and personnel size); whereas operational decisions are problems related to daily operations (e.g. relocation operations, personnel shift assignments). The first two levels of the decisions (i.e. strategic and tactical) are taken into consideration together in the first chapter of this part. A multi-objective (different objectives are applied for the users and the operator of the system) mixed integer linear programming formulation with its relaxation is proposed and solved for different scenarios. The model is applied on a real case, a car-sharing service in the city of Nice.

In the second chapter of this part, the operational problem of one-way car-sharing systems are handled. A mixed integer linear programming formulation is proposed for the problem

which decides on initial vehicle locations, relocation operations and assignment of relocation personnel to the shifts. In order to keep the decisions robust, flexible and applicable, some extra soft constraints are added. We aim to improve flexibility of the service from the users point of view, it is assumed that the users might pick-up and drop-off vehicles earlier or later than their reservations occasionally. In order to cope with these situations, set of soft constraints are introduced to the model that keeps vehicles and empty spots at the right place at the right time. This chapter is still work in progress but is included in the dissertation to represent preliminary results.

In the last part of the thesis, we propose a tool for a parsimonious travel time estimation for the first two parts of the thesis. We mostly benefit from Variational theory and macroscopic fundamental diagram literature. We start with extending the prior work on the travel time estimation works on homogeneous networks and apply similar methods to the systems with heterogeneous system characteristics; i.e. link lengths, offset between traffic signals and incoming turns. The research is conducted for both unimodal and multimodal networks. More specifically, in this part of the research, we explore the effect of network parameters on the two key characteristics of macroscopic fundamental diagram: (i) the network capacity and (ii) the density range when the network capacity is maximum.

Although scarce data do not enable us to do, in the process of improving the tool we have reached to some conclusions that are applicable not only to travel time estimation but also to traffic in urban networks. A closed analytical formulation, that utilizes system characteristics, to calculate the density range of a homogenous network is proposed and proven. Then, the effects of the changes on the system characteristics are investigated for both homogeneous and heterogeneous networks. In addition, the effect of the incoming turns is modeled and its intensity is explored. In an extended research, similar investigations are conducted for the multimodal networks, i.e. networks with public transportation buses. These tools can be utilized for the development of hierarchical control strategies for large scale congested transport networks.

Keywords: emergency response, spatial queues, hypercube queueing models, location models, one-way car-sharing, electric mobility, multi-objective optimization, integer programming, vehicle relocation, scheduling, Variational theory, macroscopic fundamental diagram, simulation, network capacity, multimodal systems

Résumé

La présente étude est motivée par la question de la planification des systèmes de transports à la demande dans les réseaux urbains à grande échelle. Nous nous intéressons particulièrement aux services d'intervention d'urgence et aux systèmes d'autopartage opérant dans les zones urbaines congestionnées. L'objectif final de l'étude est d'améliorer la qualité de service et de réponse de ces deux types de systèmes grâce aux outils de la recherche opérationnelle.

Dans la première partie de cette thèse, nous travaillons sur une méthode qui permet d'évaluer la performance des systèmes avec files d'attentes distribuées spatialement et qui utilisent différents taux de service fournisseur-client. Plus précisément, nous proposons deux nouveaux modèles de files d'attente spatiales dans lesquels les taux de service dépendent à la fois du fournisseur et du bénéficiaire du service. On peut considérer ces deux modèles comme des extensions des modèles de files d'attente hypercube connus. Ces modèles ont des applications dans beaucoup de domaines dont les systèmes d'intervention d'urgence (par exemple pour les ambulances, la police ou encore le dépannage d'urgence) et de transports (par exemple les taxis, les services de transport à la demande ou les services de transport adaptés (para-transit)) en ville. La première contribution à la littérature dans cette partie consiste en le développement de deux modèles hypercube dans lesquels on applique différents taux de service selon la distance entre le fournisseur du service et le client. Pour que les problèmes abordés restent analytiquement solubles et aisément maniables, nous supposons qu'il existe deux taux de service différents pour chaque fournisseur : l'un lorsqu'il fournit le service dans sa zone (c'est-à-dire un service intra district) et l'autre lorsqu'il le fournit en-dehors (c'est-à-dire un service interdistrict). La seconde contribution est de proposer une approche qui permette d'obtenir une solution pour les problèmes qui ne sont pas analytiquement maniables du fait de leur taille. Nous testons ces deux méthodes au travers de plusieurs heuristiques efficaces afin de montrer leur applicabilité pour des algorithmes traitant des problèmes réels.

Dans la seconde partie de cette thèse, nous travaillons à améliorer les systèmes d'autopartage, principalement ceux qui utilisent des véhicules électriques et dont les conditions de fonctionnement sont différentes du free-floating. Nous considérons ce problème à trois niveaux : stratégique, tactique et opérationnel. Les décisions stratégiques concernent l'infrastructure (par exemple la localisation et la taille des stations) ; les décisions tactiques concernent les véhicules et le personnel (par exemple la taille de la flotte de véhicules et le nombre d'employés nécessaire) alors que les décisions opérationnelles concernent le fonctionnement du système au jour le jour (par exemple la redistribution des véhicules ou le tournus et les horaires du personnel). Les deux premiers aspects (stratégique et tactique) sont traités conjointement

dans le premier chapitre de cette partie. Une formulation en programmation linéaire mixte avec variables entières et à fonctions objectif multiples (différentes fonctions objectifs sont appliquées pour l'opérateur du système et l'utilisateur) est proposée et utilisée pour obtenir la solution dans différents scénarii. Le modèle est appliqué à un cas réel, celui d'un service d'autopartage à Nice.

Dans le deuxième chapitre de cette partie, c'est le problème opérationnel pour ces systèmes d'autopartage qui est traité. Une formulation en programmation linéaire mixte avec variables entières est proposée pour le résoudre. Il décide du positionnement initial des véhicules, des opérations de redistribution, de l'emploi du temps ainsi que des missions de redistribution du personnel. Afin de garantir la robustesse, la flexibilité et l'applicabilité des décisions issues du modèle, quelques contraintes supplémentaires sont ajoutées. On cherche à améliorer la flexibilité du point de vue des usagers : il est alors supposé qu'ils puissent occasionnellement prendre ou déposer un véhicule plus tôt ou plus tard que prévu dans leurs réservations. Pour gérer de telles situations, un ensemble de contraintes est introduit dans le modèle afin d'avoir des véhicules et des places de parc libres au bon moment et au bon endroit. Ce chapitre est en cours d'avancement mais est tout de même inclus dans la dissertation pour présenter les résultats préliminaires qui ont été obtenus.

Dans la dernière partie de cette thèse, nous proposons un outil économe en ressources pour estimer les temps de parcours dans le cas des problèmes traités précédemment. Nous nous appuyons principalement sur la littérature existante traitant de la théorie variationnelle et du diagramme fondamental de zone. Nous étendons d'abord les travaux existants sur l'évaluation du temps de trajet dans des réseaux homogènes à des réseaux aux caractéristiques hétérogènes (par exemple longueur des routes et artères, offset entre les feux tricolores ou encore intensité des flux de trafic tournant). Nous menons cette recherche à la fois sur les réseaux multimodaux et sur les réseaux unimodaux, c'est-à-dire comprenant un et un seul mode de transport. Plus spécifiquement, dans cette partie de la recherche, nous évaluons les effets des paramètres du réseau sur les deux caractéristiques clés du diagramme fondamental de zone : la capacité et la gamme de densité atteinte lorsque le réseau est à capacité.

Malgré le peu de données à disposition, nous sommes parvenus, au cours du processus d'amélioration des outils développés, à certaines conclusions qui sont applicables non seulement à l'estimation du temps de parcours mais aussi au trafic dans les réseaux urbains. Une solution en forme close qui repose sur les caractéristiques du système et qui donne la gamme de densité à capacité pour un réseau homogène a été proposée et prouvée. Ensuite, les effets de la modification des caractéristiques du système sont étudiés à la fois pour les réseaux homogènes et les réseaux hétérogènes. En outre, l'effet des flux de trafic tournant est modélisé et leur impact en fonction de leur intensité étudié. La recherche sur ce sujet est étendue aux réseaux multimodaux, particulièrement ceux dans lesquels sont déployés des services de bus réguliers. Les outils développés dans cette partie peuvent être utilisées pour créer des stratégies de contrôle hiérarchisées dans des réseaux de transports congestionnés à grande échelle.

Mots-clés : intervention d'urgence, files d'attentes spatialement distribuées, modèles de file

d'attente hypercube, modèles de localisation, autopartage, mobilité électrique, optimisation multi objectifs, programmation en nombres entiers, redistribution de véhicules, planification, théorie variationnelle, diagramme fondamental de zone, simulation, capacité du réseau, systèmes multimodaux.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context and Motivation

At the beginning of the 20th Century, the world population was around 1.7 billion and not more than 20% of the whole world population was living in the cities. In 2010, first time in history, total urban population has passed the rural population. The 7 billion barrier has reached on October 2011 and in the year 2016, it is expected to have an urban population of 4 billions (UN, 2011). This dramatic population increase in the urban areas increases the pressure on the urban services. In order to keep the level of service in acceptable levels under high demand scenarios, operations need to be planned and handled in the highest efficiency. In this thesis, we are dealing with the modeling and optimization of two different urban systems, emergency response and car-sharing services. Our aim is to increase their efficiency while keeping them economically feasible. The models provided for both systems are applicable to different demand responsive systems. In both models we are dealing with location and dispatching decisions. These systems are beneficial to the society, environmental- and user-friendly.

In the first part of the thesis, we are dealing with spatial queueing models. These models can be regarded as useful tools for modeling distributed urban service systems such as emergency services (e.g. ambulance, police, fire, emergency repair), door-to-door delivery services (e.g. mail and parcel delivery), neighborhood service centers (e.g. clinics, libraries) and transportation services (e.g. bus, subway, taxi and para-transit services) (Larson and Odoni, 1981). More specifically we build our research on extending the prior work on 2^n hypercube queueing model proposed by Larson (1974). 2^n hypercube queueing model assumes that compared to time spent on the scene, travel time is negligible in services. In other words, in this model, for any given server, service time is not effected by the location of the incident. In our research, we proposed two new 3^n hypercube queueing models (one detailed and one aggregated) that enable to use different service rates for different server-incident pairs. This way, effect of distance between the server and the incident is also taken into account in calculating system performance of these systems. The developed framework allows to model and optimize medium to large-scale systems, which would not be possible with the classical

hypercube model.

The second part of the thesis is about (non-floating) one-way (electric) car-sharing systems. Car-sharing is a new model of car rental that enables users to rent for short period of time. Depending on the type of the system, they can be regarded as complementary services to rigid, scheduled public transportation for the society. They contain both the efficiency of the public transportation for the society and the flexibility of owning a vehicle for the users at the same time. It can be regarded as the missing link for continuous choice of people's demand, the way to answer this problem with minimum cost.

In general, there are two types of car-sharing systems mostly used in the system. Two-way and free-floating one-way. In two-way, vehicles should be returned where they have been picked-up. They are easy to maintain by the operators and enable users to do early-reservations. On the other hand, free-floating one-way systems allow users to drop-off vehicles in designated areas. However, early-reservations are not directly applicable since vehicles locations are not predictable. Our recommendation is something that takes strong parts of both systems: reservation capability of two-way systems and drop-off flexibility of non-floating one-way systems. In the system we have proposed, vehicles are picked-up and dropped-off into designated parking areas. Users should state their origin and destination stations and, pick-up and drop-off locations in their reservation requests. The system is designed with the objective to serve a large number of requests while keeping the cost for the operator at reasonable levels through a multi-objective optimization. Designated parking spots are also crucial for electric vehicles since they need to be charged regularly, especially when utilization rates are high.

In Chapter 3, we deal with the strategic (e.g. location, size and count of the stations) and tactical decisions (e.g. fleet and relocation personnel size) of one-way car-sharing systems. In Chapter 4, we handle the operational decisions (e.g. vehicle and personnel relocations) with a model that enables flexibilities to pick-up and drop-off times.

The third part of the thesis (chapters 5 and 6) is related to travel time estimation in congested urban areas. We have mostly benefited from literature related to macroscopic fundamental diagram and Variational theory. We have built on the previous works on Daganzo and Geroliminis (2008) and Geroliminis and Daganzo (2008). These two papers showed the existence of relationship between the density and the flow of urban networks. In addition, the former proposed moving observer method to calculate flow and density in urban networks with homogeneous characteristics. The cuts generated with the help of this method bounds the flow-density diagram from above.

The two main contributions of this part of the thesis is to find a closed formulation for the density range of the maximum capacity of the urban networks. Second, we have applied the method to the networks with heterogeneous system characteristics (i.e. link length, offset and distance between traffic signals). We have also added the effect of left turn to our results. Last but not least, multimodal networks are taken into consideration with different system characteristics. In this part of the thesis, our ultimate aim was to estimate speed with the

help of flow and density. Although, lack of data prevent us to implement such tools for our work, this approach has potential to be applied in the operational research literature. It is also currently integrated for large scale urban control for single mode and multimodal traffic systems.

1.2 Thesis Contributions

We can categorize the contributions of the thesis into three groups: *hypercube queueing models*, *one-way car-sharing systems* and *estimation of network capacity in congested urban systems*.

1.2.1 Hypercube Queueing Models

- We propose two new 3^n hypercube queueing models.

  The previous hypercube queueing model proposed by Larson (1974) assumes that the service rate of a server is not a function of the customer (or the incident) characteristics it is serving. In other words, he assumes that the distance between the customer and the server has no effect on the service rate. We propose two new 3^n hypercube queueing models which enable servers to have different service rates for different customer groups.

- We implement a new approximation algorithm for medium to large size cases to solve 3^n hypercube queueing model.

  The size of the state space of hypercube queueing models grows exponentially. There is a need for an algorithm that can handle these problems. For this purpose, we develop an approximation algorithm that integrates partitioning to smaller (and faster) problems and then merging the larger ones with another 3^n aggregate hypercube queueing model.

- We devise the 3^n hypercube queueing models inside optimization frameworks.

  Hypercube queueing models are descriptive models and only enable to analyze. In this research, we implement them inside optimization frameworks and use them as tools to improve spatial queueing systems.

1.2.2 One-Way Car-Sharing Systems

- We propose a decision support tool for strategic and tactical decisions for one-way electric car-sharing systems.

  In non-floating one-way car-sharing systems, users pick-up and drop-off vehicles into designated parking areas. Different than the two-way systems, users have freedom to park vehicle into a different parking space than the spot they have picked-up. In

this research, we propose a decision support tool to have better strategic and tactical decisions, e.g. locations and capacities of the stations, fleet size and relocation personnel count. A multi-objective mathematical programming formulation containing both the operator's and the users' objective functions is developed to model general electric car-sharing systems. Various scenarios are used in order to keep results robust.

- A mathematical model for operational decisions of one-way electric car-sharing systems is developed.

  We regard daily vehicle and personnel relocations, and personnel assignments to designated shifts as operational decisions. We develop a mixed integer linear programming formulation for these operational decisions. The model takes reservations of the users and layout of the car-sharing system as input and finds the decisions which minimizes total cost of operations. In order to keep outcomes of the model robust for the users' reservations, we add some soft constraints that assumes users' stated pick-up and drop-off times might deviate from the reality. With the help of these soft constraints, we keep vehicles and empty spots at the right place at the right time.

1.2.3 Estimation of Network Capacity in Congested Urban Systems

- We propose and prove new closed formulation for finding the density range of the maximum capacity of the large scale homogeneous urban networks.

  The formulation given in Daganzo and Geroliminis (2008) is altered for the moving observer method. This new formulation is also supported with a set of lemmas and their proofs.

- We expand the method to networks with heterogeneous characteristics and left turns.

  The work of Daganzo and Geroliminis (2008) and Geroliminis and Daganzo (2008) are applied on networks with homogeneous characteristics. We apply similar approaches on networks with heterogeneous characteristics in terms of topology and signal settings. The effect of incoming turns in the network capacity are also investigated.

- The network capacity for multimodal systems that involve service related stops (buses, taxis etc) is also investigated.

  In this research, we investigate the effect of public buses on the urban networks. For different frequency of bus arrivals and bus-stop locations, the change in capacity and density range of the maximum capacity are investigated. While bus operations might decrease the vehicular capacity, there is a significant increase in the passenger flows in a network. These findings can be integrated in an optimization approach for redistribution of urban space among different modes.

1.3 Thesis Outline

We organize the thesis in three parts.

Part I focuses on hypercube queueing models.

Chapter 2 deals with 3^n hypercube queueing models and approximation algorithm including partitioning model. An exact and an approximate algorithms are developed to have steady state properties of 3^n hypercube queueing problems. This chapter is presented as:

- B. Boyacı and N. Geroliminis. Extended hypercube models for large scale spatial queueing systems. In *91th Annual Meeting of the Transportation Research Board*, Washington D.C., 2012c
- B. Boyacı and N. Geroliminis. Facility location problem for emergency and on-demand transportation systems. In *12th Swiss Transport Research Conference*, Monte Veritá, 2012b
- B. Boyacı and N. Geroliminis. Extended hypercube queueing models for stochastic facility location problems. In *25th European Conference on Operational Research (EURO XXV)*, Vilnius, 2012a
- B. Boyacı and N. Geroliminis. Extended hypercube models for location problems with stochastic demand. In *2nd Symposium of the European Association for Research in Transportation*, Stockholm, 2013

This chapter is submitted to the journal of *Tranportation Research Part B: Methodological* for publication.

Part II is related to (non-floating) one-way (electric) car-sharing problems.

Chapter 3 deals with strategic and tactical decisions of generic one-way car-sharing systems. The optimization framework takes demand, current system state (e.g. stations in operation) and cost components (e.g. fixed and variable cost of stations) as an input and returns an efficient frontier of the solutions.

This chapter is submitted to the *European Journal of Operational Research*. It is under the second round reviews. This chapter is also presented as:

- B. Boyacı, N. Geroliminis, and K. Zografos. A generic one-way multi-objective car-sharing problem with dynamic relocation. In *Proceedings of the Eighth Triennial Symposium on Transportation Analysis (TRISTAN VIII)*, San Pedro de Atacama, 2013c
- B. Boyacı, N. Geroliminis, and K. Zografos. Developing and solving an integrated multi-objective model for supporting strategic and tactical decisisions for one-way

car-sharing systems. In *26th European Conference on Operational Research (EURO INFORMS MMXIII)*, Rome, 2013a

- B. Boyacı, N. Geroliminis, and K. Zografos. An optimization framework for the development of efficient one-way car-sharing systems. In *13th Swiss Transport Research Conference*, Monte Veritá, 2013b

Chapter 4 contains the mathematical model for operational decisions (e.g. vehicle and personnel relocation) of generic one-way car-sharing systems. This part of the thesis is still research in progress. Preliminary results are shared in the dissertation.

Part III is about estimation of the network capacity in the congested urban networks.

Chapter 5 contains research on estimation of networks capacity in urban networks with traffic signals. Variational theory and macroscopic fundamental diagrams are used to investigate the capacity and the density range of the maximum capacity for different network characteristics. A preliminary version of this chapter is presented as:

- B. Boyacı and N. Geroliminis. Exploring the effect of variability of urban systems characteristics in the network capacity. In *90th Annual Meeting of the Transportation Research Board*, Washington D.C., 2011b

The same chapter is published as:

- N. Geroliminis and B. Boyacı. The effect of variability of urban systems characteristics in the network capacity. *Transportation Research Part B: Methodological*, 46(10):1607–1623, 2012

Chapter 6 is an extension of work on network capacity to multimodal networks. We add public buses to the network to see the effect of multimodality on the capacity and the density range of the maximum capacity. This chapter is published and presented as:

- B. Boyacı and N. Geroliminis. Estimation of the network capacity for multimodal urban systems. In *6th International Symposium on Highway Capacity and Quality of Service*, 2011a. 6th International Symposium on Highway Capacity and Quality of Service

A journal publication of the chapter is under preparation.

Hypercube Queueing Models Part I

2 Approximation Methods for Large-Scale Spatial Queuing Systems

Different than the conventional queueing systems, in spatial queueing systems (SQS) the service rate for each customer-server pairs differs and the server that intervenes for a specific customer is not known a priori, depending on the availability of servers at the moment a request was made. These features make the SQS computationally expensive (almost intractable for large scale) but at the same time more suitable for real-life problems with high reliability expectations. Emergency response and on-demand transportation systems are two similar systems that can be modeled with the spatial queueing systems.

Emergency response system location problems are one of the first problems immensely dealt in the optimization literature. In most approaches, the instances of this problem are modeled as either set covering or transportation models which disregard stochastic nature of the problem.

In this chapter, we aim to solve facility location problems as SQS with stochastic demand and service time. The stochasticity concerned here is temporal and spatial, that emerges from the uncertainty in the demand and service time. In order to tackle this problem Larson (1974)'s 2^n hypercube queueing model (HQM) is extended to 3^n HQM. In our model, there are two different possible service types for each server: (i) service for locations in the proximity of a server (area of responsibility) and (ii) service for other locations where the first responsible server is busy during this event. In addition, to decrease the dimension of the problem, which is intractable due to their size, a 3^n aggregate hypercube queueing model (AHQM) is also developed that treats group of servers (bins) in a similar manner by considering interactions among bins. An efficient graph partitioning algorithm is proposed to cluster servers in groups with an objective to minimize the interactions among groups. Both exact and approximate approaches are integrated inside two optimization methods (i.e.variable neighborhood search and simulated annealing) to find server locations that improve system performance. Computational experiments showed that both developed models are applicable to use inside optimization algorithms to find good server locations and to improve system performance measures of SQS.

2.1 Introduction

Location-allocation of *emergency response systems* is one of the oldest problems in the operations research literature. Locating ambulances, fire brigades and police-beats were the pioneer problems mathematically modeled and solved. Although there are quite a few number of works on the subject, many of them disregards the stochastic nature of the problem and find solutions with deterministic assumptions. However, this specific property is the one that differs emergency response system location-allocation problems from the other types of location-allocation problems. This randomness (in demand rates, service times and servers' intervention) creates unexpected congestion and eventually causes losses. While stochasticity in demand and service rates have been included in many researches, the choice of the server based on the state of the system (location of request and availability of other servers) has been addressed in only a few instances for small-scale systems.

There are different methods in the literature dealing with locating emergency response systems. One of the models that was proposed by Larson (1974) models this problem as a spatial queueing model which is also known as 2^n hypercube queueing model (HQM). In 2^n HQM, each emergency response unit is regarded as a server on an Euclidean space and each of them has two states, available and busy generating 2^n possible states for the system (where n is the number of servers); these are the vertices of a hypercube.

Larson (1974) assumed that since the time spent on the way to scene is negligible compared to the service time on scene, the region that is served has no effect on the service time, i.e. for a specific server, service time is the same for any region. This may be acceptable for some systems like fire brigades but not for ambulances. In this research, our aim is to alter Larson (1974)'s 2^n HQM in such a way that enables the model to use different service rates for different server-region pairs. For this purpose a new 3^n HQM is proposed. In 3^n HQM, each server has three states: available, busy inside primary service area (intradistrict) and busy outside primary service area (interdistrict), which creates an intractable hypercube for even medium size problems (with more than 8 servers). In order to tackle larger problems an aggregate method, namely 3^n aggregate HQM (AHQM) is also developed. Instead of estimating each server state separately, 3^n AHQM keeps the number of servers at each of the 3 states (i.e. available, busy with intradistrict, busy with interdistrict) at each *bin* (i.e. set of servers). To identify bins, AHQM is integrated inside mix aggregate hypercube queueing algorithm (MHQA). In a nutshell, MHQA bi-partitions the whole problem area into tractable HQM subproblems and merges the solutions of each subproblems' in the reverse order of the bi-partitioning with AHQM. Both methods are used to find better locations for emergency vehicles to improve a real and an experimental regions with the help of two optimization algorithms, the variable neighborhood search (VNS) and the simulated annealing (SA).

The remainder of the chapter is organized as follows: In the next section, Section 2.2, we describe significant literature about locating emergency response systems. We start with the early location-allocation models and extend it to very recent emergency response systems

literature. Section 2.3 describes the two new models, i.e. 3^n HQM and 3^n AHQM, and their comparison with Larson (1974)'s 2^n HQM. Section 2.4 contains the definition and steps of MHQA with the mathematical model of the partitioning algorithm. In Section 2.5, we share the computational results of our two algorithms, 3^n HQM and MHQA. This part contains both the accuracy of the two models compared to the simulation of the real system and the results from the optimization algorithms, VNS and SA. In the last section, we discuss the conclusions and the future research directions.

2.2 Literature Survey

The earliest models dealing with the location of emergency response systems assume deterministic demand and service. They disregard the stochastic nature of the problem and model the problem by median and coverage models.

The first *median problem* was created by Fermat in the 17th Century: Given a triangle, find the median point in the plane such that the sum of the distance from each point of the points to the median point is minimized. Weber (1909) extended the problem with more than three points with weights and objective minimizing total weighted distance. Both Weiszfeld (1937) and Hakimi (1964) proposed methods to solve the *Fermat-Weber problem* optimally, the former gives the optimal location in the Euclidean space whereas the later for networks. Cooper (1963, 1964, 1972) modeled the existing Fermat-Weber problem with more than one facilities and proposed efficient heuristic methods. Calvo and Marks (1973) grouped the population into groups and introduced facilities of these groups. Weaver and Church (1985) proposed the *vector assignment p-median problem* which aims to minimize total transportation cost while forcing the demand nodes to have service from k closest facilities with predefined ratios.

Coverage models are used to locate facilities (i.e. emergency response systems) in such a way to maximize coverage and/or minimize number of facilities. The first two models, the *location set covering problem* (LSCM) aims to minimize number of facilities to cover all demand (Toregas et al., 1971) and the *maximal location set covering problem* (MCLP) aims to maximize total coverage with limited number of facilities (Church and ReVelle, 1974). Schilling et al. (1979) proposed a model that aims to have multiple coverages with different types of facilities. Daskin and Stern (1981) promoted the multiple coverage as a secondary objective. Aly and White (1978), Hogan and Revelle (1986), ReVelle and Hogan (1989), Marianov and ReVelle (1992), Ball and Lin (1993), Gendreau et al. (1997) have extended the works given above in different aspects.

In addition to the two main types of models given above, there are also dynamic models proposed in the literature. The main idea in these models is to relocate the facilities (e.g. ambulances, fire brigades) when one or more of the facilities are dispatched for an incident. Kolesar and Walker (1974) proposed a model for the fire brigades. Recently, with the increase in computational power, relocation model applications for ambulances have also emerged. Gendreau et al. (2001) proposed a parallel tabu search heuristic to solve relocation model

efficiently. Gendreau et al. (2005) and Schmid and Doerner (2010) are the two recent models on dynamic facility location problems for emergency response systems where the latter is the multistage approach of the former one. Andersson and Värbrand (2006) proposed a decision support tool for a similar aim.

Larson (1974) proposed a *hypercube queueing model* (HQM) which is the first model that embeds the queueing theory in facility location-allocation literature. This model analyzes systems such as emergency services, door-to-door pickup and delivery services, neighborhood service centers and transportation services which has response district design and service-to-customer mode (Larson and Odoni, 1981). The solution of the model provides state probabilities and other related system performance measures (e.g. workload, average service rate, loss rate) for any given server locations. Nevertheless, it is a descriptive model that only allows to analyze scenarios Galvão and Morabito (2008). HQM models the current configuration as a continuous-time Markov process and does not determine the optimal configuration.

The first model proposed by Larson (1974) assumed that the service time is not a function of the locations of the calls for service and the dispatched unit. This argument was supported with the claim that the time spend on the way to scene is negligible compared to the service time on scene. Since, even with this simplification, as the number of servers (n) is increased, number of states grows exponentially; Larson (1975) proposed a heuristic method. Jarvis (1985) altered this heuristic for the systems with both server and customer dependent service times. Atkinson et al. (2008) proposed a partial 3^n model in which each region takes service from two servers with different service rates.

Takeda et al. (2007) showed the benefits of decentralization of emergency response systems with the help of hypercube models in Brazil. Iannoni and Morabito (2007); Iannoni et al. (2008) embedded hypercube in a genetic algorithm framework to locate emergency vehicles along a highway. They extended the problem to enable multiple dispatch (i.e. more than one server may be needed for an incident). Geroliminis et al. (2009, 2011) integrated the location and districting decisions in the same optimization and solve the problem by using steepest descent and genetic algorithm respectively.

As it is stated before, there is an extensive literature on location and coverage literature. The more interested readers can apply to Hale and Moberg (2003), Owen and Daskin (1998), Brotcorne et al. (2003) and Laporte et al. (2009) for broader survey for the related topic. The former two are surveys generally on location and coverage literature whereas the latter two focus more on the ambulance location-relocation models.

2.3 Hypercube Queueing Models

This section provides a description of the existing 2^n hypercube model of Larson (1974) and formulates two new models to deal more accurately with interactions of servers.

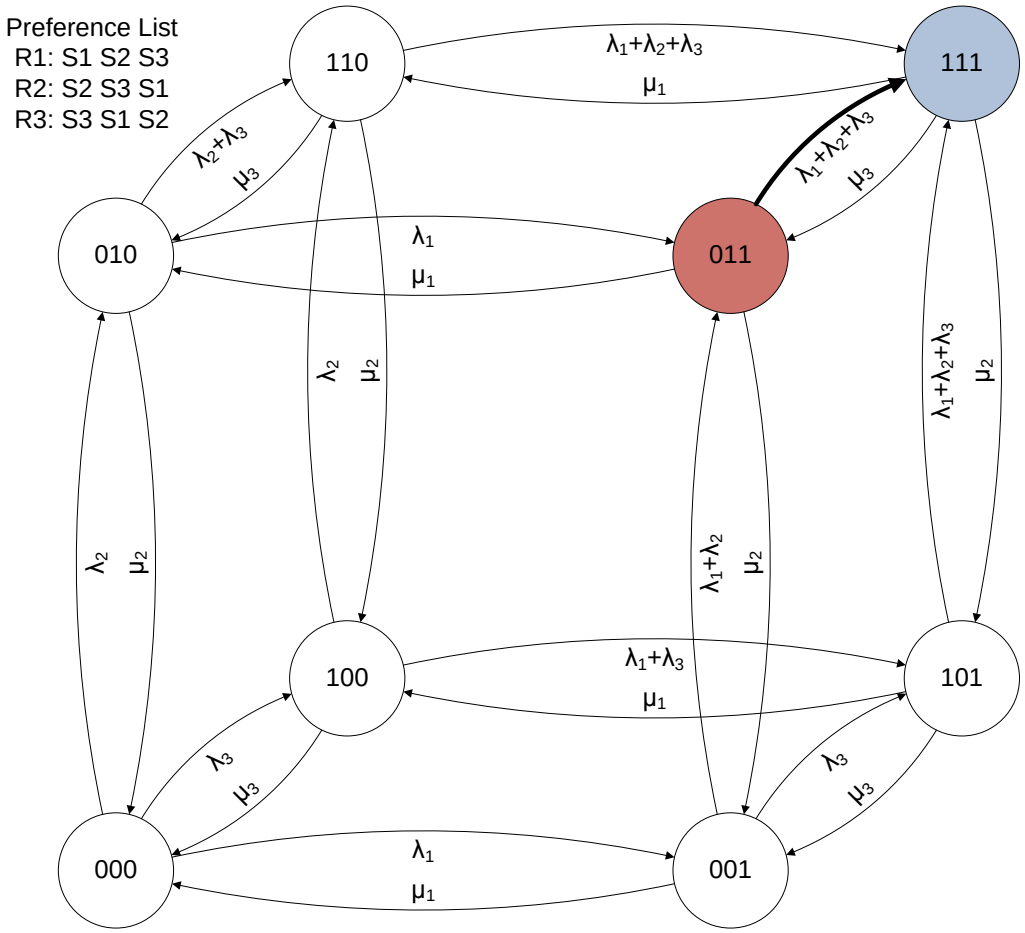


Figure 2.1: Larson (1974)'s 2^n HQM for three servers with equal intra and interdistrict service rates (μ_i). State "011", "111" and the transition connecting them is shown with different colors.

2.3.1 A Note on 2^n Hypercube Queueing Model

The HQM proposed by Larson (1974) includes *hypercube* in the name since the transition graph of the continuous time Markov chain representing this queueing structure has a hypercube structure when the number of servers is more than three. The state variables contain n binary variables for n servers showing if server i is available (0) or busy (1). Each state is a number in base 2 and each digit shows the state of the corresponding server. For each region, which is called *atom* (j) in HQM literature, there exists a priority list of servers. Incidents at each atom are dispatched to the available server with the highest priority for this atom. If there does not exist any available server that can serve the atom, either the call is lost (i.e. call for ambulance may be dispatched by a backup system) or joins a queue to be served (i.e. customers are asked to wait until there is an available server) depending on the assumptions. Service requests arrive from each atom according to an independent Poisson process with rate λ_j and servers have exponential service rates of μ_i for any atom served. The transition graph of 2^n HQM with three servers can be seen in Figure 2.1. It is seen on that simple graph that as the system gets

congested, the burden on the free server(s) increases. For instance in state "011" all the servers but the third are busy. That is why the next incident in any region will be served by the third server and transition rate from "011" to "111" is $\lambda_1 + \lambda_2 + \lambda_3$. Such a model does not consider different service rates for inter and intradistrict responses.

2.3.2 3^n Hypercube Queueing Model

In this research, we first develop the transition equations for a 3^n HQM. If different rates for intra and interdistrict responses are applied, each server has three possible states: available (0), busy with intradistrict (1) and busy with interdistrict (2). Figure 2.2 is a transition graph of a system of the same type with three servers. Given the large number of transitions and states the illustration for a system with more servers would be difficult to visualize. For instance "210" represents the state in which first server is available, the second intervenes an incident inside its own region and the third intervenes an incident outside its own region (state reads from right to left). Since there are three possibilities for each server, the number of states also increases to 3^n.

Note that, the server has always priority for the incidents inside its own intradistrict area. When the system is empty, the first assignment should be an intradistrict assignment. However, this does not prevent having states such as "222". Although, practically rare for lightly congested systems, it has a non-negative probability in all systems.

The general transition equation for the states of 3^n HQM can be stated as:

$$\begin{aligned} &\mathbb{P}_r \left[\mathbb{1}\left(\exists i : T(r,i) = 0\right) \sum_j \lambda_j + \sum_{i:T(r,i)=1} \mu_i + \sum_{i:T(r,i)=2} \mu'_i \right] \\ = &\sum_{q,i:D(q,r,i)=1} \mathbb{P}_q \mu_i + \sum_{q,i:D(q,r,i)=2} \mathbb{P}_q \mu'_i + \sum_{q,i:D(r,q,i)=1} \mathbb{P}_q \sum_{j \in R_i} \lambda_j \\ &+ \sum_{q,i:D(r,q,i)=2} \mathbb{P}_q \sum_{j \in S(r,i)} \lambda_j \end{aligned} \tag{2.1}$$

in which q and r are states which can be regarded as numbers in base 3 and, $\mathbb{P}_q$ is the steady-state probability of state q. i and j represent servers and atoms respectively. $T(q,i)$ is the condition of server i in state q (i.e. i^{th} digit of q), R_i is the set of atoms in the intradistrict area of server i, $S(q,i)$ is the set of atoms that have interdistrict response from server i if there is an incident during state q which is generated by priority lists. $\mathbb{1}(*)$ is an indicator function. It is equal to 1, if $*$ is true and 0 otherwise. $D(q,r,i)$ is another function defined as:

$$D(q,r,i) = \begin{cases} c, & \text{if } d(q,r) = 1, T(r,i) = 0, T(q,i) = c \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

where $d(q,r)$ is the Hamming distance between states q and r (i.e. minimum number of transitions to reach from q to r). $D(q,r,i)$ simply shows state pairs with the same server

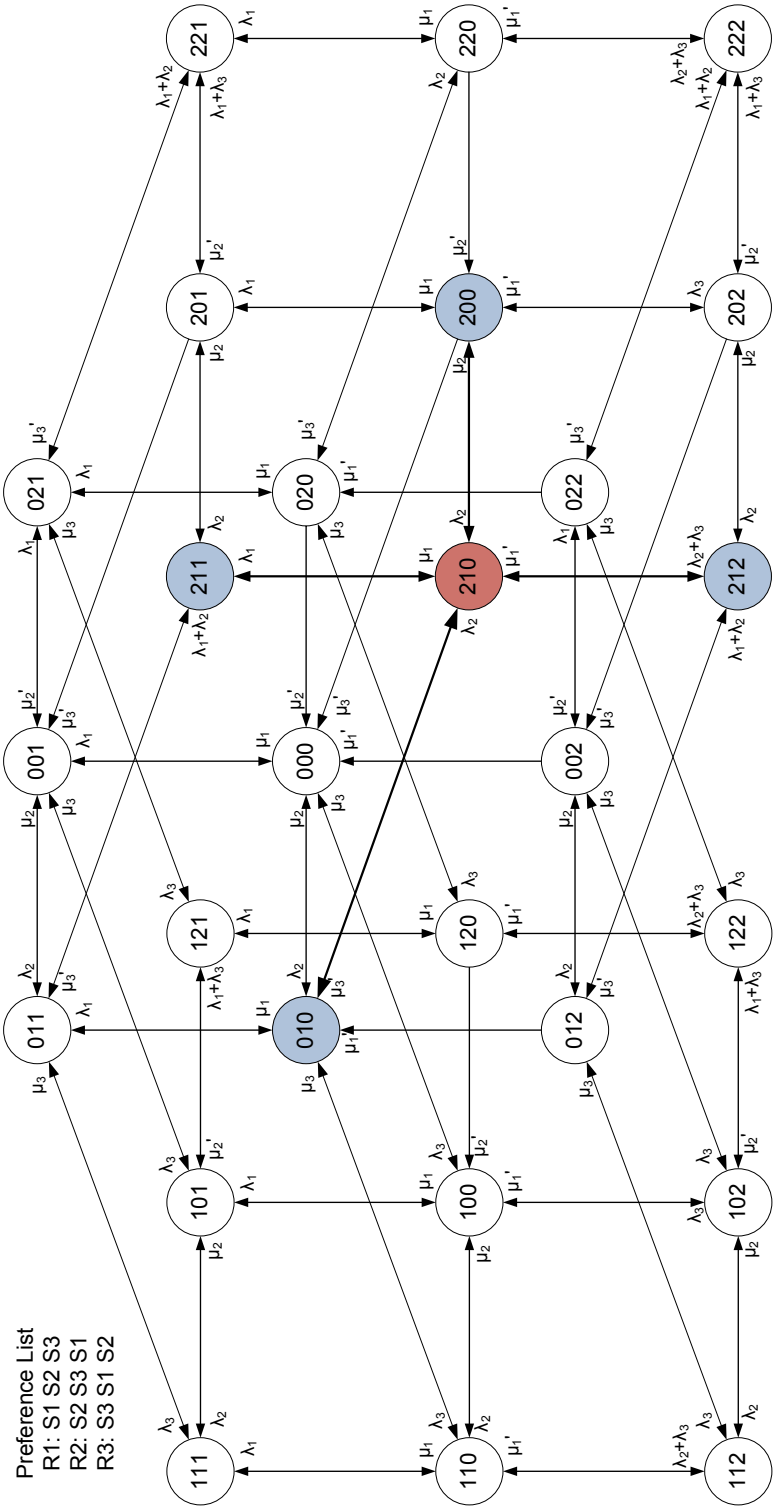


Figure 2.2: 3^n HQM model for two servers with different intra (μ_i) and interdistrict (μ_i') service rates. State "210", states directly connected to it and transitions are colored differently.

conditions except server i. If server i is available in state r and busy in state q, condition of the server in state q is the output of the function. Memoryless arrivals and service rates are simplifying the size of transitions as only states with Hamming distance equal to 1 are connected. Such an approximation is reasonable for different types of queueing systems. Real data can further investigate this assumption.

In building Equation 2.1, LHS is equal to the rate of leaving state r whereas RHS is equal to the rate of entering state r. On the LHS of Equation 2.1, the total rate leaving state r is multiplied with steady state probability of r (i.e. $\mathbb{P}_r$). To keep the formulation simpler, we assume that, every atom is reachable by any server. In addition, we are formulating a system without queue, incident from any atom is lost if the system is full. In other words, upper transition (i.e. transition to a state with one more busy server) is possible only if there exists an available server. That is the reason, the term $\exists i : T(r,i) = 0$ exists in front of the term$\sum_j \lambda_j$.

The second and third summations multiplied with $\mathbb{P}_r$ in the LHS of Equation 2.1 are for lower transitions (i.e. transition to a state with one less busy server). The transition rate differs if a server is in intradistrict ($T(r,i) = 1$) or interdistrict ($T(r,i) = 2$) respond.

The RHS of the Equation 2.1 is composed of transitions to state r from the states that are one Hamming distance away. The first two summation terms are from upper transitions. They show transitions from an upper state q after an intradistrict (first term) or interdistrict (second term) service. The former is multiplied with μ_i and the latter is with μ'_i.

The last two summations on the RHS of 2.1 are from lower transitions to state r. If the condition of server i in state r is intradistrict service, then there exists a state q that forms a transition to state r. Note that, in state q and r all servers have the same condition except state i. Server i is available in state q and busy with intradistrict in state r. The rate of the transition is the total arrival rate for the atoms in the intradistrict area of server i. This is formulated with the third summation on the RHS of 2.1. Similarly, if the condition of server i in state r is interdistrict service, then we can state there is a transition between state q and r in which the only difference between two states is the condition of server i: It is available in state q and busy with interdistrict in state r. The fourth summation deals with the cases in which server i is the available server with the highest priority for atom j in state q and atom j is not in the intradistrict area of server i. This is formalized with the last summation on the RHS of 2.1.

To give an illustrative example, the transition equation for red painted state "210" in Figure 2.2 can be written as:

$$\begin{aligned}\mathbb{P}_{210}\left(\lambda_1+\lambda_2+\lambda_3+\mu'_3+\mu_2\right)\\ = \mu_1\mathbb{P}_{211}+\mu'_1\mathbb{P}_{212}+\lambda_2\mathbb{P}_{200}+\lambda_2\mathbb{P}_{010}\end{aligned} \tag{2.3}$$

2.3.3 3^n Aggregate Hypercube Queueing Model

Although 3^n HQM is more accurate than 2^n, the increase in the number of states is more and not applicable for real life cases. For instance a system with 20 servers needs more than three billion states in 3^n HQM. In order to cope with that, we develop a 3^n aggregate HQM (AHQM). In this new model, a new concept called *bin* is used to represent servers. It is assumed that, each bin (b) has a capacity as it consists of a group of servers and each state consists of 2 values for each bin, which show the number of busy servers with intra and interdistrict responses at each bin. For intra and interdistrict service rates μ_b and μ'_b for bin b and demand rate λ_j for atom j, transition diagram for two bins with two servers in each bin can be depicted as seen in Figure 2.3. Note that, each row of the state name shows condition of each bin. Given that the number of servers per bin (capacity) is known, the state of each bin is described with only two values. The values on the left and right are number of servers occupied by intra and interdistrict responses respectively.

The general transition diagram for 3^n AHQM can be written as:

$$\begin{aligned}
&\mathbb{P}_r \Bigg[\mathbb{1}\left(\exists n : \tilde{T}(r,b,\text{free}) \neq 0\right) \sum_j \lambda_j + \sum_b \tilde{T}(r,b,\text{intra})\mu_b \\
&+ \sum_b \tilde{T}(r,b,\text{inter})\mu'_b \Bigg] = \sum_{\substack{q,b: \\ \tilde{D}(q,r,b,\text{intra})=1}} \mathbb{P}_q \mu_b + \sum_{\substack{q,b: \\ \tilde{D}(q,r,b,\text{inter})=1}} \mathbb{P}_q \mu'_b \\
&\quad + \sum_{\substack{q,b: \\ \tilde{D}(r,q,b,\text{intra})=1}} \mathbb{P}_b \sum_{j \in R_b} \lambda_j + \sum_{\substack{q,b: \\ \tilde{D}(r,q,b,\text{inter})=1}} \mathbb{P}_b \sum_{j \in S(r,b)} \lambda_j
\end{aligned} \tag{2.4}$$

in which in addition to definitions used in Equation 2.1, b is index for bins. $\tilde{T}(r,b,*)$ shows number of servers in bin b in the condition of $*$ (i.e. "inter", "intra", "free") in state r and $\tilde{D}\left(q,r,b,*\right)$ can be defined as:

$$\tilde{D}\left(q,r,b,*\right) = \begin{cases} 1, & \text{if } d(q,r) = 1, \tilde{T}(q,b,*) = \tilde{T}(r,b,*) + 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

where $*$ stand for conditions "inter" and "intra".

We can see how transition equations are generated on a specific state. In Figure 2.3, in state "10/01", there is a busy server in bin 1 with intradistrict response which is shown with the first line of the state name. In bin 2 there is also a busy server which is in interdistrict response. The transition equation for this state can be written as Equation 2.6. Note that, in Figure 2.3,

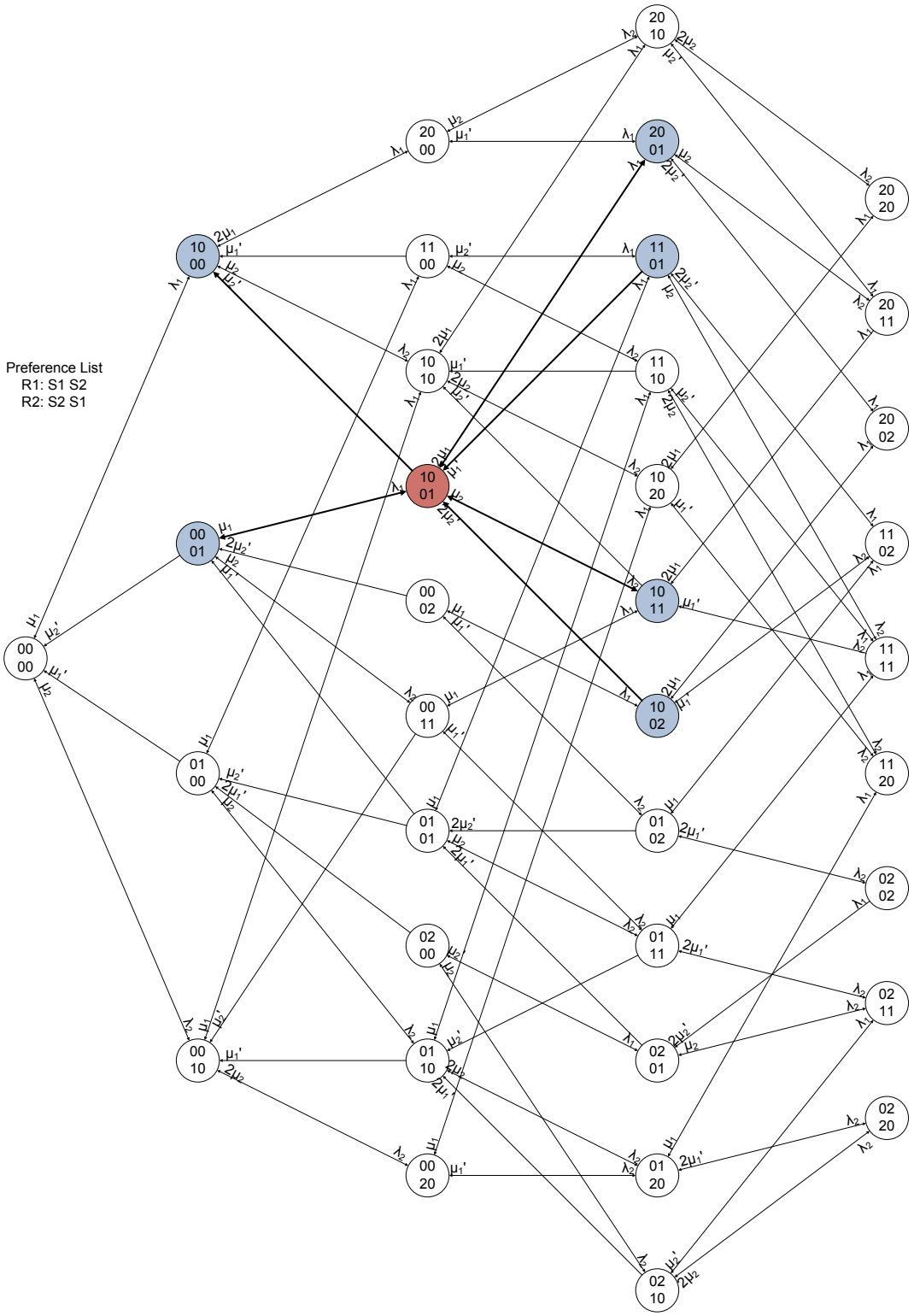


Figure 2.3: 3^n AHQM for two bins containing two servers in each bin with different intra (μ_b) and interdistrict (μ'_b) service rates, and primary demand areas (λ_j). State "10/01" and states connected to it are filled with different colors to show an example of transition equations.

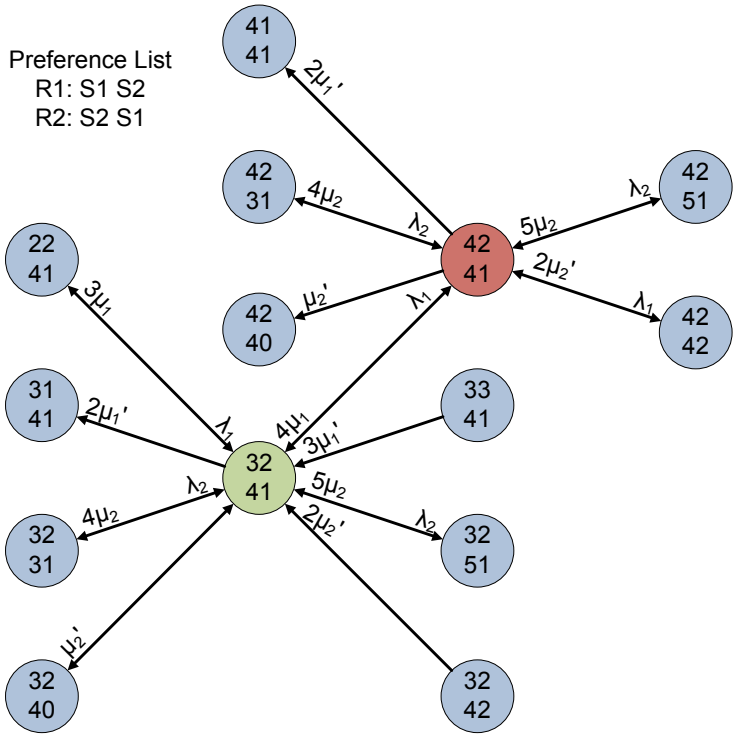


Figure 2.4: Part of the transition diagram of a 3^n AHQM for two bins containing six servers in each bin with different intra (μ_b) and interdistrict (μ'_b) service rates, and primary demand areas (λ_j). States "32/41" (green), "42/41" (red) and their neighbor states (blue) are depicted to help to visualize transition equations of the former two states.

state "10/01" and its neighbor states are colored with red and blue respectively:

$$\begin{aligned}&\mathbb{P}_{\substack{10\\01}}\left(\lambda_1+\lambda_2+\mu_1+\mu'_2\right)\\&\qquad=2\mu_1\mathbb{P}_{\substack{20\\01}}+\mu_2\mathbb{P}_{\substack{10\\11}}+\mu'_1\mathbb{P}_{\substack{11\\01}}+2\mu'_2\mathbb{P}_{\substack{10\\02}}+\lambda_1\mathbb{P}_{\substack{00\\01}}\end{aligned}\tag{2.6}$$

As for the case of 3^n HQM a visualization of all states in a figure for a large number of servers is difficult. To further clarify with a more complex example, in Figure 2.4, some specific states of a 3^n AHQM with two bins of 6 servers each can be seen. The complete transition diagram of the system in Figure 2.4 has 784 states. We just depict two states "32/41" (green), "42/41" (red) and their neighbor states (blue) to show how transition equations are calculated. For states "32/41" and "42/41", transition equations can be written as follows:

$$\begin{aligned}&\mathbb{P}_{\substack{32\\41}}\left(\lambda_1+\lambda_2+3\mu_1+4\mu_2+2\mu'_1+\mu'_2\right)\\&=4\mu_1\mathbb{P}_{\substack{42\\41}}+5\mu_2\mathbb{P}_{\substack{32\\51}}+3\mu'_1\mathbb{P}_{\substack{33\\41}}+2\mu'_2\mathbb{P}_{\substack{32\\42}}+\lambda_1\mathbb{P}_{\substack{22\\41}}+\lambda_2\mathbb{P}_{\substack{32\\31}}\end{aligned}\tag{2.7}$$

$$\begin{aligned}\mathbb{P}_{\substack{42\\41}}\left(\lambda_1+\lambda_2+4\mu_1+4\mu_2+2\mu_1'+\mu_2'\right)\\ =5\mu_2\mathbb{P}_{\substack{42\\51}}+2\mu_2'\mathbb{P}_{\substack{42\\42}}+\lambda_1\mathbb{P}_{\substack{32\\41}}+\lambda_2\mathbb{P}_{\substack{42\\31}}\end{aligned} \tag{2.8}$$

For a 3^n AHQM, if C_b is the maximum number of servers in bin b, total number of states equals $\prod_b \frac{(C_b+2)(C_b+1)}{2}$. For this estimation, each bin can be regarded independently in this calculation. The number of servers at each bin is separated into three groups: available, busy with intradistrict, busy with interdistrict. Since the number of states each bin is independent, the total number of states of each bin can be multiplied with each other.

As described above, 3^n AHQM is a solution for larger spatial systems. It applies different service rates for intra and interdistrict responses, and has less number of states: $\prod_b \frac{(C_b+2)(C_b+1)}{2}$ is far less than 3^n. For most of the cases with two bins, this value is even less than 2^n (i.e. for the cases with 8 or more servers). For instance, the system with 16 servers has 65536 states in 2^n and over 43 million states in 3^n HQM whereas a 3^n AHQM of two bins with 8 servers each has only 2025 states. In the next section, we describe the MHQA that utilizes the two new 3^n models, i.e. 3^n HQM and 3^n AHQM, defined in the current section.

2.4 Mix Aggregate Hypercube Queueing Algorithm

The exponential increase in the number of states makes 3^n HQM not applicable to real life instances. For this purpose, we propose 3^n AHQM which has less states. However, the way 3^n AHQM is applied, is also important for the efficiency of the method and accuracy of the results. In this section, the details of this procedure will be described. Simply, the method we propose contains an iterative approach that partitions the whole problem area into *subregions* which is followed by an iterative solve for each partition and merge scheme for the pairs of partitions. In our approach during the different steps of partitioning and merging we consider interactions between groups of servers, i.e. "bins", that are important especially in systems with many busy servers. In the following two subsections firstly the iterative solution procedure and then the partitioning algorithm are described. An illustration of the procedure is provided in Figure 2.5. Figure 2.5a shows the whole region with the location of all servers (red dots). Dark color represents atoms of high demand and lighter color the ones of lower demand. Figure 2.5b shows the primary areas of responsibilities of each server estimated with a Voronoi approach based on Euclidean distance. Figures 2.5b-d shows the results of the sequential partitioning method which is described later in more details.

After sequential partitioning we end up with a number of core subregions, see for example the outcome of Figure 2.5e. These subregions are modeled as 3^n HQMs. With 3^n HQMs, any needed performance measures can be calculated. Service rate for the number of servers, availability of each server, loss rate of each atom and percentage of time each server spends at intra or interdistrict responses are found for the algorithm. Then, in the sequence of merging,

an inverse process of partitioning is followed to estimate performance measures for the whole area of study (moving from Figure 2.5e to Figure 2.5a now).

The core subregions are merged to larger *compounds* subregions, which are modeled as 3^n AHQM. Compound subregions are also merged to larger subregions with 3^n AHQM until the whole area is covered. Note that at each merging step, only two subregions (core or compound) are merged to a larger compound region. For intradistrict service rate of each bin, the service rate calculated from the previous step is used. Loss rate of each atom in sibling subregion (i.e. subregions which are the pairs of each other in sequential partitioning) and availability of each servers are used to calculate interdistrict service rates. We assume that servers are busy most of the time within their core region responses. The availability of a server can be assumed to be proportional to one minus the occupancy in its core region. Since, in the next steps, all servers are merged and regarded as servers inside bins, we use the availability of each server in a bin to calculate the probability if an interdistrict call can be served or not. The formulation we have conducted for this purpose can be seen in the next subsection.

2.4.1 Bin Interactions

If a 3^n HQM is applied for each partition of Figure 2.5D without considering any interactions, the system performance measures would be consistently less accurate. In our experiments, we see that the interaction between bins is 5-50% (see Figure 2.11). However, this relationship also needs a correction, we cannot assume every atom can be served by any bin. Since, bins are not physically at the same location, it is not possible to assign a single location for them. There is no straightforward way to calculate distance between an atom and a bin. However, we need the distance between the bin and the atom to decide if the atom can be served by this bin or not. In order to cope with that, we assume, with some probability, some atoms might not be served even though there is at least some servers available in a bin of the neighbor subregion.

It is easy to illustrate how we apply methodology on a toy example before a formal mathematical notation is introduced, shown in Figure 2.6. Assume that, all servers in region 2 are busy. Consider now 3 servers from region 4 (servers 14, 16 and 17) and the areas within their accessibility distance (shown with the 3 circles). Let us assume that a new request for service arrives from an atom located in the dark or light blue squares, close to the boundary of the two regions. In such a state, if an incident occurs in light and dark blue atoms, these incidents can either be served with servers in region 4 or they are lost. There is only one server (i.e. server 14) that can serve light blue atom and three servers (servers 14, 16 and 17) that can serve dark blue atom from region 4. In the AHQM, servers from the same bin cannot be differentiated after merging. We can only calculate number of servers available or busy with either intra or interdistrict responses at each state. So, in order to approximate the probability of being served, we use the formulation given in Equation 2.9. For instance, if all 6 servers available in region 4, we can assume that both light and dark blue atoms can be served by the servers in region 4, as there is always a server available that can reach both atoms. If there are only

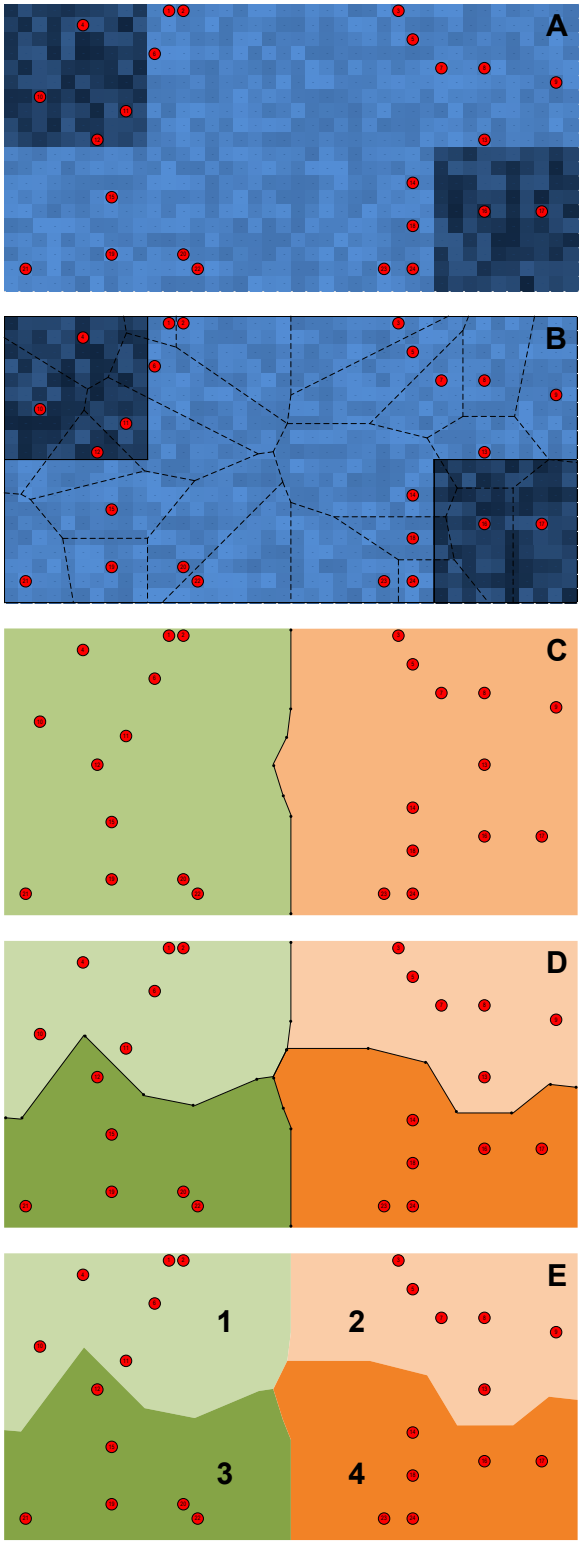


Figure 2.5: An illustration of the partitioning approach.

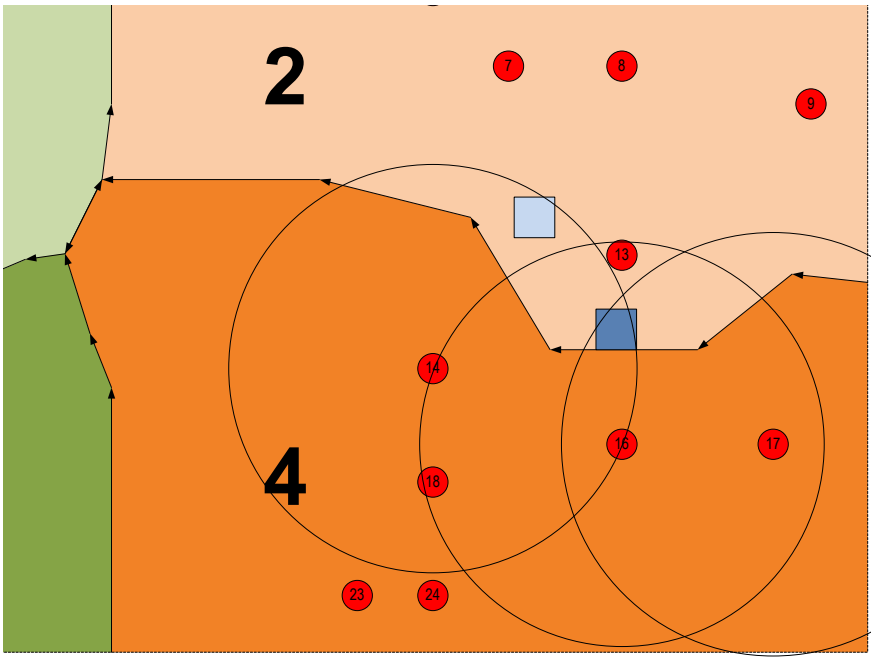


Figure 2.6: An illustration for bin interactions. Dark and light blue squares represent demand requests and red circles show the server locations in two subregions 2 and 4. Larger circles centering servers show their accessibility area. If there is no available server in region 2, for an incident happening in light and dark blue atoms, we need to use servers in region 4.

4 servers available in region 4, we can still assume dark blue atom is served with probability 1 because in the worst case, one of the three servers that can reach dark blue atom should be available. Nevertheless, if there are not more than 3 servers available, then with some probability both light and dark blue regions cannot be served. But, we can state that, dark blue region can be served with a higher probability because light blue atom can be reached by only server 14, whereas dark blue atom is reachable by servers 16 and 17 in addition to 14.

Let us assume there are n servers in a bin, first m of them can reach to the atom of the sibling subregion and k servers are busy. When $m = 0$ the probability of serving the atom by a server from this bin equals 0. When $k < m$ the incident in this atom is served with probability 1, since even in the worst case, there has to be a server available to serve it. However if $k \geq m$, the probability that none of the available servers can reach the atom is approximated as:

$$\mathbb{P}(\text{not served}) = \frac{\prod_{i=1}^{m} \mathbb{P}_i \left[\sum_{\substack{\forall L \in \mathscr{P}(\mathbb{N}_{m+1}^{n}): \\ |L|=k-m}} \left(\prod_{\forall i \in L} \mathbb{P}_i \right) \right]}{\sum_{\substack{\forall L \in \mathscr{P}(\mathbb{N}_{1}^{n}): \\ |L|=k}} \left(\prod_{\forall i \in L} \mathbb{P}_i \right)} \tag{2.9}$$

where $\mathscr{P}(*)$ represents the power set of $*$, $\mathbb{N}_a^b$ is an inclusive sequence of integers starting from a to b and $\mathbb{P}_i$ is the probability that server i is busy. In Equation 2.9, the denominator is equal to the sum of the all cases' probabilities with k busy servers. The numerator is equal to the sum of the probability of cases where all servers that can reach the atom are busy. As a result Equation 2.9 gives conditional probability that, there does not exist any available server which can reach the atom given k busy servers. Note that, the same combination term ($\mathbb{C}_k^m$) canceled out each other both in the numerator and the denominator.

Equation 2.9 is an approximate value for the probability of an incident happening in the selected atom does not get an interdistrict service even though there is an available server in the bin. In order to calculate loss rate caused by this phenomena, the demand that is not served in the previous steps of the algorithm is multiplied with the probability in Equation 2.9 and assumed to be lost. Total demand of the atom minus calculated loss rate is assigned as the demand of the atom. This is done for each atom and the new 3^n AHQM is modeled with these demand values. After solving the model, loss rate calculated by 3^n AHQM and assumed loss rates are summed up and assigned as total loss rates of each atom.

An informal pseudocode for our algorithm is given in Figure 5.3 where $\bar{T}_l^*(n)$ and $\bar{R}_l^*(n)$ stands for average service time and rate for bin composed of servers of subregion l for the intra and interdistrict responses. $T(i,j)$ is the total service time needed for a response to atom j by server i. J_l are the atoms in subregion l and, intra_i and inter_i are atoms in intra- (i.e. atoms closest to server i) and interdistrict (i.e. atoms serviceable but not closest to server i) area of server i. If a 3^n HQM was solved for each partition without considering any interactions, the system performance measures would be consistently less accurate for semi-congested systems (medium to high demand).

2.4.2 Partitioning Model

As stated before, the size of 3^n HQM grows exponentially with the number of servers and is applicable only for problems with limited number of servers. In order to cope with that, we develop an aggregate approach that devices both 3^n HQM and 3^n AHQM. In order to have accurate results in efficient time we need a partitioning algorithm that partitions the whole problem area to subregions with an objective to minimizes the total demand that is served by more than one bins. Even if in Section 2.4.1 we develop a framework to deal with interactions, given that AHQM does not keep track of individual servers, this estimation contains some level of error. For fixed server locations, by minimizing interaction between subregions' (i.e. minimizing services provided by servers of other subregions), we decrease the error of approximation algorithm. The partitioning algorithm will also determine if for some instances of specific problems (e.g. with no common regions), HQMs can be solved independently. We should also fulfill the following properties:

1. The number of servers in each partition should be the parameter of the partitioning algorithm.

1. Partition the whole problem area into subregions. A binary tree structure is created with partitioning. Leaves of the binary tree are core subregions, and the rest of the nodes are compound subregions composed of two (core or compound) subregions.

2. Iterate each node with depth first search

   - If selected subregion l is core subregion

   (a) Calculate average intra and interdistrict service time $\left(\bar{T}_i^{\text{intra}}, \bar{T}_i^{\text{intra}}\right)$, and calculate average service rates $\left(\bar{R}_i^{\text{intra}}, \bar{R}_i^{\text{intra}}\right)$ for each server i in selected core subregion:
   $\bar{T}_i^{\text{intra}} \leftarrow \frac{\sum_{j \in \text{intra}(i)} d_i T(i,j)}{\sum_{j \in J_l \cap \text{intra}(i)} d_i}, \bar{T}_i^{\text{inter}} \leftarrow \frac{\sum_{j \in J_l \cap \text{inter}(i)} d_i T(i,j)}{\sum_{j \in \text{inter}(i)} d_i}, \bar{R}_i^{\text{intra}} \leftarrow \frac{1}{\bar{T}_i^{\text{intra}}}, \bar{R}_i^{\text{inter}} \leftarrow \frac{1}{\bar{T}_i^{\text{inter}}}$

   (b) Solve 3^n HQM with $\bar{R}_i^{\text{intra}}$, $\bar{R}_i^{\text{inter}}$ and d_j where the former two stand for inter and intradistrict response rates for each server i and latter for demand rate of each atom j in the selected core subregion l.

   (c) Calculate probability of server i busy (P_i), loss rate of atom j (loss_j), and average service rate for the number of servers busy $\bar{S}_l^{\text{intra}}(n)$ where $n \in \mathbb{N}$.

   - If selected subregion l is a compound subregion with children regions l_1 and l_2 /*Assume bins b_1 and b_2 are composed of servers of subregions l_1 and l_2 respectively.*/

   (a) Calculate average interdistrict service times ($\bar{S}_l^{\text{inter}}$) of bins in child subregions of l:
   $\bar{S}_{l_1}^{\text{inter}} \leftarrow \frac{\sum_{i \in I_{l_1}} \sum_{j \in J_{l_2} \cap \text{inter}(i)} \text{loss}_j T(i,j)}{\sum_{i \in I_{l_1}} \sum_{j \in J_{l_2} \cap \text{inter}(i)} \text{loss}_j}, \bar{S}_{l_2}^{\text{inter}} \leftarrow \frac{\sum_{i \in I_{l_2}} \sum_{j \in J_{l_1} \cap \text{inter}(i)} \text{loss}_j T(i,j)}{\sum_{i \in I_{l_2}} \sum_{j \in J_{l_1} \cap \text{inter}(i)} \text{loss}_j}$

   (b) Calculate average interdistrict service rates of each child subregion l:
   $\bar{R}_l^{\text{inter}}(n) \leftarrow \frac{1}{\bar{S}_l^{\text{inter}}} \quad \forall l = \{l_1, l_2\}$ and $n \in \mathbb{N}$.

   (c) Generate a 3^n AHQM with given intra and interdistrict service rates of each subregion (or equivalently bin) $\bar{R}_l^{\text{intra}}(n)$ and $\bar{R}_l^{\text{inter}}(n)$ respectively for $l = \{l_1, l_2\}$.

   (d) For each state, find the loss rate because of server unavailability $\left(\text{loss}'_j\right)$ by multiplying loss_j and Equation 2.9. Update demand of each atom with $\left(d_i - \text{loss}'_j\right)$ for corresponding states.

   (e) Solve generated 3^n AHQM.

   (f) Calculate loss rate of each atom j with the equation: $\text{loss}_j \leftarrow \text{loss}_j^{\text{AHQM}} + \text{loss}'_j$.

   (g) Calculate average service rate for the number of servers busy $\bar{S}_l^{\text{intra}}(n)$ for $n \in \mathbb{N}$

3. Return the values calculated for the subregion in the root node (which is equivalent to the whole problem area).

Figure 2.7: Pseudocode for the MHQA

We need to set the number of servers in each subregion. There is a maximum size that is efficiently solvable with both hypercube models and over partitioning (i.e. using more partitions than needed) decreases the accuracy of the final result.

2. Servers in the same partition should be adjacent to each other. This prevents disconnected atoms and helps to have connected subregions which improves accuracy of the method.

3. Partitioning should be sequential in order to apply the approximation algorithm.

4. Partitioning should be efficient. Our aim in developing an approximation algorithm is to evaluate instances in an optimization framework. To do that, we need efficient algorithms in all steps of the evaluation.

For this purpose, we have developed an algorithm that generates "cuts" on the problem area and creates subregions. This algorithm first utilizes a Voronoi diagram for server locations. Afterwards, one or more network flow problems are solved on the line segments of the Voronoi diagram. Flows in these problems start from and end at the vertices on the borders and flows on the inner edges of Voronoi diagram. The set of flows are regarded as the cuts we need to create subregions. To obtain a partitioning with the objectives stated above, for the following indices and sets:

$i \in I$	servers
$j \in J$	atoms
$l \in L$	partitions
$v \in V$	vertices on Voronoi diagram
$e \in E$	edges on Voronoi diagram
$k \in K$	paths on Voronoi diagram

parameters:

$V / \bar{V}$	all/outer vertices on Voronoi diagram
E_v	edges connected to vertex v
I_e^1 / I_e^2	servers that are separated by edge e
J_i	atoms that are accessible by server i
d_j	weight of atom j
G^{kl}	$\begin{cases} 1 & \text{if partition } l \text{ is on the selected side of path } k \\ 0 & \text{otherwise} \end{cases}$
V_e	vertices connected to edge e
S^l	number of servers in partition l

and variables:

s_v^k	number of times an inner vertex v is visited in path k
s_v^k	$\begin{cases} 1 & \text{if outer vertex } v \text{ exists in path } k \\ 0 & \text{otherwise} \end{cases}$
z_e^k	$\begin{cases} 1 & \text{if edge } e \text{ exists in path } k \\ 0 & \text{otherwise} \end{cases}$

$$x_i^k \quad \begin{cases} 1 & \text{if server } i \text{ is on the selected side of path } k \\ 0 & \text{otherwise} \end{cases}$$

$$q_i^l \quad \begin{cases} 1 & \text{if server } i \text{ belongs to partition } l \\ 0 & \text{otherwise} \end{cases}$$

$$y_j^l \quad \begin{cases} 1 & \text{if atom } j \text{ is accessible by one of the servers} \\ & \text{that belongs to partition } l \\ 0 & \text{otherwise} \end{cases}$$

$$t_v^k \quad \begin{cases} 1 & \text{if vertex } v \text{ exists in path } k \\ 0 & \text{otherwise} \end{cases}$$

we develop a binary integer programming problem (BIPP).

$$\min \sum_l \sum_j d_j y_j^l \tag{2.10}$$

$$\text{s.t.} \sum_{v \in \bar{V}} s_v^k = 2 \qquad \forall k \in K \tag{2.11}$$

$$s_v^k = 2 t_v^k \qquad \forall k \in K; \forall v \in V \setminus \bar{V} \tag{2.12}$$

$$\sum_{e \in E_v} = s_v^k \qquad \forall k \in K; \forall v \in V \tag{2.13}$$

$$\sum_i x_i^k = \sum_{l: \text{ if } G^{kl}=1} S^l \qquad \forall k \in K \tag{2.14}$$

$$x_{I_e^1}^k - x_{I_e^2}^k \le z_e^k \qquad \forall k \in K; \forall e \in E \tag{2.15}$$

$$x_{I_e^2}^k - x_{I_e^1}^k \le z_e^k \qquad \forall k \in K; \forall e \in E \tag{2.16}$$

$$x_{I_e^2}^k + x_{I_e^1}^k \ge z_e^k \qquad \forall k \in K; \forall e \in E \tag{2.17}$$

$$2 - \left(x_{I_e^2}^k + x_{I_e^1}^k \right) \ge z_e^k \qquad \forall k \in K; \forall e \in E \tag{2.18}$$

$$x_i^k \le \sum_{l: \text{ if } G^{kl}=1} q_i^l \qquad \forall k \in K; \forall i \in I \tag{2.19}$$

$$1 - x_i^k \le \sum_{l: \text{ if } G^{kl}=0} q_i^l \qquad \forall k \in K; \forall i \in I \tag{2.20}$$

$$\sum_i q_i^l = S^l \qquad \forall l \in L \tag{2.21}$$

$$y_j^l \ge q_i^l \qquad \forall l \in L; \forall j \in J; \forall i \in J_i \tag{2.22}$$

$$\sum_l q_i^l = 1 \qquad \forall i \in I \tag{2.23}$$

$$t_v^k \in \{0,1\} \quad s_v^k \in \{0,1,2\} \qquad \forall v \in V \setminus \bar{V}; k \in K \tag{2.24}$$

$$s_v^k \in \{0,1\} \qquad \forall v \in \bar{V}; k \in K \tag{2.25}$$

$$z_e^k \in \{0,1\} \quad x_i^k \in \{0,1\} \qquad \forall k \in K; \forall e \in E; \forall i \in I \tag{2.26}$$

$$q_i^l \in \{0,1\} \quad y_j^l \in \{0,1\} \qquad \forall l \in L; \forall i \in I; \forall j \in J \tag{2.27}$$

In this BIPP, Equation 2.10 minimizes the total demand that is covered by servers of each partition. As a result, with this objective, for fixed locations of servers, the total demand that

is served by more than one partitions' servers is minimized. The physical reasoning of this objective is that as a detailed 3^n HQM will be solved for each core subregion, the influence of demand outside this core subregion should be minimal, as it is not considered at this stage. Constraints 2.11 force the model to have only two different outer vertices visited for each path with constraints 2.25. With these constraints, we are creating paths which start and end in the borders of the Voronoi diagram, similar to cuts dividing the whole Voronoi diagram into two pieces. Constraints 2.12 require that if an inner vertex is visited, it should be visited twice with the help of the second part of constraints 2.24. Note that, the mathematical model selects both edges and vertices, and if an edge is selected, vertices that belongs to this edge should also be visited. Different than outer vertices, inner vertices should be visited twice, to have a (continuous) path made of edges.

Constraints 2.13 work for inner and outer vertices differently because of constraints 2.11 and 2.12: If an inner vertex v is visited (i.e. $s_v^k = 2$), two of the edges that are adjacent to vertex v have to be visited as well. However if v is an outer vertex, if it is visited (i.e. $s_v^k = 1$) only one of the adjacent edges has to be visited.

Constraints 2.14 satisfy that the partition sizes that are in the selected side of path k, should sum up to the number of servers in the selected side of path k. Note that when the model is generated, each S_l is calculated in a way that each partition has a predetermined number of servers. For instance, for the partitioning in Figure 2.5, we set $S^l = 6$ where $l = 1, ..., 4$. Binary variables G^{kl} are set in a way that each path has equal number of partitions in both of their sides: $G^{1,1} = G^{1,3} = G^{2,1} = G^{2,2} = 1$ and the rest of the $G^{kl} = 0$, which means that path $k = 1$ (the vertical path) creates partitions $l = 1$ and $l = 3$, and path $k = 2$ (the horizontal path) creates partition $l = 1$ and $l = 2$ in their selected sides. As a result, we end up with 4 partitions with 6 servers that are formed by two paths. If for example, we prefer to make 3 partitions composed of 8 servers for the same example, we again need 2 paths but with three partitions. In this case, we should set $S^l = 8$ for $l = 1, 2, 3$ and $G^{1,1} = G^{1,2} = G^{2,1} = 1$ and the rest of the $G^{kl} = 0$. This way, we will make two paths in a way that, path $k = 1$ has two partitions ($l = 1$ and ($l = 2$) and path $k = 2$ has one partition ($l = 1$) on their selected sides.

If the edge between two servers does not belong to path k (i.e. $z_e^k = 0$), constraints 2.15 and 2.16 ensure that the servers on the opposite side of edge e belongs to the same side of path k. Similarly, if the edge between two servers belongs to path k (i.e. $z_e^k = 1$), constraints 2.17 and 2.18 forces the servers to be at the opposite side of path k.

Constraints 2.19 and 2.20 assign each server into their groups by checking in which side of the paths they belong. In the case depicted in Figure 2.5, partition $l = 1$ is in the selected side of both paths. If a server is at the same side, it is assigned to partition $l = 1$. On the other hand, if a server belongs to not selected sides of both paths, it is assigned to partition $l = 4$ because this partition is selected as such by setting $G^{1,4} = G^{2,4} = 0$. In a similar way, the rest of the partitions' servers are assigned.

Constraints 2.21 sets the number of servers in each partition to the partition's size. Constraints 2.22 associates atoms with partitions: If an atom can be accessed by a server, one of the servers that belongs to this partition can access to the atom. Constraints 2.23 place each server into one and only one partition.

Constraints 2.24-2.27 sets variables as binary variables except for the variables s_v^k in which v represents an inner vertex of the Voronoi diagram. The reason of this assignment is, as stated above, to satisfy that inner vertices should be visited twice in a continuous path starting from and ending at the outer vertices of the Voronoi diagram. Physical topological boundaries (e.g. rivers, mountains etc) that do not allow interactions between specific servers and atoms, can be easily integrated in the above formulation as further constraints.

In experiments, we have observed that adding cuts iteratively is much more efficient than solving the whole BIPP at once. As a result, we develop and use a heuristic that generates a single cut at each step. This is also consistent with the step of merging, which will follow inverse iterations of the partitioning. Steps of this heuristic for an instance can be seen in Figure 2.5. Figure 2.5a shows the problem area: demand intensity (darker color represents higher demand) and locations of servers (red circles). The first step of the algorithm is to generate a Voronoi diagram (Figure 2.5b). The steps afterwards are iteratively adding cuts to whole problem area. In our case, first a vertical then a horizontal cut is added (Figure 2.5c-d).

2.5 Computational Results

In this section, we first evaluate the accuracy of the models described in Section 2.3 and 2.4 (i.e. 3^n HQM and MHQA) by comparing them with the results of a discrete event simulation. Then, we utilize our models to evaluate instances in two approximation heuristics with an objective to identify close-to-optimum locations that optimize specific performance measures (loss rates and service times): variable neighborhood search (VNS) (Mladenović and Hansen, 1997) and simulated annealing (SA) (Kirkpatrick et al., 1983). Finally, we provide different performance measures of the optimization results that highlight the importance of the developed models. All of the algorithms in this work are developed under C# .NET environment. For partitioning algorithm IBM ILOG Cplex 12.4 is used through Concert user interface. MATLAB 7.9 through MATLAB Automation Server interface is used for matrix operations. All experiments are conducted on a PC with Intel Core2 Quad 3.00 Ghz processor.

In order to test the method, we use two different networks for demand distribution: Central Athens network with demand for tow-away services for bus operations (taken by Geroliminis et al. (2011)) (top) and an experimental network (bottom) which are given in Figure 2.8. In these figures, each square shows a 1km^2 area. The value inside each square shows 10^4 times the ratio of arrival rate to total arrival rate. Euclidean norm is used to calculate distance. Total service time is the sum of on-scene service time and the total travel time to reach incident atom and coming back. We did not test networks with physical boundaries but it can be easily integrated to our hypercube models (by applying exact travel times for each atom-

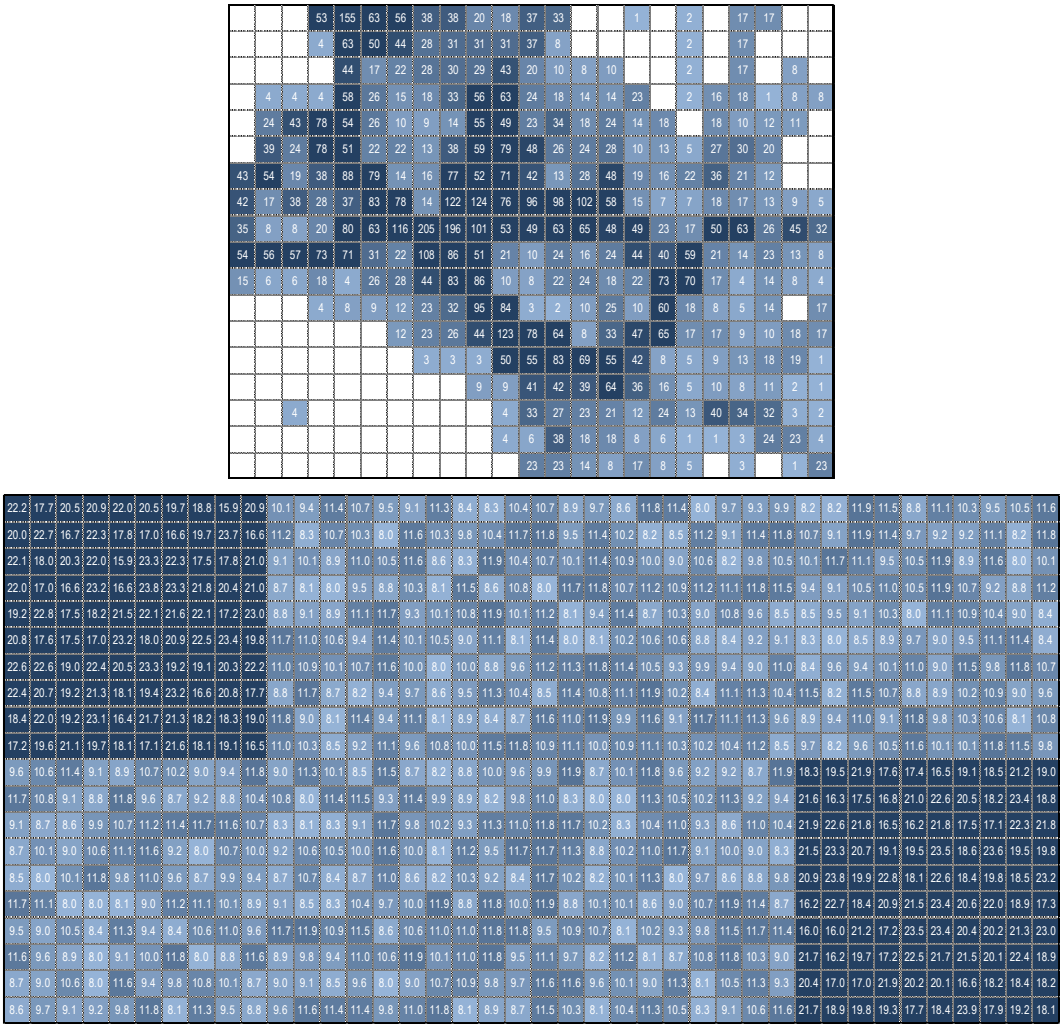


Figure 2.8: The demand distribution of the two networks used in our experiments: Central Athens (top) and experimental (bottom).

server pairs) and partitioning algorithm (by generating an additional edge in the Voronoi diagram). Different distance metrics (e.g. rectilinear, squared Euclidean) can also be applied to them. The only difference may be in the partitioning algorithm. Voronoi diagrams work for Euclidean distance metric. For another metric, we need different set of vertices and edges but the mathematical model used in the partitioning algorithm is still applicable.

2.5.1 Accuracy of 3^n HQM

In this part, the lost rate computed by MHQA is compared with the solution of the discrete event simulation for a case with 12 servers. We tested following instances with three demand (5, 15, 45 instance/hour), average on-scene service time (5, 10, 20min) and accessibility distance (10, 15, 20kms) for each demand distribution which makes 27 scenarios in total

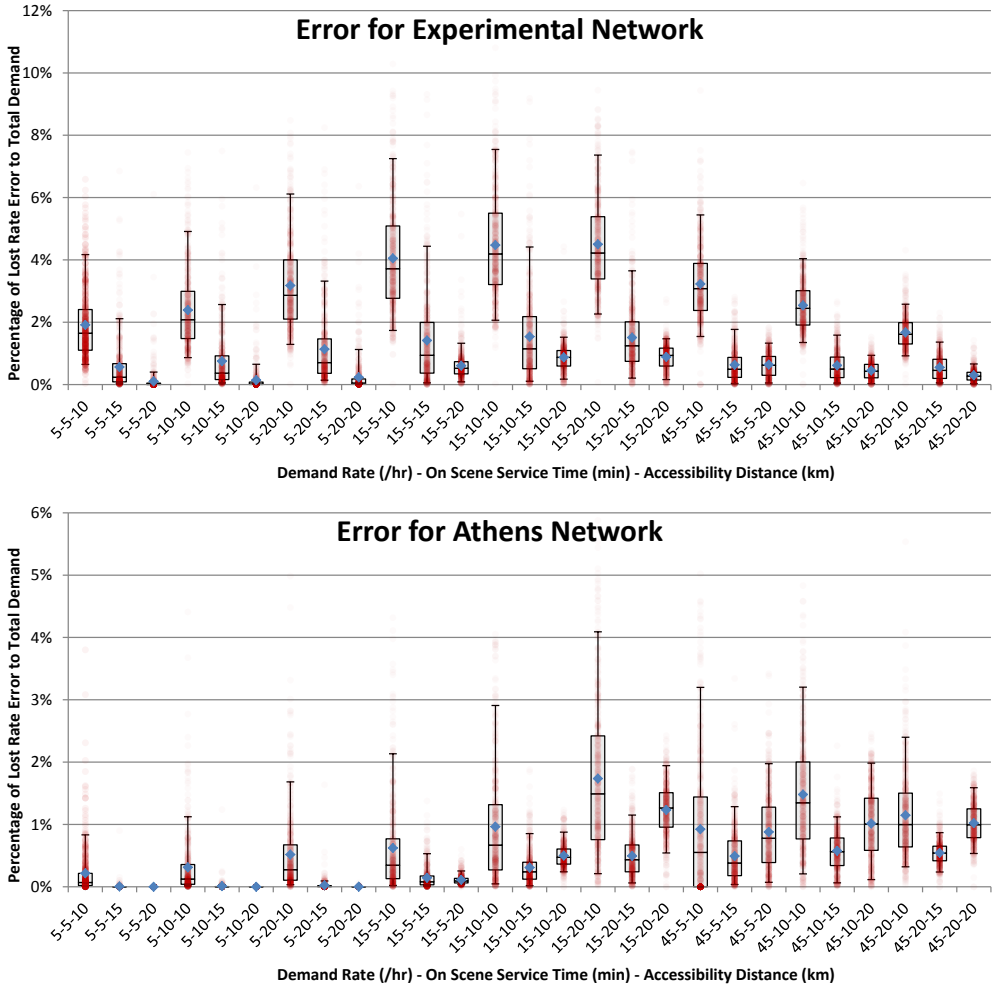


Figure 2.9: The ratio of difference between the loss rates calculated by simulation and the approximation method proposed.

for each instance. It is assumed that each server travels with a speed of 60km/h. On-scene service time and inter-arrival times are distributed exponentially. In simulation, a random value is generated whereas in approximation method it is assumed that total service time is exponentially distributed with the sum of travel time and on scene service time. We need such an assumption for the Markovian property. We generated 500 random instances with 12 facilities. In approximation algorithm, the whole problem area is partitioned into two core subregions with 6 servers each. Then the algorithm described in Figure 5.3 is applied. Both simulation and our method are run over these networks. The percentage of error in loss rates are reported in Figure 2.9. We calculate the errors by comparing the values of MHQA with simulation values.

To ensure that simulations have reached steady states, 25 parallel simulation instances were created with 11 batches simulating length of 50 days each. The first batch of each run was

discarded and mean of the rest of the batches of all 25 parallel simulations are reported. Length of the simulations are selected in a way that calculated values have tight enough confidence intervals to guarantee steady state.

For each instance, simulation took around 25s for both networks whereas our method took around 3s for Athens network and 7s for the experimental network on average. The comparison showed that compared to simulation, our method gives results with acceptable error (less than 5% error on average and 10% in the worst case) in less run time (12-28% of simulation run time). This error gets even less for increased server range, for the scenarios with the range of 20km, average error is less than 1% and in 95% of the cases error is less than 2%. Furthermore, simulation might need longer run times to have accurate results if we want to calculate more detailed performance measures (e.g. loss rate per number of busy servers) or for larger instances of the problem.

2.5.2 Heuristics for Better Location of Servers

In this part, we have tested our exact 3^n HQM (for cases with less than or equal to 8 servers) and mix aggregate hypercube (for cases with more than 8 servers) algorithms inside two heuristic approaches to identify close-to-optimum locations of servers: variable neighborhood search (Mladenović and Hansen, 1997) and simulated annealing (Kirkpatrick et al., 1983). Both methods are initialized with the maximum expected coverage location model (MEXCLP) (Daskin, 1983). MEXCLP is selected because it is fast and gives good results. In VNS algorithm, we assume that if two instances' all but one servers are in different locations, they are neighbors. In other words, in every iteration, a randomly selected server's location is changed. We use the same neighborhood structure in SA algorithm. We set the starting "temperature" coefficient to 1 and increase it by 10% in every 20 iterations. Temperature is assumed to be the division of temperature coefficient with the average lost rate in every iteration. We have applied 3^n HQM for cases with 6,7 and 8 servers for total arrival rates of 5, 8, 10, 15 and 20 requests/hour. For cases with 12 and 16 servers, arrival rates are doubled (i.e. 10, 16, 20, 30, 40 requests/hour), problems are solved by approximation algorithm with two partitions of equal size. Run times for the former three (i.e. 6, 7 and 8-server) cases are set to one hour, whereas the latter two (i.e. 12 and 16-server) cases are run for four hours. We have applied two different on scene service times: 5 and 20 minutes. For all cases, maximum accessibility distance is set to 30km. Found minimum lost rate and percent lost rate improvements after MEXCLP for Athens and experimental networks (Figure 2.8) by both heuristics (i.e. VNS and SA) for cases with realistic lost ratios (ratio of lost rate to the total arrival rate) can be seen in tables 2.1 and 2.2 respectively.

For both Athens and experimental networks, it is observed that there is an improvement of more than 20% on average for the lost rates over MEXCLP. Average lost rate improvement gets around 50% for the Athens network with short on scene service times (i.e. 5 minutes). We have not observed any significant difference between VNS and SA. Since our primary goal in this research is to test the applicability of hypercube models inside search algorithms, we have not

# servers	on scene serv. time	demand	MEXCLP	VNS		SA	
				value	% impr.	value	% impr.
6	5	8	0.129	0.080	38.28	0.080	37.95
		10	0.422	0.266	37.07	0.268	36.50
	20	5	0.174	0.138	20.61	0.138	20.52
		8	1.197	0.989	17.36	0.989	17.34
7	5	8	0.031	0.018	41.98	0.018	42.48
		10	0.136	0.079	42.32	0.081	40.78
	20	5	0.059	0.046	22.30	0.045	23.18
		8	0.653	0.521	20.19	0.528	19.12
8	5	10	0.044	0.021	52.53	0.023	46.62
		15	0.651	0.335	48.60	0.393	39.60
	20	5	0.019	0.013	29.13	0.014	27.18
		8	0.340	0.248	26.87	0.260	23.46
12	5	20	0.077	0.016	79.45	0.018	76.40
		30	1.824	0.690	62.17	0.692	62.05
	20	10	0.039	0.020	49.23	0.020	48.62
		16	0.993	0.726	26.86	0.691	30.43
16	5	30	0.121	0.058	51.88	0.025	79.35
		40	1.666	1.113	33.2	0.412	75.27
	20	16	0.102	0.059	41.74	0.045	55.40
		20	0.645	0.374	42.00	0.368	43.00

Table 2.1: The best lost rate found by MEXCLP, VNS and SA algorithms for the Athens network given in Figure 2.8a.

searched for parameters that may give better final results for both VNS and SA.

From the results in tables 2.1 and 2.2 one can also observe that the lost rates dramatically increase with the increase of on scene service time. Small increase in demand has also considerable influence on the lost rate. Queueing systems are unpredictably complex and need custom-built algorithms to be tested. We also observe that the performance of SA and VNS against MEXCLP get better with the increase in the number of servers. Last but not least, careful readers might realize that for the same value we have calculated different percent improvements. This is the consequence of showing results with limited precision. We also noticed (not shown here) that even after 30 minutes (instead of 4 hours) the VNS and SA methods provide similar improvement with a 4 hour run.

2.5.3 Performance Measures of Hypercube Models

In this subsection, we conduct further analysis to some location instances improved by the optimization heuristics. In the first analysis we investigate the effect of accessibility range

# servers	on scene serv. time	demand	MEXCLP	VNS		SA	
				value	% impr.	value	% impr.
6	5	5	0.076	0.055	26.72	0.055	26.67
		8	0.725	0.594	18.09	0.594	18.09
	20	5	0.432	0.377	12.63	0.377	12.67
		8	2.074	1.926	7.11	1.926	7.11
7	5	5	0.022	0.013	39.27	0.013	39.27
		8	0.343	0.252	26.29	0.252	26.29
	20	5	0.207	0.167	19.34	0.168	19.15
		8	1.436	1.267	11.81	1.27	11.62
8	5	8	0.129	0.098	23.94	0.098	23.94
		10	0.483	0.387	19.82	0.387	19.82
	20	5	0.084	0.065	22.77	0.068	19.71
		8	0.894	0.776	13.21	0.798	10.73
12	5	16	0.263	0.054	79.36	0.054	79.36
		20	1.272	0.472	62.90	0.472	62.9
	20	10	0.200	0.095	52.49	0.099	50.18
		16	2.680	2.062	23.03	2.064	22.99
16	5	20	0.068	0.008	87.92	0.008	87.92
		30	2.423	0.786	67.58	0.786	67.58
	20	16	0.551	0.282	48.85	0.290	47.37
		20	2.203	1.458	33.84	1.448	34.29

Table 2.2: The best lost rate found by MEXCLP, VNS and SA algorithms for the experimental network given in Figure 2.8b.

and demand on servers' workload (fraction of time a server is busy) and intradistrict service ratio for fixed locations, with the 3^n HQM for 8 servers. These fixed locations are estimated through the optimization heuristics for some level of demand and they do not recalculated when demand changes. In the second analysis, the effect of demand increase on the number of busy servers and the ratio of the time spend on intradistrict service in bin level is analyzed. A larger instance with 12 servers solved by 3^n AHQM is considered (2 bins of 6 servers each).

In order to see the effect of accessibility range and demand on servers' utilities, ten instances (5 demand levels and 2 accessibility ranges) with 8 servers on the experimental network are analyzed. To have a better insight, the location of the servers are fixed in all instances. We set the best locations found by the heuristic for demand 5 requests/hour and on scene service time equal to 20 minutes (see Table 2.2). The fraction of time each server is busy (lines) and the fraction of busy time each server is in intradistrict response (columns) are reported in Figure 2.10a and b. Demand levels (5, 8, 10, 15 and 20 requests/hour) are shown with different colors. Two different accessibility ranges (15 and 30min) are used in two separate graphs, 2.10a and b respectively. The lost rate for different demand levels are reported in the legends of each

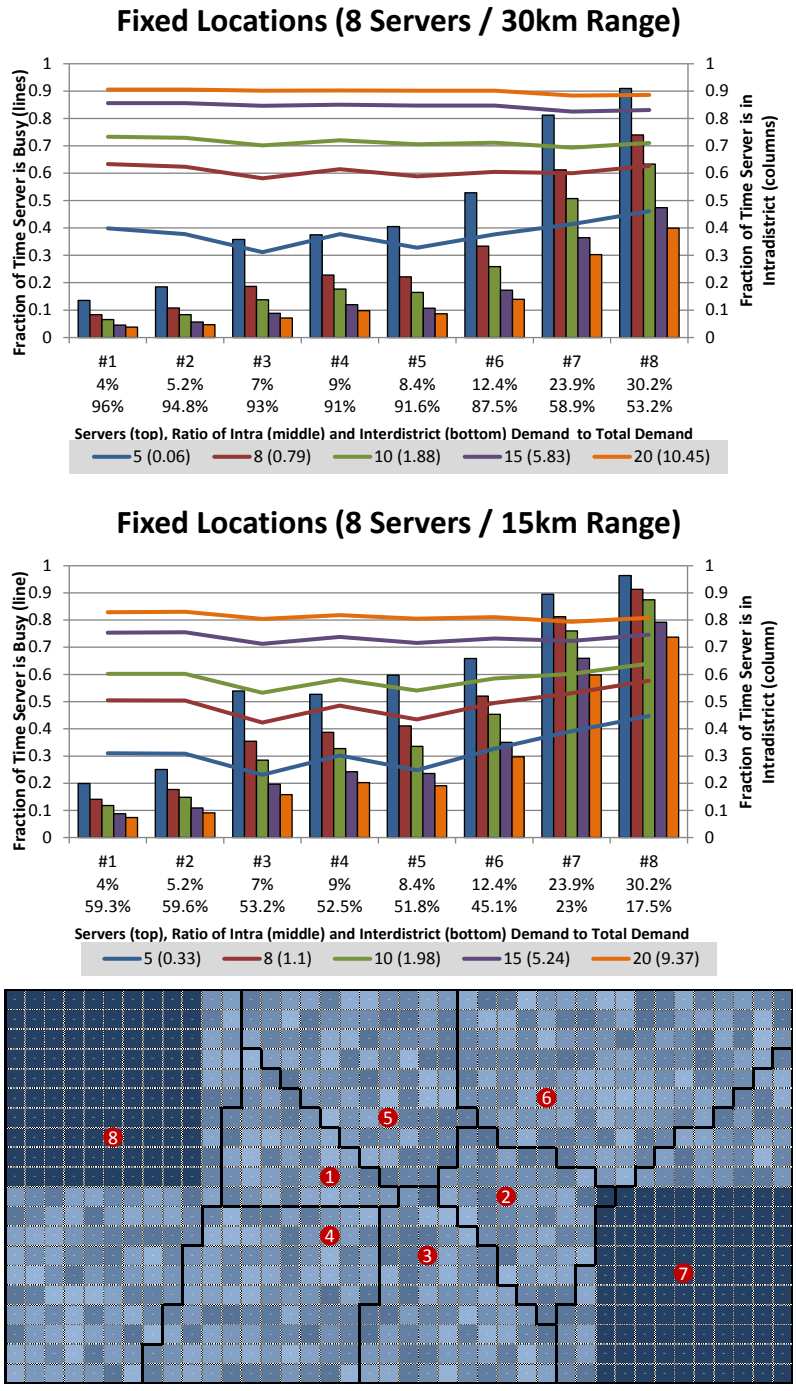


Figure 2.10: In the top two graphs (a and b), the effect of increased demand and accessibility range is shown for incidents of 8 servers (with 3^n HQM). The same server locations are selected in all incidents, which can be seen in the map given below the two graphs (c). Both the fraction of time each server is busy (line) and the fraction on intradistrict response (column) are shown in a and b. The total demand at each servers' primary and secondary service area are shown under the x-axis respectively. Note that the values in parenthesis in the legend are the loss rates.

figure inside the parenthesis. The fraction of the total demand each server has in their intra and interdistrict area are also reported under the x-axis. Note that the first line of percentages sum up to 1, which means that all atoms are reachable by at least one server. The second line has much higher values because it considers the accessibility of interdistrict responses. The locations of each server (#1 to #8 are shown in the x axis of the graphs) can also be seen in the map reported in Figure 2.10c. Note that, the intensity of the blue color shows the level of demand; darker the blue, higher the demand.

One of the striking outcomes of these graphs is the ratio of the time spend in intradistrict responses. Although, the probability of being busy fluctuates a little among servers (5-10%), the ratio of intradistrict responses is quite variant for different servers (5-95%). This fact shows the importance of using a model which differentiates intra and interdistrict responses. The main reason to have such fluctuations in intradistrict response ratios is the difference between the responsibility areas of each server. Servers with more intradistrict demand spend more on intradistrict service. This is expected. However, seeing such a huge difference between time spent on intradistrict service among servers is an interesting observation. This high intradistrict fraction motivates to investigate the effect of the accessibility distance in the results.

In addition, by comparing Figure 2.10a and b, we can see the effect of accessibility range and the importance of dispatching policy. In this research, advanced dispatching policies are not investigated. In every state, the available server with the minimum service time is dispatched at all times. For low demand levels, higher accessibility range works better than the low. For demand 5 requests/hour, we get a lost rate of 0.06 /hour with a 30km accessibility range whereas this value is 0.33 /hour for an accessibility range of 15km. With the increase in demand, lower accessibility range improves more and gives better results. For demand level 20, lower accessibility range has a smaller loss rate than the higher accessibility range because, requesting from a server to intervene in an incident far from his responsibility area under high demand conditions, might result in loss demand for intradistrict requests.

There is something more to add to these analysis. Although, compared to high accessibility range, low accessibility range always provides a higher intradistrict ratio for all instances and servers, for low demand levels, it does not give better loss rates. This is probably because of the congestion in the system. By lowering accessibility range, although we encourage the model to have more intradistrict responses, we are also decreasing the secondary service areas of each server. In other words, we are lowering the number of servers that can be dispatched at each incident and obviously specific to our instance, this declines the performance of the system for lower demand and ameliorates for higher demand. Note that, even for small accessibility range, all demand is accessible by at least one server.

In our second analysis, we investigate performance measures from a larger instance with 12 servers solved by the MHQA (2 bins with 6 servers each). We take best instances we have found with our heuristics for 12 servers and 20min on scene service time in the experimental

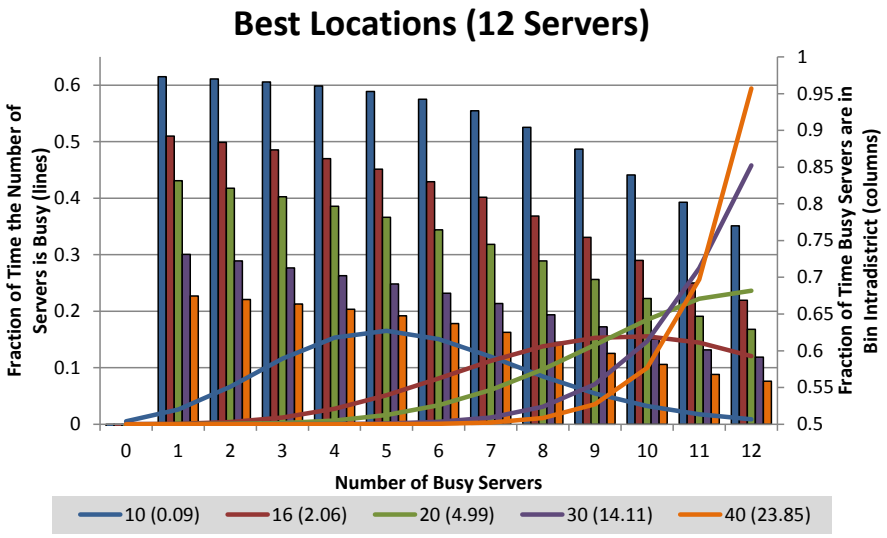


Figure 2.11: The effect of increased demand on 5 best instances of different demand levels. In all instances, the number of servers is 12, on scene service time is 20min and accessibility range is 30km. The fraction of time for different busy servers' count (line) and the ratio of servers in intradistrict response (columns) can be seen in the figure. Lost rates for each demand level is given in parenthesis in the legend, next to the related demand level.

network. Different than the previous analysis, we investigate the probability of having specific number of servers busy in the system all of which can be seen in Figure 2.11. These values are shown with columns. With lines, the percentage of time servers busy with bin instradistrict responses (i.e. dispatching inside bin) are depicted. We use 5 different demand levels and each of which are shown with a different color. Again, loss rates are shown inside parenthesis in the legend next to the related demand level.

The first interesting result that can be observed from this figure is the effect of increased demand on the system efficiency. The increase in demand results in less time spend in intradistrict (equivalently more time in interdistrict) responses. The increase in the number of busy servers has also similar effect. With the increase in busy server count, intradistrict response ratio decreases.

One of the other important findings observed in graphs in Figure 2.11 is the interaction between subregions. As a simpler approach, we can take each core subregion as separate problems and do calculations regarding this assumption (e.g. sum loss rate of each subregions to find total loss rate). However, results in Figure 2.11 show that, amount of time spent for interdistrict responses is significant. Because of this, disregarding interdistrict responses between subregions and estimate performance measures with 3^n HQM without the step of merging two compound regions, might create inaccurate results and does not seem to be a right approach.

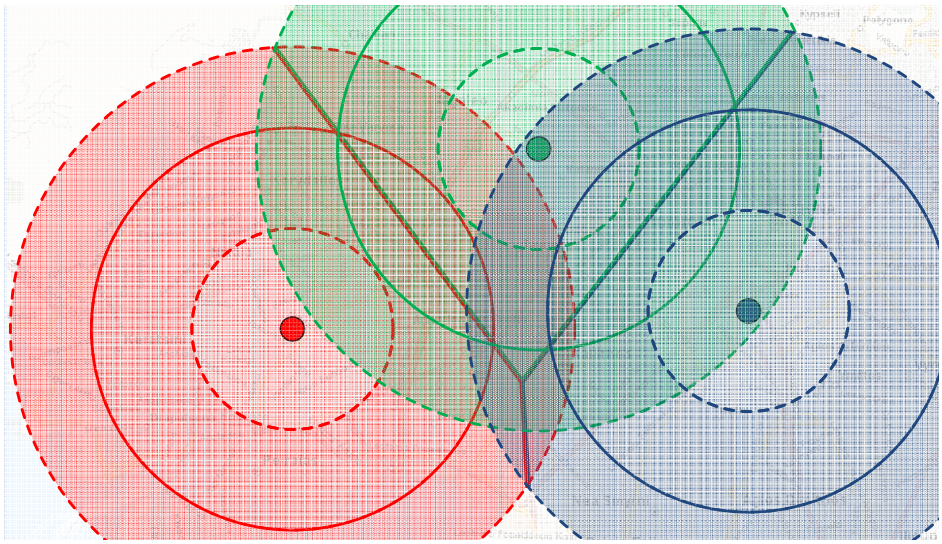


Figure 2.12: Illustration of a system with three servers and three service range belts for each server.

2.6 Concluding Remarks

In this research, we have proposed two new 3^n hypercube models and two algorithms that utilize these two methods. We first compare our approach with the system with real service time and show the accuracy of our approach for different parameters on two different networks: one real (central Athens) and one experimental with different spatial distribution of demand. In order to show the applicability of our two algorithms inside an optimization framework, the two methods are implemented with the variable neighborhood search and simulated annealing. From these experiments, it is seen that, although hypercube queueing models are not optimization models, they can be used inside such frameworks. Then, we have investigated the percentage of time the servers spend on intra and interdistrict responses. The percentage of interdistrict responses is not negligible.

As future research, our aim is to change the way we define intra and interdistrict response areas. In hypercube queueing models, service rate and service priority are different parameters. We can still define intra and interdistrict areas with service priority, instances can be served by the closest available server. However, we can arrange different service rates for different distance ranges from the servers. We can create two or more service range belts and we can apply different service rates for each belt (see Figure 2.12). Another direction to further investigate is the partitioning algorithm. With some additional cuts, we might create a tighter convex hull which in the end might give a more efficient partitioning algorithm. We are also interested in to work on some efficient heuristic methods to replace the partitioning algorithm. Last but not least, different dispatching algorithms are also worth investigating. For the same server locations, intelligent dispatching schemes may give better results.

One-Way Car-Sharing Systems Part II

3 An Optimization Framework for One-Way Car-Sharing Systems

Electric vehicle sharing systems have been introduced to a number of cities around the world as a means of increasing mobility, reducing congestion, and pollution. Electric vehicle sharing systems can offer one or two-way services. One-way systems provide more flexibility to users since they can be dropped-off at any stations. However their modeling involves a number of complexities arising from the need to relocate vehicles accumulated at certain stations. The planning of one-way electric vehicle-sharing systems involves a host of strongly interacting decisions regarding the number, size and location of stations, as well as fleet size.

In this chapter we develop and solve a multi-objective MILP model for planning one-way vehicle-sharing systems taking into account vehicle relocation and electric vehicle charging requirements. For real world problems the size of the problem becomes intractable due to the extremely large number of relocation variables. In order to cope with this problem we introduce an aggregate model using the concept of the virtual hub. This transformation allows the solution of the problem with a branch and bound approach.

The proposed solution generates the efficient frontier and allows decision makers to examine the trade-off between operator's revenues and users' net benefits. The capabilities of the proposed approach are demonstrated on a large scale real world problem with available data from Nice, France. Extensive sensitivity analysis was performed by varying demand, station accessibility distance and subsidy levels. The results provide useful insights regarding the efficient planning of one-way electric vehicle sharing systems.

3.1 Introduction

Car-sharing (also known as shared-use vehicle) systems have attracted considerable attention with multiple implementations worldwide due to their potential to improve mobility and sustainability (Shaheen and Cohen, 2013). These systems provide benefits both to their users and the society as a whole. Reduced personal transportation cost and mobility enhancement have been cited as the two most notable benefits to individual users. Reduction in the parking space

requirements is one of the benefits that society have. Recent studies show that, depending on the type of the system car-sharing also decrease average vehicle kilometers traveled and, likely to decrease congestion (Crane et al., 2012) and emissions (Shaheen and Cohen, 2013). Provision of affordable mobility to economically disadvantaged groups with on-demand and public transportation systems is one of the other societal benefits (Duncan, 2011).

The attractiveness of car-sharing systems is determined by the level of service offered and the cost associated with the use of the system. The level of service is influenced by the accessibility of vehicle stations by the potential users, i.e. (i) the distance between user's origin and destination from pick-up and drop-off vehicle stations respectively, and (ii) the availability of vehicles at stations. On the other hand, station number and size, as well as fleet size and availability of vehicles, at the "right time" at the "right station", influence the cost of establishing and operating a car-sharing system.

The car-sharing systems can be classified into flexible "one-way" and the more restricted "two-way" types, according to whether the users should return the rented vehicle at a different or at the location they picked it up. The "one-way" systems are also classified as "free-floating" and "non-floating" according to parking spot restrictions. The former refers to a system without restricted parking spots. Users can pick-up or drop-off vehicles in a restricted area with some borders. The latter is used for defining systems in which pick-up and drop-off locations of the vehicles should be designated parking spots. In "free-floating" models, reservation is not possible whereas "non-floating" models provide users both reservation and freedom of one-way trips. In reality, most of the non-floating one-way systems in operation works with partial reservations (e.g. without destination reservation). Although full reservation needs further planning, we believe that, system with such a property may provide a better service and attract more customers. This work focuses on one-way systems with reservations both at the origin and destination.

The problem of ensuring vehicle availability and fulfilling reservation becomes more prominent when the vehicles can be rented and used on a one-way basis in non-floating systems. The one-way operation of the vehicles coupled with the imbalance of demand for vehicles, both at the origin of the trip (pick-up station) and at the destination (drop-off station), may result a situation where the vehicles are accumulated to stations where they are not needed, while at the same time there is vehicle shortage at the stations where more vehicles are needed (Barth et al., 2004).

Vehicle relocation, i.e. transfer of vehicles from stations with high vehicle accumulation to stations where shortage is experienced, is a technique that has been proposed to improve the performance of one-way car-sharing systems (e.g. Kek et al., 2006; Cucu et al., 2009; Jorge et al., 2013). The lack of efficient vehicle relocation coupled with the need to guarantee a given level of vehicle availability may lead to an unnecessary increase of the fleet size and vehicle under-utilization. The efficient and cost-effective strategic planning, and the operation of one-way car-sharing systems require models that will determine the number and location of

the service stations, the fleet size, and the dynamic allocation of vehicles to stations optimally. These models should assist decision makers to strike an optimum balance between the level of service offered and the total cost (including vehicle relocation costs) for implementing and operating the car-sharing system.

However, the literature currently lacks a model that can consider simultaneously decisions related to the determination of station location, size and number, and fleet size, while taking into account the dynamics of vehicle relocation and balancing. Existing models (Lin and Yang, 2011; Correia and Antunes, 2012) either look at station locations without due consideration to vehicle relocation decisions (Lin and Yang, 2011), or consider station locations assuming that only the demand in the catchment area of opened stations needs to be serviced (Correia and Antunes, 2012). In the case where vehicle relocation is modeled (Correia and Antunes, 2012), the relocation of the vehicles and the associated costs are considered only at the end of the operating period (usually a day), and therefore they are influencing the fleet size.

The objective of this chapter is twofold: (i) to develop and solve a mathematical model for determining the optimum fleet size, and the number and location of the required stations of non-floating one-way car-sharing systems by taking into account the dynamic repositioning (relocation) of vehicles, and (ii) to apply the proposed model for planning and operating a one-way electrical car-sharing system in the city of Nice, France.

The remainder of this chapter is organized as follows. Section 3.2 provides an overview of previous related work and further elaborates on the arguments justifying the need for the proposed model, Section 3.3 presents the formulation and the solution approach of the proposed model, Section 3.4 describes the application of the proposed model for planning and operating a one-way electrical car-sharing system in Nice, France while Section 3.5 discusses the research conclusions and provides recommendations for future research.

3.2 Previous Related Research

Models related to the planning and operation of car-sharing systems can be classified into the following two broad categories: i) models addressing strategic planning decisions, and ii) models supporting operational decisions.

3.2.1 Models for Strategic Planning Decisions

Strategic planning decisions seek to determine the number, size and location of stations, and the number of the vehicles that should be assigned to each station, in order to optimize a measure or a combination of measures of system performance. Station location models have been developed to locate bicycle stations (Lin and Yang, 2011) and car stations (Correia and Antunes, 2012). Although the focus of our work is on electrical car-sharing systems, we also review models that address the station location of shared-use bicycles, given some similarities

of the two systems.

The problem of locating stations for shared-use bicycles has been studied recently (Lin and Yang, 2011). This paper presents a model for determining the number and location of bicycle stations and the structure of the network of bicycle paths that should be developed to connect the bicycle stations. The problem is formulated as a non-linear integer model. The objective function used expresses the total yearly cost encountered by the operator and the users. A small scale example was used to illustrate the model and a branch and bound algorithm was used to solve it. This model does not consider the daily variation of demand and the problems arising from the dynamic accumulation/shortage of bicycles due to the variation of demand in time and space.

The optimization of vehicle depot locations and the definition of the number of parking spaces (size) for each depot has been also addressed (Correia and Antunes, 2012). The number of parking spaces at each depot is determined by the maximum number of vehicles that are allocated to each station throughout an operating day. Vehicle relocation (and the associated relocation cost) is considered only at the end of the entire operating period (i.e. day). Thus, this model does not treat explicitly the dynamic imbalance created by the one-way operation and therefore it does not rebalance the vehicles at the end of each operating sub-interval (e.g. hour). This model assumes that the vehicle imbalance problem is by-passed through the optimum depot location and size. This assumption makes this model applicable to the systems with low/balanced demand and/or high station capacity. For high and unbalanced demand, and limited capacity of stations, its performance is debatable. The objective function of the model seeks to maximize the profit of the operating agency and takes into account the depreciation, maintenance and relocation (at the end of the operating period) costs of the vehicles, the maintenance cost of the depots, and the revenues generated by the system operations. This model makes the assumption that only trips associated with open stations need to be served. Thus, the demand (trips) that falls outside the catchment area of open stations associated with the stations that are not open is ignored. As a consequence, this model does not consider the access and egress cost of the potential users to/from the candidate station locations. A direct implication of this assumption is that, the proposed model cannot be used to study the trade-off between station accessibility cost and system benefits. Finally, this model does not consider the dynamic relocation of vehicles throughout the operating period. The proposed model was used to analyze a case study in Lisbon and an optimizer based on branch-and-cut algorithms was used to solve the problem.

A recent work also models one-way car-sharing problem with an MILP considering relocation through out the day (Jorge et al., 2013). Similar to Correia and Antunes (2012), the model exogenously associates trips to stations. Different than the previous work, the model enables relocation at any time of the day. The objective function maximizes the profit of the operating agency. The model is tested on three different scenarios and the results are supported with a simulation. In simulation, cost of relocation is minimized with minimum cost flow algorithm. Results on different scenarios show that, with dynamic relocation, car-sharing system modeled

on the demand in Lisbon, Portugal starts profiting.

The problem of determining the fleet size and the distribution of vehicles among the stations of a car-sharing system was studied in relation to the Personal Intelligent City Accessible Vehicles (PICAVs). This system uses a homogeneous fleet of eco-friendly vehicles and allows one-way trips (Cepolina and Farina, 2012). The stations are parking lots that offer vehicle recharging services and are located at inter-modal transfer points and near major attraction sites within a pedestrian area. The number, location and capacity of stations are not determined by the model, hence constitute inputs to the simulated annealing process. To cope with the imbalance of vehicle accumulation of the one-way system, this model introduces the concept of supervisor. The task of the supervisor is to direct users that are flexible in returning the vehicle to alternative stations, as to achieve a balanced operation and fulfill a maximum waiting time constraint. The objective function of this model includes the minimization of the daily system and user costs subject to a maximum waiting time constraint. The value of the objective function of the model was estimated through micro-simulation. A simulated annealing approach was used for determining the fleet size and for allocating vehicles among system stations.

Models for evaluating the performance of a network of car-sharing stations has been introduced in the literature (Fassi et al., 2012; George and Cathy, 2011). This problem arises when the demand for car-sharing services changes (increases) and as a consequence the network of stations should be adapted to serve better the emerging demand profile. In response to this need a decision support tool was developed which allows decision makers to simulate alternative strategies leading to different network configurations. Such strategies include opening and/or closing stations, and increasing the capacity of stations. This tool is based on discrete event simulation and seeks to maximize the satisfaction level of the users and to minimize the number of vehicles used (Fassi et al., 2012). This model does not address vehicle relocation as it is based on a system that does not allow one-way use of vehicles. Performance analysis for shared-use vehicles systems has been proposed in the literature using a closed queuing network model (George and Cathy, 2011). In this approach, both exact and approximate solution methods are proposed to evaluate the bike sharing system Vélib operating in Paris, France with over 20000 bicycles and 1500 locations.

3.2.2 Operational Decisions

A major decision associated with the operation of one-way car-sharing systems is how to relocate vehicles. The vehicle relocation problem arises from the imbalanced accumulation of vehicles at stations when the car-sharing system allows their one-way use. Different strategies and models have been proposed in the literature to cope with the vehicle relocation problem.

The relocation of shared vehicles can be realized by using operating staff (Kek et al., 2006) or it can be user based (Barth et al., 2004). Two user-based relocation strategies namely, trip-joining and trip-splitting have been proposed (Barth et al., 2004). The trip-joining strategy is used

when two users have common pick-up and drop-off stations and there is a shortage of vehicles at the pick-up station. In this case, the users are asked to share the ride. The trip-splitting strategy is used when there is a surplus of vehicles at the pick-up station and there are users that are traveling as a group. Under this condition, the users are asked to use separate vehicles when there is a shortage of vehicles at their destination (Barth et al., 2004). The trip-joining and the trip-splitting strategies have been analyzed using data collected from a car-sharing system operated at a university and through simulation. The results of the simulation model suggest that the need for vehicle relocations can be decreased by 42% by using these strategies (Barth et al., 2004).

Shortest time, and inventory balancing strategies have been used (Kek et al., 2006) for staff based vehicle relocation. The shortest time strategy relocates vehicles from other stations to minimize the travel time needed for a staff member from his/her current location to the station where the vehicle is available plus the travel time needed from the station that the vehicle is available to the station where the vehicle is needed. The inventory balancing strategy relocates vehicles from stations with over-accumulated vehicles to stations that experience vehicle shortages. Both strategies were tested through a simulation model which was validated using data from an operational car-sharing system (Kek et al., 2006). An optimization-trend-simulation decision support system (Kek et al., 2009) is proposed which uses the same simulation model. In this three-phase decision support system, the effectiveness of different relocation policies are evaluated according to zero-vehicle time (duration of the vehicle shortage), full-port time (shortage of empty parking space when needed) and number of relocation.

The dynamic allocation of vehicles among the stations of a car-sharing system to maximize profit has been modeled in (Fan et al., 2008). The fleet size, the location of stations, and the demand for trips for a given planning horizon are known in advance. Penalties associated with unserved trip requests are not considered. A multistage stochastic linear model with recourse has been proposed to address this problem. A stochastic optimization method based on Monte Carlo simulation was used to solve the proposed model (Fan et al., 2008). This model considers only the vehicle relocation decisions. Furthermore, vehicle relocation is performed at the end of the day.

Chance constraint modeling has been used to study fleet redistribution (Nair and Miller, 2011). This model assumes that system configuration, current inventory of each station, costs and demand at each station are known in advance. The model aims to find the minimum cost fleet redistribution plan for the demand expected in the near future. The chance constrained model with reliability p (CCM-p) is constructed and solved by utilizing a special technique involving p-efficient points (PEPs) (Prékopa, 2003). The model is applied on the Intelligent Community Vehicle System in Singapore, a one-way system with 14 stations, 202 parking spaces and 94 vehicles.

Relocation operations in bike-sharing systems are also investigated in the OR literature. Asymmetric demand creates problem of imbalance for bike-sharing systems. This results in increase

in the number of users (i.e. who try to rent bikes from empty stations or to leave bikes to full stations) who cannot utilize the system properly. As a result redistribution of bikes becomes inevitable. The literature contains solutions for both static and dynamic balancing problems. Static balancing problem disregards customer demand and assumes the system does not operate during redistribution (e.g. during the night). Whereas in dynamic balancing problem, demand varies with time and redistribution operations are performed accordingly. The static balancing problem has been modeled as a single vehicle one-commodity capacitated pickup and delivery problem (SVOCPDP) and was solved with an exact algorithm based on column-generation (Chemla et al., 2012). Additional formulations of the static balancing problems have been proposed in Raviv et al. (2013). Dantzig and Wolfe (1960) and Benders (1962) decompositions have been also used to solve the dynamic balancing problem (Contardo et al., 2012).

In the literature, there are also other types of problems that share common structures with the one-way car-sharing problem. The multiple depot vehicle scheduling problem with time windows (MDVSPTW) is one of the examples (Desaulniers et al., 1998). In the MDVSPTW, each customer has a request of tight time windows with a precise start and end time of operations, and a fleet of vehicles serves these customers one at a time. Each vehicle in the fleet belongs to a depot and the vehicles have to return to their depot at the end of the service. The objective of the problem is to minimize the number of vehicles and empty trips.

The literature review revealed that existing modeling efforts make a sharp separation between strategic and tactical decisions. This means that strategic decision-making models do not integrate in their structure aspects of tactical and operational decisions (e.g. vehicle relocation, fleet size) which, as we demonstrate in this chapter have a significant bearing on the cost and performance of the car-sharing system. On the other hand, operational models are focused on the detailed modeling of different types of relocation strategies, assuming that the location, number, and station and fleet size are exogenously defined.

In reality, strategic, tactical, and operational decisions are interweaved and therefore there is a strong interaction between the three decision making levels. Strategic decisions are primarily related to the definition of the location, number, and size of stations and interact with the tactical decision of fleet size determination. In turn the fleet size is affected by vehicle relocation which is an operational decision. Here it is important to stress the fact that both fleet size and vehicle relocation influence the strategic level decisions. The above discussion suggests that there is a need for a model that will be able to address the strategic and tactical decisions by taking into account (at a macroscopic level) the impact of vehicle relocation. Figure 3.1 illustrates these interactions. The above discussion suggests that there is a need for a model that will be able to address the strategic and tactical decisions by taking into account the impact of vehicle relocation. In what follows we are presenting such a model. In this work, although we regard pricing as an input to our model, different pricing policies can be generated with different price parameters and solved in separate runs to see their effect to the model. Elasticity of demand due to pricing is beyond the scope of this work and can be

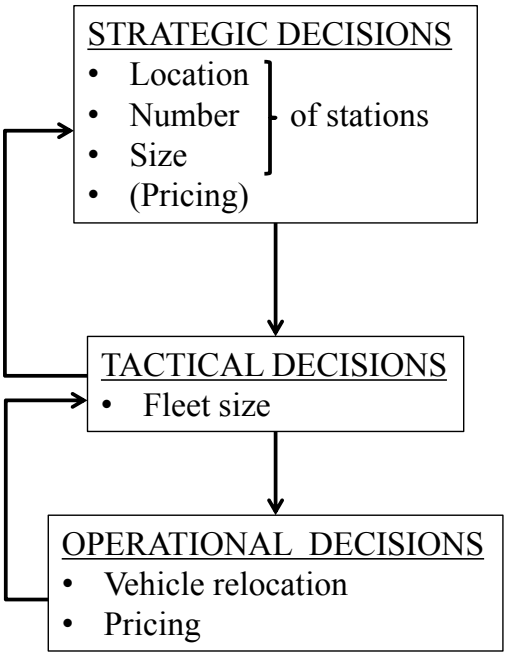


Figure 3.1: Relationship between strategic, tactical and operational decisions

investigated in the future.

3.3 Model Description

The proposed model is motivated from the planning of electrical one-way car-sharing system. Shared-use electric vehicles are used to serve trips within a given geographical area. The system operates on the basis of reservations and therefore the origin-destination matrix for the planning period is known in advance. Stochastic and seasonal demand variations are also considered in the optimization process. In what follows we provide a description of the system in terms of its demand and supply characteristics.

3.3.1 System Characteristics

i. Vehicles: A homogeneous fleet of electric vehicles is used to provide the services. Any type of trip request can be accommodated by any available vehicle.

ii. Stations: Vehicles are picked-up and dropped-off at designated stations. Stations have the necessary infrastructure for parking and recharging the vehicles. Each station provides a specific number of parking places which defines the station size. Station size varies among stations and the size of each station determines its capacity.

iii. Time Intervals: An operating day is divided into time intervals (not necessarily equally long) and each operation (i.e. rental, relocation, charging) starts at the beginning and finishes

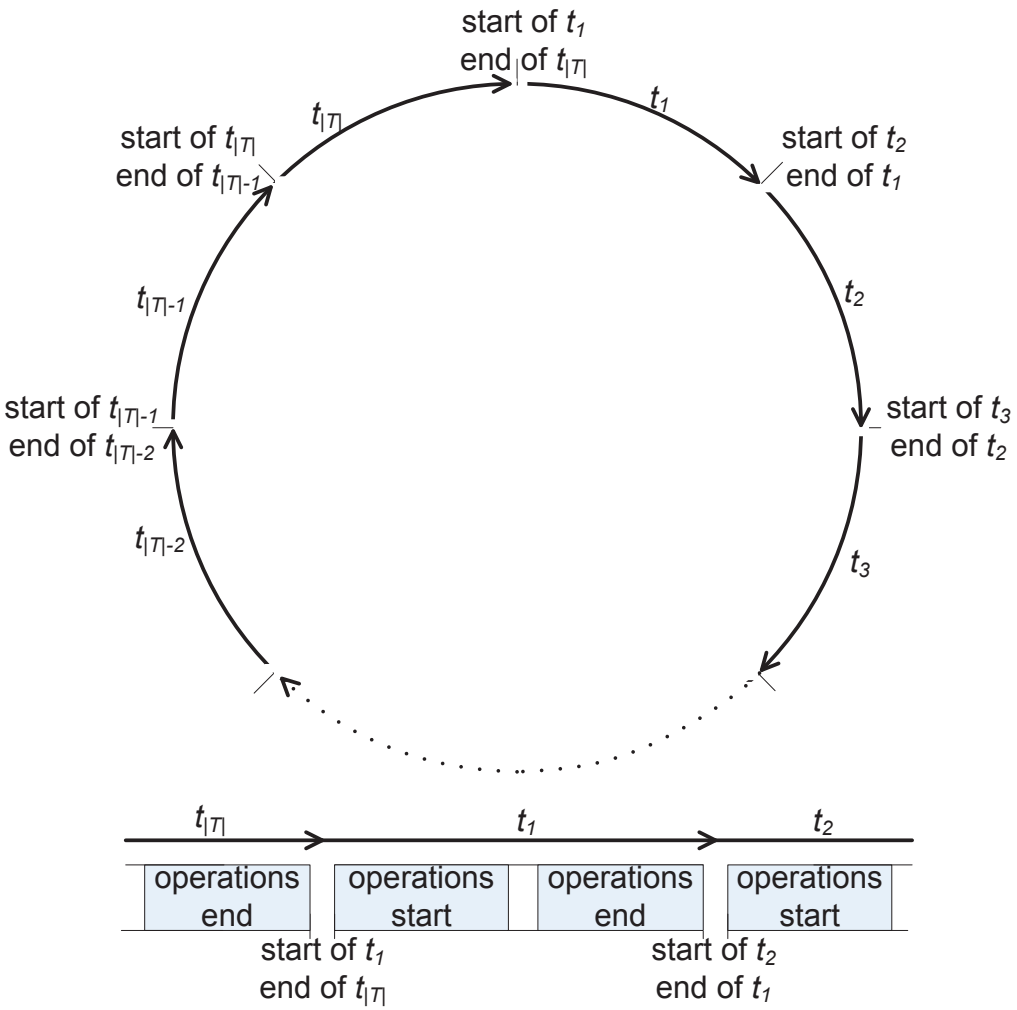


Figure 3.2: The relationship between time intervals and operations where $T = \{t_1, t_2, ..., t_{|T|}\}$ is the set of time intervals.

at the end of a time interval. The model assumes that demand is cyclic and it repeats itself on a daily basis for a given time horizon (e.g. season, day of the week) and the first time interval of a given day starts after the last time interval of the previous day (Figure 3.2).

iv. Operations: The system involves three types of operations: rental, relocation and charging.

a. Rental: The system operates on the basis of reservations and allows one-way rental of vehicles. Reservations are made in advance of the pick-up time. Origin and destination locations, and pick-up and drop-off times are also known. Vehicles are picked-up/dropped-off from/at a station that is accessible to the initial origin/destination location of the respective user at pre-specified (when reservation is made) periods. It is assumed that each rental starts at the beginning of a time interval and ends at the end of the same or a subsequent time interval (Figure 3.2).

b. Relocation: The system allows one way rental of vehicles. As a result, there might be accumulation and/or shortage of vehicles at stations. Relocation is used to rebalance the system resources, i.e. vehicles. Relocations can last more than one time interval (Figure 3.2). During relocation, the vehicle is not available with the exception of extremely closely located stations (i.e. less than 2kms) in which case rental and relocation can take place at the same time interval. The total time spend for relocation operations during a time interval cannot exceed the total available time of the staff assigned to a working shift.

c. Charging: The system modeled in this chapter uses electric vehicles. In order to model the electric vehicles charging period, it is assumed that after a vehicle is returned from a rental operation, it has to stay in the station for a fixed period of time which represents the charging period of the vehicle.

v. Working Shift: A set of consecutive time intervals defines a working shift. Working shifts are used to model the personnel needed for relocation operations.

vi. Centers: In the model, centers represent demand points that can be served by the same set of (candidate) stations. To illustrate how the centers are defined we are using the example shown in Figure 3.3. Figure 3.3a depicts the origin and destination of demand and the station locations. Figure 3.3b shows the stations that are accessible by different origin and destination locations. Please note that more than one station may be accessible from a given origin/destination point. The origin/destination points that can access the same set of stations are clustered together and constitute a center. Figure 3.3c illustrates two centers (shaded areas) and trips (demand) associated with these centers. The grouping of demand into centers decreases the number of variables since the trips with the same origin and destination centers are grouped together and allows the solution of larger instances of problems. The distance between a center and a station is the average of all distances defined by the demand points of a given center and the associated station.

vii. Demand: Demand represents an aggregation of trip reservations (orders) of rentals that are associated with the same set of origin and destination centers and have common departure and arrival time intervals. In order to satisfy an "order" (i) a vehicle from a station that is accessible from the origin location (or equivalently center) at the beginning of the departure time interval, and (ii) a parking space at a station that is accessible from the destination location (or equivalently center) at the end of the arrival time interval have to be available. Note that, "orders" do not have to be assigned to the closest stations but to accessible ones.

viii. Atoms: An atom represents a small geographical area with known population. The atoms are used to model the population coverage of the car-sharing system. In our model, we assume that there is a maximum distance that determines if an atom is covered. Thus, if there is an open station closer than the predefined maximum value (coverage distance), the atom is covered. Atoms used in population coverage of example problem can be seen in Figure 3.4.

ix. Costs and Revenues: The model includes two objective functions expressing the objectives of the users and the operator. The operator's benefits include vehicle rental revenues and

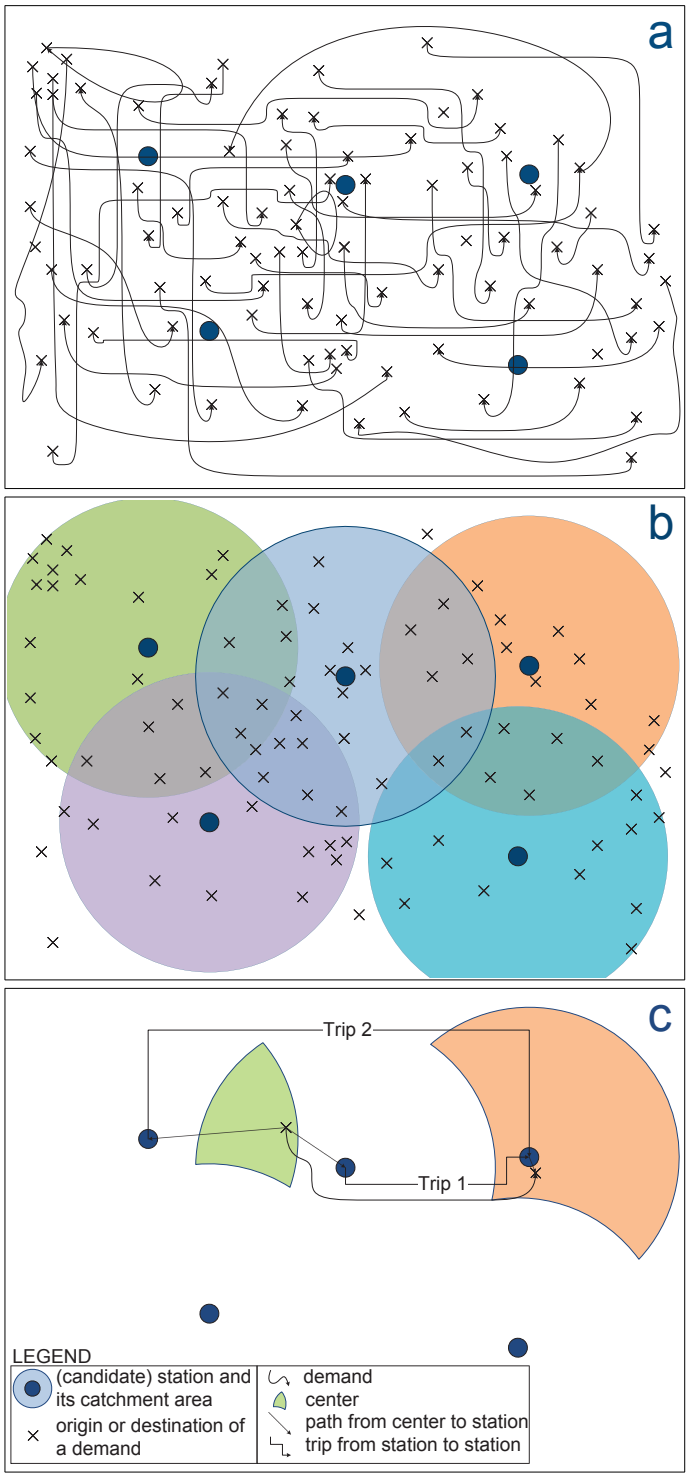


Figure 3.3: (a) Location of stations and historical trips generated between origins and destinations; (b) Origins and destinations are grouped according to the set of accessible (candidate) stations; (c) Based on this aggregation, a specific demand can be served in two different ways (trip 1 and 2)

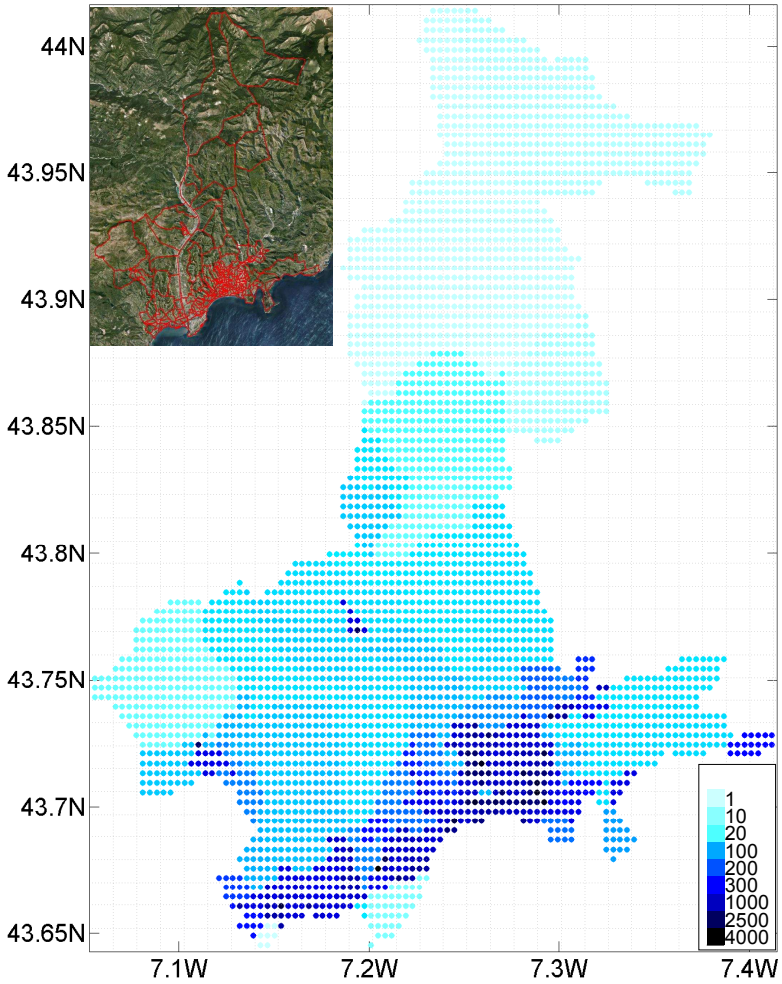


Figure 3.4: Atoms used in population coverage

subsidies, while costs include maintenance, operation and relocation of vehicles, and station opening costs. Users' net benefit is calculated as the difference between the utility gain in terms of monetary value, and the sum of vehicle rental and accessibility costs. In what follows (see items a to h below) we define all these terms.

a. Vehicle Rental Cost: The amount paid by the users to the operator to rent a vehicle expressed in €/unit time. This value is predefined and regarded as an input to the model.

b. Subsidy: It represents money paid directly to the operator, by public agencies, to cover revenue deficits per rental in €/unit time.

c. Fixed Vehicle Cost: The cost encountered by the operator expressed in €/day (e.g. depreciation, insurance)

d. Variable Vehicle Cost: The cost of the operator per km vehicle rented (e.g. cost of energy, maintenance cost due to wear-and-tear).

e. Vehicle Relocation Cost: The cost related to the relocation operations of the vehicles. It has two components: the relocation personnel cost (per shift) and the cost for driving vehicles between stations.

f. Station Operating Cost: The cost of operating a station. It is a function of the number of operating parking spaces.

g. User Utility: The monetary value of the utility gained by the users by each satisfied trip expressed in €/unit time.

h. Accessibility Cost: The monetary value of time of the users required to reach a station from their origin and from stations to their destination expressed in €/distance.

x. Scenarios: Alternative scenarios are defined by varying the input parameters of the model (e.g. weekdays, weekends). Scenarios are used to obtain a more representative average system performance.

xi. Scenario Groups: The set of scenarios which addresses the same strategic decisions and parameters (e.g. number of vehicles, relocation personnel cost) belongs to the same scenario group. In order to account for daily variation within the same season (e.g. summer, autumn, winter), each season is set as a scenario group and more than one scenarios is generated according to day of the week (e.g. weekdays, weekends).

3.3.2 Mathematical Model

In this part, we represent the mathematical structure of the proposed model. We first define the sets and indices used to describe the model as well as the functions, variables and parameters in Section 3.3.2. In Section 3.3.2, the detailed multi-objective mathematical model is given and its objective functions and constraints are described in details. The aggregate model and the rational for to have an aggregate model are presented in Section 3.3.2.

Inputs

Sets and Indices:

$i, k \in I$**:** center indices

$j, l \in J$**:** (candidate) station indices

$t, u, w \in T$**:** time interval indices

$f \in F$**:** working shift index

$a \in A$**:** atom index

$s \in S$**:** scenario index

$g \in G$**:** scenario group index

Functions:

next$(t,\#)$**:** time interval that is # intervals after time interval t

cover(a)**:** set of stations that are accessible from atom a

btwn(t,u)**:** set of time intervals from t to u

close$\left(j\right)$**:** set of stations that relocation with station j is possible during the same time interval

Parameters:

$\mathbf{SOC}_j$**:** cost for establishing station j

$\mathbf{PSC}_j$**:** cost per parking space available at station j

$\mathbf{VFC}^g$**:** fixed vehicle cost per vehicle-day in scenario group g

$\mathbf{VOC}^{stu}_{jl}$**:** operating cost of a vehicle rented at time interval t from station j to reach station l at time interval u in scenario s

$\mathbf{VRC}^{gt}_{jl}$**:** relocation cost of moving a vehicle from station j to l starting at time interval t in scenario group g

$\mathbf{AC}^{gt}_{ij}/\overline{\mathbf{AC}}^{gt}_{ij}$**:** accessing/egressing cost from/to center i to/from station j at time interval t in scenario group g

$\mathbf{RPC}^{g}_{f}$**:** cost of relocation personnel for working shift f in scenario group g

$\mathbf{RC}^{gtu}_{jl}/\mathbf{SA}^{gtu}_{jl}$**:** rental charge/subsidy when a vehicle is rented at time interval t from station j to reach station l at time time interval u in scenario group g

$\mathbf{UG}^{stu}_{jl}$**:** user utility when a vehicle is rented at time interval t from station j to reach station l at time time interval u in scenario s

$\mathbf{CAP}_j$**:** maximum number of available parking spaces for station j

COV: minimum percentage of population need to be covered by open stations

$\mathbf{PR}_a$**:** percent of population inhabiting in atom a

$\mathbf{OD}^{stu}_{ik}$**:** number of orders starting at the beginning of time interval t from center j ending at the end of time interval u at center k for scenario s

$\mathbf{RI}^{gt}_{jl}$**:** time intervals needed to relocate a vehicle from station j to l starting at the beginning of time interval t in scenario group g

$\mathbf{LRI}_{jl}^{gt}$: last time interval of relocation if a vehicle is relocated from station j to l starting at the beginning time interval t in scenario group g

$\mathbf{SI}_{f}^{g}$: time intervals included in working shift f in scenario group g

$\mathbf{RT}_{jl}^{gt}$: time spend to relocate a vehicle from station j to l at the beginning of time interval t in scenario group g

$\mathbf{WH}^{gt}$: total available working hours for a shift operating during time interval t in scenario group g

$\mathbf{SW}^{s}$: weight of the net benefit of scenario s in the objective function

$\mathbf{CT}_{jl}^{stu}$: charging periods of vehicles rented at time interval t from station j to reach station l at time interval u in scenario s

N: maximum number of open stations

$S(g)$**:** scenarios belonging to scenario group g

$G(s)$**:** scenario group of scenario s

Decision Variables:

x_j**:** binary variable showing if (candidate) station j is open or not

c_j**:** number of parking spaces operating in station j

v^g**:** number of vehicles used in scenario group g

d_a**:** binary variable showing if atom a is covered by a station or not

h_f^g**:** number of relocation personnel needed during shift f in scenario group g

Auxiliary Variables:

n_j^{st}**:** number of available vehicles in station j at the beginning of time interval t in scenario s

y_{ikjl}^{stu}**:** number of trip orders satisfied from center i renting vehicle from station j to make a trip at the beginning of time interval t to reach center k through station l at the end of time interval u in scenario s

z_{jl}^{stu}**:** number of vehicles rented from station j at the beginning of time interval t to reach station l at the end of time interval u in scenario s

m_{ik}^{stu}**:** number of unserved orders of OD_{ik}^{stu}

$p_{ij}^{st}/\overline{p}_{ij}^{st}$: number of vehicles rented/left from/to station j at the beginning/end of time interval t to/from center i in scenario s

$q_{j}^{st}/\overline{q}_{j}^{st}$: number of vehicles rented/left from/to station j at the beginning/end of time interval t in scenario s

b_{t}^{s}: number of vehicles rented before time interval t which are still rented during time interval t in scenario s

e_{t}^{s}: number of vehicles being relocated during time interval t for which their relocation started before t in scenario s

r_{jl}^{st}: number of vehicles relocated from station j to l starting from the beginning of time interval t in scenario s

Detailed Model

$$\max \sum_{(s,j,l,t)} \mathrm{SW}^s \left[\overbrace{\sum_u \left(\mathrm{RC}_{jl}^{stu} + \mathrm{SA}_{jl}^{stu} - \mathrm{VOC}_{jl}^{stu} \right) z_{jl}^{stu}}^{\text{rental charge + subsidy - vh. operating costs}} - \overbrace{\mathrm{VRC}_{jl}^{G(s)t} r_{jl}^{st}}^{\text{vh. relocation cost}} \right]$$
$$- \sum_g \sum_{s \in S(g)} \mathrm{SW}^s \left(\overbrace{\sum_f \mathrm{RPC}_f^g h_f^g}^{\text{personnel cost}} + \overbrace{\mathrm{VFC}^g v^g}^{\text{vehicle maintenance cost}} \right) - \overbrace{\sum_j \left(\mathrm{SOC}_j x_j + \mathrm{PSC}_j n_j^* \right)}^{\text{st. operating and parking costs}} \tag{3.1}$$

$$\max \sum_s \mathrm{SW}^s \left[\overbrace{\sum_{(j,l,t,u)} \left(\mathrm{UG}_{jl}^{stu} - \mathrm{RC}_{jl}^{stu} \right) z_{jl}^{stu}}^{\text{utility - rental charge}} - \overbrace{\sum_{(i,j,t)} \left(\mathrm{AC}_{ij}^{G(s)t} p_{ij}^{st} + \overline{\mathrm{AC}}_{ij}^{G(s)t} \overline{p}_{ij}^{st} \right)}^{\text{accessibility cost}} \right] \tag{3.2}$$

$$\text{s.t. } c_j \le \mathrm{CAP}_j x_j \quad \textbf{(a)} \quad n_j^{st} \le c_j \quad \textbf{(b)} \qquad \forall j \text{ and } \forall s,j,t \tag{3.3}$$

$$\sum_j x_j \le \mathrm{N} \tag{3.4}$$

$$c_j \ge x_j \quad \textbf{(a)} \quad \sum_{(s,t)} n_j^{st} \ge x_j \quad \textbf{(b)} \qquad \forall j \tag{3.5}$$

$$d_a \le \sum_{j \in \mathrm{cover}(a)} x_j \quad \textbf{(a)} \quad \sum_a \mathrm{PR}_a d_a \ge \mathrm{COV} \quad \textbf{(b)} \qquad \forall a \tag{3.6}$$

$$\sum_{(j,l)} y_{ikjl}^{stu} + m_{ik}^{stu} = \mathrm{OD}_{ik}^{stu} \qquad \forall s,i,k,t,u \tag{3.7}$$

$$\sum_{(i,k)} y_{ikjl}^{stu} = z_{jl}^{stu} \qquad \forall s,j,l,t,u \tag{3.8}$$

$$\sum_{(k,l,u)} y_{ikjl}^{stu} = p_{ij}^{st} \quad \textbf{(a)} \quad \sum_{(k,l,u)} y_{kilj}^{sut} = \overline{p}_{ij}^{st} \quad \textbf{(b)} \qquad \forall s,i,j,t \tag{3.9}$$

$$\sum_{(i,k,l,u)} y^{stu}_{ikjl} = q^{st}_j \quad \textbf{(a)} \qquad \sum_{(i,k,l,u)} y^{sut}_{iklj} = \overline{q}^{st}_j \quad \textbf{(b)} \qquad \forall s,j,t \quad (3.10)$$

$$q^{st}_j \le n^{st}_j - \sum_l r^{st}_{jl} + \sum_{l \in \text{close}(j)} r^{st}_{lj} \qquad \forall s,j,t \quad (3.11)$$

$$n^{st}_j - q^{st}_j + \overline{q}^{st}_j - \sum_l r^{st}_{jl} + \sum_{\substack{(l,u):\\ \text{LRI}^{G(s)u}_{jl}=t}} r^{su}_{jl} = n^{\text{snext}(t,1)}_j \qquad \forall s,j,t \quad (3.12)$$

$$\sum_{\substack{(j,l,u,w):\\ t \in \text{btwn}(u,w)\setminus u}} z^{suw}_{jl} = b^{st} \qquad \forall s,t \quad (3.13)$$

$$\sum_{\substack{(j,l,u):\\ t \in \text{btwn}\left(u,\text{LRI}^{G(s)t}_{jl}\right)\setminus u}} r^{su}_{jl} = e^{st} \qquad \forall s,t \quad (3.14)$$

$$\sum_j n^{st}_j + b^{st} + e^{st} \le v^{G(s)} \qquad \forall s,t \quad (3.15)$$

$$\sum_{\substack{(j,l,u):\\ t \in \text{RI}^{G(s)u}_{jl}}} \text{RT}^{G(s)t}_{jl} r^{su}_{jl} \le \text{WH}^{G(s)t} h^{G(s)}_f \qquad \forall s,f,t \in \text{SI}^{G(s)}_f \quad (3.16)$$

$$r^{st}_{jl} \le c_j \quad \textbf{(a)} \qquad r^{st}_{lj} \le c_j \quad \textbf{(b)} \qquad \forall s,j,l,t \quad (3.17)$$

$$n^{st}_j \ge \sum_{(s,j,l,u,w): t \in \text{CT}^{suw}_{lj}} z^{suw}_{lj} \qquad \forall s,j,t \quad (3.18)$$

$$x_j, d_a \in \{0,1\}; \quad n^{st}_j, q^{st}_j, c_j, \overline{q}^{st}_j, b^{st}, e^{st}, v^s \in \mathcal{N} \qquad \forall j,a,s,t \quad (3.19)$$

$$p^{st}_{ij}, \overline{p}^{st}_{ij}, r^{st}_{jl}, m^{stu}_{ik}, z^{stu}_{jl}, y^{stu}_{ikjl}, h^g_f \in \mathcal{N} \qquad \forall s,i,k,j,l,t,u,g,f \quad (3.20)$$

The problem formulation is described in equations 3.1-3.20. To summarize, equations 3.1-3.2 are the two objectives of the model. Constraints 3.3-3.6 and 3.17 are related to station opening and capacities. Constraints 3.7-3.10 and 3.14 establish relations between decision and auxiliary variables. Constraints 3.11-3.12 satisfies feasibility of station capacities during operations. Constraints 3.15-3.16 satisfy vehicle and personnel restrictions respectively. Constraints 3.18 are for charging restrictions and can be omitted for combustion vehicle systems.

The first objective function (Equation 3.1) expresses the maximization of the net revenue for the operator. Net revenue is calculated as the difference between the sum of total rental revenue and subsidy minus station, vehicle and relocation costs. Note that all of the values in both objective functions except station opening cost are weighted analogous to the number of days (e.g. five for weekdays, two for weekends) of each scenario (SW^s). This is due to the fact that the location of the stations and the number of parking spaces are regarded as strategic decisions and therefore have to be the same in all scenarios. However the rest of the parameters are scenario (e.g. the number of vehicles) specific. The net revenue for the trip starting from station j to station l from the beginning of time interval t to time interval u in scenario s of given type equals the rental charge per trip $\left(\text{RC}^{stu}_{jl}\right)$ plus subsidy $\left(\text{SA}^{stu}_{jl}\right)$ minus operating cost $\left(\text{VOC}^{stu}_{jl}\right)$ times the number of trips of the same type served $\left(z^{stu}_{jl}\right)$.

The relocation cost has two components: (i) The vehicle cost related to the total km driven to relocate and (ii) the labor cost associated with the cost of the personnel used to relocate the vehicles. The total vehicle relocation cost is equal to the expenses of all the relocation operations. The vehicle relocation cost for the relocation starting from station j at time interval t to station l in scenario s is equal to the sum per relocation $\left(\text{VRC}_{jl}^{G(s)t}\right)$ times the number of relocations $\left(r_{jl}^{st}\right)$. Similarly, the relocation personnel cost equals the sum of all personnel costs. The total personnel cost for shift f in scenario group g equals the unit personnel cost $\left(\text{RPC}_{f}^{g}\right)$ times the number of staff hired for this shift $\left(h_{f}^{g}\right)$.

The fixed vehicle cost depends on the total number of vehicles operating in the system. For scenario s, this cost is equal to the product of the unit fixed vehicle cost $\left(\text{VFC}^{g}\right)$ and the number of vehicles in the system $\left(v^{g}\right)$ in scenario group g. Note that, for scenarios belonging to the same (scenario) group, the number of vehicles is the same, since we regard the number of vehicles as a tactical decision.

The station operating and parking space costs are the costs dedicated to station operations. There is a fixed cost for operating a station $\left(\text{SOC}_{j}\right)$ and a variable cost $\left(\text{PSC}_{j}\right)$ for each parking space $\left(n_{j}^{*}\right)$ operating at given station j.

The second objective (Equation 3.2) expresses the maximization of the users' net benefit. UG_{jl}^{stu} can be defined as the monetary value (i.e. €) of the utility gain for each realized trip starting from station j to station l from the beginning of time interval t to time interval u in scenario s of the same type. Similarly, the rental fee is the money paid to the operator for the rental of vehicles by the users $\left(\text{REV}_{jl}^{stu}\right)$ and total rental charge equals the sum of them. The accessibility cost is the cost associated with the access or egress of a station from a center.

Constraints 3.3 restrict the number of parking spaces (station capacity constraint), and the number of available vehicles for each time interval and station. If a station is not open in a candidate station location, the station capacity is set to zero. If the station is open, then there is an upper bound (CAP_{j}) for its capacity. Constraint 3.4 limits the total number of operating stations. Constraints 3.5a and 3.5b require that if a station is open, at least one parking space and an operation (i.e. rental, relocation) from this station should be assigned as well. These constraints are essential in order to guarantee the coverage of the demand by an open station that has at least a capacity of one parking space. Constraints 3.6a and 3.6b are the atom coverage constraints, i.e. if an atom is covered or not, and population coverage constraints, i.e. the car-sharing system is accessible by a given percentage of the population, respectively. Constraints 3.7 ensure that the total number of orders is equal to the sum of the satisfied demand and unserved (lost) orders.

Constraints 3.8 postulate that the total satisfied demand that is assigned to a trip starting from center i at the beginning of time interval t by a vehicle from station j, ending at the end of time interval u in center k through station l in scenario s over origin/destination center pairs (i,k), is equal to the trip starting from station j to station l from the beginning of time interval

t to time interval u in scenario s. Similarly, constraints 3.9a indicate that the total number of the trips defined above from station j is equal to the number of vehicles rented from station j at the beginning of time interval t to serve demand from centers i. Constraints 3.9b do the same as constraints 3.9a for the vehicles originating from center i left at station j at the end of period t. Constraints 3.10a and 3.10b are equivalent to Constraints 3.9a and 3.9b and ensure respectively the same conditions for the vehicles that are rented/left from/to a station j. Thus, Constraints 3.8, 3.9a, 3.9b, 3.10a and 3.10b establish the functional relationship between the variables y, and z, p ($\overline{p}$) and q ($\overline{q}$) respectively. Please note that, variables z express vehicle assignments independent of the center to which originate/end their movement, variables p and $\overline{p}$ indicate customer movements from centers to stations and from stations to centers respectively, and variables q and $\overline{q}$, signify the number of vehicles rented from and left to a station respectively.

Constraints 3.11 require that the number of vehicles leaving a station (due to rental and relocation) at the beginning of interval t cannot exceed the number of vehicles available at that stations at the same time interval. Constraints 3.12 are the "vehicle conservation" constraints for each station.

Constraints 3.13 and 3.14 are used to establish the functional relationship between variables b, e, and z, r respectively. Variables b and e are used in Constraints 3.15 to determine the total number of vehicles (fleet size) of the system. Constraints 3.16 are introduced to ensure the per shift availability of the workforce needed to perform vehicle relocations.

Constraints 3.17a and 3.17b set an upper bound to relocation from and to every station respectively. This upper bound equals to the number of operating parking spaces in related station. For a station which is not open, the number of relocations from and to this station are set to zero with the same constraints respectively.

Constraints 3.18 are restrictions specific to electric-car-sharing systems. These constraints force the vehicles to stay and be charged, after each rental operation, at the station they arrived. These constraints require that the number of vehicles in the station should be greater than or equal to the number of vehicles requiring charging.

Aggregate Model

$$\text{Equations } 3.2 - 3.10,\ 3.13,\ 3.15,\ 3.18 - 3.20 \tag{3.21}$$

$$\max \sum_{(s,j,l,t)} \mathrm{SW}^s \left[\sum_u \left(\mathrm{RC}_{jl}^{stu} + \mathrm{SA}_{jl}^{stu} - \mathrm{VOC}_{jl}^{stu} \right) z_{jl}^{stu} \right] - \overbrace{\sum_{(s,j,t)} \mathrm{SW}^s \mathrm{VRC}_j^{G(s)t} \left(r_j^{st} + \overline{r}_j^{st} \right)}^{\text{new vh. relocation cost}} \tag{3.22}$$

$$- \sum_g \sum_{s \in S(g)} \mathrm{SW}^s \left(\sum_f \mathrm{RPC}_f^g h_f^g + \mathrm{VFC}^g v^g \right) - \sum_j \left(\mathrm{SOC}_j x_j + \mathrm{PSC}_j n_j^* \right)$$

$$q_j^{st} \leq n_j^{st} - r_j^{st} + \sum_{l \in \text{close}(j)} r_l^{st} \qquad \forall s, j, t \qquad (3.23)$$

$$q_j^{st} + \sum_{l \in \text{close}(j)} q_l^{st} \leq n_j^{st} + \sum_{l \in \text{close}(j)} n_l^{st} \qquad \forall s, j, t \qquad (3.24)$$

$$n_j^{st} - q_j^{st} + \overline{q}_j^{st} - r_j^{st} + \sum_{(j,u):t=\text{LRI}_j^{G(s)u}} \overline{r}_j^{su} = n_j^{s\text{next}(t,1)} \qquad \forall s, j, t \qquad (3.25)$$

$$\sum_{(j,t):t=\text{LRI}_{jl}^{G(s)u}} r_j^{su} = \sum_j \overline{r}_j^{st} \qquad \forall s, j, t \qquad (3.26)$$

$$r_j^{st} \leq c_j \quad \textbf{(a)} \qquad \overline{r}_j^{st} \leq c_j \quad \textbf{(b)} \qquad \forall s, j, t \qquad (3.27)$$

$$\sum_{(j,u):t \in \text{btwn}\left(u, \text{LRI}_j^{G(s)t}\right) \setminus u} r_j^{su} + \overline{r}_j^{su} = e^{st} \qquad \forall s, t \qquad (3.28)$$

$$\sum_{(j,l,u):t \in \text{RI}_j^{G(s)u}} \text{RT}_j^{G(s)t} \left(r_j^{su} + \overline{r}_j^{su}\right) \leq \text{WH}^{G(s)t} h_f^{G(s)} \qquad \forall s, f, t \in \text{SI}_f^{G(s)} \qquad (3.29)$$

$$r_j^{st} \geq 0 \qquad \overline{r}_j^{st} \geq 0 \qquad \forall s, j, t \qquad (3.30)$$

In real life instances, the model described by equations 3.1-3.20 may result in problem sizes that are not possible to be efficiently solved. Although for most of the variables, we only generate those that have positive values and construct the corresponding constraints accordingly, we do not have this opportunity for the relocation variables r_{jl}^{st}. As the relocations can happen between any station pairs, we need to generate $|J|^2|S||T|$ number of variables which renders the case of Nice, France impossible to solve. An instance of 142 candidate stations, 12 scenarios and 15 time intervals needs more than 3.6 millions variables of type r_{jl}^{st} only. In order to cope with this issue, we assume that the relocated vehicles are firstly accumulated in an imaginary hub and then distributed from that hub to the stations. For this issue, two new variables, r_j^{st} and $\overline{r}_j^{st}$ are defined as that the number of vehicles relocated from/to station j starting from the beginning/finishing at the end of time interval t in scenario s. With this change, the number of variables of type r decreases to $2|J||S||T|$ which means 51120 variables instead of over 3.6 millions.

In addition, we substitute the constraints 3.11, 3.12, 3.14 and 3.16 and 3.17, with the following constraints 3.23-3.30. Moreover, the vehicle relocation cost part of the operator's objective function (Equation 3.1) is replaced with Equation 3.22. Note that, parameters LRI_j^{gt}, VRC_j^{gt} and RT_j^{gt} shows the last time interval, the vehicle relocation cost and time spent when a vehicle is relocated from/to station j to/from hub starting in time interval t in scenario group g respectively. The aggregate model is expressed by equations 3.21-3.30.

Constraints 3.23 and 3.24 replace constraints 3.11. Constraints 3.23 postulate that the total number of trip starting from station j at the beginning of time interval t in scenario s cannot be more than the number of available vehicles at the beginning of the time interval t; minus the number of relocations from station j; plus the number of relocations from the stations that are close enough to station j to have relocations at the same time interval. Constraints

3.24 set an upper bound for each station group close enough to have relocations to the same station. For each set of stations, the total number of trips started from the corresponding set of stations cannot be more than the total number of available vehicles at these stations.

Constraints 3.25 replace constraints 3.12 of the first model. Constraints 3.26 require that the total number of relocations from stations to the imaginary hub ending in time interval t should be equal to the number of relocations to the stations from the imaginary hub starting in time interval t. This is applicable for each time interval and scenario.

Constraints 3.27a and 3.27b replace constraints 3.17a and 3.17b. They set the number of relocations to the number of operating parking spaces. Constraints 3.28 and 3.29 work the same as constraints 3.14 and 3.16 respectively. The former constraints calculate the number of vehicles under relocation whereas the latter constraints decide on the manpower need for each time interval in each scenario.

3.4 Model Application

The model presented in Section 3.3.2 was applied to plan a one-way electric-car-sharing system in Nice, France. The study area is 294.19km^2 and has a population 327188 inhabitants between ages 15-64 with a density 1112 persons/km^2. The area under consideration consists of 210 regions. The population of each region was obtained from 2009 census data (INSEE, 2010). We assume that the population is uniformly distributed inside regions and calculate the population of each atom accordingly. The atoms and their population can be seen in Figure 3.4.

The whole model is implemented in C# .NET environment. IBM ILOG Cplex Version 12.5 with Concert Technology is used for solving MILPs. To cope with the enormous number of relocation variables, the aggregate model (Section 3.3.2) is used. The exact and aggregated relocation costs will be compared later. For each station, half of the average distance of closest n stations is calculated and regarded as the distance of the same station to the imaginary hub. This approach generates values that closely approximate real relocation distances. To further investigate the performance of the approximation, a simulation environment that compares average real and hub relocation distance for 1000 cases is generated with different n values. In Figure 3.5, the error for different values of the number of relocations (n) are compared. We use n = 20 which results to an average minimum error. In other words, when distance for relocation is calculated, the distance from a station to the hub is assumed half of the average distance of 20 closest (candidate) stations. Note that in the aggregate model a relocation is composed of two legs: relocating vehicle from its old station to the imaginary hub and to its new destination from the hub. A similar approach is used for the second leg. The number of relocations per personnel has values between seven and 15 which results in error not more than 0.7km per relocation. Since distance per relocation observed is around 4km and the total cost of relocation is not more than 20% of the objective function value of each case (see in figures 3.9 and 3.10), this relaxation might not create an error more than 3.5%. Also

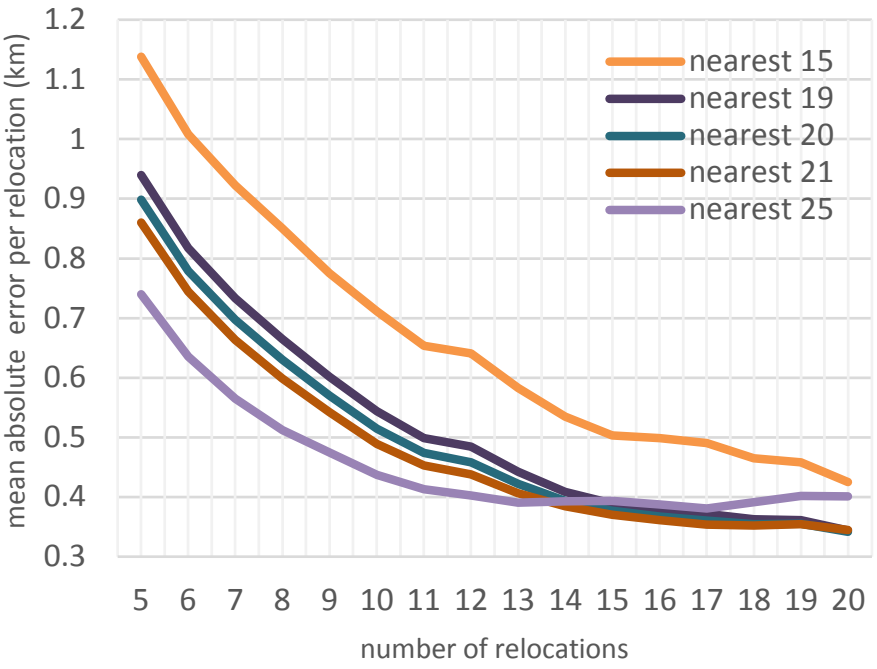


Figure 3.5: Average absolute error of imaginary hub usage in relocation for different number of relocations. Different n values are compared in order to find the most suitable value for our case.

post-analysis showed that the difference between the cost of relocation operations calculated by the aggregate model and the exact model is less than 2% of the operator's revenue on average. In order to deal with the extremely large size of the problem, we take advantage of the sparsity of the matrices of the variables and we do not generate the variables that have zero value. This decreases the number of variables of aggregate model in order of magnitude from 10 to 5.

To guarantee generation of feasible solutions in reasonable time, extra cuts are generated with CPLEX. The runs are taken on a computer with 3.00 Ghz Intel Core 2 Quad CPU and 8 GB of RAM. All runs are realized as single threaded programs and every run is terminated when either they reach 2% optimality gap or 9 hours run time. Most of the runs that are represented here were terminated in less than three hours and all of the runs had an optimality gap less than 8%.

The summary of the methodology for the entire approach can be seen in Figure 3.6 where w_{operator} and w_{users} stands for weights of operator and users benefit respectively. The terms *superior* and *inferior* used in finding candidate station section refers to superiority and inferiority in coverage respectively. If a candidate station covers one more origin or destination location in addition to another candidate station's covered locations, the former candidate station is superior to the latter and latter is inferior to the former.

1. Data reading and parameter creation

(a) Read population data and create atoms

(b) Read historical demand data

(c) Conversion from two-way demand data to one-way

i. For each historical demand datum

A. If waiting time is greater than predefined value, split the historical demand and create two new demands

(d) Time Interval Selection

i. Set working shifts

ii. Find time intervals consistent with working shifts that minimizes the variation of demand count in each time interval

2. Finding candidate locations

(a) Finding all candidate locations for set covering model

i. Iterate over all demand origin and destination locations

A. If current location is *superior* to any previously added location, remove previously added *inferior* solution

B. If current location is not inferior to any previously added location, add current location to locations for candidate locations for set covering model

(b) Finding candidate locations for aggregate model

i. Set $i = N$

ii. While $|J|$ < maximum number of stations

- Run *maximal set covering problem* Church and ReVelle (1974) with number of sets = i
- Add candidate locations found in the solution to $|J|$
- Set $i \leftarrow i+1$

3. Mathematical model

(a) Select predefined number of days from historical demand and create scenarios

(b) Set values of $w_{\text{operator}} > 0$ and $w_{\text{users}} > 0$

(c) Variable creation

i. Create variables d_a, x_j, n_j, c_j and b_t^s

ii. For each demand

- Create (or increment upper bound if already created) variables y_{ikjl}^{stu}, m_{ik}^{stu} and z_{jl}^{stu}
- Create (or increment upper bound if already created) variables p_{ij}^{st}, $\overline{p}_{ij}^{st}$, q_j^{st} and $\overline{q}_j^{st}$

iii. Create variables r_{jl}^{st}, $\overline{r}_{jl}^{st}$ and e_t^s (if $\neq 0$)

iv. Create variables h_f^g and v^g

(d) Constraint creation

i. For each $j \in J$: Create constraints 3.3a, 3.5a and 3.5b

ii. For each (s, j, t): Create constraint 3.3b

iii. For each $a \in A$: Create constraint 3.6a

iv. Create constraints 3.4 and 3.6b

v. For each demand: Create constraints 3.7-3.10

vi. For each (s, t): Create constraints 3.13, 3.15 and 3.28

vii. For each (s, j, t): Create constraints 3.18, 3.23-3.27

(e) Create objective function with the multiplication of w_{operator} and w_{users}, and 3.22 and 3.2 respectively

(f) Solve the model and, calculate net users' and operator's benefit

Figure 3.6: Summary of the methodology for the entire approach with the weights w_{operator} and w_{users} for the users' and operator's benefit respectively

3.4.1 Car-Sharing System in Nice

The current system operating in Nice is a two-way car-sharing system (no need for relocation operations). However, the proposed model deals with the case of one-way car-sharing, which makes the implementation more demanding. Therefore, there was a need to convert the existing two-way car-sharing data into one-way. This conversion was achieved by looking at the current database and creating one-way data by splitting the trips into one-way legs when the idle time of the rented vehicle at a given location was exceeding one hour, and the location was accessible from a station (i.e. the distance between the location and the stations is less than 500m). The problem formulation and solution procedure of Section 3.3 are not affected by this conversion and other methods could be utilized to generate the one-way demand (e.g. population surveys) (Efthymiou et al., 2013).

We use the origin and destination locations of the real demand in two steps. First, we solve a *maximal set covering problem* (Church and ReVelle, 1974) to identify the candidate station locations for the aggregate model. For each origin and destination, the (existing or candidate) stations that are accessible (the distance between two points is less than the maximum accessibility distance) are found. In addition to existing 42 stations, the model was forced to choose 100 new candidate locations for the stations. Second, we group the locations into centers. This grouping was done according to the (existing or candidate) stations that are accessible to them. The locations with the same accessible stations were assigned to the same centers. The accessibility distance between a center and a station is calculated by taking the average of the distance between the elements of the center and the station (Figure 3.3). The graph showing the locations of the origin and destination of the trips (crosses), the operating (blue) and candidate (red, gray and black) stations' locations (dots) and their catchment areas (circles with the same colors) can be seen in Figure 3.7 in which x-axis shows the longitude and y-axis shows the latitude values. Note that, the covered origin and destination locations by already operating and/or selected candidate stations have dark gray color, and each grid is a square with sides of 1 km.

After solving set covering problems, the set of candidate locations for the aggregate model (defined in Section 3.3.2) is produced. The aggregate model is solved with different weights (of users' and operator's benefit) in order to generate an efficient frontier for the given case. A total of 8 different scenarios of four seasons (spring, summer, autumn, winter) for two different day groups (weekdays, weekends) were selected. A working shift is assigned for each time interval. It was also assumed that the number of operating vehicles and relocation personnel for the same season is the same. This is because the fleet and crew size decisions are considered tactical and do not change within the same season. Each scenario was constructed by using two days of the real demand of the same day group in the same season. The capacity of each station was set to five vehicles and the model was asked to choose 28 more stations (from a set of 100 candidates) in addition to 42 stations that are already operating. Each day was divided into 15 time intervals. The time intervals are generated in such a way that the total duration of rental time (vehicle-hours) in each time interval in the historical demand are almost equal.

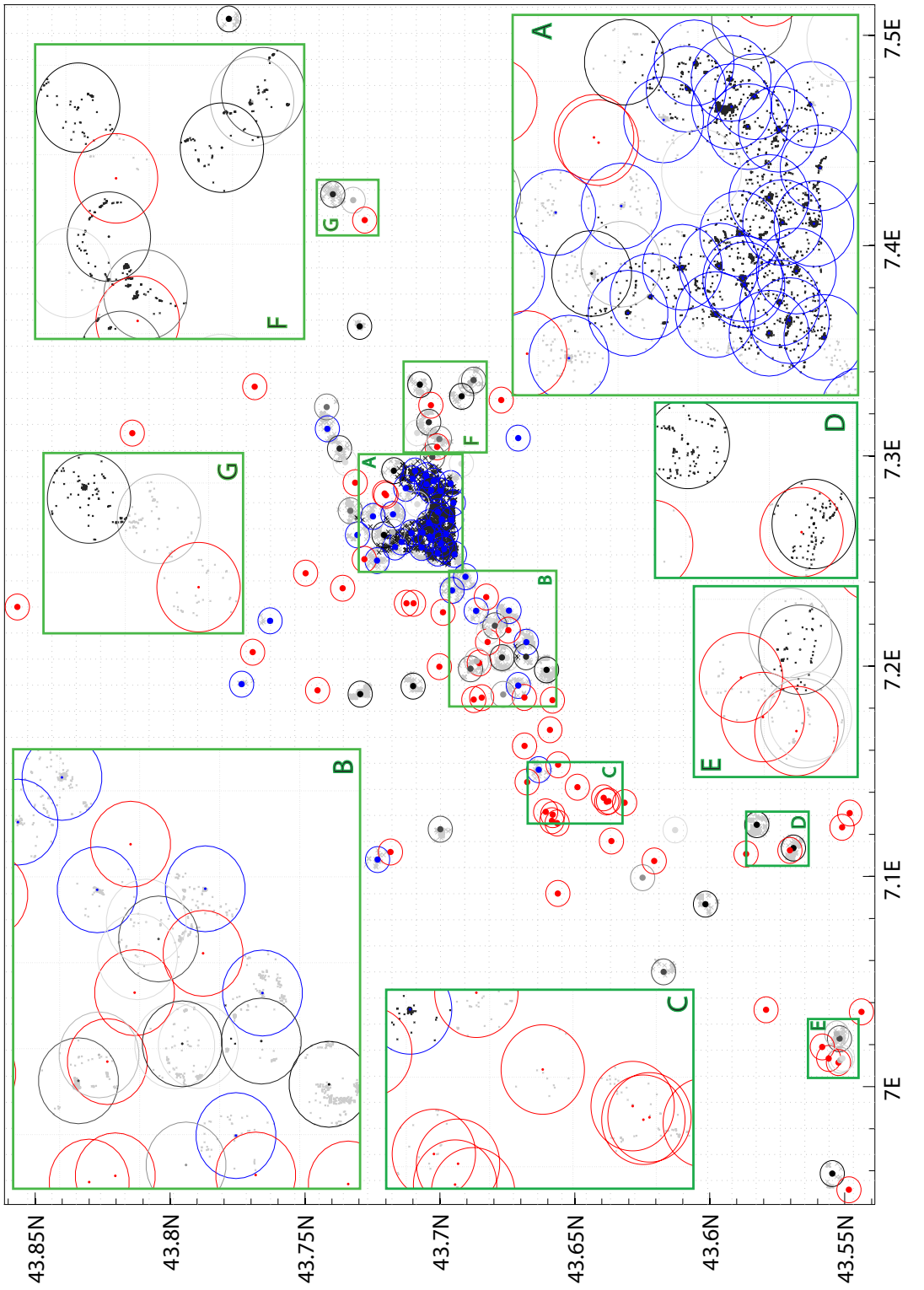


Figure 3.7: The origin and destinations of the divided trips, the operating (blue) and candidate (gray, black and red) stations and their catchment areas

fixed vehicle cost (€/day):	20
vehicle operating cost (€/km):	0.01
average number of trips per scenario:	155.2
average trip length (km):	30
max accessibility distance (km):	0.5
minimum coverage:	20%
subsidy (€/hour):	5
revenue per time interval (€/hour):	8
accessibility cost (€/km):	5
utility (€/hour):	12
relocation speed (km/h):	30
relocation personnel cost (€/h):	18

Table 3.1: Values of the parameters used in the model

Given that: i) each vehicle has a maximum range of 120km, ii) the average trip length is 30km, and iii) it takes 8 hours to fully charge an empty battery, it follows that each vehicle should be charged at least for 2 hours before it becomes operationally available. An average value for the charging duration is utilized for all trips as the operator is not aware of the distance that will be traveled by the driver at the beginning of the trip. A more detailed model can be solved in the operational problem, where uncertainty in the duration of charging can be considered. The values for some of the other parameters applied in the model are presented in Table 3.1. The fuel cost is low because the system is operating with electric vehicles. Note that, the stated values have been properly modified to ensure data confidentiality.

Using the parameters presented in Table 3.1, we solved the model and generated the efficient frontier provided in Figure 3.8 by using weighted sum method Cohon (2004). The selected candidate stations can also be seen in Figure 3.7. The candidate stations shown with red color are the candidates that are not selected, the ones with gray and black are the stations selected at least once. The intensity of the color given to the selected candidate stations increases as the frequency of their appearance in the efficient frontier increases. For instance, black means the candidate station appears in all the efficient solutions whereas the lightest gray suggests that it appeared in only one of them. The circles around each station shows the stations' accessibility area which is a circle with 500m of radius.

As it can be seen in Figure 3.7, although the part of the data used to create the efficient frontier composed of 16 of the 421 days, selected candidate stations manage to cover locations with high demand. For instance, there is an accumulation of demand around the coordinates 43.73N-7.19E and the model selects to operate a station there in all efficient solutions.

From the efficient frontier shown in Figure 3.8, it can be seen that the operator should sacrifice some of its net revenue in order to improve total users' benefit and vice versa. Although the revenue and subsidy of the served demand is higher when more demand is served, the rate of increase of the operational costs (e.g. vehicle operating cost, relocation cost) is higher than

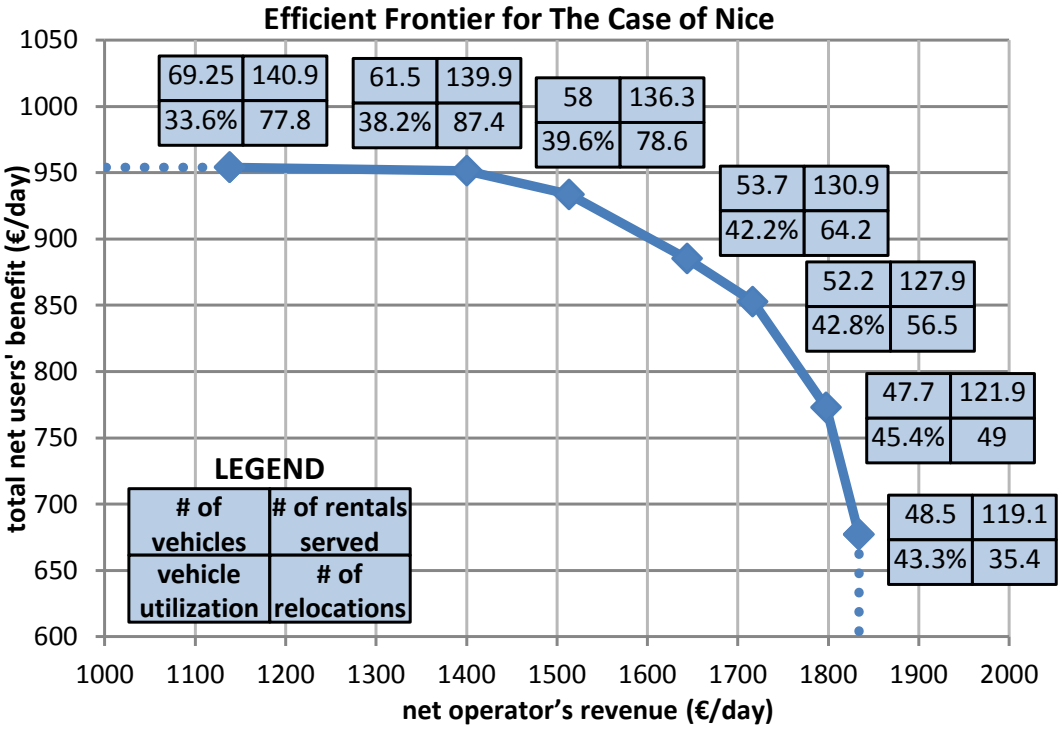


Figure 3.8: The efficient frontier for the case of Nice, France.

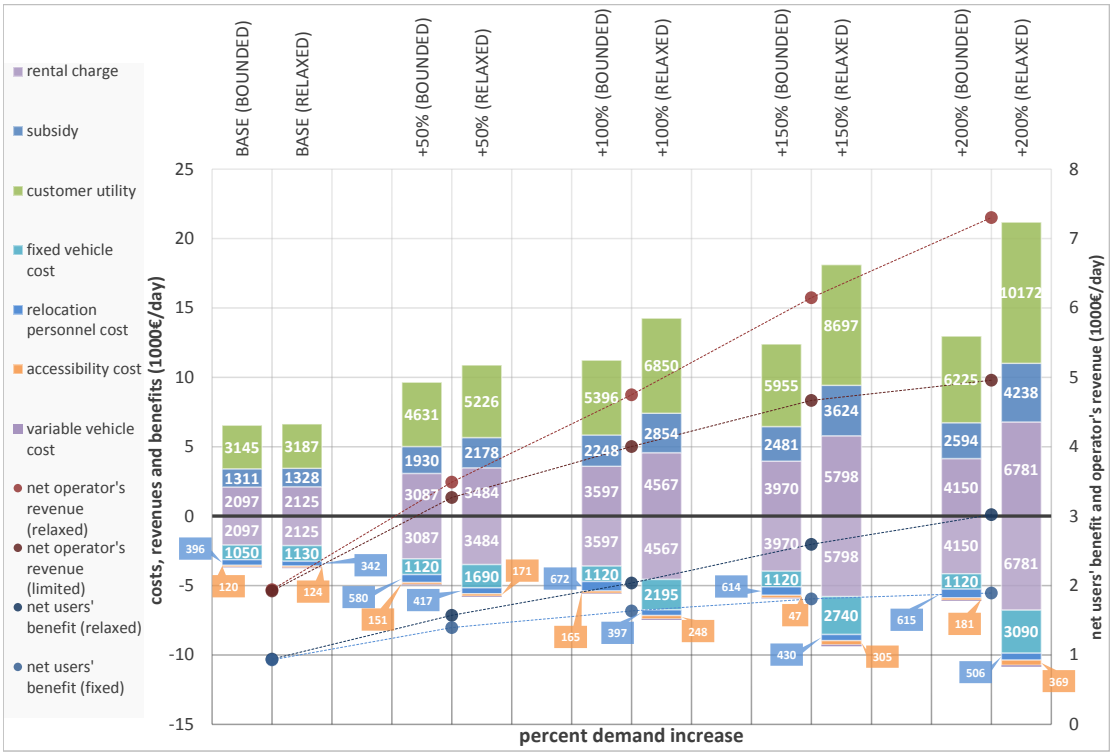
the rate of increase of the associated benefits. Both the number of vehicles in the system and the increase of relocation operations decrease the utilization of the vehicles.

Another interesting result is associated with the selection of common stations in determining the efficient frontier. It is observed that (in addition to 42 already operating stations) all seven efficient solutions select stations among a set of 46 candidate locations. More specifically, 13 of these stations appear in all solutions; 5, 7 and 3 in six, five and four solutions (out of seven) respectively. This result suggests that from station location point of view, the efficient station locations are not in conflict when considering the user and the operator objectives and the solution is robust. Since there is no conflict in station locations, these 28 stations are assumed to be operating stations in addition to already operating 42 stations in the further analysis.

After deciding about the number and location of the stations (strategic decision), we perform further analysis in order to explore if different demand levels, coverage distances and subsidy amounts influence the solution.

3.4.2 Effect of Demand

Firstly, we examine the effect of demand by using five different levels and equal weight for the users' and the operator's objectives. The results of these runs are demonstrated in Figure 3.9. In Figure 3.9, there are two sets of bar charts for each level of demand. These bar



average # of lost demand	32.9	29.0		88.7	39.0		151.8	48.4		238.4	64.3		311.2	78.9						
average # of served demand	122.3	126.3		152.2	201.9		156.7	260.1		156.4	330.5		156.2	388.5						
average # of demand	155.2			240.9			308.5			394.8			467.4							
average lost demand ratio	21.2%	18.7%		36.8%	16.2%		49.2%	15.7%		60.4%	16.3%		66.6%	16.9%						

Figure 3.9: The costs, benefits and revenues with the increased demand

charts correspond to different number of available vehicles, bounded vs. relaxed. Bounded is referred to the cases where the number of vehicles is forced to be less than or equal the corresponding number of the baseline scenario. Please note that, in the relaxed case there is no such constraint. Moving from left to right we generate for both cases (bounded and relaxed), alternative demand levels by increasing the baseline demand by 50% up to the level of 200%. The table at the bottom of the graph, summarizes the total number of trip requests, the number of lost demand and their percentage.

For the relaxed case, the operator's benefits for increasing levels of demand are increasing faster than the users' benefits. In the bounded case we observe the same pattern. As the demand increases, net benefits are increasing since the model can select to serve the most profitable customers from a larger pool of candidate customers. In the bounded case, the slope of users' and operator's benefits curves decreases as the demand increases. This is because of the limitation on the number of vehicles. This is an expected result since the model does not penalizes lost demand while at the same time increases the value of the objective function from the served demand. Note that, this increase of demand results to a higher density of demand, a fact that gives more flexibility to the model to select customers leading to improved objective function values. For the 50% increased demand, the benefit lost for both the operator

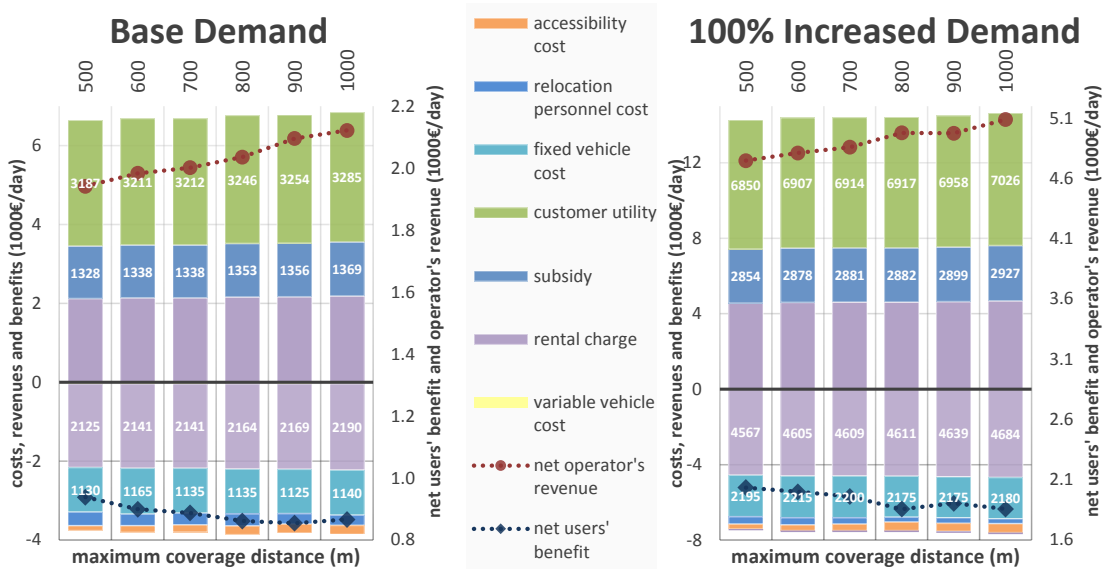


Figure 3.10: The costs, benefits and revenues for different maximum accessibility distances

and the users are almost imperceptible. However, the difference between the relaxed and bounded cases becomes significant with a demand increase of 100%. This means that in case of significant increase in demand, the system has to be redesigned in some aspects to improve the quality of service and revenues.

Another important finding is the relationship of costs, benefits and revenues. Since the rental fee is a cost for the users and a benefit for the operator, it has no effect in our objective function for this specific example since equal weights are used for the users' and the operator's objectives. The subsidy and the users utility are the only two values contributing to the increase of the value of the objective function and consequently more orders (customers) are served.

In the calculation of the required relocation personnel, it is observed that relocation cost is not significantly affecting operator's income. In the most congested system, not more than 35 hours of relocation personnel are required which corresponds to a cost of €615, about 12% of the rental charge. This finding suggests that relocation operations donot significantly increase the operator's cost.

The accessibility cost is not significant because both the accessibility cost per km (5 €/km) and maximum accessibility distance (0.5 km) are substantially lower compared to other costs (e.g. utility: 20€/h, relocation personnel cost: 18€/h).

Another important finding is related to the change in the percentage of unserved requests. The unsatisfied demand is increasing with the increase of the total number of trips. In the "relaxed" cases the percentage of lost demand is decreasing until +100% demand. This may be due to the fact that the cost of unserved demand due to shortage of vehicles is less than the cost of

acquiring extra vehicles to serve the lost demand. However, we observe an increase in the percentage of lost demand when demand is more than doubled. From a detailed observation of the results, it can be inferred that the concentration of demand during specific intervals at specific geographical locations is high. During these intervals the model either prefers not to serve additional "orders", since the cost is more than the benefit or cannot manage to serve extra demand since it reaches its limitations in busy time intervals. On the other hand in bounded cases, when demand is increased more than 100%, the number of demand served does not change. A careful look at the results show that, the bounded system reaches its limitations and cannot serve more customers without increasing system resources (e.g. the number of vehicles).

3.4.3 Effect of Accessibility Distance

The effect of maximum accessibility distance was also investigated for two different levels of demand (e.g. base and +100%). Six different accessibility distances from 500 to 1000m in every 100m intervals were tested. The demand generated for the 500m accessibility distance is used for all 6 cases to test only the effect of flexibility. Figure 3.10 shows the value of the objective function components (left axis) and the operator's and users' net benefits (right axis) as a function of maximum coverage distance.

In both graphs shown in Figure 3.10, it can be seen that the maximum accessibility distance decreases the net users' benefit slightly (around 1%) while operator's revenue is improved 1-4% for each accessibility distance increment. However, the same trend is not followed by the demand served. These two results are the consequence of the flexibility introduced to the system. The average number of accessible stations for the covered origin or destination points increases from 2.30 to 6.65. The increase of the number of accessible stations, results to an expanded feasible region and leads to an improvement of the operator's revenue. Since accessibility cost is low (5€/km) compared to operational costs of the operator, the model leads to choice that decrease the operational cost when accessibility distance is increased.

This analysis shows the importance of station accessibility. In our model, the effect of other public transportation systems to accessibility distance is not taken into consideration. It is assumed that the users can reach stations that are close enough to walk, while they might be more options in multimodal transport network. This underlines the nature of the car-sharing systems that work as systems complimentary to public transportation, which contribute to the improvement of the overall mobility.

3.4.4 Effect of Subsidy

The effect of subsidy on car-sharing system performance was also studied. Three different levels of subsidy (0, 2.5 and 5€/hour) were investigated for three different levels of demand (50% decreased, base and 100% increased demand). Alternatively, if an exact model of demand

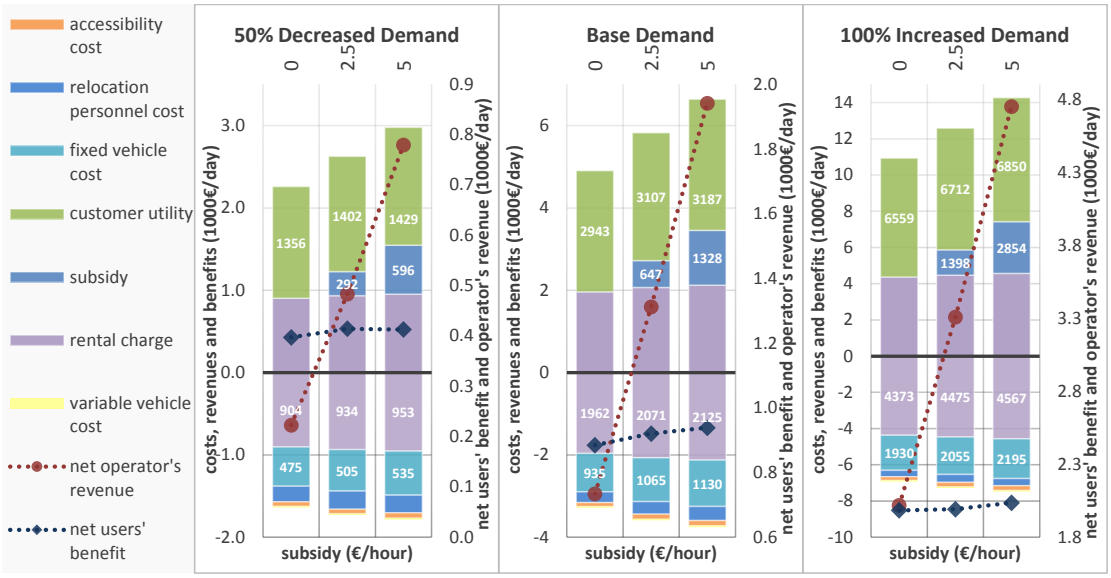


Figure 3.11: The costs, benefits and revenues for different subsidy levels

sensitivity to pricing exists, a similar analysis could be made. The results of this analysis are shown in Figure 3.11. The value of the objective function components (left axis) and the operator's and users' net benefits (right axis) are shown for different levels of subsidy.

The results of this analysis suggest that, the percent of demand served increased by 4-11%. Unprofitable demand in low or no-subsidy becomes profitable for the operator. Although 5-15% increase of the operator's cost (fix and variable vehicle, and relocation personnel costs) is required, the extra revenues generated outweigh the extra costs.

Note that, the increase in subsidy results in increase in the number of vehicles. However, it is not the case for the relocation personnel. Since increased subsidy enables operator to have more vehicles, the system becomes less dependent on relocation operations.

Another important finding of the analysis of subsidy levels relates to the effect of demand balance between demand level and subsidy on net revenues. If the net revenues of the operator for the same subsidy amount with different demand levels are compared, it can be observed that the increase in the profit is faster than the increase in demand. The operator earns more than double with double demand. It is something expected: Increase in demand makes the system more efficient and profitable as a result the level of subsidy can decrease.

3.5 Concluding Remarks

A multi-objective model for supporting strategic and tactical planning decisions for car-sharing systems was developed and tested in a large scale real world setting. The model considers simultaneously the net benefits of both the operator and the users. The proposed

model closes a gap in the existing literature by considering simultaneously decisions associated with the allocation of strategic assets, i.e. stations and vehicles of car-sharing systems and the allocation of personnel for relocation operations (tactical decision). The model provides decision makers with ample opportunities to perform sensitivity analysis for relevant model parameters. This feature is particularly useful for cost values that are difficult to establish empirically (e.g. utility gain of satisfied customers, population coverage, station accessibility cost). Furthermore, the multi-objective nature of the model allows the decision maker to examine the trade-off between operator's profit and users' level of service. This last feature is of particular importance if we consider that car-sharing systems are subsidized with public funds. The results obtained from the application of the model to a case resembling real world decision making requirements, provides useful information regarding the system performance.

Although the model provides satisfactory results for the case under consideration, it should be pointed out that the results are dependent on the model parameters used and cannot be directly generalized. However the proposed model can be utilized in different settings without difficulty. The value of the research presented herein stems from the innovative model proposed and its use for supporting strategic and tactical decision for car-sharing systems.

Research work under way involves the integration of the proposed model with a simulation model that will provide a more realistic representation of the relocation operation costs by looking on operational decisions. Modeling the operational problem and assigning the vehicle rosters while taking their electrical charge level into consideration is another future work directions. A field implementation of the proposed framework for one-way car-sharing is under preparation. Operational problem will consider different sources of uncertainties, such as last minute reservations, deviations from scheduled pick up and drop off times, level of charging and others. An operational model can also influence or redirect demand with pricing strategies, by giving for example the flexibility to choose the exact station or location (multiple stations) to the users.

4 Operational Framework for One-Way Car-Sharing Systems

Car-sharing (also known as shared-use vehicle) system is a new model for car rental in which people rent cars for short periods of time. It is an attractive alternative to the people who make only occasional use of a vehicle because of its cost-effectiveness and environmental-friendliness. In addition to that, it has several benefits to the society such as reduction in congestion, pollution, demand for parking spaces and incentive to drive less. In most of the real-world applications of car-sharing systems works in two ways, i.e. the vehicle should be returned to where it is rented from. Although there are some examples of one-way car-sharing systems in real-world, they are not preferred by the operators because of their operational difficulties e.g. relocations of vehicles.

In this research we aim to propose an optimization framework for the operational decisions of one-way electric car-sharing systems. Specifically we work on a model that decides on the relocation of the vehicles and the schedules of the personnel responsible from scheduling in general one-way car-sharing systems. For this purpose a mathematical model is formulated and solved. Since the work is still in progress, only preliminary results will be shown.

4.1 Introduction

As stated in Chapter 3, car-sharing systems have recently become an alternative for car ownership in the big cities. It attracts more people and have various benefits to not only individuals that uses the system but also to the society (Duncan, 2011).

The level of service offered by car-sharing system has two important components: The accessibility of the stations by the users and the availability of vehicles at the stations. The higher the accessibility of the stations and availability of the vehicles, the higher the level of service offered to the potential users, and hence the higher the attractiveness and potential utilization of the system. On the other hand, the station number and size, the fleet size and availability of vehicles and similar decisions influence the cost of establishing and operating a car-sharing system.

Ensuring the availability of the vehicles is a more serious problem when vehicles can be used on a one-way basis, i.e. when a vehicle picked-up at a station is not necessary to be returned back to the same station. It is highly probable to see imbalance of demand for cars, both at the origin of the trip (pick-up station) and at the destination (drop-off station). This phenomenon might create shortages of vehicles and empty spots. Vehicle relocation of vehicles is one of the solutions to these problems (Barth et al., 2006; Kek et al., 2006; Cucu et al., 2009; Fassi et al., 2012). However, inefficient vehicle relocation strategies might create high cost for the operators.

The efficient and cost-effective relocation planning and scheduling the operations of the relocation personnel require the use of models that will determine the movements of both the vehicles and the personnel. These models should assist decision makers to strike an optimum balance between the level of service offered and the total cost (including vehicle relocation costs) for operating the car-sharing systems.

In this research, we are dealing with the operational problems of non-floating one-way electric car sharing-systems. Although in reality, most of the one-way non-floating systems works with partial reservation (e.g. does not allow destination reservation), we believe that providing full reservation will increase service quality and attract more customers to the car-sharing systems. Here, we are dealing with the operational problem of full reservation and assume that every reservation provides origin and destination stations in addition to start and end time of the rentals with some certainty. Our aim is to propose a mathematical model that will help us to decide on the relocation of vehicles and schedules of the relocation personnel with this input. In the following section, mathematical model and its description is proposed. In the next section, recent experimental results are shared. We end up the chapter with conclusions and future work directions.

4.2 Mathematical Model

The mathematical model of the operational problem has similarities with the strategic model which is discussed in Chapter 3. We can list important differences between the strategic model and operational model as follows:

1. The strategic model aims to decide on the locations of the parking spots and their capacities. In the operational model, they are taken as parameters.
2. The strategic model decides on the number of vehicles in the system. The arrivals and departures of the vehicles to the system are predetermined in the operational model.
3. In the strategic model, multiple scenarios are used to have a robust model. In the operational problem, we are dealing with a single scenario since our decisions are daily operations but not future strategies.
4. In the strategic model, relocation personnel is used as a resource. In the operational model,

we take them as part of the system and model their movements.

5. In the strategic model, in order to keep variable size in tractable limits, long time intervals are utilized. In the operational model, time intervals are shorter.

6. In the strategic model, users are assigned to the stations that are reachable from the users' origins and destinations of their trips. In the operational model, users are assigned to the stations that are closest to origins and destinations of the trip.

7. In the strategic model, separate objectives are used for the users and the operator. In the operational model, we deal with a single objective function.

8. In the strategic model, we model just stations whereas in the operational model we (plan to) hubs and intermediary nodes.

Before stating the mathematical model, we can briefly describe the system characteristics. In the mathematical model, the operating day is divided into time intervals.Each entity (i.e. vehicles and personnel) can do at most one operation at each interval. In other words, it is assumed that, every operation (i.e. rental, relocation, charging, movement of personnel) takes at least one time interval.

We model the infrastructure with nodes and arcs. We assume that the vehicles and the relocation personnel moves over these arcs and end up their trips on nodes. Each arc has a distance and travel time with/without a vehicle. Travel time with a vehicle is used in relocation operations. Travel time without a vehicle applies to movements of relocation personnel when they use other means of transport (e.g. walking). Although the model supports using different travel times for different arc pairs and means of transport, in our experiments we use two different average speeds for with/without a vehicle trips and Euclidean distance to calculate the distance between nodes. As expected, travel time with a vehicle is shorter than a travel without one. The vehicles and personnel can travel to the all adjacent nodes and the nodes that are reachable within one time interval duration.

Nodes are composed of not only stations but also hubs and intermediary nodes. Hubs are parking spots that allows to keep vehicles but not rentals. Intermediary nodes are the nodes that are the nodes which connects nodes with each other. We introduce intermediary nodes to decrease number of variables in the system.

In order to satisfy a demand, a vehicle at the origin station at the starting time interval of the trip and an empty spot at the destination station at the ending time interval should be available. In addition, if a vehicle has returned from a trip, it needs to be charged proportional to the travel distance of the trip it has already finished.

We force but not oblige the system to serve all demand in order not to have infeasible solutions. Unserved demand is penalized with a high coefficient. As a result model gives a solution which serves as much trip as possible.

In this mathematical model, both the vehicles and relocation personnel are regarded as discrete flows. In other words, vehicles are not differentiated from each other. This assumption applies for personnel as well. When a vehicle arrives from a trip, with some constraints, we force the vehicle to stay at the station for some time. We apply this restriction by keeping a number of vehicles at the station which is a function of the trips served. In addition, in order to keep solution robust to the users' arrival and departure times, extra vehicles and empty spots are kept at the stations where they are needed and possible. We have implemented soft constraints that enables model to keep extra vehicles before the start time of the trips at the origins and after the end time of the trips at the destinations. Similar soft constraints are added to the model to have extra empty spots after the start time of the trips at the origin and before the end time of the trips at the destinations.

Sets and indices, parameters and variables can be seen below. Afterwards, mathematical model will be presented with a brief description of each constraint.

Sets and Indices

$i \in I$**:** trips

$j, l \in J$**:** nodes (stations, hubs, intermediary nodes)

$k \in K$**:** vehicle arrivals/departures

$s \in S$**:** personnel shifts

$t, u, w \in T$**:** time interval indices

Parameters

Origin$_i$**/Dest**$_i$**:** origin/destination station of trip i

Start$_i$**/End**$_i$**/C.End**$_i$**:** starting/ending/charging end time intervals of trip i

CAP$_j$**:** vehicle capacity of node j

start$_s$**/end**$_s$**:** start/end time interval of personnel shift s

cost$_s$**:** cost of personnel shift s

DA$_j$**/MA**$_j$**:** list of nodes that are accessible from node j in driving/moving

DEnd$^t_{jl}$**:** ending time interval of a vehicle relocation started at time interval t from node j to l

MEnd$^t_{jl}$**:** ending time interval of a personnel movement started at time interval t from node j to l

VC: maximum number of personnel that can be carried with a vehicle in any relocation

$\mathbf{W}^t_{ij}/\overline{\mathbf{W}}^t_{ij}$: weight of the safety factor of vehicle/empty spot for trip i, node j and time interval t

$\mathbf{RBP}^t_i/\mathbf{RAP}^t_i$: ratio of the safety factor before/after pick-up for trip i, the origin station of trip i and time interval t

$\mathbf{RBP}^t_i/\mathbf{RAP}^t_i$: ratio of the safety factor before/after drop-off for trip i, the destination station of trip i and time interval t

n^0_j: number of vehicles fully charged in node j at the beginning of first time interval

Variables

n^t_j: number of vehicles in node j at time interval t

h^t_j: number of personnel in node j at time interval t

$r^{t(u)}_{jl}$: number of vehicles relocated from node j to $l \in \mathrm{DA}_j$ starting at time interval t and ending at time interval $u = \mathrm{DEnd}^t_{jl}$

$p^{t(u)}_{jl}$: number of personnel driving from node j to $l \in \mathrm{DA}_j$ starting at time interval t and ending at time interval $u = \mathrm{DEnd}^t_{jl}$

$\overline{p}^{t(u)}_{jl}$: number of personnel moving from node j to $l \in \mathrm{DA}_j$ starting at time interval t and ending at time interval $u = \mathrm{DEnd}^t_{jl}$

z_i: binary variable showing if trip i is served or not

f^t_{ij}: fulfilled vehicle availability goal at time interval t for trip i

$\bar{f}^t_{ij}$: fulfilled empty spot availability goal at time interval t for trip i

v_s: number of personnel used from shift s

$a^t_j/\bar{a}^t_j$: number of personnel started/ending working at time interval t in node j

For the given indices, parameters and variables defined above, we can write the mathematical model for the operational problem of one-way electric car-sharing systems as follows:

$$\min \quad \overbrace{\sum_i \mathrm{L}_i (1 - z_i)}^{\text{lost customer penalty}} + \overbrace{\sum_s \mathrm{PC}_s v_s}^{\text{personnel cost}} + \overbrace{\sum_{j,l,t:l \in DA_j} \mathrm{RC}_{jl}^{t(u)} r_{jl}^{t(u)}}^{\text{relocation cost}} - \overbrace{\sum_{i,j,t} \left(\mathrm{W}_{ij}^t f_{ij}^t + \overline{\mathrm{W}}_{ij}^t \overline{f}_{ij}^t \right)}^{\text{safety bonus}} \quad (4.1)$$

$$\text{st. } n_j^{t+1} = n_j^t + \sum_{\substack{\text{end}_i = t, \\ \text{dest}_i = j}} z_i - \sum_{\substack{\text{start}_i = t, \\ \text{origin}_i = j}} z_i + \sum_{l,u} \left[r_{lj}^{u(t)} - r_{jl}^{t(u)} \right] \qquad \forall j, t \quad (4.2)$$

$$n_j^t \leq \mathrm{CAP}_j + \min \left\{ \sum_{l,u} r_{lj}^{u(t)}, \sum_{l,u} r_{jl}^{t+1(u)}, h_j^t \right\} \qquad \forall j, t \quad (4.3)$$

$$n_j^t \geq \sum_{\substack{i:\text{start}_i = t, \\ \text{origin}_i = j}} z_i + \sum_{l,u} r_{jl}^{tu} + \sum_{\substack{i:\text{dest}_i = j, \\ \text{end}_i < t \leq \text{c.end}_i}} z_i \qquad \forall j, t \quad (4.4)$$

$$r_{jl}^{tu} \leq p_{jl}^{tu} \qquad \mathrm{VC} r_{jl}^{tu} \geq p_{jl}^{tu} \qquad \forall j, l \in \mathrm{DA}_j, t \quad (4.5)$$

$$h_j^{t+1} = h_j^t - \sum_{l,u} p_{jl}^{t(u)} - \sum_{l,u} \overline{p}_{jl}^{tu} - \overline{a}_j^t + \sum_{l,u} p_{lj}^{u(t)} + \sum_{l,u} \overline{p}_{lj}^{u(t)} + a_j^t \qquad \forall j, t \quad (4.6)$$

$$h_j^t \geq \sum_{l,u} p_{jl}^{t(u)} + \sum_{l,u} \overline{p}_{jl}^{tu} + \overline{a}_j^t \qquad \forall j, t \quad (4.7)$$

$$\sum_{\text{first}_s = t} v_s = \sum_j a_j^t \qquad \sum_{\text{last}_s = t} v_s = \sum_j \overline{a}_j^t \qquad \forall s \quad (4.8)$$

$$\sum_i f_{ij}^t \leq n_j^{t+1} - \sum_{\substack{i:\text{end}_i < t \leq \text{CEnd}_i, \\ \text{dest}_i = j}} z_i \qquad \forall j, t \quad (4.9)$$

$$\sum_i \overline{f}_{ij}^t \leq \mathrm{CAP}_j - n_j^{t+1} \qquad \forall j, t (4.10)$$

$$f_{i\,\text{origin}_i}^t \leq \mathrm{RBP}_i^t z_i^t \qquad f_{i\,\text{dest}_i}^t \leq \mathrm{RAD}_i^t z_i^t \qquad \forall i, t (4.11)$$

$$\overline{f}_{i\,\text{origin}_i}^t \leq \mathrm{RAP}_i^t z_i^t \qquad \overline{f}_{i\,\text{dest}_i}^t \leq \mathrm{RBD}_i^t z_i^t \qquad \forall i, t (4.12)$$

$$z_i \in \{0,1\}; \;\; v_s, r_{jl}^{tu}, n_j^t, h_j^t, p_{jl}^{tu}, e^t, b^t \in \mathcal{N}; \;\; f_{ij}^t, \overline{f}_{ij}^t \in \mathcal{Z}^+ \qquad \forall i, j, t, u, l, s (4.13)$$

The objective function, given with Equation 4.1 contains four different cost functions. A penalty is applied for every lost customers. Personnel and relocation costs are the two expenses for the operational decisions. Former stands for the relocation personnel whereas latter is for the fuel consumed. The part named as safety bonus in the objective function is for making the model flexible for the users' pick-up and drop-off times. With the safety factors (W_{ij}^t and $\overline{\mathrm{W}}_{ij}^t$) and variables stating fulfilled vehicle and empty spot availability goals (f_{ij}^t and $\overline{f}_{ij}^t$), model is directed to a solution that keeps as much vehicles and empty spots as possible where needed.

Constraints 4.2 keep flow conservation for each node at each time interval. Number of vehicles at node j at time interval $t+1$ is equal to the number of vehicle at the same node at the previous interval plus the number of vehicles arrived from rental and relocation minus the number of vehicles left for rental and relocation.

Constraints 4.3 are used to update the "real" capacity of each node. In our network design, in order to a have less number of nodes, we prefer not to implement separate nodes for stations. In other words, when a vehicle is under a long relocation, it might visit some station nodes in its intermediary steps. However, we do not want to count these vehicles as vehicles parked to the stations. More precisely, we do not want these vehicles to decrease the capacity of the stations. Because of this reason, we are assuming that, capacity of a station at any time interval is increased with the minimum of the number of vehicles arrived from relocation during the same time interval, the number of vehicles depart for relocation during the next time interval and the number of personnel at the same time interval. In order to keep mathematical model neat, constraints 4.3 are written with a minimum function. Note that, this constraint can be implemented without any extra variables.

Constraints 4.4 ensure that number of vehicles leaving the station is not more than available number of vehicles at the same station. The same constraints keeps vehicles at the station that need to be charged as well. More specifically, since we do not differentiate the vehicles, these constraints ensure that, the number of vehicles at the station is at least the number of vehicles that need to be charged at a given station at a given time. With these constraints we satisfy that, for any station at any time interval, number of vehicles at the beginning of the time interval should be greater than or equal to the number of vehicles need charging, the number of vehicles rented and the number of vehicles that are relocated from this station at the same time interval.

Constraints 4.5 are used to connect the relocation operations with the personnel. Constraints on the left assure that there is at least one personnel driving the vehicle when it is relocated. Constraints on the right prevent to transport personnel more than the capacity of the vehicle.

Constraints 4.6 are flow conservation equations for the personnel for each node, at each time interval. Total number of personnel at the beginning of the next time interval is equal to the total number of personnel at the beginning of the previous time interval plus the net number of personnel started and ended working and plus the net number of personnel arrived to and departed from the node at the same time interval with or without a vehicle.

Constraints 4.7 assure, the number of personnel at node j at time interval t is always at least the number of personnel departed from the node plus the number of personnel ended working.

Constraints 4.8 keep relationship between the number of personnel used from the personnel shifts and the number of personnel started/ended working. The constraints on the left assure that the total number of personnel started working at any node at time interval t should be covered by the shifts that are starting at time interval t. Similarly, the constraints on the right assure, the total number of personnel ended working at any node at time interval t should be covered by the shifts that are ended working at time interval t.

Constraints 4.9 and 4.10 are for the safety factor of cars and empty spots respectively. The former constraints ensure that the number of vehicles that are available to rent from node j at the beginning of the time interval t should cover as much fulfilled vehicle availability goal as possible. In a similar way, the latter constraints make the same relationship between the number of empty spots and the fulfilled empty spot availability goal.

Constraints 4.11 and 4.12 are the upper bounds on the fulfilled vehicle and empty spot availability goals respectively. If a trip is accepted to be served, availability goals can take values not more than some predefined values which are a function of time intervals and trips.

Last but not least, it is worth to note here that, although this model is constructed for systems with electric vehicles it can be applied for the systems with combustion vehicles as well. By setting, charging end time interval to end time interval of each trip, the model can be used for the latter systems without any problems.

4.3 Experimental Results

As it is stated above, this part of the thesis is still work in progress. However, in order to validate the model, some experiments are conducted with the data of Veolia, which is operating in Nice. In our preliminary analysis, we use the 53 stations that are planned to be operated as a one-way electric car-sharing system. Only the stations are taken as the nodes of the system and no hubs or intermediary nodes are added to the model. We created an experimental network by connecting the adjacent nodes with arcs. The network of the system can be seen in Figure 4.1.

For relocation of vehicles and movement of personnel, Euclidean distance and two different speed is used. For relocation operations with a vehicle, average speed is set to 30km/h whereas it is assumed the the relocation personnel move with a speed of 10km/h without a vehicle. The mathematical model is solved for a period of 16-hour with the time interval length of 15 minutes which makes in total 64 time intervals.

The demand is generated from the realized two-way car-sharing trips. If a trip is posed for more than 30 minutes in a location which is closer than 500m to one of the current operating two-way stations, it is assumed that this trip can be divided into two or more one-way trips. These generated one-way trips are populated in a list and 9 sets of demand are created randomly with different number of demands. We make sets of demand with the size of 50, 100 and 200.

Three different shifts with a duration of 4 hours are given as a parameter to the model. It is assumed that, cost of a relocation personnel is 72€/shift. It is assumed that a vehicle with an empty battery is fully charged within 8 hours and the range of the electric vehicles are 120km with a full battery. Distance traveled at each trip are set proportional to the length of the rental. Vehicle capacity in relocation operations are set to 4. For each station, we set the capacity to three. Two fully charged vehicles are located at each station at first time interval. Penalty for

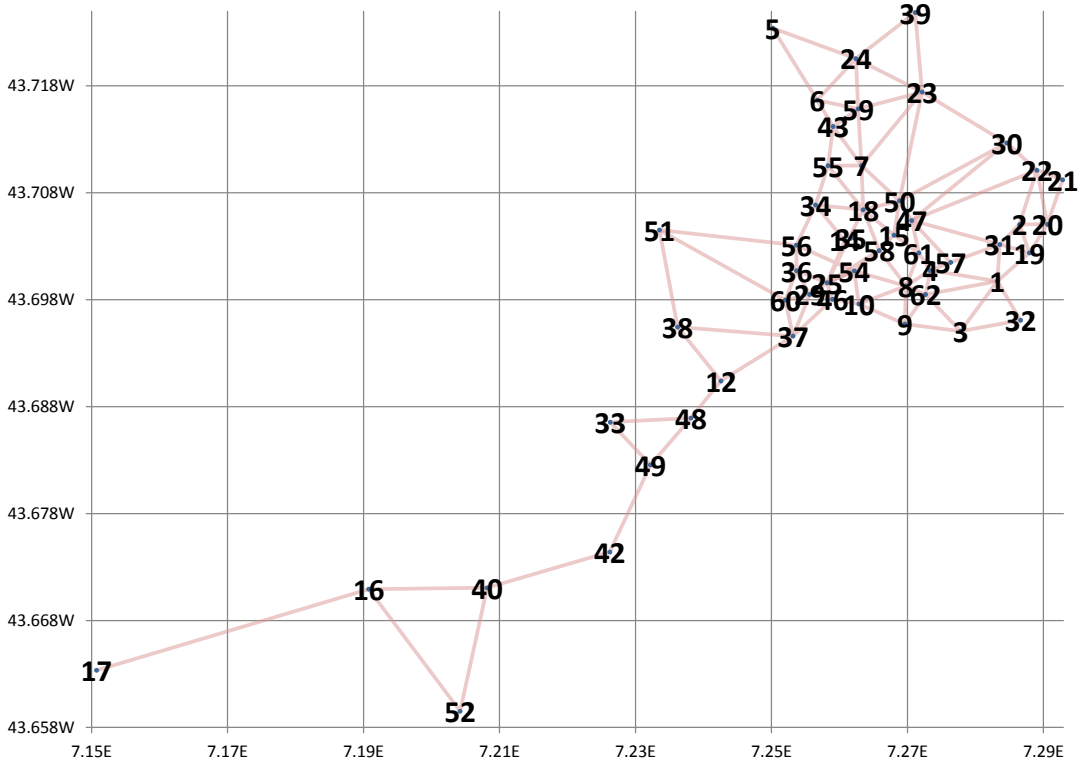


Figure 4.1: The network of Nice utilized in the operational model

lost demand is set to €10000. During relocation operations, we assumed that there is a fuel cost which is 0.02€/km. We set different safety factor values (RBP_i^t and RAP_i^t ranging from 0.05 to 0.5. Weight of the safety factors (W_{ij}^t and $\overline{W}_{ij}^t$) have values between 2 and 10.

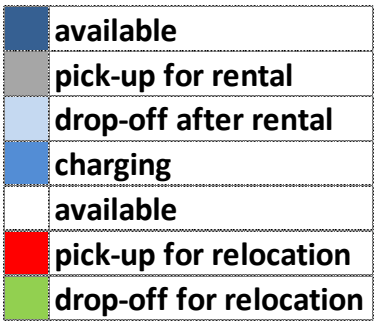


Figure 4.2: Legend for figures 4.3 and 4.4

We conduct two different analysis. In the first analysis we validate our model with the help of a Gannt chart prepared for one of the congested instances with a demand of 200. The relocation operations and the personnel movements are traced to validate the mathematical model. The whole Gannt chart of the operations can be seen in Figure 4.3. Note that, in Figure 4.3 rows are the spots belonging to each operating stations and columns are time intervals of 15minutes.

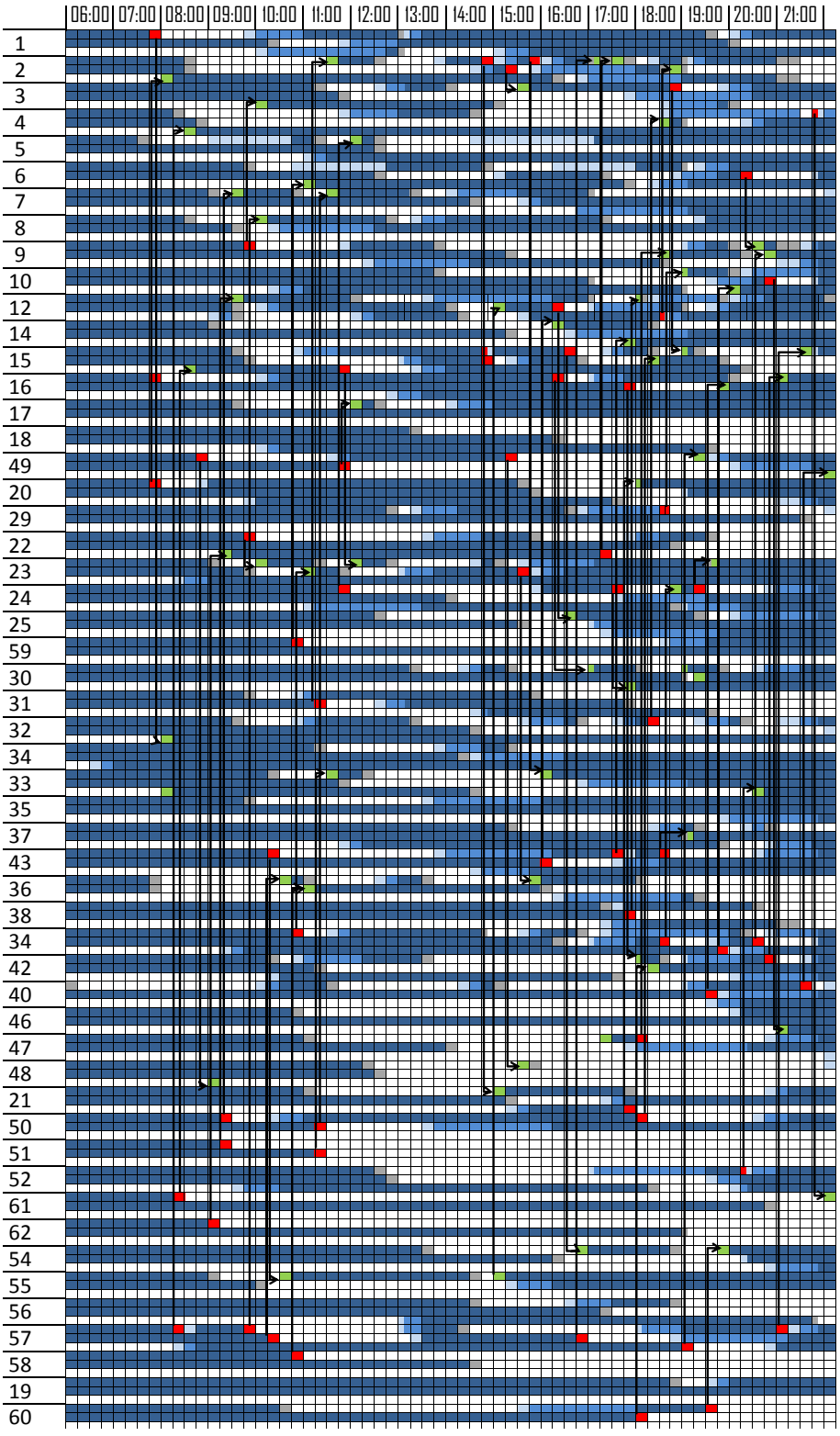


Figure 4.3: A Gannt chart including all time intervals and stations for an instance with 200 trip requests

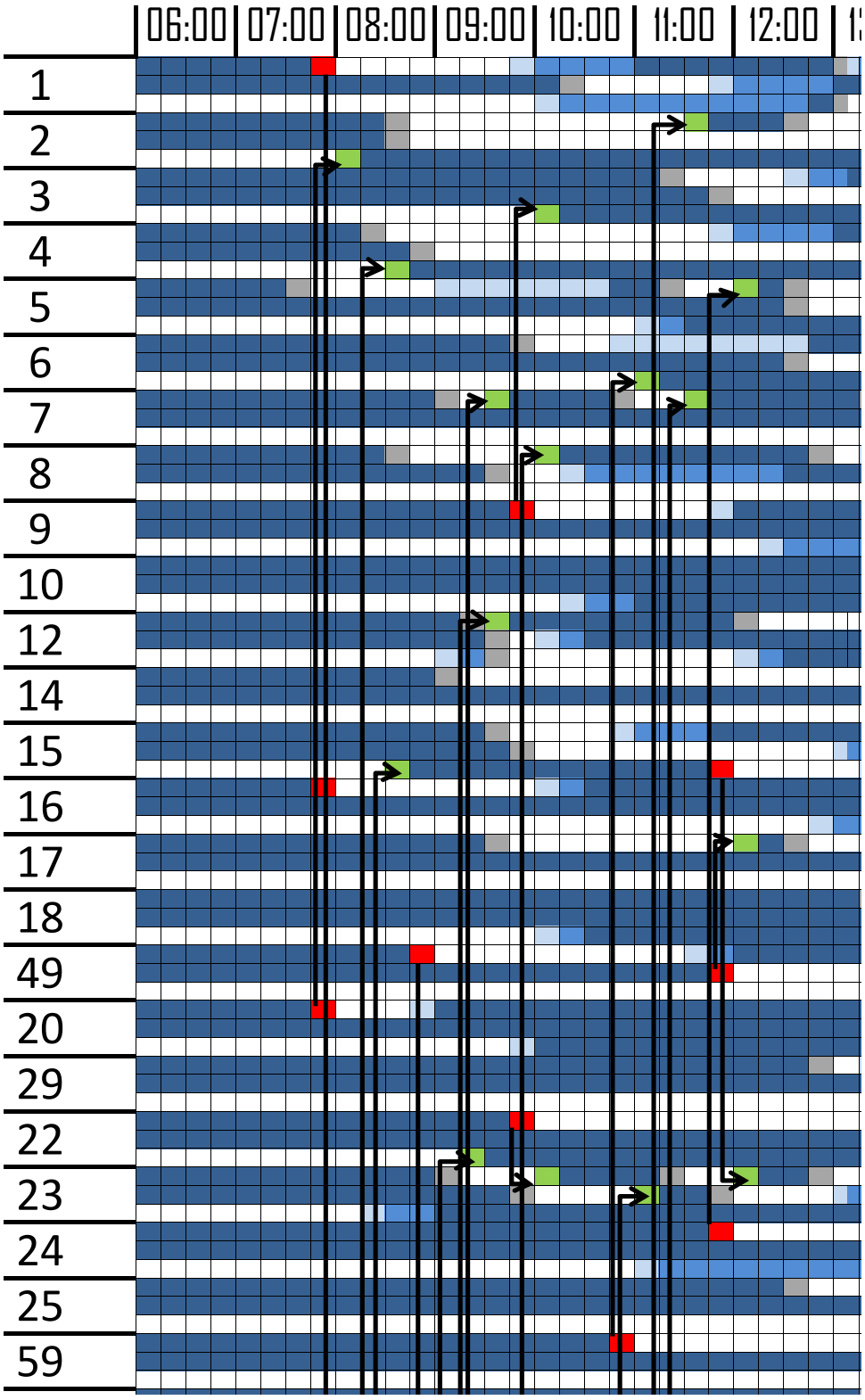


Figure 4.4: A smaller area of the Gannt chart given in Figure 4.3

demand	scenarios	with relocation			without relocation		
		I	II	III	I	II	III
50	run time (s)	175	234	238	6.2	6.8	7.1
	objective (€)	8993	49082	-953	108882	178909	148882
	personnel cost (€)	144	216	216	N/A	N/A	N/A
	# lost customer	1	5	0	11	18	15
	% served	98	90	100	78	64	70
100	run time (s)	473	473	298	6.3	6.3	6.7
	objective (€)	78342	78342	48186	357802	357765	327755
	personnel cost (€)	648	648	504	N/A	N/A	N/A
	# lost customer	8	8	5	36	36	33
	% served	92	92	95	64	64	67
200	run time (s)	649	1099	26822	6.5	6.3	7.0
	objective (€)	236084	216090	176258	825539	675651	725642
	personnel cost (€)	576	504	648	N/A	N/A	N/A
	# lost customer	24	22	18	83	68	73
	% served	88	89	91	58.5	66	63.5

Table 4.1: Comparing one-way car-sharing systems with and without relocation

If there is a vehicle at the spot during the selected time interval, it is shown with a color of blue. Light blue shows that, the vehicle on the spot has just returned from a trip. Darker blue states that, the vehicle on the spot needs to be charged for a new operation. Spots that are indicated with the darkest blue indicates that the vehicle on that spot is fully charged and ready for a new operation. If a vehicle is rented, the box in the Gannt chart showing the state of the spot during the time interval that the user supposed to pick-up the vehicle is painted with gray. Green and red shows the relocation operations between stations. Picked-up vehicles for relocations are shown with red. If a vehicle is dropped-off to a station after relocation, then the color of that spot at the end of relocation is green. Relocation operations can also be followed with the black arrows. Each origin and destination of relocations are connected with these arrows. The colors and their corresponding states can be seen in Figure 4.2.

In a smaller area of the Gannt chart of Figure 4.3 can be seen in Figure 4.4. In this figure, we can see the relocation operations and the states of the spots at some stations. For instance, Station 4 starts with a two available vehicle. During time intervals starting with 08:15 and 8:45, both vehicles are rented by the users. However, at 8:30-8:45 time interval, the third spot is filled with a vehicle that is relocated from the first spot of Station 20. This relocated vehicle stays in the spot until the end of the time horizon we can observe in this figure.

In addition to validation of the model with a Gannt chart, we compare the same instances with and without relocation operations. The results of these experiments can be seen in Table 4.1. Note that fuel cost for relocations is not reported since it is too low compared to other costs. In these experiments, we have observed that the percentage of served demand for the instances with relocation increase with the increase in demand. However, the effect is dramatic in the

cases without relocation. Almost half of the demand (around 60%) could not be served, if relocation operations are not done for the instances with the demand of 200. This value is around 65-70% for lower demand values. Relocation improves this value and serves 90% of the demand on average. Please also note, without relocation, vehicles might accumulate in some specific stations which results in even higher unserved demand ratios in longer time horizons (e.g. one week). Although the results are from preliminary runs, it is obvious that continuous relocation operations are needed to have a decent level of service in one-way car-sharing systems.

4.4 Conclusions and Future Research Directions

In this short chapter, we describe a mathematical model that can be utilized to improve the efficiency of relocation operations in one-way (electric) car-sharing systems. First, Gannt chart prepared for an instance is described. We prepare this chart to see the flow of vehicles and personnel in the system. Second, we compare systems with and without relocation operations. Although, we need to have broader analysis, preliminary results show that, relocation operations are considerably improving the quality of the service.

As stated above, we have recently started to work on the operational problem and still working. We recently have developed a simulation platform to simulate the demand in different car-sharing systems. We have also started to implement simple heuristics to have simple rules to handle the changes in the system.

We also plan to implement some changes in the mathematical model. As an example, in our mathematical model, since we do not differentiate vehicles from each other, we only make vehicles available when they are fully charged. However, this creates high rates of demand loss. We want to solve this problem by enabling the vehicles not fully charged to be rented. However, at the same time we will direct the system to use primarily fully charged vehicles by putting penalties on utilizing not fully charged vehicles. We also want to implement hubs in our mathematical models. We think that, a depot with higher capacity than the stations could increase flexibility of the relocation operations which will help us to increase the availability of the service with less expenses.

Part III
Estimation of Network Capacity in Congested Urban Systems

5 The Effect of Variability of Urban Systems in The Network Capacity

Recent experimental analysis has shown that some types of urban networks exhibit a low scatter reproducible relationship between average network flow and density, known as the macroscopic fundamental diagram (MFD). It has also been shown that heterogeneity in the spatial distribution of density can significantly decrease the network flow for the same value of density. Analytical theories have been developed to explore the connection between network structure and an MFD for urban neighborhoods with cars controlled by traffic signals. However these theories have been applied only in cities with deterministic values of topological and control variables for the whole network and by ignoring the effect of turns. In our study we are aiming to generate an MFD for streets with variable link lengths and signal characteristics and understand the effect of variability for different cities and signal structures. Furthermore, this variability gives the opportunity to mimic the effect of turning movements. Route or network capacity can be significantly smaller than the capacity of a single link, because of the correlations developed through the different values of offsets. The above analysis would not be possible using standard traffic engineering techniques. This will be a key issue in planning the signal regimes in a way that maximizes the network capacity and/or the density range of the maximum capacity.

5.1 Introduction

Recent studies (Geroliminis and Sun, 2011a; Mazloumian et al., 2010; Daganzo et al., 2011) have shown that networks with heterogeneous distribution of link density exhibit network flows smaller than those that approximately meet homogeneity conditions (low spatial variance of link density), especially for congested conditions. Also, note that the scalability of flows from a series of links to large traffic networks is not a direct transformation. For example, route or network capacity can be significantly smaller than the capacity of a single link as this is expressed by sG/C (s is the saturation flow, G and C are the durations of green phase and cycle). This is because of the correlations between successive arterial links, the creation of spillback queues and the effect of offsets (Daganzo and Geroliminis, 2008). In case of long links, these effects are negligible and the propagation of traffic is much simpler. Nevertheless,

congestion often occurs in the city centers with dense topology of short links.

At the link scale, traffic flows can be unpredictable or chaotic when a network is critically congested because of different driving behavior patterns, the effect of route choice, the fast dynamics of link travel times and origin-destination tables and the computational complexity (too many particles/cars). These observations make the development of global traffic management strategies, to improve mobility for a large signalized traffic network with a microscopic analysis, intractable. An alternative is a hierarchical control structure, where a network can be partitioned in homogeneous regions (with small spatial variance of link density distribution) and optimal control methodologies can identify the inter-transfers between regions of a city to maximize the system output, as expressed by the number of trip endings. These policies can change the spatial distribution of congestion in such a way that the network outflow increases. This is a challenging task that requires knowledge on how the network flow for a region of a city changes as a function of topology, control and level of congestion.

The physical tool to advance this research is the macroscopic fundamental diagram (MFD) of urban traffic, which provides for some network regions a unimodal, low-scatter relationship between network vehicle density (veh/km) and network space-mean flow (veh/hr). The first theoretical proposition of such a physical model was developed by Godfrey (1969), while similar approaches were also initiated by Herman and Prigogine (1979) and Daganzo (2007). The physical model of MFD was observed with dynamic features through empirical data in congested urban networks in Yokohama (Geroliminis and Daganzo, 2008). Other empirical or simulated analysis for MFDs with low or high scatter can be found in Buisson and Ladier (2009), Daganzo et al. (2011), Geroliminis and Sun (2011b), Gayah and Daganzo (2011), Courbon and Leclercq (2011), Ji et al. (2010), Saberi and Mahmassani (2012) and others. Nevertheless, it is not obvious whether the MFDs would be universal or network-specific. More real-world experiments are needed to identify the types of networks and demand conditions, for which invariant MFDs with low scatter exist.

To evaluate topological or control-related changes of the network (e.g. a re-timing of the traffic signals or a change in infrastructure), Daganzo and Geroliminis (2008) and, Helbing (2009) have derived analytical theories for the urban fundamental diagram, using a density-based and a utilization-based approaches respectively. The first reference proved, using variational theory (Daganzo, 2005), that an MFD must arise for single-route networks with a fixed number of vehicles in circulation (periodic boundary conditions and no turns). The same reference also gives explicit formulae for the single-route MFD with deterministic topology, control and traffic characteristics (i.e. all intersections have common control patterns, the length of its links and their individual fundamental diagrams are all the same. The reference conjectured that these MFD formulae should approximately expected to hold for homogeneous, redundant networks with slow-changing demand. The methodology estimates the average speed and the maximum passing rate (rate that cars can overpass him) for a large number of observers moving forward or backward and stopping only at traffic lights. Then by considering that each observer can create a “cut” in the MFD, its shape is estimated as the lower envelope of all these

cuts.

In this chapter we provide several extensions and refinements of the analytical model for an MFD. We explore how network parameters (topology and signal control) affect two key characteristics of an MFD, (i) the network capacity and (ii) the density range for which the network capacity is maximum. We first provide an analytical proof that simplifies the estimation of the density range for which the network capacity is maximum by utilizing only spatial and control parameters of the network. We also investigate how sensitive are these two characteristics in small changes of the parameters. Afterwards, we relax the deterministic character of the parameters and investigate how variations in the signal offsets and the link lengths affect network capacity and density range. These results can be utilized to develop efficient control strategies for a series of signalized intersections as these variations can describe not only differences in network parameters, but also different characteristics in driver behavior. Later, we imitate the effect of incoming turns in a long arterial and we show that these turns can significantly decrease the network capacity even if vehicle density remains unchanged. To precisely describe all the above phenomena we initially provide some analytical proofs for a simplification of the variational theory approach and then we develop a simulator to study the non-deterministic effects.

5.2 A Note on Variational Theory

Daganzo and Geroliminis (2008) developed a moving observer method to show that the average flow-density states of any urban street without turning movements must be bounded from above by a concave curve. The section also shows that, under the assumptions of variational theory, this curve is the locus of the possible (steady) traffic states for the street; i.e., it is its MFD.

Their method builds on a finding of Daganzo (2005), which showed that kinematic wave theory traffic problems with a concave flow-density relation are shortest (least cost) path problems. Thus, the centerpiece of variational theory (as is the fundamental diagram for kinematic wave theory) is a relative capacity ("cost") function (CF), $r(u)$, that describes each homogeneous portion of the street. This function is related to the known fundamental diagram (FD) of kinematic wave theory Q. Physically, the CF gives the maximum rate at which vehicles can pass an observer moving with speed u and not interacting with traffic; i.e., the street's capacity from the observer's frame of reference. Linear CFs correspond to triangular FDs. Daganzo (2005) assumed a linear CF characterized by the following parameters: k_0 (optimal density), u_f (free flow speed), κ (jam density), w (backward wave speed), s (capacity), and r (maximum passing rate). CF line crosses points $\left(u_f, 0\right)$, $(0, s)$, $(-w, r)$ and has a slope equal to $-k_0$. Other applications of variational theory in modeling traffic phenomena can be found in Laval and Leclercq (2008) and Daganzo and Menendez (2005).

A second element of variational theory is the set of "valid" observer paths on the (t, x) plane starting from arbitrary points on the boundary at $t = 0$ and ending at a later time, $t_0 > 0$. A

path is "valid" if the observer's average speed in any time interval is in the range $[-w, u_f]$. If $\mathscr{P}$ is one such path, $u_{\mathscr{P}}$ be the average speed for the complete path, and $\Delta(\mathscr{P})$ is the path's cost which is evaluated with $r(u)$, $\Delta(\mathscr{P})$ bounds from above the change in vehicle number that could possibly be seen by observer $\mathscr{P}$. Thus, the quantity:

$$R(u) = \lim_{t_o \to \infty} \inf_{\mathscr{P}} \{\Delta(\mathscr{P}) : u_{\mathscr{P}} = u\} / t_o \tag{5.1}$$

is an upper bound to the average rate at which traffic can overtake any observer that travels with average speed u for a long time. Note that (5.1) is a shortest path problem, and that $R(0)$ is the system capacity. Building on Equation (5.1), Daganzo and Geroliminis (2008) proved that a ring's MFD with periodic characteristics in time and space (traffic signals every L meters with common green duration G and cycle C and no turns), $Q = Q(k)$, is concave and given by Equation (5.2), i.e., it is not an upper bound, but a tight relation. Figure 5.1a illustrates that Equation (5.2) is the lower envelope of the 1-parameter family of lines on the (k, q) plane defined by $q = ku + R(u)$ with u as the parameter. Note this equation also describes the passing rate of an observer moving with constant speed u in a stationary traffic stream with flow q and density k. The main difference is that traffic signals create non-stationary conditions as vehicles stop at traffic signals and this relation does not apply in all cases. We call these lines "cuts" because they individually impose constraints of the form: $q \leq ku + R(u)$ on the macroscopic flow-density pairs that are feasible on a homogeneous street.

$$q = \inf_{u} \{ku + R(u)\} \tag{5.2}$$

Because evaluating $R(u)$ in Equation (5.2) for all u can be tedious, Daganzo and Geroliminis (2008) proposed instead using three families of "practical cuts" that jointly bound the MFD from above, albeit not tightly. It has been shown (Daganzo, 2005) that for linear CF's, an optimal path always exists that is piecewise linear: either following an intersection line or else slanting up or down with slope u_f or $-w$. The practical cuts are based on observers that can move with only 3 speeds: u_f, 0, or w and stop at intersections during red times and possibly during green periods as well. Recall that an observer's cost rate (maximum passing rate) is $q_B(t)$ if the observer is standing at intersection with capacity $q_B(t) \leq s$ and otherwise it is given by a linear CF, i.e. it is either 0, s or r depending on the observer's speed.

Figure 5.1b-c provide an explanation how the "practical cuts" are estimated for a series of intersections with common length and signal settings (green G, cycle C and offset δ). Offset is the time difference between the starting of the green phase for two consecutive signals. If the green phase of the signal in the downstream starts later than the green phase of the upstream then δ is positive for the downstream traffic signal and vice versa. Case $(4, F)$ shows the fastest moving observer, who runs with speed u_f and stops only at red phases once every 4 signals. No vehicles are passing him, the first "cut" crosses the $k-q$ plane at $(0,0)$ and has a slope equal to the average speed of the observer. The 2nd observer $(3, F)$ still runs at u_f but stops during the green period every third signal. Thus, he has a smaller average speed and

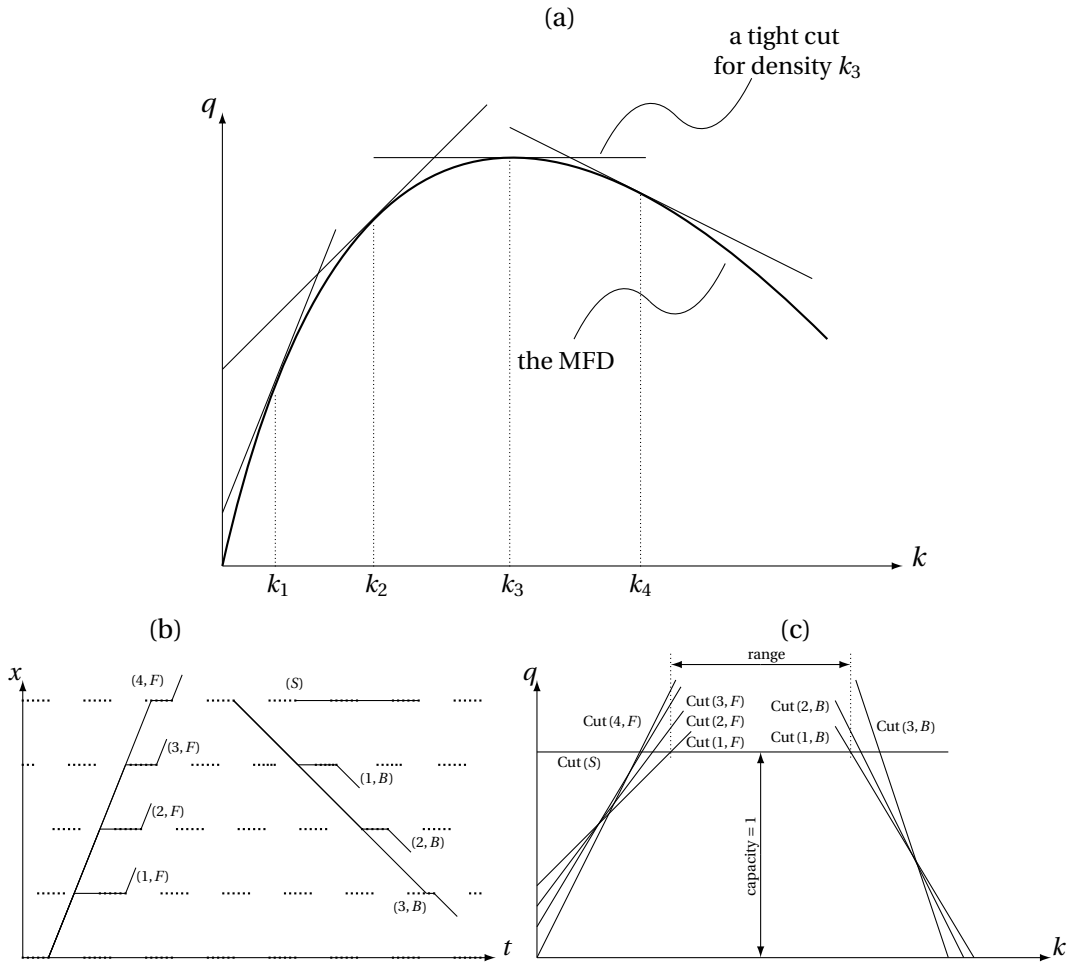


Figure 5.1: (a) The MFD defined by a 1-parameter family of "cuts" (Daganzo and Geroliminis, 2008) and both forward, backward and stationary observers in a time-space (b) and their associated "cuts" in a network flow-density diagram (c).

vehicles are passing him at rate s when waiting in green phases. This passing rate is shown in the second "cut" as the constant of the line that crosses axis q at $q > 0$. The 3rd observer $(2,F)$ stops in every two signals whereas the fourth $(1,F)$ stops in every signal, while the fifth observer named as (S) is the stationary observer with zero speed. Cases $(3,B)$-$(1,B)$ show the backward moving observers, who are passed at rate r when moving in the opposite direction and at rate s when waiting in green phases. The lower envelope of all these "cuts" produces the MFD.

The authors conjectured that the following regularity conditions should ensure a good analytical approximation of the MFD: (i) a steady and distributed demand; (ii) a redundant network ensuring that drivers have many route choices and that most links are on many desirable routes; (iii) a homogeneous network with similar links; (iv) links with an approximate FD that is not significantly affected by turning movements when flow is steady. Conditions (i)-(iii)

should create a near-equilibrium as in Wardrop (1952) with similar travel times on all links; and, since the links are similar, with similar densities too. Condition (iv) implies that the variational theory method, applied to a single link with many efficient cuts, yields a tight MFD. The estimated MFD (despite its simplistic approach) fits well empirical and simulated data for Yokohama and San Francisco, two networks that only roughly meet the regularity conditions.

In this work, we focus on two important parameters of the MFD, produced by the family of cuts. The first one is q_{max}, the maximum value of the dimensionless network capacity, normalized by the maximum capacity of a single link sG/C, where G is the green duration during a cycle C. The second one is r_d, the range of density for which this maximum value occurs. A negative range does not have any physical meaning, except that capacity is less than 1. A large value of positive range can be considered as a metric of robustness for the system, which can retain its maximum capacity values for multiple congestion levels. We investigate these two parameters for different types of networks with homogeneous or heterogeneous characteristics of topology and signal settings.

5.3 Homogeneous Networks

The methodology of moving observers in Section 5.2 could have a potential application in real cities in cases data from loop detectors are not available or do not cover the whole region. Nevertheless, the large number of cuts, would make the application not straightforward. In this section we provide an analytical proof, which significantly decreases the number of required cuts for the estimation of the range r_d. We prove that in case of homogeneous networks, the slowest forward and backward and the stationary observers suffice to estimate the same value of r_d, by running all observers described in Daganzo and Geroliminis (2008). This proof provides us the ability to obtain closed form analytical equations for r_d. We start with Lemmas 1 and 2, which prove some mathematical operations, necessary for the next steps. In Lemmas 3 (forward) and 4 (backward) we prove that the slowest moving observers provide tighter cuts than all the non-fastest observers. Lemmas 5 and 6 complete the proof by comparing with the fastest observers, who run at u_f or w and stop only at real red phases. In the end of the section, Corollary 1 provides the closed form of analytical equations of r_d for different cases. The reader can omit the proofs without lack of continuity for the rest of the chapter.

In Daganzo and Geroliminis (2008), for an observer who stops every γ signals for extended red phase, delay at each stop d_γ and average speed u_γ are defined as:

$$d_\gamma = C\left(\left\lceil \frac{\gamma L/u_f - \delta}{C} \right\rceil - \frac{\gamma L/u_f - \delta}{C}\right) \tag{5.3}$$

$$u_\gamma = \frac{\gamma L}{d_\gamma + \gamma L/u_f}. \tag{5.4}$$

$\gamma_{\max}$ can be calculated as:

$$\gamma_{\max} = 1 + \max\left\{\gamma : \gamma\left(L/u_f - \delta\right)/C - \left\lfloor \gamma\left(L/u_f - \delta\right)/C \right\rfloor \leq G/C\right\}. \tag{5.5}$$

The time spend in extended red phase, f_γ equals to:

$$f_\gamma = \frac{d_\gamma - C + G}{\gamma L/u_f + d_\gamma} \text{ for } \gamma = 1, 2, ..., \gamma_{\max} - 1. \tag{5.6}$$

Similar formulations can be applied for backward moving observer, by changing u_f with w and δ with $\delta_w = C - \delta$. To differentiate them from the values given for forward moving observers, we define the same values for backward moving observers with $d^b_\gamma, u^b_\gamma, \gamma^b_{\max}$ and f^b_γ respectively.

In these set of lemmas and proofs, we want to show that, slowest forward and backward moving observers' cuts always have the tightest cut on stationary moving observers cut. We define, $\text{Cut}(\gamma, F)$ and $\text{Cut}(\gamma, B)$ as forward and backward moving observers' cuts which stops in an extended red or red (if $\gamma = \gamma_{\max}$ or $\gamma^b_{\max}$) phase every γ traffic signal. $\text{Cut}(S)$ is the cut of stationary moving observer. In order to facilitate proofs' readability, let us introduce two types of dimensionless variables Φ_γ and Φ^b_γ for forward and backward moving observers such that:

$$\Phi_\gamma = \left\lceil \frac{\gamma L/u_f - \delta}{C} \right\rceil \text{ and } \Phi^b_\gamma = \left\lceil \frac{\gamma L/w - \delta_w}{C} \right\rceil. \tag{5.7}$$

With the help of these variables, one can find k-coordinates of the intersection points of moving observers' cuts with stationary observers cut as follows:

$$\left\{\text{Cut}(\gamma, F) \cap \text{Cut}(S)\right\}_k = s\left[\frac{C-G}{\gamma L}\left(1 - \Phi_\gamma\right) + \frac{1}{u_f} + (C-G)\left(-\frac{\delta}{CL}\right)\right] \tag{5.8}$$

$$\left\{\text{Cut}(\gamma_{\max}, F) \cap \text{Cut}(S)\right\}_k = s\left[\frac{G}{\gamma_{\max} L}\Phi_{\gamma_{\max}} + \frac{G\delta}{CL}\right] \tag{5.9}$$

$$\left\{\text{Cut}(\gamma, B)\,\text{Cut}(S)\right\}_k = s\left[\frac{C-G}{\gamma L}\left(\Phi^b_\gamma - 1\right) + \frac{C-G}{L}\left(1 - \frac{\delta}{C}\right) + \frac{1}{u_f}\right] \tag{5.10}$$

$$\left\{\text{Cut}\left(\gamma^b_{\max}, B\right) \cap \text{Cut}(S)\right\}_k = s\left[-\frac{G}{\gamma^b_{\max} L}\Phi^b_{\gamma_{\max}} - \frac{G}{CL}(C-\delta) + \frac{1}{w} + \frac{1}{u_f}\right] \tag{5.11}$$

Lemma 1 *For $f(x) = \frac{1}{x}(\lceil tx \rceil - 1)$ where $t \in \mathbb{R}$, $f(1) \leq f(x)$ for $x \in \mathbb{Z}^+$.*

Proof 1 *Assume $n < t \leq n+1$ for $n \in \mathbb{Z}$. Note that, this assumption includes all real values of t.*

From this assumption:

$$n < t \leq n+1 \Rightarrow nx < tx \leq nx+x \Rightarrow \lceil tx \rceil \leq nx+x \Rightarrow f(x) = \frac{1}{x}\left(\lceil tx \rceil - 1\right) \leq n+1-\frac{1}{x}. \quad (5.12)$$

If we calculate $f(1)$

$$n < t \leq n+1 \Rightarrow \lceil t \rceil = n+1 \Rightarrow f(1) = n, \quad (5.13)$$

which means $f(1) \leq f(x)$ *for* $x \in \mathbb{Z}^+$.

Lemma 2 $\frac{\gamma_{\max}(L/u_f-\delta)}{C} \notin \mathbb{Z}$ *and* $\frac{\gamma^b_{\max}(L/w-\delta_w)}{C} \notin \mathbb{Z}$.

Proof 2 *Assume* $\frac{\gamma_{\max}(L/u_f-\delta)}{C} \in \mathbb{Z}$. *Then* $\frac{\gamma_{\max}(L/u_f-\delta)}{C} - \left\lfloor \frac{\gamma_{\max}(L/u_f-\delta)}{C} \right\rfloor > G/C$ *from Inequality 5.5. But this is a contradiction since* $\frac{\gamma_{\max}(L/u_f-\delta)}{C} \in \mathbb{Z}$, $\frac{\gamma_{\max}(L/u_f-\delta)}{C} = \left\lfloor \frac{\gamma_{\max}(L/u_f-\delta)}{C} \right\rfloor$ *and Inequality 5.5 does not hold. So* $\frac{\gamma_{\max}(L/u_f-\delta)}{C} \notin \mathbb{Z}$. *Similar proof can be applied for the second part of the Lemma by changing* w *with* u_f *and* δ_w *with* δ.

Lemma 3 *The slowest forward moving observer is the observer who has the tightest cut on stationary observers cut among all the other non-fastest forward moving observers.*

Proof 3 *Let us define* k_γ *as the* k*-coordinate of the* $Cut(\gamma, F) \cap Cut(S)$ *which is given in Equation 5.8. We are looking for the value of* γ *that maximizes* k_γ, *i.e. this cut will intersect with the stationary observer at the rightest possible point, creating the tightest cut. Then our problem can be defined as:*

$$\arg\max_\gamma \{k_\gamma\} = \arg\max_\gamma \left\{ s\left[\frac{C-G}{\gamma L}\left(1-\Phi_\gamma\right) + \frac{1}{u_f} + (C-G)\left(-\frac{\delta}{CL}\right)\right]\right\}. \quad (5.14)$$

Since the last two elements of the RHS of Equation 5.14 is not a function of gamma and $s\frac{C-G}{L}$ *is always positive, we can reduce the problem to:*

$$\arg\max_\gamma \{k_\gamma\} = \arg\max_\gamma \left\{ \frac{1}{\gamma}\left(1-\Phi_\gamma\right)\right\} = \arg\max_\gamma \left\{ \frac{1}{\gamma}\left[1-\left\lceil \frac{\gamma\left(L/u_f-\delta\right)}{C}\right\rceil\right]\right\}. \quad (5.15)$$

$$\text{Let } n = \frac{L/u_f-\delta}{C}, \text{ then } \arg\max_\gamma \{x_\gamma\} = \arg\max_\gamma \left\{ \frac{1}{\gamma}\left(1-\lceil n\gamma \rceil\right)\right\} \quad (5.16)$$

$$= \arg\min_\gamma \left\{ \frac{1}{\gamma}\left(\lceil n\gamma \rceil - 1\right)\right\} \quad (5.17)$$

From Lemma 1, $\gamma = 1$ is the solution.

Lemma 4 *The slowest backward moving observer is the observer who has the tightest cut on stationary observers cut among all the other non-fastest backward moving observers.*

Proof 4 *Let us define k_γ^b as the x-coordinate of the $Cut(\gamma, B) \cap Cut(S)$, which is given in Equation 5.10. Now the tightest cut is the one that minimizes the value of k_γ^b, as the backward moving observers' cuts have a negative slope in the $k-q$ plane. Then our problem can be defined as:*

$$\arg\min_\gamma \left\{k_\gamma^b\right\} = \arg\min_\gamma \left\{ s\left[\frac{C-G}{\gamma L}\left(\Phi_\gamma^b - 1\right) + \frac{C-G}{L}\left(1-\frac{\delta}{C}\right) + \frac{1}{u_f}\right]\right\} \tag{5.18}$$

$$= \arg\min_\gamma \left\{ \frac{1}{\gamma}\left[\left\lceil \frac{\gamma\,(L/w - \delta_w)}{C}\right\rceil - 1\right]\right\} \tag{5.19}$$

since RHS of the function is not a function of γ and $s\frac{C-G}{L}$ is always positive both are removed.

$$\text{Let } n = \frac{L/w - C + \delta}{C},\ \text{then } \arg\min_\gamma \left\{k_\gamma^b\right\} = \arg\min_\gamma \left\{\frac{1}{\gamma}\left(\lceil n\gamma\rceil - 1\right)\right\} \tag{5.20}$$

From Lemma 1, $\gamma = 1$ is the solution.

Lemma 5 *The slowest forward moving observer is the observer who has the tightest cut on stationary observers cut among all the other forward moving observers.*

Proof 5 *In Lemma 3 it is proven that, slowest forward moving observer has the tightest cut among all the other non-fastest forward moving observers. So it is enough to show $\epsilon \geq 0$ where ϵ is defined as:*

$$\epsilon = \{Cut(1,F) \cap Cut(S)\}_k - \left\{Cut\left(\gamma_{\max}, F\right) \cap Cut(S)\right\}_k \tag{5.21}$$

$$= s\left[\frac{C-G}{L}\left(1-\Phi_{\gamma=1}\right) + \frac{1}{u_f} - \frac{\delta}{L} - \frac{G}{\gamma_{\max} L}\Phi_{\gamma_{\max}}\right]. \tag{5.22}$$

Let $n \in \mathbb{Z}$ that satisfies

$$n - 1 < \frac{L/u_f - \delta}{C} \leq n \Rightarrow \left\lceil \frac{\left(L/u_f - \delta\right)}{C}\right\rceil = n, \tag{5.23}$$

$$\Rightarrow \epsilon = s\left[\frac{C-G}{L}(1-n) + \frac{1}{u_f} - \frac{\delta}{L} - \frac{G}{\gamma_{\max} L}\left\lceil \frac{\gamma_{\max}\left(L/u_f - \delta\right)}{C}\right\rceil\right]. \tag{5.24}$$

Since $\gamma_{\max}$ is the smallest value that does not satisfy Inequality 5.5, we have:

$$\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}-\left\lfloor\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}\right\rfloor>\frac{G}{C} \tag{5.25}$$

$$\Rightarrow\left\lfloor\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}\right\rfloor<\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}-\frac{G}{C}. \tag{5.26}$$

By Lemma 2 and Inequality 5.26 we can write:

$$\left\lceil\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}\right\rceil-1=\left\lfloor\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}\right\rfloor<\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}-\frac{G}{C}. \tag{5.27}$$

By combining Inequality 5.27 with Equation 5.24:

$$\epsilon>\left[s\left(1-\frac{G}{C}\right)\right]\left[\frac{1}{u_f}-\frac{\delta}{L}-\frac{G}{\gamma_{\max}L}+(1-n)\frac{C}{L}\right]. \tag{5.28}$$

We know the first multiplicative in the RHS of Inequality 5.28 is positive. It is sufficient to show the second one is positive too. If we multiply both sides of Inequality 5.23 by γ_{max} and apply floor function on both sides:

$$\left\lfloor\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}\right\rfloor\geq\gamma_{\max}(n-1). \tag{5.29}$$

If Inequality 5.27 is combined with Inequality 5.29 then,

$$\frac{\gamma_{\max}\left(L/u_f-\delta\right)}{C}-\frac{G}{C}-\gamma_{\max}(n-1)>0. \tag{5.30}$$

After multiplying both sides of Inequality 5.30 by a positive number $\frac{C}{\gamma_{\max}L}$ we will have:

$$\frac{1}{u_f}-\frac{\delta}{L}-\frac{G}{\gamma_{\max}L}+(1-n)\frac{C}{L}>0, \tag{5.31}$$

which is the second multiplicative on the RHS of Inequality 5.28. Since $\epsilon>0$, the slowest forward moving observer always has the tightest cut on stationary observer.

Lemma 6 *The slowest backward moving observer is the observer who has the tightest cut on stationary observers cut among all the other backward moving observers.*

Proof 6 *We will use a method similar to Proof 5. In Lemma 4, it is proven that, the slowest backward moving observer has the tightest cut among all the other non-fastest backward moving*

observers. So it is enough to show $\epsilon \geq 0$ *where* ϵ *is defined as:*

$$\epsilon = \left\{ Cut\left(\gamma^b_{\max}, B\right) \cap Cut(S) \right\}_k - \{Cut(1,B) \cap Cut(S)\}_k \tag{5.32}$$

$$= s\left(-\frac{G}{\gamma^b_{\max} L} \Phi^b_{\gamma=1} - \frac{G}{L} + \frac{1}{w} + \frac{\delta}{L} - \frac{C-G}{L} \Phi^b_{\gamma^b_{\max}} \right). \tag{5.33}$$

Let $n \in \mathbb{Z}$ *that satisfies*

$$n - 1 < \frac{L/w - C + \delta}{C} \leq n \Rightarrow \left\lceil \frac{L/w - \delta_w}{C} \right\rceil = n, \tag{5.34}$$

$$\Rightarrow \epsilon = s\left[-\frac{G}{\gamma^b_{\max} L} \Phi^b_{\gamma^b_{\max}} - \frac{G}{L} + \frac{1}{w} + \frac{\delta}{L} - \frac{Cn}{L} + \frac{Gn}{L} \right]. \tag{5.35}$$

Since $\gamma^b_{\max}$ *is the smallest value that does not satisfy Inequality 5.5, we get:*

$$\left\lfloor \frac{\gamma^b_{\max}(L/w - \delta_w)}{C} \right\rfloor < \frac{\gamma^b_{\max}(L/w - \delta_w)}{C} - \frac{G}{C}. \tag{5.36}$$

By combining Lemma 2 with Inequality 5.36 we can write:

$$\left\lceil \frac{\gamma^b_{\max}(L/w - \delta_w)}{C} \right\rceil - 1 = \left\lfloor \frac{\gamma^b_{\max}(L/w - \delta_w)}{C} \right\rfloor < \frac{\gamma^b_{\max}(L/w - \delta_w)}{C} - \frac{G}{C}. \tag{5.37}$$

If Inequality 5.37 is combined with Equation 5.35

$$\epsilon > \left[s\left(1 - \frac{G}{C}\right) \right] \left(\frac{1}{w} + \frac{\delta}{L} - \frac{Cn}{L} - \frac{G}{\gamma^b_{\max} L} \right). \tag{5.38}$$

We know the first multiplicative in the RHS of Inequality 5.38 is positive. It is sufficient to show the second one is also positive. If we multiply both sides of Inequality 5.34 by γ^b_{max} *and apply floor function on both sides:*

$$\left\lfloor \frac{\gamma^b_{\max}(L/w - \delta_w)}{C} \right\rfloor \geq \gamma^b_{\max}(n-1). \tag{5.39}$$

If Inequality 5.37 is combined with Inequality 5.39 then,

$$\frac{\gamma^b_{\max} L}{wC} - \frac{\gamma^b_{\max}(C-\delta)}{C} - \frac{G}{C} - \gamma^b_{\max}(n-1) > 0. \tag{5.40}$$

After multiplying both sides of this inequality by a positive number $\frac{C}{\gamma^b_{\max} L}$ *we will have:*

$$\frac{1}{w} - \frac{C}{L} + \frac{\delta}{L} - \frac{G}{\gamma^b_{\max} L} - (n-1)\frac{C}{L} > 0, \tag{5.41}$$

which is the second multiplicative of Inequality 5.38. Since $\epsilon > 0$*, the slowest backward moving observer always has the tightest cut on stationary observer.*

Corollary 1 *The density range of maximum capacity can be calculated by using slowest forward moving, slowest backward moving and stationary observers. For four different cases, the range can be formulated as:*

$$\begin{matrix}\textit{many forward}\\ \textit{many backward}\end{matrix} : Range_{M\text{-}M} = s\frac{C-G}{L}\left(\left\lceil \frac{L/w - \delta_w}{C}\right\rceil + \left\lceil \frac{L/u_f - \delta}{C}\right\rceil - 1\right) \tag{5.42}$$

$$\begin{matrix}\textit{one forward}\\ \textit{one backward}\end{matrix} : Range_{1\text{-}1} = s\left(\frac{1}{w} + \frac{1}{u_f}\right) - s\frac{G}{L}\left(\left\lceil \frac{L/w - \delta_w}{C}\right\rceil + \left\lceil \frac{L/u_f - \delta}{C}\right\rceil + 1\right) \tag{5.43}$$

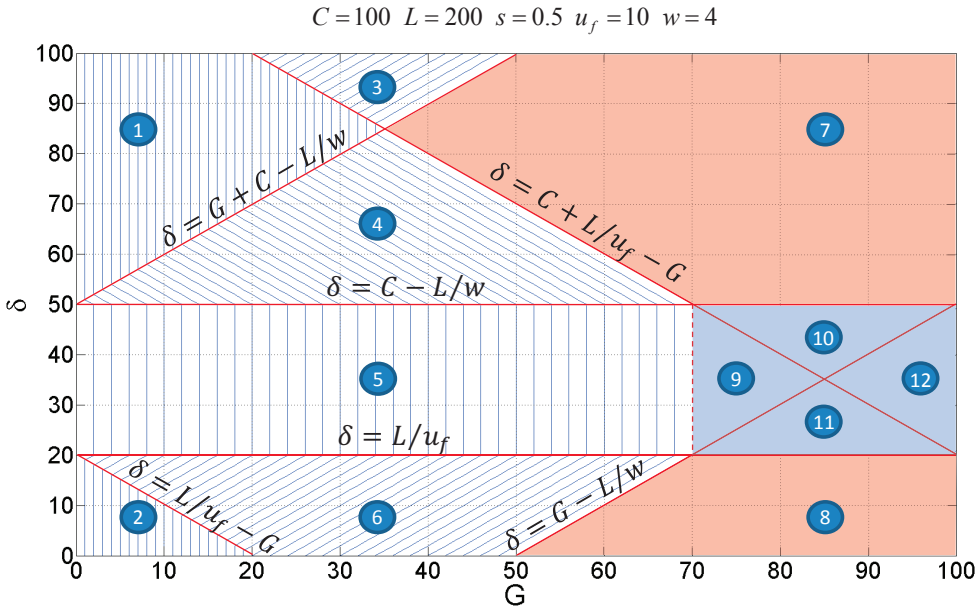
$$\begin{matrix}\textit{many forward}\\ \textit{one backward}\end{matrix} : Range_{M\text{-}1} = s\left(\frac{C-G}{L}\left\lceil \frac{L/u_f - \delta}{C}\right\rceil - \frac{G}{L}\left\lceil \frac{L/w - \delta_w}{C}\right\rceil - \frac{C-\delta}{L} + \frac{1}{w}\right) \tag{5.44}$$

$$\begin{matrix}\textit{one forward}\\ \textit{many backward}\end{matrix} : Range_{1\text{-}M} = s\left(\frac{C-G}{L}\left\lceil \frac{L/w - \delta_w}{C}\right\rceil - \frac{G}{L}\left\lceil \frac{L/u_f - \delta}{C}\right\rceil - \frac{\delta}{L} + \frac{1}{u_f}\right) \tag{5.45}$$

Equations 5.42-5.45 provide a closed form description of range. In case range has a negative value, this means that the network capacity is smaller than 1, i.e. smaller than sG/C. Figure 5.2 provides an illustration of the estimated range for $C = 100sec$ and $L = 200m$ and analytical formulae for all regions. Note that the highlighted area with blue for G between 70 and $100sec$ shows (δ, G) pairs with negative range (capacity strictly less than one) and the highlighted area with red shows (δ, G) pairs with zero capacity (i.e. the tightest forward and backward moving observers' cuts intersect at the same point on the stationary observers cut). Even if there is not an analytical proof for the required cuts to estimate the value of capacity when range is negative, not all cuts are necessary. We run numerous simulations for different values of C, L, δ and G/C and we find out that by utilizing only the slowest and fastest, forward and backward moving observers (in total 4), the error in the estimation of capacity when range is negative is small. A few more observers are needed in case of small L/C values. A description of cases with smaller capacity is provided in the following sections.

5.4 Simulation Framework

In the previous section we assumed a homogeneous network with deterministic values of all parameters (link lengths, green durations, offsets) and no turns. However, real life networks contain some variability in the network parameters and also drivers' decisions contain



#	observers	range	#	observers	range
1	1 F - 1 B	$s\left(\frac{1}{w}+\frac{1}{u_f}-\frac{2G}{L}\right)$	7	many F - many B	0
2	1 F - 1 B	$s\left(\frac{1}{w}+\frac{1}{u_f}-\frac{2G}{L}\right)$	8	many F - many B	0
3	many F - 1 B	$s\left(\frac{1}{w}+\frac{G-\delta}{L}\right)$	9	1 F - 1 B	<0
4	1 F - many B	$s\left(\frac{1}{u_f}+\frac{C-G-\delta}{L}\right)$	10	many F - 1 B	<0
5	1 F - 1 B	$s\left(\frac{1}{w}+\frac{1}{u_f}-\frac{G}{L}\right)$	11	1 F - many B	<0
6	many F - 1 B	$s\left(\frac{1}{w}+\frac{\delta-G}{L}\right)$	12	many F - many B	<0

Figure 5.2: Regions and formulations of each region according to Corollary (1) for $L/u_f + L/w < C$.

stochasticity. By introducing a degree of variability, analytical solutions are not anymore obtainable. Thus, we develop a simulation platform to estimate the passing rates and average speeds of forward and backward moving observers running a series of many intersections with variable characteristics. By analyzing the results of the simulation, we can identify the effect of heterogeneity in the topology and signal settings at the network capacity and density range.

While variational theory allows for changes in the network parameters, it does not give the ability to introduce drivers with different characteristics (free flow speed, capacity headway, etc.) and turning movements. But, we can mimic the effect of driver stochasticity and incoming turns by adjusting offsets and green durations. For example, consider an arterial's signal plan, which has been designed for a perfect progression, a "green wave", with offsets equal to L/u_f. By introducing some randomness in the offset, e.g. $L/u_f \pm 5$ sec we can imitate the variability

in the free flow travel time of the first vehicles in the platoon. Also, we will show how incoming turns affect the network capacity in cases that network density remains unchanged.

The simulation platform includes a time-space diagram with many links (~ 1000) to produce more rigorous results. The network parameters (lengths, offsets and greens) are specified from the user in the beginning of the simulation in column or matrix forms. After creating the simulation environment, we initiate different types of observers at the start time of a green phase running at the free flow speed from the upstream in the direction of flow and with the backward wave speed from downstream in the direction against flow as described by the theory of Section 5.2. When they are moving all observers on the same direction have equal speeds (free flow or backward wave) but they have different behavior than the deterministic case at (normal or extended) red phases, as there is not a repetitive deterministic pattern due to signal settings and link lengths (e.g. we cannot say that an observer stops every 3 signals).

In simulation, this pattern becomes stochastic by giving probabilities that an observer will stop if he meets an extended red phase. According to Daganzo and Geroliminis (2008) an extended red phase is used to make observers stop every $1, 2, ..., \gamma_{max} - 1$ traffic signals. Each observer in the simulation is assigned with a probability of stopping each time he meets a green phase. Faster observers are assigned with smaller probability and slower observers with higher one. For instance, if the probability assigned to this observer is 0, this observer will only stop when he hits red whereas the observer with probability 1 will stop in every signal even though he hits green. If we consider an observer with probability p, he will pass on green phases with probability $1 - p$. For each observer we have to estimate two values, the average speed and the average passing rate. Average speed can be calculated by dividing the sum of link lengths to the total travel time. Similarly, we track the number of passing/passed vehicles for each observer during the simulation and divide them by the total travel time. For the same value of p, a number of iterations is performed (~ 10) and different paths can be constructed because of the stochastic behavior of moving observers. A lower envelope of cuts is estimated to be consistent with VT. Nevertheless, we have noticed that all iterations for the same value of p and different random seeds give almost identical results. An informal pseudocode for a single forward moving observer can be seen below. The one for the backward moving observer is the same. The only difference is the speed, the start and the direction of the movement.

Note that in the pseudocode given above, green phase matrix $\mathbf{G}$ is represented as a set of unions of real number intervals $\mathbf{G}_i$ to be consistent with mathematical notations. They represent the same parameter set. Figure 5.3b shows a time space diagram with the forward (F) and backward (B) moving observers with different values of stopping probabilities (value is shown on the top of the figure).

input: n (number of traffic signals), $\mathbf{l}$ (link length vector), $\mathbf{G}_i$ (green set for signal i), p (extended red phase probability), u_f (free flow speed) and s (capacity)
output: v (average speed) and PR (average passing rate)

1. Set time t to first green signals start time t_0 and total passing rate TPR to 0.
2. Repeat if current traffic signal counter $i < n$

(a) Advance the time by travel time on that link: $t \leftarrow t + \frac{L[i]}{u_f}$.

(b) Update the current traffic signal counter $i \leftarrow i + 1$.

(c) If the trajectory hits a green ($t \in \mathbf{G}_i$)

i. If it is an extended red phase $\left(p > \text{RAND}\,(0,1)\right)$

A. Find extended red phase duration k.

B. Update total passing rate: $TPR \leftarrow TPR + ks$.

C. Advance time t to next green.

ii. If it is not an extended red phase $\left(p \leq \text{RAND}\,(0,1)\right)$

A. Do nothing.

(d) else if hits a red ($t \notin \mathbf{G}_i$)

i. Advance time t to next green.

3. Calculate the average speed $v \leftarrow \frac{\sum_{i=1}^{n} L_i}{t - t_0}$ and the average passing rate $PR = \frac{TPR}{t - t_0}$.

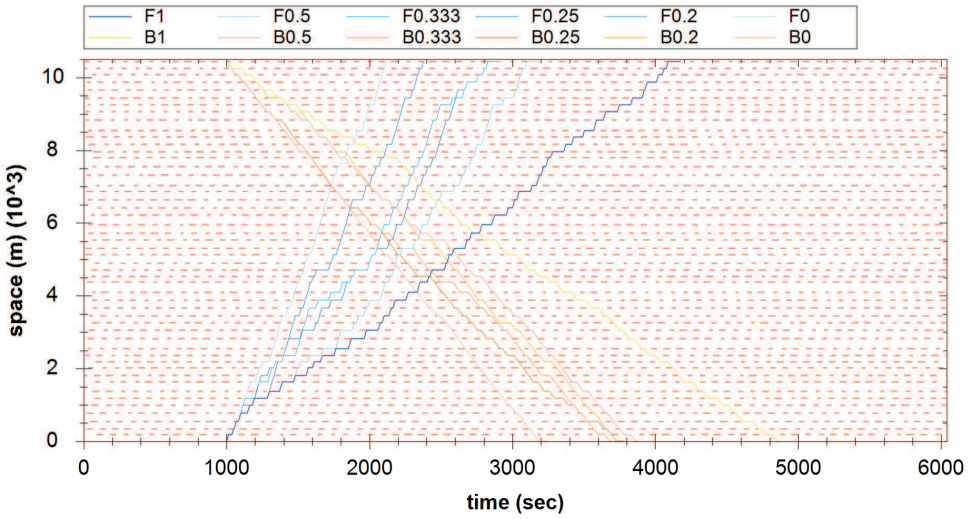


Figure 5.3: Simulation platform: (a) pseudocode (b) time-space diagram

5.4.1 Incoming Turns

Daganzo and Geroliminis (2008) showed that an MFD must arise for single-route networks with a fixed number of vehicles in circulation (i.e., periodic boundary conditions and no turns).

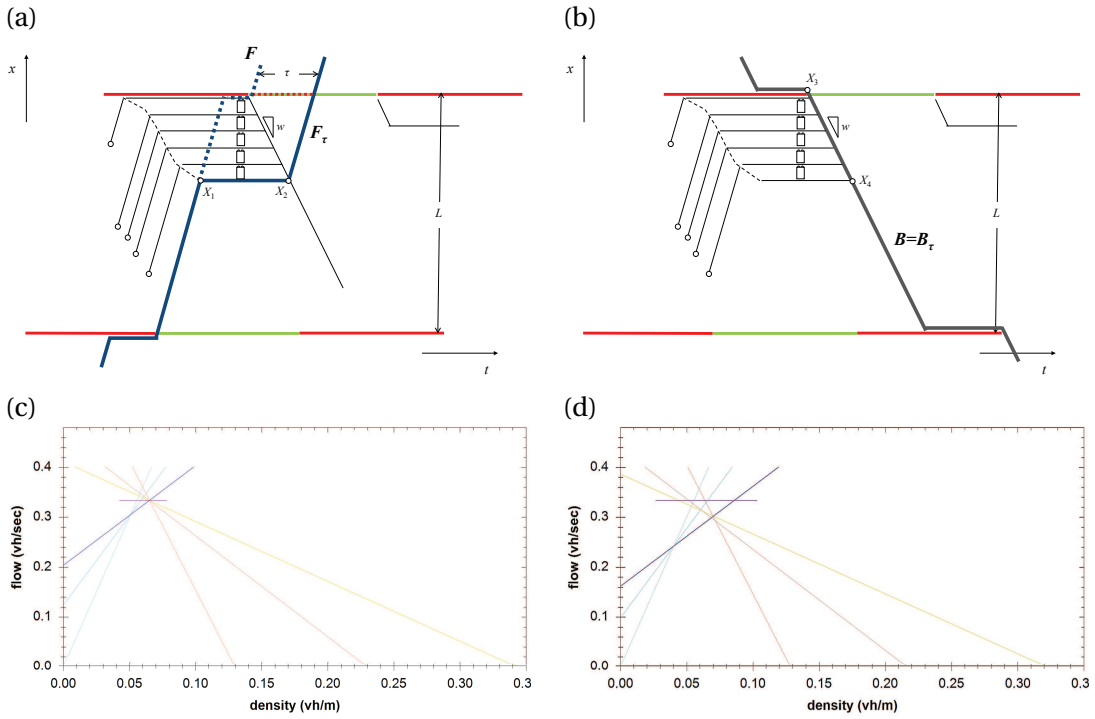


Figure 5.4: Integrating the effect of incoming turns within variational theory: time-space diagrams for forward (a) and backward (b) moving observers with ($\boldsymbol{F}$ and $\boldsymbol{B}$) and without turns ($\boldsymbol{F}_\tau$ and $\boldsymbol{B}_\tau$), flow-density diagrams without (c) and with (d) turns.

The authors also conjecture that the MFD formulae should apply to a network of intersecting routes if the numbers of vehicles in these routes are similar and roughly constant over time. We now address the effect of incoming turns in a single-route network by introducing bottlenecks of variable capacity in the proximity of the traffic signals.

Incoming turns from cross streets can significantly decrease the performance of a signalized intersection as they (i) interrupt the progression of green waves for properly timed signals and (ii) decrease the available storage capacity of the link and can cause the occurrence of spillbacks. This problem is difficult to solve in the general case because inflow needs to be separated in two classes (upstream through and incoming turns) and the “cuts” approach cannot directly identify the mixing of the vehicle origins. We address the above phenomena by changing the signal and cost function characteristics for forward and backward moving observers. To be consistent with variational theory, these turns should not significantly change the link density from one link to another, i.e. the incoming turns are considered as local link phenomena, for example there is a similar number of outgoing turns in every link to keep density constant. We also assume that incoming turns are served first because they enter the link when the through approach is in a red phase. This might not be exact when residual queues still exist or incoming turns might occur from an unsignalized intersection or a parking lot (internal source).

Consider now a queue of incoming vehicles from cross streets, Q, which entered when the signal was red for the through movement (Figure 5.4a). If these vehicles did not exist, a forward moving vehicle would follow trajectory $\boldsymbol{F}$ (Figure 5.4a) and would stop for some time at the traffic signal stop line. Because of the queue of incoming turns, the upstream vehicle needs to stop between points X_1 and X_2, follow trajectory $\boldsymbol{F}_\tau$ and cross the intersection τ seconds later, where $\tau = sQ$. According to variational theory the cost (passing rate) of a forward moving observer who stops in the middle of a link between X_1 and X_2 is st where t is the duration of stop. But, in reality no vehicles can overpass this observer while stopping, because in front of him there is a queue of vehicles entered from a cross street. This can be shown if one estimates the change between points X_1 and X_2 in the Moskowitz function $N(t,x)$ which expresses the cumulative number of vehicles in the $t-x$ plane. It is also known that this change is the same for all possible paths between A and B (Daganzo, 2005). Moskowitz function value changes during this extended red phase by an amount τs. Thus, we imitate this effect by increasing the red phase of the signal by τ, only for the forward moving observers. So, these observers, will follow trajectory $\boldsymbol{F}_\tau$ instead of $\boldsymbol{F}$. But, this extended red in the beginning of the green has passing rate zero, not s. Nevertheless, the stationary observer in front of the traffic signal continues to count vehicles for the whole duration of the real green phase as the incoming turns are served in the first τ seconds of green.

For the backward moving observer, $\boldsymbol{B}_\tau$ (Figure 5.4b) the approach is slightly different. This observer does not need to stop in the extended red phase of τ seconds. But when traveling backwards between points X_3 and X_4, its passing rate is not r, but zero. Thus, the queued vehicles from cross street, give the ability to the observer to travel in this queue with zero cost.

Based on the above, for the forward moving observers, both the speed and passing rate decreases; for the backward moving observers, only the passing rate decreases; while for the stationary observer there is no change. Thus, tighter cuts are created which can decrease both the range and the capacity of the MFD. An example is shown in Figure 5.4c and d with and without turning effects. These transformations in the trajectories of the observers are in accordance with the Lagrangian variational principles, as expressed by Daganzo (2005) and Leclercq et al. (2007) and they are valid even when the Moskowitz function is not continuous and experiences step-jumps in the time-space profile (e.g. because of incoming turns in our case). Analysis of the results is provided in the next section.

In reality, there is often a spatial correlation between incoming turns, i.e. a series of successive intersections load the main route with net positive inflows from cross streets. In this case the proposed approach lacks of methodological correctness. The reason is that these turns might increase the density in the aforementioned links of the main route and result in an inhomogeneous distribution of density, while our approach considers links with similar density. This case is not addressed in the current work, but future work can utilize our results by partitioning the route in sub-routes and estimate individual fundamental diagrams with variational theory. A research priority should be how to integrate serial or parallel routes of MFDs, with different average densities to describe the dynamics of traffic flow.

5.5 Results

In this part, we firstly investigate the deterministic cases which is solved by the analytical formulae given in "homogeneous networks" section and then continue with the "simulation" results. In order to generate isoquants either the analytical formulation (Corollary 1) or the simulation method (Figure 5.3) is used for finite number of (x, y) values which are elements of Cartesian product of number of predefined points on x and y axis of the given graph. For example for Figure 5.5a $(x, y) \in \{0.10, 0.11, ..., 0.90\} \times \{0, 1, ..., 60\}$, for Figure 5.7c $(x, y) \in \{0.10, 0.11, ..., 0.90\} \times \{\mathrm{U}(180, 180), \mathrm{U}(179, 181), ..., \mathrm{U}(80, 280)\}$ where $\mathrm{U}(*)$ represents uniform distribution.

As it is defined in Figure 5.1c, dimensionless capacity (values between 0 and 1) is the ratio of the maximum flow q_{max} to the flow observed by the stationary observer which equals to sG/C. Range is the the difference between maximum and minimum density which yields the maximum flow q_{max}. Since the value of the negative range does not mean anything and capacity equals to 1 if the range is nonnegative, it is possible to merge contour lines for both range and capacity at the same graph. These merged graphs can be seen in Figures 5.5-5.11. Note that, blue stands for the range whereas red stands for the capacity.

5.5.1 Deterministic Network Parameters

Figures 5.5 and 5.6 summarizes how range and capacity change for different values of offsets, green and cycle durations and link length. Figure 5.5a-c plot range and capacity with δ and G/C for three different cases (i) $C = 60sec$, $L = 110m$ (short link, small cycle), (ii) $C = 90sec$, $L = 225m$ (long link, medium cycle) and (iii) $C = 120sec$, $L = 180m$ (medium link, long cycle). The white area between the blue and red isoquants in graphs represents scenario with capacity 1 and range zero, i.e. tightest forward and backward observers intersect with the stationary one at the same point. Note that for a range of δ (e.g. $8 - 38sec$ in Figure 5.5a), density range is invariant with offset. Note also that by increasing G/C after some value, capacity decreases (remember that the graphs show dimensionless capacity). Note also that the effect of bad coordination in the capacity of a short link is much more significant in case of short links and when the cycle is longer (compare Figure 5.5b with Figure 5.5c). Notice that not only perfect coordination in offset (L/u_f) but also the values between L/u_f and $C - L/w$ gives the maximum range for any given G/C ratio. Furthermore, in this region, range is independent of the offset which was also proved in Section 5.3. Certainly, the positive range region is larger either when L is larger or C is smaller. This fact can be expected from the Equations (5.42-5.45).

Figure 5.5d shows how combinations of green duration and green over cycle ratio (G vs. G/C) affect range and capacity for medium size links with $L = 180m$ and bad offset (first vehicle arrives $13.5sec$ before the beginning of green phase). Note there is a boundary line which shows under what cases there is a capacity drop. Note also that this drop occurs for a larger range of green duration when the cycle is longer. This is an intuitive observation, as longer cycles create longer queues that can spillback and decrease the output of intersections. For

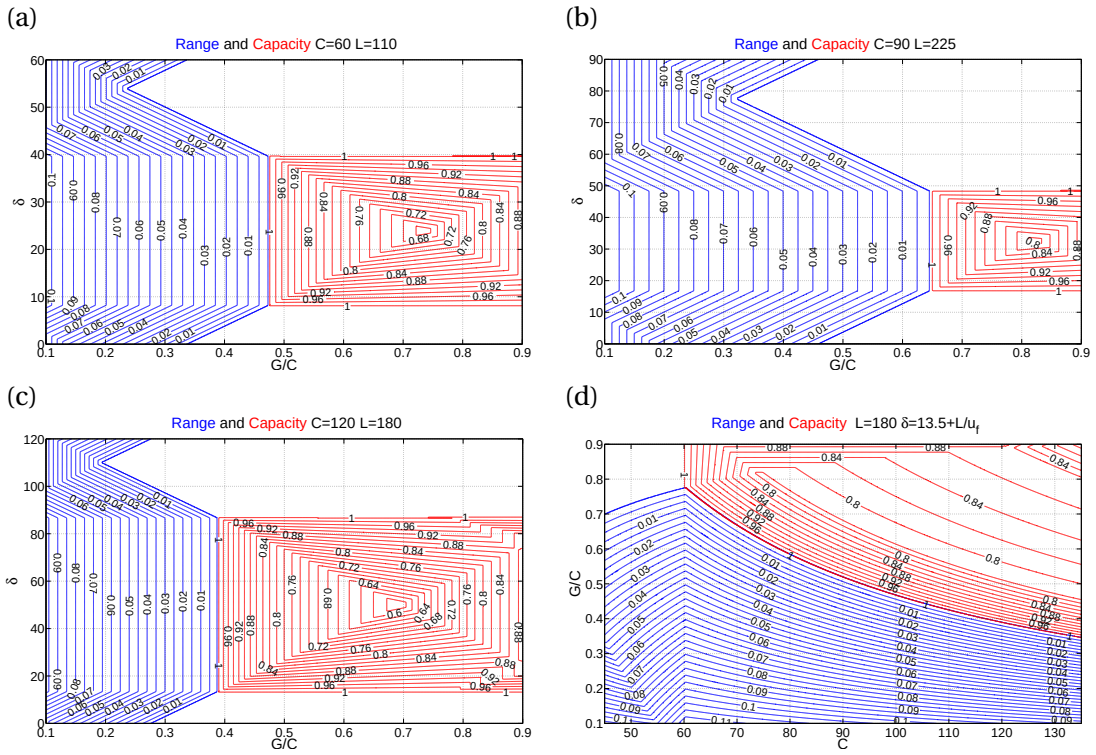


Figure 5.5: Deterministic cases (Homogeneous networks). Range and Capacity for different values of topological and signal characteristics (part 1).

$C = 90sec$, by increasing G/C from 0.5 to 0.7 (40% increase) the improvement in the maximum number of vehicles that can be served is too small (10% increase), $0.25vh/sec$ vs. $0.28vh/sec$ (the values have been obtained by multiplying the numbers of the graph with sG/C). This possibly means that the additional G/C can be utilized to serve cross streets with less delays. Similar graphs can be produced for different values of offsets and link lengths.

Figure 5.6e-f show the effect of the link length on range and capacity. Range is increasing as the link length increases in both graphs. Note that for the same value of offset, as increase in G causes a decrease in range, as the stationary observer has a higher passing rate value. An interesting observation is that routes with higher green durations (mainly the ones that carry a lot of traffic) can experience smaller capacities ($q_{max} < 1$) for a wider range of link lengths (critical length is $80m$ for $G = 24$ and $140m$ for $G = 36sec$). When L is large, capacity is always 1 for any value of offset. In this case, we can choose offset in a way to maximize the range, as this means that the signal can operate at capacity for a wide range of densities. Also, values in the white regions might not be stable as small changes in demand can create spillbacks or capacity drops.

In Figure 5.6g-h, we investigate the effect of length and green phase. Note that the boundary line for capacity less than 1, is a piecewise linear function of L and G/C. The locus of points that produce capacity drop is much larger set in case of bad offsets (Figure 5.6g). Note also

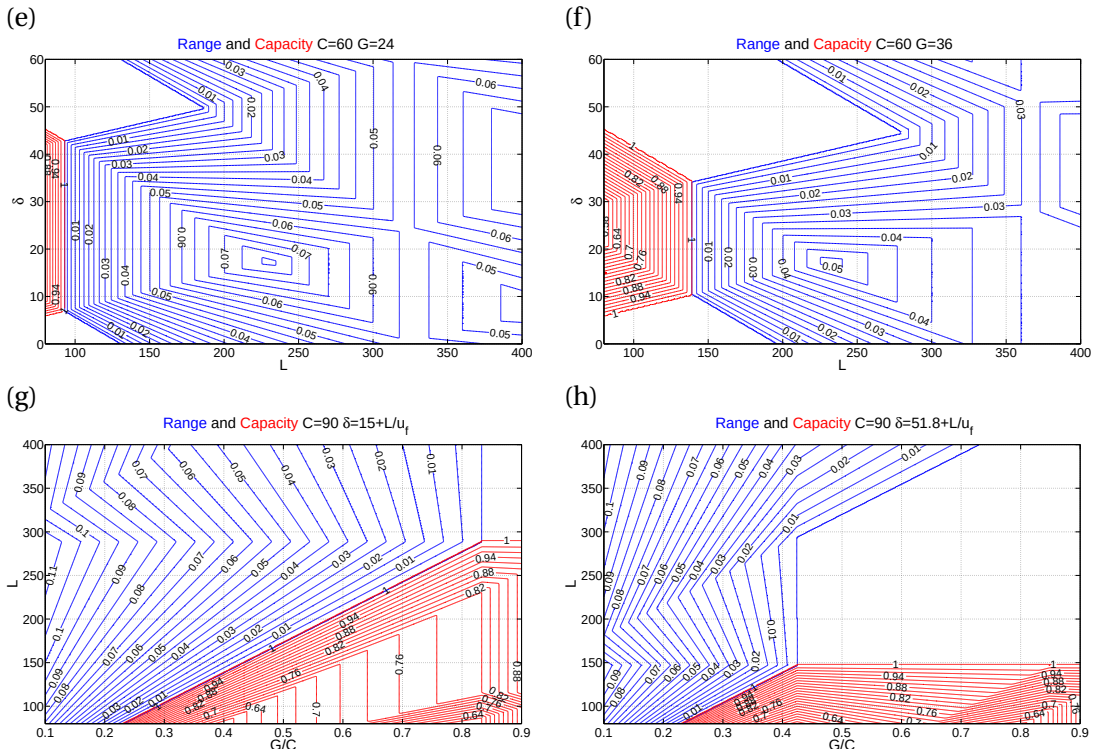


Figure 5.6: Deterministic cases (Homogeneous networks). Range and Capacity for different values of topological and signal characteristics (part 2).

that for $\delta = 51.8sec$ there is a large number of pairs with capacity 1 and range zero, i.e. all tight cuts are intersecting at the same point. In these cases heterogeneity in the distribution of congestion in the network might create a significant capacity drop (e.g. Mazloumian et al., 2010). For example for $L = 150m$ and $\delta = 15sec$ (bad offset where vehicles have to stop in every signal), G/C ratio greater than 0.43, will cause not full utilization of signal capacity while for the case of Figure 5.6h there is no capacity drop for any value of G/C. These graphs show the importance of the described methodology in estimating the effect of signal characteristics and link length in the network capacity. All of the above analysis would not be possible using standard traffic engineering techniques (e.g. the Highway Capacity Manual).

5.5.2 Stochastic Network Parameters

We now utilize the simulation platform to identify the effect of variability when compared with the deterministic cases described before. The results presented assume a uniform distribution for offsets and link lengths, $U(min, max)$. One can apply different distributions if needed. We analyze two sets of variations for the different topological and signal parameters: (i) the mean value is constant and range of the variable changes and (ii) the range is constant and the mean of the variable changes. The next three subsections present results for variations in the effect of link lengths, offsets and incoming turns.

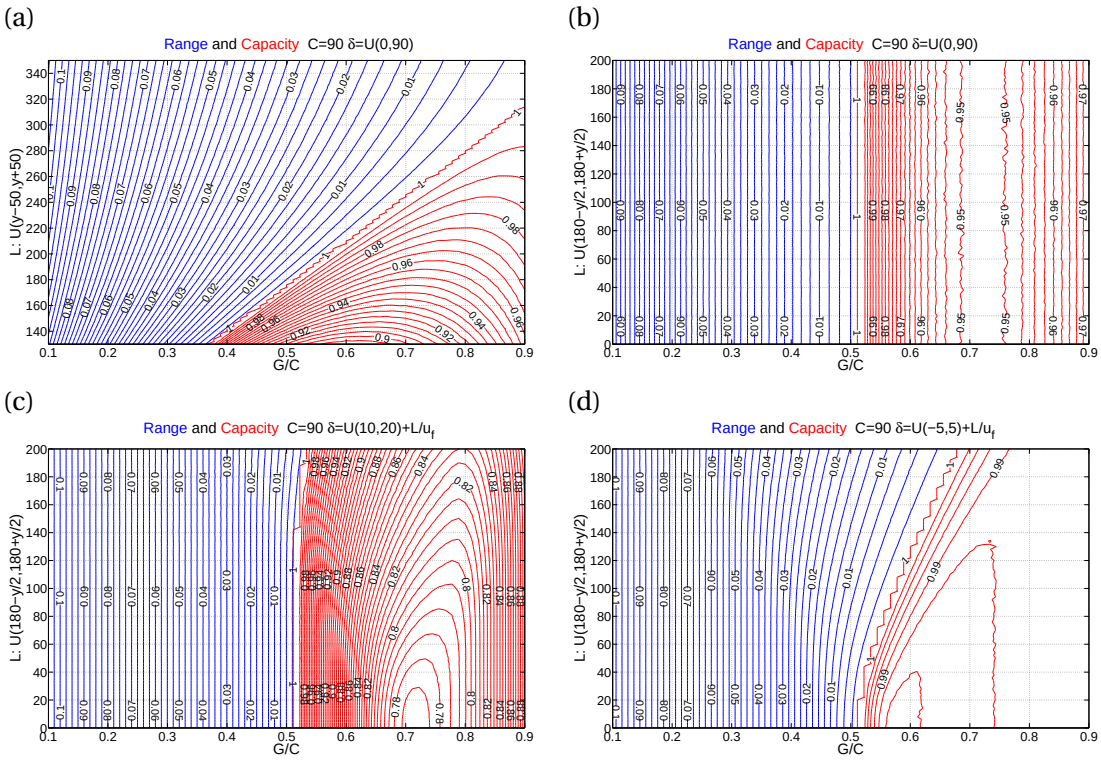


Figure 5.7: Stochastic L (part 1)

Variations in Link Length

In the graphs given in Figures 5.7 and 5.8, both the change in the mean (for a given variance) and the variance (for a given mean) of the distribution of L are analyzed. In Figure 5.7a-b, offsets are random at every intersection (varying uniformly between 0 and cycle length) and link length L has a uniform distribution. In Figure 5.7a, L has a constant range of $100m$ and variable mean, while in Figure 5.7b, L has a constant mean (medium length link with $E[L] = 180m$) and variance varies. In the first case we see that the critical link length for which a capacity drop occurs increases with the G/C ratio (almost linearly). For $G/C \leq 0.3$ capacity is always fully utilized, while for larger values we might observe a drop up to 10% for $G/C = 0.6$. In other words, shorter link lengths are more sensitive to green ratio. Once compared with the deterministic case of Figure 5.6g (bad offset), we observe that the capacity drop is less intense in case of random offsets. Figure 5.7b is intuitive to understand, as vertical isoquants mean that increase in the length variability have no effect in capacity or range for medium length links and random offsets. Thus, we are interested in identifying the critical length variability which changes the deterministic results.

In the Figure 5.7c-d, the effect of G/C ratio and variance of link length is analyzed for links with average length $E[L] = 180m$ in the case of almost perfectly (first vehicle from upstream arrives in the beginning of green) and badly coordinated signals (first vehicle arrives $15sec$ before

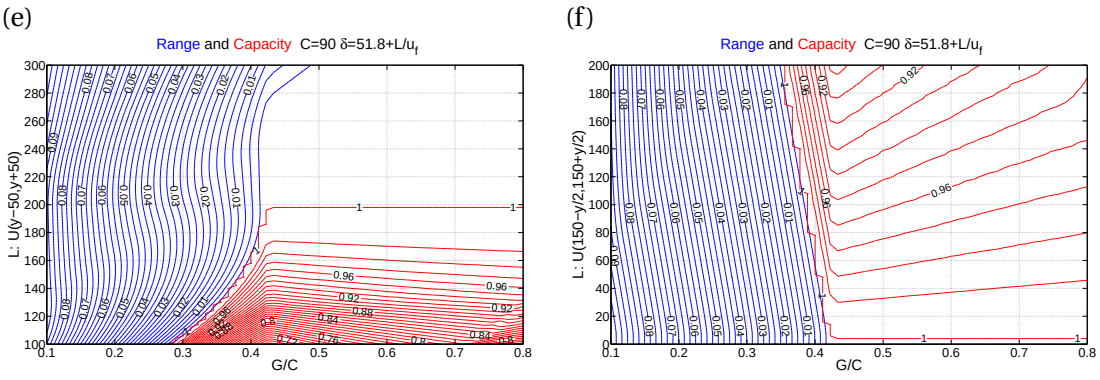


Figure 5.8: Stochastic L (part 2)

the end or red phase). A small variation $\pm 5sec$ has been introduced for good and bad offsets. Results show that length variability increase has no effect for values of G/C smaller than 0.45 (capacity is always 1 for these values). But, when capacity is less than 1 there is a range of G/C where significant increase in capacity is observed for the case of bad offsets (about 15% change for $G/C = 0.65$ once comparing deterministic L and highly variable L between 80 and $280m$). Good offsets are not significantly affected (change not more than 2%).

Figure 5.8e-f have the same structure as Figure 5.7a and Figure 5.7b but the offset is deterministic at a value of $51.8sec$, and the mean length is $150m$ in Figure 5.8f . This is the value for which capacity is equal to 1 in the case of Figure 5.6h. It is clear that for small values of G/C there is a positive range (capacity is always 1) and neither variability in length, nor change of the mean have an effect. But for $G/C \geq 0.4$ the situation is different. In Figure 5.8e, increase of G/C does not significantly change the capacity for a given length structure (note the horizontal lines). But, as expected the longer the mean the longer the capacity.

Figure 5.8f has some additional interesting characteristics. We note that when capacity is less than 1 ($G/C \geq 0.4$), increase in the length variability decreases the value of capacity (maximum drop about 6%). This is the opposite once compared with Figure 5.7c, where in case of bad offsets, increase in length variability creates an increase in the value of the capacity. To explain this, we need to look at the deterministic graphs for the specific values of offsets (Figure 5.6g-h). For example, let's focus on $G/C = 0.7$. For $L = 150m$ (Figure 5.6h), an increase in the variability will result in many short links with smaller capacity, while the longer links will have same value of capacity (equal to 1). For $L = 180$m (Figure 5.6g), as variability increases, most of the short links will have capacity around 0.76 (constant) while capacity will increase for longer links. Note that this explanation provides is a qualitative insights as the stochastic case (Figures 5.7, 5.8 and 5.9) cannot be reproduced by a linear combination of the deterministic examples.

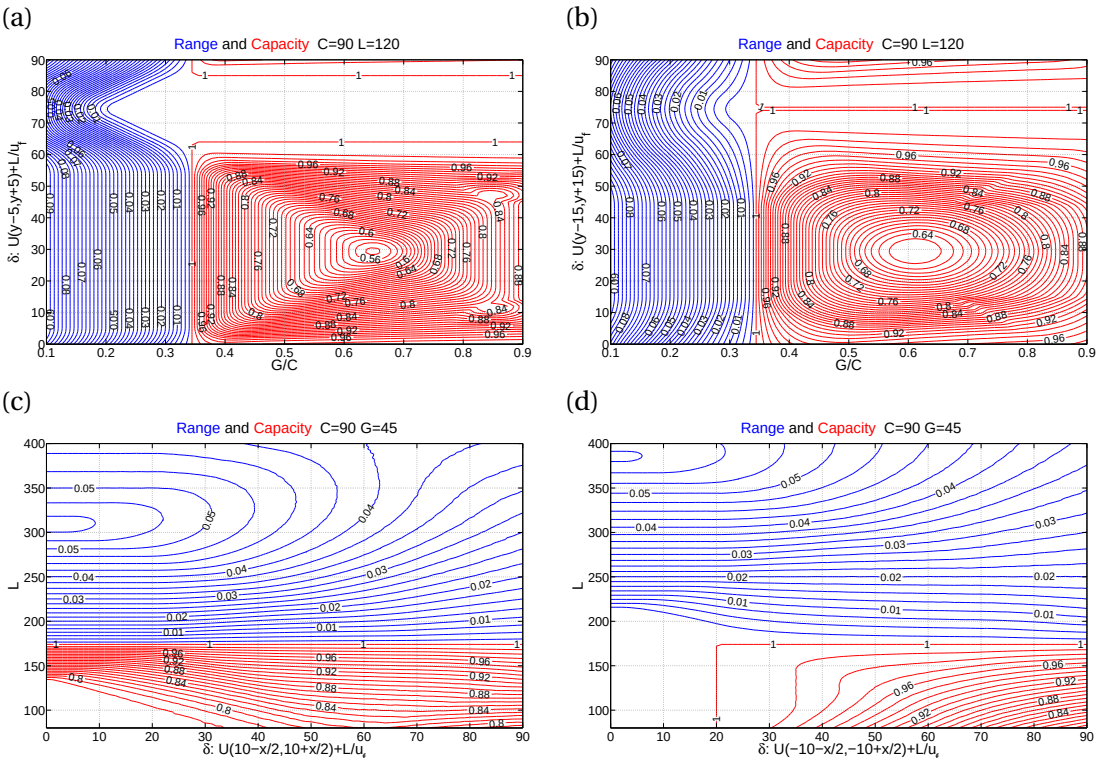


Figure 5.9: Stochastic δ

Variations in Offset

We first investigate the effect of small ($\pm 5sec$) and large ($\pm 15sec$) variations in offset by changing mean offset and G/C. The results are summarized in Figure 5.9a-b. If these graphs are compared with the deterministic graphs (Figure 5.5a-c), we can say that small offset variations have no significant effect both in capacity and range. This means that small differences in drivers' characteristics (e.g. free-flow speed, reaction time) cannot decrease the performance of traffic signals. However, in case of large offset variations (in case of poorly designed signals) the effect can be significant, especially in regions with capacity less than 1. The choice of appropriate offsets is more critical for routes or subnetworks which carry high demand and have large G/C values, e.g. high directional flows in the morning peak with small cross-street flows. Note that capacity can decrease by an amount of 20-30% for G/C in the range of 60-70%.

In the remaining two graphs, the range of the offset is investigated for good (the first vehicle from upstream arrives $10sec$ after the beginning of green) and bad coordination (the first vehicle arrives $10sec$ before the end of red). All the parameters except the means of the offsets are the same in (c) and (d). When capacity is smaller than 1 (for $L < 140m$), higher offset variability improves the capacity value in case of bad offsets and has a negative effect in case of good offsets. This result is intuitive as more (less) vehicles will hit the green phase during bad (good) coordination. For long links, offset variability decreases the capacity as for a given

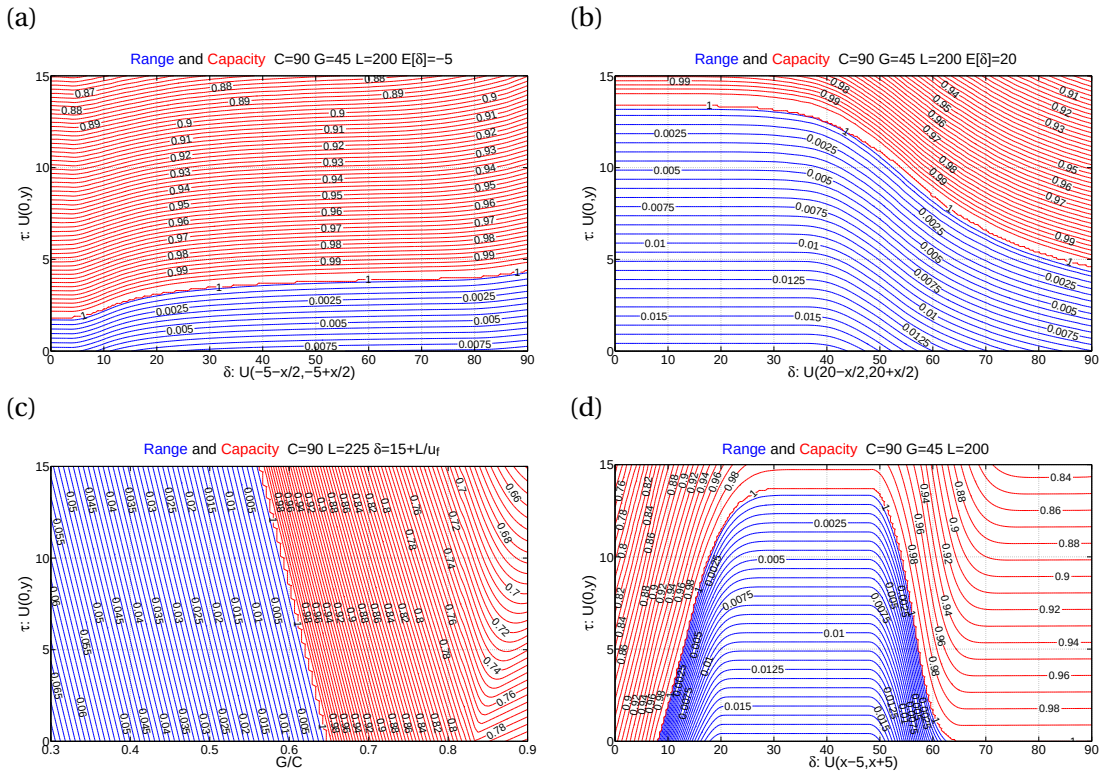


Figure 5.10: The effect of incoming turns in capacity and range (part 1).

L and G/C range is maximized when there are only one forward and one backward observers (Figure 5.5a-c).

The Effect of Incoming Turns

We now show that incoming turns from cross streets can significantly decrease the performance of a signalized intersection in some cases as they interrupt the progression of properly timed signals and decrease the available storage capacity of the link.

In all graphs of Figures 5.10 and 5.11 the vertical axes is the amount of incoming turns (expressed as the extended red phase $\tau = sQ$, which is assumed uniform between 0 and an increasing value). The first two graphs on the top show the effect of turns as the variability of offsets increases for good (left graph) and bad (right graph) offsets. In case of almost perfect offsets the effect of turns is very significant because the value of range for zero turns is very small. On the other hand, the bad offsets can absorb a high number of turns without capacity decrease. Notice that the values of the two graphs coincide for $\delta = 90sec$ as this represents the case of random offsets. A signal timing with bad offsets can absorb up to 13 seconds of turning (6.5 vehicles), while even one incoming vehicle can create problems for good coordination. But even in case of bad offsets, large G/C is problematic as range is smaller even for zero turns. Thus, in case of incoming turns, the signal plans should be chosen in a way that maximizes the

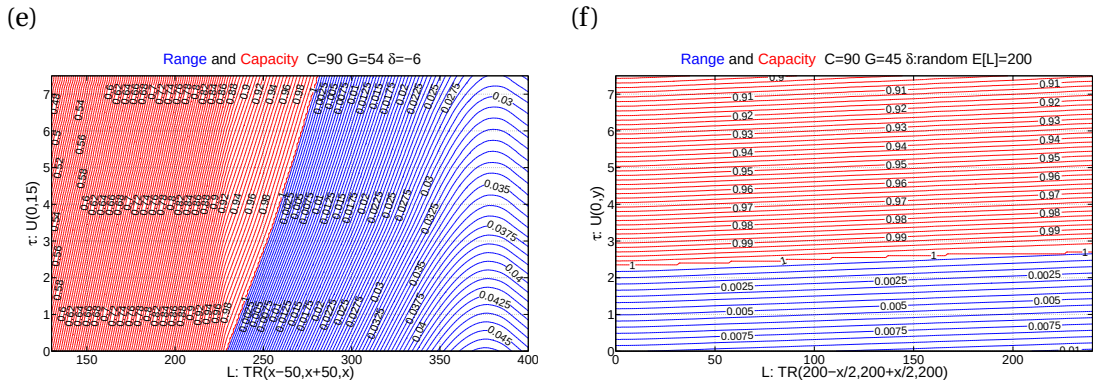


Figure 5.11: The effect of incoming turns in capacity and range (part 2).

range as capacity can be significantly decrease. For example in Figure 5.10d one can see that a bad offset with higher range is much more robust than a good offset with small range. Of course if signals are undersaturated, they will operate in values much less than capacity and the effect of turns will be minimal. But, in this work we mainly investigate signal performance in high demand conditions. Also, from Figure 5.11e-f it is clear that as length increases the effect of incoming turns becomes smaller because it is more difficult to have queue spillbacks.

5.6 Conclusions

In this chapter we have provided several extensions and refinements in the variational theory of traffic flow, which provides analytical formulae for the macroscopic fundamental diagram of urban networks. In our study we investigated the effect that have in the MFD, different degrees of variability in link lengths and signal characteristics for different city topologies and signal structures. We have integrated the effect of incoming turns in the estimation of the MFD and we showed that in many cases network capacity can significantly decrease. The scalability of flows from a series of links to large traffic networks is not a straightforward transformation. Route or network capacity can be significantly smaller than the capacity of a single link, because of the correlations developed through the different values of offsets. The above analysis would not be possible using standard traffic engineering techniques (e.g. the Highway Capacity Manual).

There is still a weak understanding on how one can characterize the breakdown dynamics and congestion spreading phenomena of traffic flows in these types of urban networks. While cascading phenomena are present in many types of physical or social systems (financing, interactions) city traffic has interesting irregularities that should be studied. We should identify the relative traffic variables that would allow a better prediction of the severity of developing and spreading of traffic congestion. This will provide clearer insights about the large variations of traffic congestion from one day to another even if demand profiles are similar.

These results can be of great importance to practitioners and city managers to unveil simple and robust signal timing planning in such a way that maximizes the network capacity and/or the density range of the capacity. The results of this chapter can be utilized to develop efficient hierarchical control strategies for heterogeneously congested cities. A network can be partitioned in homogeneous regions (with small spatial variance of congestion distribution) and optimal control methodologies can identify the inter-transfers between regions of a city to maximize the system output, as expressed by number of trip endings (see for example Ji and Geroliminis (2011) for partitioning and Daganzo (2007), Geroliminis and Daganzo (2007) or Haddad and Geroliminis (2012) for optimal control). The main logic of the strategies is that they try to decrease the inflow in regions with points in the decreasing part of an MFD and high demand for trip destinations. Given the estimated values from this task, the analysis of the current work can identify signal parameters in the individual regions of a city to move traffic smoothly at the desired flows, without concentrating a large number of vehicles at the boundaries of the regions. By restricting access to congested cities, a city manager can significantly improve system output, highlighting the importance of a reliable estimator of subnetwork/route capacity. While there are vast contributions in traffic control problems for freeways through ramp metering, the area of control for large urban regions or mixed networks still remains a challenge. Our research provides tools to shed some light towards this direction.

Current extensions of this research are investigating the network capacity and MFD patterns for cities with more complex structure (multiple modes of traffic competing for the same urban space). A difficult problem to address is how the redistribution of urban space between cars and more efficient modes can improve passenger network flows.

6 Estimation of the network capacity for multimodal urban systems

As more people through different modes compete for the limited urban space that is set aside to serve transport, there is an increasing need to understand details of how this space is used and how it can be managed to improve accessibility for everyone. Ultimately, an important goal is to understand what sustainable level of mobility cities of different structures can achieve. Understanding these outcomes parametrically for all possible city structures and mixes of transport modes would inform the decision making process, thereby helping cities achieve their sustainability goals. In this chapter we focus on the network capacity of multimodal systems with motorized traffic and extra emphasis in buses. More specifically, we propose to study how the throughput of passengers and vehicles depends on the geometrical and operational characteristics of the system, the level of congestion and the interactions between different modes. A methodology to estimate a macroscopic fundamental diagram and network capacity of cities with mixed-traffic bus-car lanes or with individual bus-only lanes is developed and examples for different city topologies are provided. The analysis is based on realistic macroscopic models of congestion dynamics and can be implemented with readily available data.

6.1 Introduction

Mobility and transportation are two of the leading indicators of economic growth of a society. As cities around the world grow rapidly and more people and modes compete for limited urban space to travel, there is an increasing need to understand how this space is used for transportation and how it can be managed to improve accessibility for everyone. This research seeks to shed some light in the macroscopic modeling of traffic flow for overcrowded cities with multimodal transport. To enhance more in this direction, we are interested in developing a macroscopic methodology to model different types of multi-modal systems, which contain buses, cars, taxis etc. with emphasis in conflicts for the same road space (e.g. mixed traffic of buses and cars; vehicles searching for parking while intervening with moving-to-destination vehicles; taxis or delivery trucks that stop to pick up passengers or goods etc.). More specifically, we propose to study how the throughput of passengers and

vehicles depends on the geometrical and operational characteristics of the system, the level of congestion and the interactions between different modes. The analysis is based on realistic macroscopic models of congestion dynamics and can be implemented with readily available data. The existence and the analytical modeling of a macroscopic fundamental diagram for multi-modal cities will be developed. Ultimately, the goal of the proposed work is to develop modeling and optimization tools which will contribute on how to allocate city space to multiple transportation modes and to understand what sustainable level of accessibility cities of different structures can achieve.

Despite the different features of these modes in terms of occupancy (number of passengers), driving behavior (speeds, acceleration and deceleration profiles, length) duration of travel, scheduled vs. non-scheduled service, a common characteristic is the following: All of these vehicles when moving to an urban environment make stops related to traffic congestion (e.g. red phases at traffic signals) and *other* stops, which also cause delays to the transportation system as a whole: buses stop at bus stops to board/alight passengers; taxis stop frequently and randomly when they search/pick up/deliver passengers; cars may stop/manoeuvre when search/find a parking spot; delivery trucks stop to pick up/deliver goods. While there is a good understanding and vast literature of the dynamics and the modeling of congestion for congestion-related stops, the effect of *service or general purpose* stops in the overall performance of a transportation system still remains a challenge. It is intuitive that the effect of these stops during light demand conditions in the network capacity is almost negligible, but nowadays city centers are experiencing high level of congestion and the frequency in time and space of the *service or general purpose* stops is significantly high. In this chapter we focus on conflicts service stops of buses, but the developed methodology can be directly applied to all the aforementioned cases.

The influence of each type of conflict have in the performance of the overall system significantly depends on the type of the network (signal and geometric characteristics), the type of service (BRT vs. regular transit vs. irregular bus services, like in under-developed countries), the operational characteristics (type of stop, passenger demand) the allocation of the road space (mixed traffic or individual bus lanes or bus and taxi lanes etc.) and the density of each mode. It strives to quantitatively estimate how each of the above parameters affect the total throughput of vehicles. Additionally, we aim to consider the effect of different occupancies between vehicles because it is important to recognize that some modes are more productive than others. Given the fact that construction of new infrastructure is not a feasible solution, we are seeking more sustainable alternatives, where more road space is allocated to public transit and "car-less cities" are organized with well-operated public transit. Our research will provide the methodological framework to quantitatively evaluate these changes with models consistent with the physics of traffic.

6.2 Literature Review

Traffic in real cities is complex, with many modes sharing streets, and congestion evolving as demand patterns change over the course of a day. Existing literature on the physics of urban mobility can be divided generally into city-scale (macroscopic) efforts and street-scale (microscopic) works. City-scale investigations have thus far looked only at the behavior of car mode without considering the interactions of different modes in traffic congestion. Studies of multiple modes, on the other hand, have only been made at the street-level scale for unrealistic time-independent scenarios. While recent findings in the macroscopic modeling and dynamics of traffic in cities have provided knowledge of single-mode, single-reservoir cities and single-mode, multi-reservoir cities (Geroliminis and Daganzo, 2007, 2008; Geroliminis and Sun, 2010), our understanding of multi-mode cities is limited. Thus, the existing body of work leaves a gap to be filled-a physically realistic time-dependent, city-scale model including multiple modes is much needed.

6.2.1 Traffic Models for Multi-Modal Transport Systems

Until the 1970s the mode was almost always the automobile, but since then some planning studies have looked at public transport on a city scale, particularly buses on idealized road networks. Making road space allocation decisions, however, requires consideration of multiple modes. To date, such considerations have been made only at the much finer street scale and still in a time-independent (unrealistic) environment. On the public transport side, city-scale modelers have looked at how systems should be designed. Wirasinghe et al. (1977) considered how to systematically design a bus transit system for an idealized city with centralized demand. They developed a model to minimize costs to users and operators by setting stop spacing and service headways, and then determining where feeder-buses to rail stations versus direct-buses operate most efficiently. However, these models have been only applied to one mode, and in the steady state. Work has also been done to look at how multiple modes can share the road, but only on the street-scale level. Sparks and May (1971) developed a mathematical model to evaluate priority lanes for high occupancy vehicles on freeways. Later, Dahlgren (1998) and Daganzo and Cassidy (2008) have studied how different modes use freeways, recognizing that if different modes serve different numbers of passengers, then analyses should not view all vehicles the same. But these works are limited to small scale systems. They looked at the effect on total passenger travel time if a lane on a specific road section were dedicated to multiple occupant vehicles. This consideration of different occupancies between vehicles is important because it recognizes that some modes are more productive than others. The importance of considering passengers rather than vehicles was further voiced by Vuchic (1981). He criticized street-scale evaluations based only on vehicle flows, because multimodal systems should not view all modes as the same. The quantitative treatment of the transit process (e.g. network route design, scheduling) is reflected in a considerable amount of effort in numerous books and publications (e.g. Ceder, 2007; Ceder and Wilson, 1986; Hickman et al., 2008), and will not be addressed in this work.

Many researchers have looked at allocating street space between more than one mode whether it be through the dedication of a freeway lane to high occupancy vehicles or a lane for buses on a city street (Radwan and Benevelli, 1983; Black et al., 1992). Like earlier studies, they focused on passenger travel time, in this case considering various degrees of mixing among modes in traffic, as well as high occupancy vehicle lanes and bus-only lanes. This method has limited applicability, however, because it assumes steady state traffic flow which ignores the fluctuations and spill-over effects that typically characterize urban traffic congestion. More recently, Currie et al. (2004) argued for a full accounting of impacts including environmental impacts in planning studies of road space allocation. That analysis is based on a disaggregate micro-simulation which relies on intensive travel data inputs that are typically unreliable or unavailable. Also, micro-simulation studies cannot systematically cover the space of possible inputs, and therefore an understanding of multimodal traffic on the city-scale remains elusive. While the above-cited methodology, to its credit, promotes accounting for a wide range of impacts, the analyses have yet to be conducted on a full city scale.

6.2.2 Macroscopic Models of Single-Mode Traffic in Urban Networks

With respect to macroscopic modeling of single-mode traffic, various theories have been proposed for the past 40 years to describe vehicular traffic movement in cities on an aggregate level. These works have attempted to predict both the average and the distribution of speed in an urban area as a function of explanatory variables that characterize the demand and the network infrastructure (e.g. Smeed, 1967; Wardrop, 1968). But these models cannot be used to describe the rush hour in a congested city as they contain monotonically decreasing relationships between average speed and flow.

The first instance of a macroscopic fundamental diagram (MFD) showing an optimum car density was presented by Godfrey (1969). Earlier studies looked for macro-scale traffic patterns in data of lightly congested real-world networks (Godfrey, 1969; Ardekani and Herman, 1987; Olszewski et al., 1995) or in data from simulations with artificial routing rules and static demand Williams et al. (1987); Mahmassani and Peeta (1993). However, the data from all these studies were too sparse or not investigated deeply enough to demonstrate the existence of an invariant MFD for real urban networks.

Support for its existence has been given only very recently (Geroliminis and Daganzo, 2007, 2008). These references showed that (1) the MFD is a property of the network itself (infrastructure and control) and not of the demand, i.e. the MFD should have a well-defined maximum and remain invariant when the demand changes both with the time-of-day and across days and (2) the space-mean flow is maximum for the same value of critical density of vehicles, independently of the origin-destination tables and (3) there is a robust linear relation between the neighborhoods' average flow and its total outflow (rate vehicles reach their destinations). These properties can be a reliable tool for decision-makers to evaluate demand-side policies for improving mobility and anticipate the results of smart traffic management policies. For

example, signals in Zürich are already controlled dynamically to maintain the speed and reliability of surface transit (Ott, 2002). Nevertheless, MFDs should not be universally expected. Recently, Geroliminis and Sun (2011a) explored the effect of inhomogeneous distribution of vehicles in space and time and studied how the spatial variability of vehicle density can affect the shape, the scatter and the existence of a well-defined Macroscopic Fundamental Diagram. Daganzo et al. (2011) investigated bifurcation and instability issues of an MFD for a two-ring network. Buisson and Ladier (2009) and Ji et al. (2010) investigated how different parameters affect the shape of the MFD with simulation and real data from Toulouse and Amsterdam. While MFDs should be expected for arterial networks that satisfy the above condition, freeway networks have topological or control characteristics that are different; non-redundant, no traffic signals, hysteresis phenomena (Geroliminis and Sun, 2011b). From the above it is clear that the existing body of work leaves a gap to be filled -a physically realistic time-dependent, city-scale model including multiple modes is much needed.

6.3 Methodological Description

In this research, for estimating the network capacity, *moving observer* method is used. With the help of *Variational theory*, MFD for a ring road with no incoming or outgoing traffic can be estimated. Theoretically, this curve is an upper bound for the estimated MFD for the real urban networks Daganzo and Geroliminis (2008). However findings from experimental and real networks have shown that, the estimated upper bound is almost tight for equally congested, redundant urban networks Geroliminis and Daganzo (2008).

The moving observer method, Variational theory and their theoretical background is described more in Section 5.2. Please refer to this section for further information.

6.3.1 Extension of VT for Multi-Modal Networks

Let's consider here a street of length L with a fixed number of lanes but any number of intersections. The intersections can be controlled by stop lines, roundabouts, traffic signals or any type of control that is time-independent on a coarse scale of observation; i.e. large compared with the signal cycles. We are interested in solutions where the flow at the downstream end of the street matches the flow at the upstream end; e.g. as if the street formed a ring, because then the average density does not change. In VT, the street can also have any number of time-invariant and/or time-dependent point bottlenecks with known capacities. The bottlenecks are modeled as lines in the (t,x) plane on which the "cost" per unit time equals the bottleneck capacity, $q_B(t)$. The only constraint is that all intersections and bottlenecks need to have periodic characteristics both in time and space, i.e. we cannot model a single accident or a non-recurrent bottleneck with VT.

The above broad applicability of VT gives us the flexibility to model many different types of conflicts in traffic movements as hypothetical traffic signals of type p, with the following

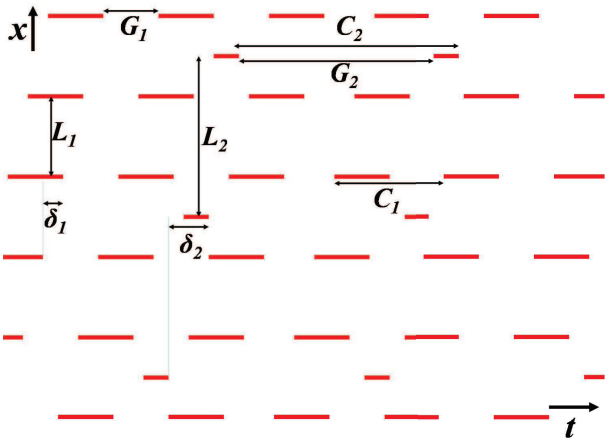


Figure 6.1: A time-space diagram for an arterial with periodical signalized intersections and bus stops

periodic characteristics: length L_p, cycle C_p, green duration G_p offset δ_p and capacity during red phase, C_r^p . For example, Figure 6.1 shows how a bus stop can be treated in VT. It shows an one-way, one-lane arterial with links of length L_1, traffic signals G_1 , δ_1, C_1 and mid-block bus stops without a curb pocket, every two links. The operational characteristics of this hypothetical signal-bus stop G_2 , δ_2, C_2 depend on the dwell times (G_2) and the frequencies of buses (C_2), while the capacity during red is C_2^r as buses block traffic when stop for this case. Differently than normal traffic signals G_2 , δ_2, C_2 have strong stochastic characteristics and fluctuations, which make this formulation more challenging.

Consideration shows that if all the blocks of our street are sufficiently long then the maximum average flow as expressed by the shortest path (SP) is a horizontal line along the trajectory of one of the intersections; and the capacity is simply: $R(0) = \min_p \{s_p G_p / C_p\}$. However, if some of the blocks are short then there could be shortcuts that use red periods at more than one intersection and the capacity can be significantly smaller. The definition of a short link does not only include its length but also the signal settings. In other words, very short red periods (e.g. when a bus stops for $10sec$ to pick up a passenger $100m$ upstream of the stopline) may not have significant effect in the MFD of a city with medium-length links, but a bus stop of $40sec$ may decrease the capacity of an MFD for the same city structure. Another significant variable in both cases is the frequency of bus stops or other conflicts.

An MFD for a single-mode is a two-dimensional plot between average network flow and average network density. In a multimodal system these definitions are vague because (i) there are more independent variables that affect the network flow (e.g. the density of each mode) and (ii) each mode has different passenger occupancy. This problem is complex because the dynamics of vehicle conflicts (as they will be estimated through VT) depend on the vehicle types and not on their occupancies, but the operational characteristics of the hypothetical traffic signals may depend on vehicle occupancies, as for example the dwell times of buses. Once completing all the challenging mathematical formulations and analyses of this work,

it is meaningful to introduce a new concept of an MFD for cities where now the network flow is expressed as passenger-kilometers per unit time, instead of the current approach of vehicle-kilometers per unit time.

The fact that different pairs of conflicts (e.g. bus-bus, bus-car, car-taxi) have a broad range of operational characteristics, emphasizes the need for careful consideration and analysis of each case individually. We now describe the characteristics of our methodology and report upon the estimation of model parameters for different types of multi-use lanes. The moving observer method of variational theory to estimate an MFD is a powerful tool that can simulate a broad range of applications.

Individual bus lanes

There are many cities around the world, where bus lines have been designed to fully utilize the dedicated space. An example of this is Oxford Street in London which is devoted to 18 bus lines, and is so fully utilized that it tends at times to be congested with buses. In case of individual bus lanes, we first need to estimate the relative capacity ("cost") function (CF), $r(u)$, that describes each homogeneous portion a street. These parameters can be directly estimated using GPS data of closely interacting buses or loop detector data. More specifically, free-flow speed u_f^b, jam density κ^b and capacity q_m^b suffice to estimate the cost function.

Special treatment should be given to model bus stops as hypothetical traffic signals. The reason is that buses do not run on schedule at all times and dwell times can vary significantly from bus stop to bus stop or for different buses at the same stop. Thus, both the cycle time and the duration of the green phase should be modeled with stochastic characteristics. While analytical solutions can be obtained under some special cases for specific moving observers using variational theory, we use a simple computer simulation, which will be described later, to see the effect of different distributions of bus arrivals and dwell times in the performance of bus lanes as expressed by an MFD. Detailed GPS data can be utilized to estimate these distributions, such as the behavior of offsets, which we expect to be totally random in many cases. The location of the bus stop measured as distance from the intersection, L^b, is also a critical variable for the system performance.

The total number of bus lanes and the different types of bus stops should also be considered. Normal bus stops intervene with main traffic and buses block a whole road lane when stopping for boarding/alighting passengers. In that case the capacity of the hypothetical signal during red is $C_b^r = 0$. Bus bays are the type with an extra lane at the stop site where buses can dwell on the extra lane. Because of the lower interfering, other vehicles can continue moving and the effect of a bus stopping is negligible. The number of births is also another key variable, which expresses the number of buses which can stop simultaneously and serve passengers. When all berths are full, waiting buses create longer red phases and can block the moving traffic even in case of bus bays. One can estimate the duration of red periods and cycles for different arrival profiles using queueing theory formulations. We expect that the maximum throughput of

two-lanes will be higher than double the throughput of one-lane bus systems.

Mixed traffic bus-lanes with cars

The efficiency of mixed traffic lanes heavily depends on the density of each individual mode and the geometric characteristics. When these lanes operate close to capacity and car queue lengths are medium to high, even a small number of buses intervening with car traffic can cause significant loss of capacity. In case of long links and bus stops located far from the intersection, the effect of bus operation may be negligible for the rest of the traffic. For given arterial structures, we will investigate the effect of buses for different densities of buses and public transit demand (related to dwell times and frequencies). Knowing the shape of a mixed-traffic MFD is a necessary tool to (i) understand the efficiency of these lanes and (ii) decide if by separating modes network passenger flow can increase. We will estimate MFDs for different values of (i) network variables, (network length in lane-km, average link length, number of lanes); (ii) link variables for 1-lane (u_f, κ, q_m); (iii) intersection variables (δ, C and G); (iv) bus related variables (type of bus stop, number of births, dwell times and frequency) and (v) car and bus densities. The above formulation also applies to mixed lanes of buses and taxis, where taxis are not allowed to start or terminate their service in the bus lanes (commonly used to many cities).

6.4 Implementation and Results

Daganzo and Geroliminis (2008) and Boyacı and Geroliminis (2011b) provided analytical solutions for VT problems with deterministic values of all parameters for single-mode systems. However, real life multi-mode networks contain some variability in the network or bus operation parameters, which make analytical solutions difficult. Thus, we develop a simulation platform to estimate the passing rates and average speeds of forward and backward observers running a series of many intersections with variable characteristics. The simulation platform includes a time-space diagram with many links (~ 1000). The network parameters (lengths, offsets and greens) are specified from the user in the beginning of the simulation in column or matrix forms. After creating the simulation environment, we send different types of observers at the start time of a green phase running, (i) at the free flow speed from the upstream in the direction of flow and (ii) with the backward wave speed from downstream in the direction against flow. Every observer on the same direction has equal (free flow or backward wave) speeds but they have different behavior at extended red phases. More specifically each observer is assigned with a probability that she stops if she meets an extended red phase. Faster observers are assigned with smaller probability and slower observers with higher one. For each observer we have to estimate the average speed and the average passing rate. Average speed can be found by dividing the sum of link lengths to the total travel time. Similarly, we track the number of passing/passed vehicles for each observer during the simulation and divide them by the total travel time. An informal pseudocode is shown in Figures (6.3) and

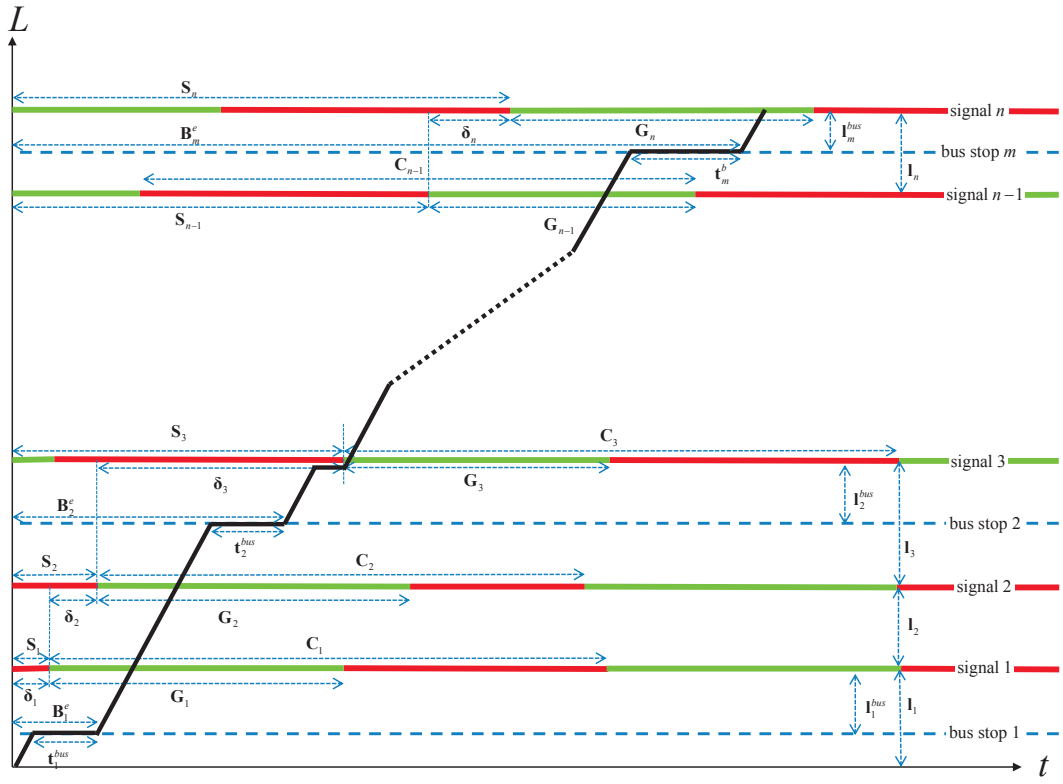


Figure 6.2: Bus stops with correlated offsets

(6.2). Bus stops are like traffic signals with a significant difference in case of multiple lanes. Normal traffic can overtake buses when they use a bus bay to stop or if there are more available lanes. Thus, an observer that stops in a "red phase" of a bus stop is overtaken with maximum passing rate equal to $(N-1)*s$, where N is the number of lanes and q_m the saturation flow. Note that the interarrival time between buses, C^{bus}, is estimated among all different bus lines traveling in the same road section and it is not the headway of a specific bus lane, which can be much larger (5-15min).

The results presented in Figure (6.4a-d), show the vehicle network capacity of multi-mode systems with different operational and topological characteristics, while figures (6.5e-h) show the passenger network capacity. The values have been normalized with the vehicle capacity of an isolated traffic signal with no buses, i.e. $s\frac{G}{C}$. To estimate passenger capacity is fair to assume that given the same operational characteristics and level of service of buses (stop frequency, headways, number of bus lines, passengers trip length etc.), the average number of passengers inside the bus, N_{pax} is proportional to the number of passengers getting in and out (consider Little's formula in queueing theory). Thus, we assume that the relationship between N_{pax} and dwell times t^{bus} is $N_{pax} = (t^{bus} - 5sec) * a$, where 5sec is the time needed to open and close the doors and a is a parameter depending on the passenger's trip length and bus headways. In our calculations we used a conservative value of 1.5, which means for

input: observer type, $\mathbf{p}$, n', $\mathbf{l}'$, $\mathbf{S}'$, $\mathbf{C}'$, $\mathbf{G}'$, $\mathbf{E}'$, $\mathbf{T}$, u_f, w, q
output: v (average speed) and PR (average passing rate)

1. If current observer is a forward moving observer:

(a) Set current time t, , state counter i and passing rate TPR: $t \leftarrow 0, i \leftarrow 1, \text{TPR} \leftarrow 0$

(b) Repeat until the end of the states: While $i \leq n'$

i. Advance the time to the next state: $t \leftarrow t + \frac{\mathbf{l}'_i}{u_f}$

ii. Find the normalized time and the time difference: $t_n \equiv t - \mathbf{S}'_i \pmod{\mathbf{C}'_i}, \Delta t = t - t_n$

iii. If the current event is not stopping and extended stop happens: If $t_n < G'_i$, $t_n > \mathbf{E}'_i$ and RAND() $< \mathbf{p}_{\mathbf{T}_i}$

A. Update the total passing rate, TPR and normalized time $t_n \leftarrow G$.

iv. If the current event in the state in stopping: If $t_n \geq G'_i$

A. Update the total passing rate, TPR and normalized time $t_n \leftarrow C'_i$.

v. Update the current time and state counter: $t \leftarrow \Delta t + t_n$ and $i \leftarrow i + 1$

(c) Calculate passing rate and average speed: PR $\leftarrow \frac{\text{TPR}}{t}$ and $v \leftarrow \frac{\sum_{i=1}^{n'} \mathbf{l}'_i}{t}$

2. If current observer is a backward moving observer

(a) Set start time to the start of the last event: $t_0 \leftarrow \mathbf{S}'_{n'}$

(b) Set current time, state counter and total passing rate: $t \leftarrow t_0$ and $i \leftarrow n' \text{TPR} \leftarrow 0$

(c) Repeat until the start of the states: While $i \geq 1$

i. Find the normalized time and time difference: $t_n \leftarrow t - \mathbf{S}'_i \pmod{\mathbf{C}'_i}, \Delta t = t - t_n$

ii. If the current event is not stopping and extended stop happens: If $t_n < G'_i$, $t_n > \mathbf{E}'_i$ and RAND() $< \mathbf{p}_{\mathbf{T}_i}$

A. Update the total passing rate, TPR and normalized time $t_n \leftarrow G$.

iii. If the current event in the state in stopping: If $t_n \geq G'_i$

A. Update the total passing rate, TPR and normalized time $t_n \leftarrow C'_i$.

iv. Update the current time and advance it to the next state: $t \leftarrow \Delta t + t_n$ and $t \leftarrow t + \frac{\mathbf{l}'_i}{w}$

v. Update state counter: $i \leftarrow i - 1$

(d) Calculate passing rate and average speed: PR $\leftarrow \frac{\text{TPR}}{t - t_0}$ and $v \leftarrow \frac{\sum_{i=1}^{n'} \mathbf{l}'_i}{t - t_0}$

3. If current observer is a stationary observer

(a) calculate passing rate and average speed: PR $\leftarrow q \frac{\sum_{i=1}^{n'} G'_i}{\sum_{i=1}^{n'} C'_i}$ and $v \leftarrow 0$

Figure 6.3: Pseudocode for finding forward and backward moving observer parameters

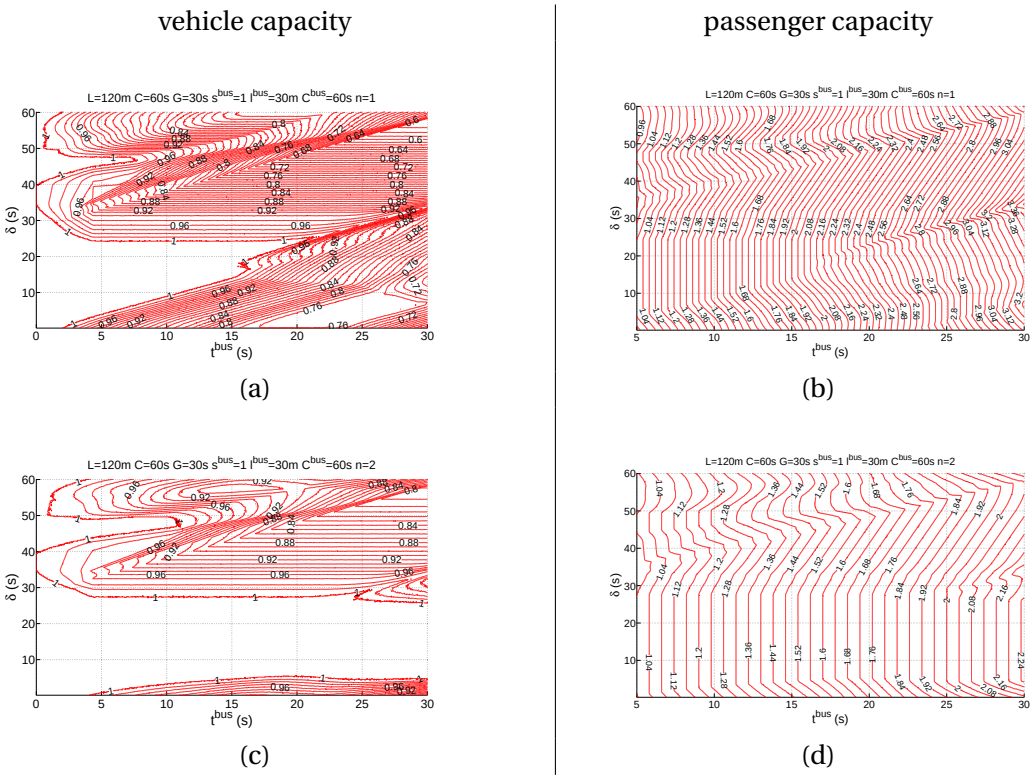


Figure 6.4: Vehicle and passenger capacities of networks with buses (part 1).

example that a passenger that stops for $15sec$, carries about 15 passengers. There are more complex formulations for bus dwell time for different parameters (e.g. Zografos and Levinson, 1986; Dueker et al., 2004; Li et al., 2006), we use a simpler approach. But different methods can be applied for the same analysis.

We analyze an example of a city structure with short links ($L = 120m$), where the effect of bus operations will be higher, given that queues can easily spillback and block upstream links. Bus stops are located close to the traffic signals ($l^{bus} = 30m$), while spatial frequency of bus stops, s^{bus}, is every 1 and 2 links. We included results in case of roads with width of one and two lanes. As mentioned before the difference is that when a bus stops in a two lane street, vehicles can overpass it from the left lane, while in one-lane street cars have to queue behind the bus. We estimated vehicle and passenger network capacity for a wide range of signal offsets and dwell times. Offset $\delta = 30sec$ is a perfect green-wave offset, while larger (smaller) values represent cases where a free-flowing car from upstream will hit the green (red) phase.

An interesting observation is that in many cases we observe horizontal lines in the graphs (e.g. for $\delta = 30-40sec$, and $t^{bus} = 10-25sec$ in Figure (6.4a)). This means that the effect of longer bus stops in the network capacity is negligible. This is for example a case where a bus stops behind a red signal to pick up passengers. For a given city topology and dwell times, city planners can choose the optimal offsets to minimize the effect of bus operations to the overall

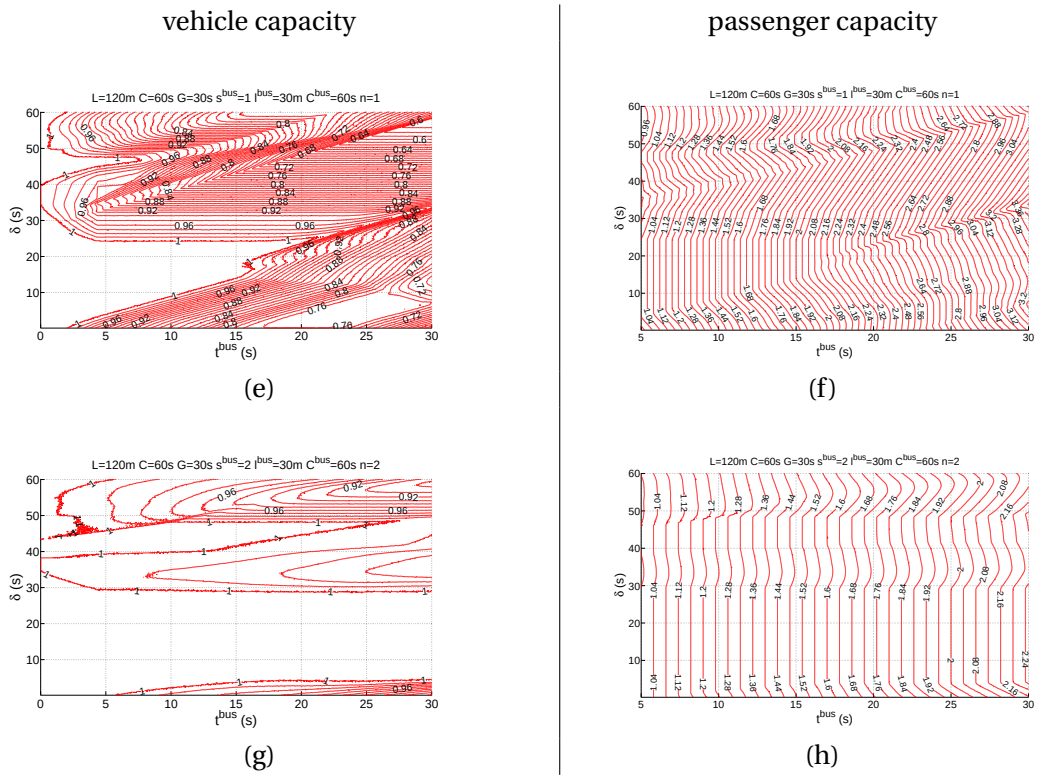


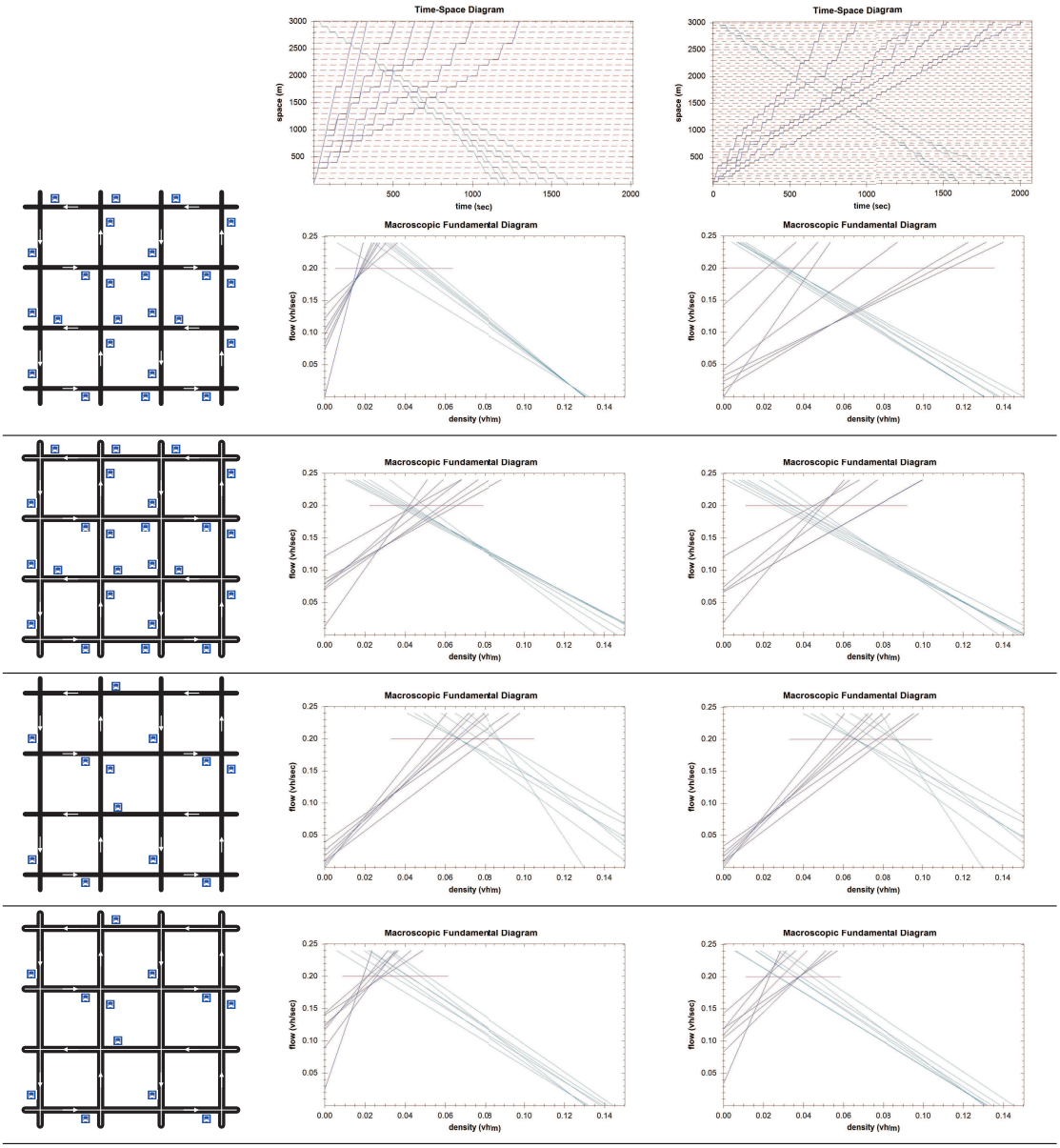
Figure 6.5: Vehicle and passenger capacities of networks with buses (part 2).

traffic. Note that in some cases, these offsets are different than the perfect green wave offsets.

While the decrease in the vehicle network capacity is small if one considers that buses are carrying more passengers, there is a significant increase in the passenger network capacity, i.e. these systems are able to serve a much higher number of passengers per unit time. This is shown in Figure (6.5e-h). But, if this is not accompanied by modal shift from private cars to public transit, then this capacity cannot be utilized as the city will operate at congested levels.

To estimate the level of congestion and the flow of passengers or cars in the network one needs to know the density of cars in the network, i.e. to know at what point of the MFD the city operates. To answer this question the dynamics of the system should be studied as a function of the modal shift. More specifically, we currently study what is the minimum level of modal shift for a given multi-modal city structure to have an increase in passenger capacity when compared with car-only networks.

We now investigate not only the capacity but also the shape of the MFD for 4 different city structures as shown in Figure (6.6). For all 4 cases the length of the block is $L = 100m$, while the number of lanes per block $n = 1$ or 2 and the spacing of bus stops is $s^{bus} = 1$ or 2 links. All bus stops are located in the middle of the block ($L^{bus} = 50m$) and the ratio $G/C = 0.4$. We investigate the shape of the MFD with and without buses for $C = 60sec$ and for buses carrying on average 10 and 20 passengers. We also summarize the passenger and network capacity



$G/C = 0.4, L = 100, L^{bus} = 50, C^{bus} = 60$															
		$n = 1$								$n = 2$					
				veh		pax						veh		pax	
$s_{bus} = 1$	$\delta = 0$	C	NB	10	20	10	20	$\delta = 50$	C	NB	10	20	10	20
		60	1.00	0.78	0.64	1.78	2.64		60	1.00	0.88	0.79	1.38	1.79
		90	1.00	0.75	0.62	1.75	2.62		90	0.73	0.73	0.73	1.23	1.73
$s_{bus} = 2$	$\delta = 35$	C	NB	10	20	10	20	$\delta = 0$	C	NB	10	20	10	20
		60	1.00	0.92	0.92	1.92	2.92		60	1.00	0.99	0.94	1.49	1.94
		90	0.72	0.72	0.72	1.72	2.72		90	0.89	0.89	0.88	1.39	1.88

Figure 6.6: Vehicle and passenger capacities of networks with and without buses for some specific cases

for C = 60 and 90sec for all of the cases. The results are summarized in Figure (6.6). For Case 1 we plot the MFDs for no bus and mixed traffic with buses carrying 20 passengers. We also show how the simulator estimates the cuts of the MFD based on the moving observers trajectories. For Cases 2-4 we plot the MFDs for the mixed traffic with buses carrying 10 and 20 passengers. Note that despite the fact that vehicular network capacity decreases, a larger number of passengers can be served per unit time. These MFDs can be utilized to estimate the level of congestion for a cities with a specific demand for cars and buses (ongoing work).

6.5 Conclusions

This work provides useful tools to develop efficient management strategies of urban congestion to minimize the effect of the general purpose stops to the rest of the traffic. Management strategies can be implemented to partition a city so that road space is deliberately allocated between competing modes. Although the allocation of this space is eminently political, it should be informed by the correct physics of traffic, which is the objective of this work. This would allow for the analysis of the performance of different modes using the same road space under different management strategies, such as mixing traffic or separating modes by special-use lanes. Further tests are needed to investigate the effect of buses in case of individual bus lanes , especially in case of high frequencies of buses, because queues of buses might be created in front of a stop with a limited number of births. Queueing analysis can be a useful tool to identify the distributions of t^{bus}, as this can be extended to large periods if consequent buses serve the same stop.

Space should be allocated taking into account spatiotemporal differences in the demand and the geometry of the road. These spatiotemporal decisions are important because, if they are made incorrectly, space could be wasted. If this wasted space could be productively used by low-occupancy vehicles without affecting the more productive modes, mobility is being restricted. For example, recent studies in Californian freeways, have questioned the effectiveness of high-occupancy lanes (HOVs) and have shown that HOV lanes are underutilized and the passenger capacity of freeways has decreased, resulting in heavier congestion levels (Chen et al., 2005).

A transportation system can be treated as an interconnected network of "reservoirs" with one or more modes moving, where each reservoir represents the streets in a neighborhood. In this extension, different parts of a city can be subject to different management strategies. Perhaps bus-only streets are allocated only in the central business district while other parts of the city allow vehicles to operate in mixed traffic. The effect of changes in one reservoir on the behavior of adjoining reservoirs will also be considered with this model. Figure (6.7) shows an example of a multi-reservoir city, with different allocations of road space in each reservoir. Our understanding of multi-mode, multi-reservoir cities is a research priority.

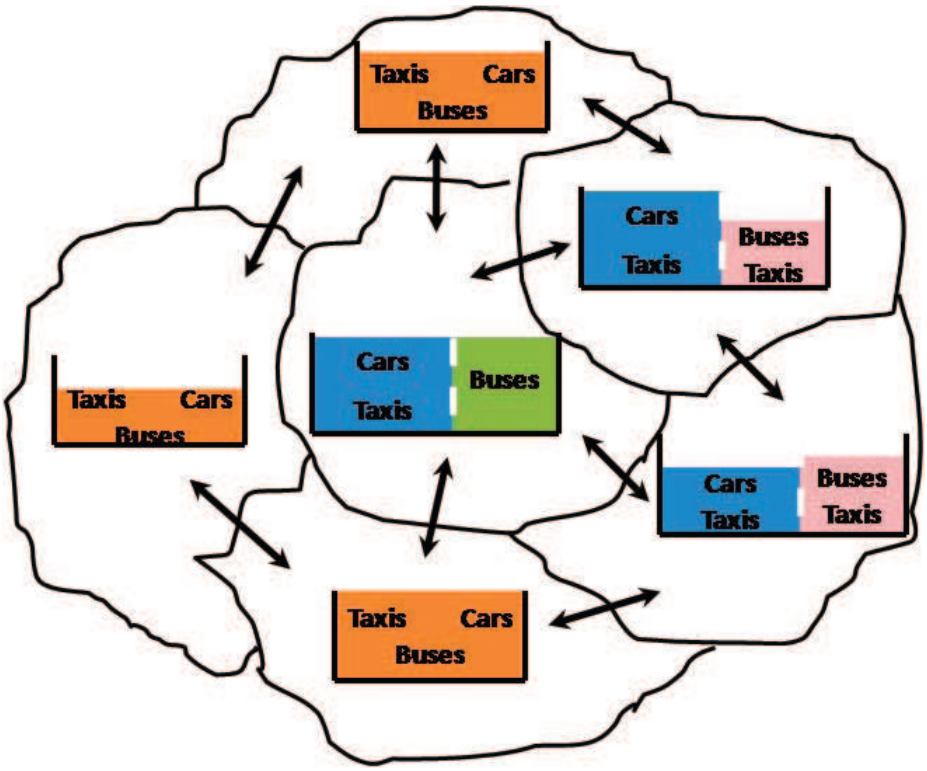


Figure 6.7: A multi-reservoir, multimodal system

7 Conclusions

This final chapter summarizes the results in this dissertation and proposes ideas for future research. This thesis deals with three complimentary subjects related to urban networks: **(i)** Hypercube queueing models are an example of spatial queueing systems and applies in a lot of problems in urban operational research literature. **(ii)** Car-sharing is a new concept to rent vehicles for short period of times; a new model that stays between public transportation and private ownership. **(iii)** Travel time, flow and density estimation in congested urban networks with the help of Variational theory is a new method that needs to be further investigated. The conclusions and contributions can be seen at the end of each chapter in details. In this final chapter, we want to state these conclusions briefly and discuss the future research directions related to all three.

7.1 Hypercube Queueing Models

The contributions of this thesis in the literature of hypercube queueing models give opportunity to model and solve queueing systems with spatial characteristics that enables service rates depending on both the server and the incident location. The models existing in the literature, do not take into consideration the incident location to decide the service rate. It is assumed that the travel time is negligible compared to service time on-scene. It might be a valid assumption for some types of queueing systems, as fire brigades. However for ambulances and taxis, travel time is as significant as on-scene service time. Unfortunately, this extension increases the problem size. In order to cope with it, we propose an approximation algorithm that combines two different hypercube models with a new mathematical model for partitioning. We also make these methods applicable inside optimization procedures and utilize them to improve system performance of such systems.

In this research, we group the incidents according to the closest servers: If an incident is served by the closest server, this service is regarded as an intradistrict service. If the same incident is served by another server that can reach the incident, this service is called interdistrict service. In our future research, we want to apply a service rate which depends only on the distance

between the server and the incident. This way, we can make some service range belts, apply different rates for each belt and calculate system performance for different server location configurations. Extensions of this work to handle larger sizes of problems is another research priority.

7.2 One-Way Car-Sharing

There are various different types of car-sharing systems. In this thesis we specifically deal with non-floating one-way (electric) car-sharing systems. In these systems, there are designated parking spots for the vehicles and rentals can be started and ended in different parking spots. In our research we mostly aim systems with reservations. We work on strategic and tactical decisions first. An optimization framework, that enables to find parking spots and fleet size is developed. We have also applied our research on an electric car-sharing system operating in Nice. A current field test in the city of Nice has taken into consideration recommendations that have been concluded through the work of this thesis. The second part of this research contains the decisions related to every day operations, e.g. relocation of vehicles and assignment of relocation personnel to shifts. The operational decision framework is still a work in progress but we share our preliminary results in this thesis in a short chapter as well.

In addition to work presented here, we are developing a simulator which will enable us to simulate not only non-floating systems with reservations but any car-sharing models. One of the other future research direction we are following contains developing heuristics for the operational problem. These heuristics will be less accurate from the operational model we present in the thesis but they will give results in short time. They will be used to "correct" our system if there is a deviation from the system that is planned at the beginning of the day by using the operational model. Extension of this work to multiple vehicle types and different types of reservations (only at the origin, with limited time constraints) will also be considered.

7.3 Estimation of Network Capacity in Congested Urban Systems

It has been recently shown that there is a relationship between network flow and density in the urban networks, known as macroscopic fundamental diagram. This relationship between flow and density enables to estimate various properties in the urban networks, e.g. travel time, capacity, flow. It is also shown that for the networks with homogeneous system characteristics (e.g. distance and offset between traffic signals) macroscopic fundamental diagrams can be generated mathematically. In our research, we have moved this research to one further step. First, we propose a closed mathematical formulation for the density range of the maximum capacity in the networks with homogeneous characteristics. Second, we apply a different approach to see macroscopic relationship on the networks with heterogeneous system characteristics. We also investigate the effect of left turn and public transportation on the network performance.

There are several future research directions in this topic. First of all, we want to see the effect of the multimodality created by other means of transport. We think that, these analysis could be useful to decide on efficient allocation of space between different modes. One of the other research directions is about partitioning the cities to have partition specific space allocations.

Bibliography

A. A. Aly and J. A. White. Probabilistic formulation of the emergency service location problem. *Journal of the Operational Research Society*, 29:1167–1179, 1978.

T. Andersson and P. Värbrand. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2006.

S. Ardekani and R. Herman. Urban network-wide traffic variables and their relations. *Transportation Science*, 21(1):1–16, 1987.

J. Atkinson, I. Kovalenko, N. Kuznetsov, and K. Mykhalevych. A hypercube queueing loss model with customer-dependent service rates. *European Journal of Operational Research*, 191(1): 223 – 239, 2008.

M. O. Ball and F. L. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1):18–36, 1993.

M. Barth, M. Todd, and L. Xue. User-based vehicle relocation techniques for multiple-station shared-use vehicle systems. *Transportation Research Record*, 1887:137–144, 2004.

M. Barth, S. A. Shaheen, F. Tuenjai, and F. Atsushi. Carsharing and station cars in asia: Overview of Japan and Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, 1986:106 – 115, 2006.

J. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–252, 1962.

J. A. Black, P. N. Lim, and G. H. Kim. A traffic model for the optimal allocation of arterial road space: a case study of Seoul's first experimental bus lane. *Transportation Planning and Technology*, 16:195–207, 1992.

B. Boyacı and N. Geroliminis. Extended hypercube queueing models for stochastic facility location problems. In *25th European Conference on Operational Research (EURO XXV)*, Vilnius, 2012a.

B. Boyacı and N. Geroliminis. Facility location problem for emergency and on-demand transportation systems. In *12th Swiss Transport Research Conference*, Monte Veritá, 2012b.

B. Boyacı and N. Geroliminis. Extended hypercube models for large scale spatial queueing systems. In *91th Annual Meeting of the Transportation Research Board*, Washington D.C., 2012c.

B. Boyacı and N. Geroliminis. Extended hypercube models for location problems with stochastic demand. In *2nd Symposium of the European Association for Research in Transportation*, Stockholm, 2013.

B. Boyacı, N. Geroliminis, and K. Zografos. Developing and solving an integrated multi-objective model for supporting strategic and tactical decisisions for one-way car-sharing systems. In *26th European Conference on Operational Research (EURO INFORMS MMXIII)*, Rome, 2013a.

B. Boyacı, N. Geroliminis, and K. Zografos. An optimization framework for the development of efficient one-way car-sharing systems. In *13th Swiss Transport Research Conference*, Monte Veritá, 2013b.

B. Boyacı, N. Geroliminis, and K. Zografos. A generic one-way multi-objective car-sharing problem with dynamic relocation. In *Proceedings of the Eighth Triennial Symposium on Transportation Analysis (TRISTAN VIII)*, San Pedro de Atacama, 2013c.

B. Boyacı and N. Geroliminis. Estimation of the network capacity for multimodal urban systems. In *6th International Symposium on Highway Capacity and Quality of Service*, 2011a. 6th International Symposium on Highway Capacity and Quality of Service.

B. Boyacı and N. Geroliminis. Exploring the effect of variability of urban systems characteristics in the network capacity. In *90th Annual Meeting of the Transportation Research Board*, Washington D.C., 2011b.

L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

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

A. B. Calvo and D. H. Marks. Location of health care facilities: an analytical approach. *Socio-Economic Planning Sciences*, 7(5):407–422, 1973.

A. Ceder. *Public transit planning and operation: theory, modelling and practice*. Butterworth-Heinemann, 2007.

A. Ceder and N. H. M. Wilson. Bus network design. *Transportation Research Part B: Methodological*, 20:331–344, 1986.

E. M. Cepolina and A. Farina. A new shared vehicle system for urban areas. *Transportation Research Part C: Emerging Technologies*, 21(1):230 – 243, 2012.

D. Chemla, F. Meunier, and R. Wolfler Calvo. Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization*, 10(2):120–146, 2012.

C. Chen, P. Varaiya, and J. Kwon. *An empirical assessment of traffic operations*, pages 105–124. Proceedings 16th International Symposium on Transportation and Traffic Theory. Elsevier, Amsterdam, Netherlands, 2005.

R. Church and C. ReVelle. The maximal covering location problem. *Papers in Regional Science*, 32(1):101–118, 1974.

J. L. Cohon. *Multiobjective programming and planning*, volume 140. Dover Publications, 2004.

C. Contardo, C. Morency, and L.-M. Rousseau. *Balancing a dynamic public bike-sharing system*. CIRRELT, 2012.

L. Cooper. Heuristic methods for location-allocation problems. *SIAM Review*, 6:37–53, 1964.

L. Cooper. Location-allocation problems. *Operations Research*, 11(3):331–343, 1963.

L. Cooper. The transportation-location problem. *Operations Research*, 20(1):94 – 108, 1972.

G. H. d. A. Correia and A. P. Antunes. Optimization approach to depot location and trip selection in one-way carsharing systems. *Transportation Research Part E: Logistics and Transportation Review*, 48(1):233 – 247, 2012.

T. Courbon and L. Leclercq. Cross-comparison of macroscopic fundamental diagram estimation methods. *Procedia - Social and Behavioral Sciences*, 20(0):417 – 426, 2011. The State of the Art in the European Quantitative Oriented Transportation and Logistics Research 14th Euro Working Group on Transportation & 26th Mini Euro Conference & 1st European Scientific Conference on Air Transport.

K. Crane, L. Ecola, S. Hassell, and S. Natarah. An alternative approach for identifying opportunities to reduce emissions of greenhouse gases. Technical report, RAND Corporation, 2012. URL http://www.rand.org/content/dam/rand/pubs/technical_reports/2012/RAND_TR1170.pdf.

T. Cucu, L. Ion-Boussier, Y. Ducq, and J.-M. Boussier. Management of a public transportation service: carsharing service. In *The 6th International Conference on Theory and Practice in Performance Measurement and Management*, Dunedin, New Zealand, 2009.

G. Currie, M. Sarvi, and B. Young. *Urban Transport X Urban transport and the environment in the 21st century*, chapter A New Methodology for Allocating Road Space for Public Transport Priority, pages 375–388. WITpress, Germany, 2004.

C. F. Daganzo. A variational formulation of kinematic waves: basic theory and complex boundary conditions. *Transportation Research Part B: Methodological*, 39(2):187 – 196, 2005.

C. F. Daganzo and M. J. Cassidy. Effects of high occupancy vehicle lanes on freeway congestion. *Transportation Research Part B: Methodological*, 42:861–872, 2008.

C. F. Daganzo and N. Geroliminis. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transportation Research Part B: Methodological*, 42(9):771 – 781, 2008.

C. F. Daganzo, V. V. Gayah, and E. J. Gonzales. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. *Transportation Research Part B: Methodological*, 45(1):278 – 288, 2011.

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

C. Daganzo and M. Menendez. A variational formulation of kinematic waves: Bottleneck properties and examples. In *Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT)*, pages 345 – 364, 2005.

J. Dahlgren. High occupancy vehicle lanes: Not always more effective than general purpose lanes. *Transportation Research Part A: Policy and Practice*, 32:99–114, 1998.

G. B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8(1):101–111, 1960.

M. Daskin. A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science*, 17(1):48–70, 1983.

M. Daskin and E. Stern. A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152, 1981.

G. Desaulniers, J. Lavigne, and F. Soumis. Multi-depot vehicle scheduling problems with time windows and waiting costs. *European Journal of Operational Research*, 111(3):479–494, 1998.

K. Dueker, T. Kimpel, and J. Strathman. Determinants of bus dwell time. *Journal of Public Transportation*, 7(1):21–40, 2004.

M. Duncan. The cost saving potential of carsharing in a US context. *Transportation*, 38: 363–382, 2011.

D. Efthymiou, C. Antoniou, and P. Waddell. Factors affecting the adoption of vehicle sharing systems by young drivers. *Transport Policy*, 29:64–73, 2013.

W. D. Fan, B. M. Randy, and N. E. Lownes. Carsharing: Dynamic decision-making problem for vehicle allocation. *Transportation Research Record: Journal of the Transportation Research Board*, 2063:97 – 104, 2008.

A. E. Fassi, A. Awasthi, and M. Viviani. Evaluation of carsharing network's growth strategies through discrete event simulation. *Expert Systems with Applications*, 39(8):6692–6705, 2012.

R. Galvão and R. Morabito. Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research*, 15(5):525–549, 2008.

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

M. Gendreau, G. Laporte, and F. Semet. Solving an ambulance location model by tabu search. *Location Science*, 5(2):75–88, 1997.

M. Gendreau, G. Laporte, and F. Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27(12):1641–1653, 2001.

M. Gendreau, G. Laporte, and F. Semet. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, 57(1):22–28, 2005.

D. George and H. Cathy. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research*, 211:198–207, 2011.

N. Geroliminis and B. Boyacı. The effect of variability of urban systems characteristics in the network capacity. *Transportation Research Part B: Methodological*, 46(10):1607–1623, 2012.

N. Geroliminis and C. Daganzo. Macroscopic modeling of traffic in cities. In *86th Annual Meeting of the Transportation Research Board*, Washington, DC, 2007. Paper No. 07-0413.

N. Geroliminis, M. Karlaftis, and A. Skabardonis. A spatial queuing model for the emergency vehicle districting and location problem. *Transportation Research Part B: Methodological*, 43(7):798 – 811, 2009.

N. Geroliminis, K. Kepaptsoglou, and M. Karlaftis. A hybrid hypercube - genetic algorithm approach for deploying many emergency response mobile units in an urban network. *European Journal of Operational Research*, 210(2):287–300, 2011.

N. Geroliminis and C. F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. *Transportation Research Part B: Methodological*, 42(9): 759–770, 2008.

N. Geroliminis and J. Sun. Properties of a well-defined macroscopic fundamental diagram for urban traffic. In *89th Annual Meeting of the Transportation Research Board*, Washington D.C., 2010.

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

N. Geroliminis and J. Sun. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. *Transportation Research Part A: Policy and Practice*, 45(9):966–979 (Paper published in the 19th International Symposium on Transportation and Traffic Theory – ISTTT), 2011b.

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

J. Haddad and N. Geroliminis. On the stability of traffic perimeter control in two-region urban cities. *Transportation Research Part B: Methodological*, 46:1159–1176, 2012.

S. Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3):450–459, 1964.

T. S. Hale and C. R. Moberg. Location science research: a review. *Annals of Operations Research*, 123(1):21–35, 2003.

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

R. Herman and I. Prigogine. A two-fluid approach to town traffic. *Science*, 204(4389):148–151, 1979.

M. D. Hickman, P. B. Mirchandani, and S. Voss. *Computer-aided systems in public transport*, volume 600 of *Lecture notes in economics and mathematical systems*. Springer, Berlin Heidelberg, 2008.

K. Hogan and C. Revelle. Concepts and applications of backup coverage. *Management Science*, 32(11):1434–1444, 1986.

A. Iannoni, R. Morabito, and C. Saydam. A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. *Annals of Operations Research*, 157:207–224, 2008.

A. P. Iannoni and R. Morabito. A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. *Transportation Research Part E: Logistics and Transportation Review*, 43(6):755 – 771, 2007. Challenges of Emergency Logistics Management.

INSEE. Résultat du recensement de la population 2009. Data File, 2010. URL http://www.insee.fr/. Institut National de la Statistique et des Études Économiques (INSEE).

J. P. Jarvis. Approximating the equilibrium behavior of multi-server loss systems. *Management Science*, 31(2):235–239, 1985.

Y. Ji, W. Daamen, S. Hoogendoorn, S. Hoogendoorn-Lanser, and X. Qian. Investigating the shape of the macroscopic fundamental diagram using simulation data. *Transportation Research Record: Journal of the Transportation Research Board*, 2161:40–48, 2010.

Y. Ji and N. Geroliminis. On the spatial partitioning of urban transportation networks. (submitted for publication), 2011.

D. Jorge, G. Correia, and C. Barnhart. Comparing optimal relocation operations with simulated relocation policies in one-way carsharing systems. In *Transportation Research Board 92nd Annual Meeting*, 13-4559, 2013.

A. G. Kek, R. L. Cheu, and M. L. Chor. Relocation simulation model for multiple-station shared-use vehicle systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1986:81 – 88, 2006.

A. G. Kek, R. L. Cheu, Q. Meng, and C. H. Fung. A decision support system for vehicle relocation operations in carsharing systems. *Transportation Research Part E: Logistics and Transportation Review*, 45(1):149–158, 2009.

S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220 (4598):671–680, 1983.

P. Kolesar and W. E. Walker. An algorithm for the dynamic relocation of fire companies. *Operations Research*, 22(2):249–274, 1974.

G. Laporte, V. Louveaux, François, F. Semet, and A. Thirion. Application of the double standard model for ambulance location. In J. A. Nunen, M. G. Speranza, and L. Bertazzi, editors, *Innovations in Distribution Logistics*, volume 619 of *Lecture Notes in Economics and Mathematical Systems*, pages 235 – 249. Springer Berlin Heidelberg, 2009.

R. Larson. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research*, 1(1):67 – 95, 1974.

R. Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23(5):845–868, September-October 1975.

R. Larson and A. Odoni. *Urban Operations Research*. Prentice-Hall, Englewood Cliffs, N.J., 1981.

J. A. Laval and L. Leclercq. Microscopic modeling of the relaxation phenomenon using a macroscopic lane-changing model. *Transportation Research Part B: Methodological*, 42(6): 511 – 522, 2008.

L. Leclercq, J. A. Laval, and E. Chevallier. The lagrangian coordinates and what it means for first order traffic flow models. *Proceedings of the 17 International Symposium on Transportation and Traffic Theory*, pages 735–753, 2007.

M.-T. Li, F. Zhao, L.-F. Chow, H. Zhang, and S.-C. Li. Simulation model for estimating bus dwell time by simultaneously considering numbers of disembarking and boarding passengers. *Transportation Research Record:*, 1971:59–65, 2006.

J.-R. Lin and T.-H. Yang. Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E: Logistics and Transportation Review*, 47(2):284 – 294, 2011.

H. S. Mahmassani and S. Peeta. Network performance under system optimal and user equilibrium dynamic assignments: implications for advanced traveler information systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1408:83–93, 1993.

V. Marianov and C. ReVelle. The capacitated standard response fire protection siting problem: deterministic and probabilistic models. *Annals of Operations Research*, 40(1):303–322, 1992.

A. Mazloumian, N. Geroliminis, and D. Helbing. The spatial variability of vehicle densities as determinant of urban network capacity. *Philosophical Transactions of the Royal Society A: Mathematical,Physical and Engineering Sciences*, 368(1928):4627–4647, 2010.

N. Mladenović and P. Hansen. Variable neighborhood search. *Computers & Operations Research*, 24(11):1097–1100, 1997.

R. Nair and E. Miller. Fleet management for vehicle sharing operations. *Transportation Science*, 45(4):524–540, 2011.

P. Olszewski, H. S. L. Fan, and Y.-W. Tan. Area-wide traffic speed-flow model for the singapore cbd. *Transportation Research Part A: Policy and Practice*, 29(4):273 – 281, 1995.

R. Ott. The Zurich experience. In *Alternatives to Congestion Charging*. Transport Policy Committee, 2002.

S. H. Owen and M. S. Daskin. Strategic facility location: A review. *European Journal of Operational Research*, 111(3):423 – 447, 1998.

A. Prékopa. Probabilistic programming. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 267–351. Elsevier, 2003.

A. E. Radwan and D. A. Benevelli. Bus priority strategy: justification and environmental aspects. *Journal of Transportation Engineering*, 109:88–106, 1983.

T. Raviv, M. Tzur, and I. A. Forma. Static repositioning in a bike-sharing system: models and solution approaches. *EURO Journal on Transportation and Logistics*, 2(3):187–229, 2013.

C. ReVelle and K. Hogan. The maximum availability location problem. *Transportation Science*, 23(3):192–200, August 1989.

M. Saberi and H. Mahmassani. Exploring properties of network-wide flow-density relations in a freeway network. In *91st Annual Meeting of Transportation Research Board*, Washington D.C., 2012.

D. Schilling, D. J. Elzinga, J. Cohon, R. Church, and C. ReVelle. The team/fleet models for simultaneous facility and equipment siting. *Transportation Science*, 13(2):163–175, 1979.

V. Schmid and K. F. Doerner. Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, 207(3):1293 – 1303, 2010.

S. A. Shaheen and A. P. Cohen. Carsharing and personal vehicle services: Worldwide market developments and emerging trends. *International Journal of Sustainable Transportation*, 7 (1):5–34, 2013.

R. J. Smeed. The road capacity of city centers. *Highway Research Record*, 169:22–29, 1967.

G. A. Sparks and A. May. A mathematical model for evaluating priority lane operations on freeways. *Highway Research Record*, 363:27–42, 1971.

R. A. Takeda, J. A. Widmer, and R. Morabito. Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research*, 34(3):727–741, 2007. Logistics of Health Care Management Part Special Issue: Logistics of Health Care Management.

C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

UN. World urbanization prospects, the 2011 revision. Data file, 2011. URL http://esa.un.org/unup/CD-ROM/Urban-Rural-Population.htm. United Nations, Department of Economic and Social Affairs, Population Division, Population Estimates and Projections Section.

V. R. Vuchic. *Urban public transportation: Systems and technology*. Prentice-Hall, 1981.

J. G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II*, 1(2):325–378, 1952.

J. G. Wardrop. Journey speed and flow in central urban areas. *Traffic Engineering and Control*, 9:528–532, 1968.

J. R. Weaver and R. L. Church. A median location model with nonclosest facility service. *Transportation Science*, 19(1):58–74, 1985.

A. Weber. *Über den standort der industrien*. JCB Mohr, 1909.

E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J*, 43(2):355–386, 1937.

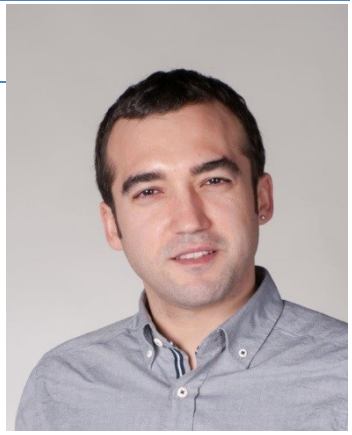
J. C. Williams, H. S. Mahmassani, and R. Herman. Urban traffic network flow models. *Transportation Research Record: Journal of the Transportation Research Board*, 1112:78–88, 1987.

S. C. Wirasinghe, V. F. Hurdle, and G. F. Newell. Optimal parameters for a coordinated rail and bus transit system. *Transportation Science*, 11:359–374, 1977.

K. Zografos and H. Levinson. Passenger service times for a no-fare bus system. *Transportation Research Record:*, 1051:42–48, 1986.

BURAK BOYACI

e-mail: burakboyaci@gmail.com



EDUCATION

01/2010 – 05/2014	Ecole Polytechnique Fédérale de Lausanne (EPFL), PhD in Civil and Environmental Engineering (ENAC) Research: Modeling and Optimization of Demand Responsive Systems and Urban Congestion Supervisor: Nikolas Geroliminis
09/2006 – 06/2009	Boğaziçi University, Master of Science in Industrial Engineering Research: Solving the capacitated multifacility Weber problem approximately Supervisor: İ. Kuban Altınel, 3rd among 22
09/2001 – 06/2006	Boğaziçi University, Bachelor of Science in Industrial Engineering Project: Distributed simulation for real time control of automated manufacturing systems Supervisor: Ümit Bilge, 9th among 58
09/1998 – 06/2001	İzmir Science High School

EXPERIENCE

01/2014 – Present	Lancaster University, Development Lecturer in Management Science Department
01/2010 – 03/2014	Ecole Polytechnique Fédérale de Lausanne, Graduate Student Researcher in Urban Transport Systems Lab
09/2006 – 01/2010	Boğaziçi University, Research Assistant in Manufacturing Management Systems Lab
01/2006 – 06/2006	Boğaziçi University, Student Assistant in Flexible Automation and Intelligent Manufacturing Systems Lab

RESEARCH INTEREST

- Linear (Continuous, Binary, Integer) and nonlinear Programming, column generation applications
- Large scale mathematical programming and multi-objective optimization
- Operations research and management, logistics, location and supply chain models
- Markov chains, queueing theory and spatial queueing systems, discrete event simulation
- Planning and scheduling, public transportation and car-sharing systems, air and freight transportation

HONORS AND FELLOWSHIPS

- 2nd rank at mathematics projects in TÜBİTAK (The Scientific and Technological Research Council of Turkey) National Science Projects Olympics, 2001
- 5th rank at mathematics (among 136669) in The National Graduate Education Entrance Exam (LES), 2005
- 40th rank (among over 400000) in Science High School Entrance Exam (LGS), 2001
- 91st rank (among over 400000) in Middle School Entrance Exam (ALGS) , 1998
- Boğaziçi University Outstanding Achievement Scholarship for Undergraduate Students (2001-2006)
- Scientific and Technological Research Council of Turkey (TÜBİTAK) Merit-Based Fellowship for Graduate Students (2006-2008)

TEACHING

ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SPRING 2012	Fundamentals of Traffic Operations and Control (lab section instructor)
AUTUMN 2012	Route choice modeling for taxi data in China (semester project supervisor)
SPRING 2012	Analysis and control of the bus bunching problem in transportation networks (semester project supervisor)
SPRING 2012	Analysis of taxi's O/D tables for a large metropolitan area (semester project supervisor)
AUTUMN 2011	Optimizing Train Design in Capacitated Railroads (semester project supervisor)
SPRING 2011	Analysis, evaluation and recommendations for the Nice-Monaco bus system (semester project supervisor)

BOĞAZİÇİ UNIVERSITY

AUTUMN 2006-9	Planning for Engineers (problem session instructor, quiz and project grader)
AUTUMN 2006-7	Supply Chain Management (problem session instructor)
SPRING 2009	Operations Research I (Linear Models) (problem session instructor, assignment grader)
SPRING 2008	Graph Algorithms and Applications (problem session instructor, assignment grader)
AUTUMN 2009	Optimization Models in Economics and Finance (problem session instructor, assignment grader)
SPRING 2006	Computer Integrated Manufacturing Systems (student assistant)

PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

J1. Boyacı B., Altınel İ.K., Aras N., 2013. Approximate solution methods for the capacitated multi-facility Weber problem. *IIE Transactions,* **45**, 97-120 (Special Issue: Scheduling & Logistics)

J2. Geroliminis N., Boyacı B., 2012. The effect of variability of urban systems characteristics in the network capacity, *Transportation Research Part B,* **46**, 1607-1623

ARTICLES UNDER REVIEW OR IN PREPARATION

R1. Boyacı B., Zografos K., Geroliminis N., under review. An optimization framework for the development of efficient one-way car-sharing systems. (submitted to European Journal of Operational Research)

R2. Boyacı B., Geroliminis N., in preparation. Approximation methods for large scale spatial queueing systems. (expected submission: May 2014)

R3. Geroliminis N., Boyacı B., in preparation. An analytical approximation of network capacity for multimodal urban systems. (expected submission: June 2014)

CONFERENCE PROCEEDINGS

C1. Boyacı B., Geroliminis N., 2013. Extended Hypercube Models for Location Problems with Stochastic Demand. In *hEART 2013 - 2nd Symposium of the European Association for Research in Transportation, Stockholm.*

C2. Boyacı B., Zografos K., Geroliminis N., 2013. Developing and Solving an Integrated Multi-objective Model for Supporting Strategic and Tactical Decisions for One-Way Car Sharing Systems. In *26th European Conference on Operational Research (EURO)*, Rome.

C3. Boyacı B., Zografos K., Geroliminis N., 2013. A generic one-way multi-objective car-sharing problem with dynamic relocation, In *Tristan VII*, San Pedro de Atacama.

C4. Boyacı, B., Geroliminis N., 2012. Extended hypercube queueing model for stochastic facility location problems. In *25th European Conference on Operational Research*, Vilnius.

C5. Boyacı B., Prem Kumar V., Binder S., Bierlaire M., 2012. Cost optimization for the capacitated railroad blocking and train design problem. In *Odysseus 5th International Workshop on Transportation and Logistics*, Mykonos.

C6. Boyacı B., Geroliminis N., 2012. Extended hypercube models for large scale spatial queueing systems. In *91th Annual Meeting of Transportation Research Board*, Washington DC.

C7. Boyacı B., Geroliminis N., 2011. Estimation of the network capacity for multimodal urban systems, *Procedia Social and Behavioral Sciences*, **16**, 803-813 (In *6th International Symposium on Highway Capacity and Quality Service*, Stockholm)

C8. Boyacı B., Geroliminis N., 2011. Exploring the effect of variability of urban systems characteristics in the network capacity. In *90th Annual Meeting of the Transportation Research Board*, Washington DC.

C9. Boyacı B., Altınel İ.K., Aras N., 2009. Computing bounds for the capacitated multifacility Weber problem by branch and price and Lagrangean relaxation, In *23rd European Conference on Operational Research (EURO)*, Bonn.

RESEARCH PROJECTS

06/2012 – 01/2014 Development and evaluation of a one-way car-sharing scheme for Veolia Auto Bleue System, Veolia Transport, Nice, France. A project to evaluate and implement a one-way car-sharing system's both operation and strategic problems and developing a simulation. Supervisor: Prof Geroliminis.

01/2011 – 01/2012 Movement conflicts in multi-modal urban traffic systems: Modeling congestion and developing more sustainable cities, Swiss National Science Foundation (SNSF) Basic Research. The utilization

of macroscopic fundamental diagram (MFD) for multi-modal networks is investigated. Supervisor: Prof Geroliminis.

01/2010 – 01/2011 Real-time hierarchical control and monitoring of urban traffic systems, Swiss National Science Foundation (SNSF) Basic Research. A closed mathematical equation is derived for homogenous networks and a discrete event simulation is implemented for the networks with stochastic parameters. Supervisor: Prof Geroliminis.

01/2009 – 07/2009 Line balancing and simulation in multi-product assembly line, OYAK Renault, Bursa, Turkey. Software is developed to balance and simulate multi-product assembly line in the factory of OYAK Renault to increase efficiency. Supervisors: Prof Bilgiç and Prof Ünal.

02/2008 – 08/2008 Large scale aircrew rostering problem by column generation approach, Hitit Computer Services, İstanbul, Turkey. Software for constructing and solving large scale airline crew scheduling problem with column generation is developed. Supervisors: Prof Aras and Prof Bilgiç. This project has received YA/EM Practice Award in the 32nd National Congress on Operational Research and Industrial Engineering in 2011.

01/2007 – 07/2007 Integrated planning and control for managing the fast and flexible manufacturing enterprise, SAP, Germany. A framework to provide a detailed architecture for the integrated planning and control of a fast and flexible manufacturing enterprise is developed. Supervisors: Prof Bilge, Prof Bilgiç, Prof Güllü and Prof Ünal.

REFEREE WORK

- European Journal of Operational Research
- Mathematical Methods of Operational Research
- EURO Journal on Transportation and Logistics
- Transportation Research Part B: Methodological
- ODYSSEUS 2012
- hEART 2012, 2013
- SIMULTECH 2012

PROGRAMMING SKILLS

C, C++, C# .Net, Java, ILOG Cplex (Concert), GAMS, Arena, MATLAB, MS Applications

LANGUAGES

Turkish (native), English (fluent), French (intermediate)

INTEREST

SCUBA and free diving, sailing, skiing

NON-PROFESSIONAL ACTIVITIES

2011 – 2013	Member of EPFL Sailing Team (Club de Voile)
2010 – 2012	Treasurer of TURQUIA 1912 – Association of Turkish Students in Switzerland
2006 – 2009	Member of Boğaziçi University Sailing Team (BÜSAIL)
2004 – Present	One star diving instructor of Confédération Mondiale des Activités Subaquatiques (CMAS)
2003 – 2006	Management Board Member and Instructor in Boğaziçi University Underwater Sports Club (BÜSAS)
2001 – 2009	Member of Boğaziçi University Underwater Sports Club (BÜSAS)

Last Updated: April 30, 2014