SMART TRANSIT DYNAMIC OPTIMIZATION AND INFORMATICS

By

Hamid Reza Sabarshad

Master of Science, Mazandaran University of Science and Technology, Iran, Sep 2006

Bachelor of Science of Engineering, Azad University, Iran, Sep 2004

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Program of

Civil Engineering

Toronto, Ontario, Canada, 2015

©Hamid Reza Sabarshad 2015

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Dissertation Title: Smart Transit Dynamic Optimization and Informatics

Degree: Doctor of Philosophy

Year: 2015

Student Name: Hamid R. Sabarshad

Program: Civil Engineering

University: Ryerson University

Abstract

With the popularity of Big Data and urban informatics, there is increasing interest in ways to use real time data to improve transportation system operations. In many real-world applications, demand is revealed dynamically over time, and consequently the routes are determined dynamically as well. In this thesis, contributions are made to several key components of a “smart” transit system framework where dynamic operations are driven by real time information.

The first component is in dynamic routing and pricing of a fleet of vehicles. A new dynamic dial-a-ride policy is introduced that features non-myopic pricing based on optimal tolling of queues to fit with the multi-server queueing approximation method. By including social optimal pricing, the social welfare of the resulting system outperforms a pricing policy based on the marginal cost increase of a passenger over a range of test instances. In the examples tested, improvements in social welfare of the non-myopic pricing over the myopic pricing were in the 20% - 31% range.

The second component is in the informatics. Effective dynamic optimization of a system (routing, scheduling, fare setting, etc.) requires effective short term prediction of traveler/customer arrival using real-time data. Several recent methods for arrival process prediction, both offline and online, are investigated using real taxi data from New York. An experiment is conducted using the same data set to draw comparisons for arrival process modeling, suggesting that the temporal seasonal factors method from Ihlers et al. (2006) is more effective as an offline approach and the IntGARCH method from Matteson et al. (2011) is more effective as an online approach.

The third component investigated is in the prepositioning of idle vehicles. Vehicles that are positioned at locations that take into account future demand can lead to reduced wait times for passengers and improved level of service. A dynamic relocation model is proposed that includes queueing delay to approximate the congestion effect of future demand. A linear problem is formulated based on Marianov and Serra's (2002) work. By varying customer arrivals, the approach provides a new managerial tool to find the optimal service level.

ACKNOWLEDGMENTS

“Let the beauty we love be what we do. There are hundreds of ways to kneel and kiss the ground.” (Rumi)

I would never have been able to finish my dissertation without the guidance of my supervisor and support from my family.

First of all, I am heartily grateful to my supervisor, Dr. Joseph Chow. His extraordinary vision, insightful guidance and high scientific standards safeguarded success of this dissertation. He is steady influence behind each step I ever made in this research. Particularly, I appreciate his consistent trust, kindness and patience.

I especially acknowledge the support given by the Canada Research Chairs program and the NSERC Discovery Grant. Without their valued support I could not have completed this thesis on schedule.

I would like to thank from my committee members including, Dr. Bhagwant Persaud, Dr. Arnold Yuan, Dr. Sharareh Taghipour, and Dr. Liping Fu for giving their important ideas and reviews.

My sincere thanks go to Dr. Kaan Ozbay from New York University for sharing NY taxicab data.

I am deeply and forever indebted to my parents for their love, support and encouragement throughout my entire life. I am also very grateful to my brothers.

Finally, I would like to thank Rachel Harpley from Department of Civil Engineering for all helps.

TABLE OF CONTENTS

Chapter Title	Page
Title Page	i
Author's Declaration.....	ii
Abstract	iii
Acknowledgements.....	v
Table of Contents.....	vi
List of Tables	x
List of Figures	xii
Chapter 1. Introduction and motivations	1
1.1. Problem definition.....	1
1.2. Research objective.....	2
1.3. Dynamic routing and pricing.....	6
1.4. Arrival process prediction	7
1.5. Prepositioning of idle vehicles	8
1.6. Outline of the Thesis	10
Chapter 2. Literature review	11
2.1. Traveling salesman problem (TSP).....	11
2.2. Vehicle routing problem.....	13

2.3. Dynamic vehicle routing problem.....	15
2.3.1. Hyytiä et al.'s model.....	23
2.3.2. Non-myopic pricing in queues	26
2.4. Prediction models: Online-offline information	27
2.4.1. The arrival process.....	28
2.4.2. Offline models	29
2.4.3. Online models.....	31
2.5. Relocation of idle vehicles	32
Chapter 3. Routing and pricing.....	37
3.1. Introduction	37
3.2. Motivational example for non-myopic pricing	39
3.3. Proposed methodology	42
3.4. Experimental study.....	50
3.4.1. Illustrative example	51
3.4.2. Social welfare comparison and algorithm evaluation.....	55
3.4.3. Sensitivity evaluation	58
3.4.4. Elasticity evaluation	58
3.5. Discussion	62
Chapter 4. Arrival process prediction	64
4.1. Introduction	64

4.2. New York City Taxi Data	65
4.3. Model specifications	73
4.3.1. Offline prediction models	73
4.3.1.1. Piecewise linear function (Nicol and Leemis, 2014)	73
4.3.1.2. Temporal seasonality factors (Ihler et al, 2006; Moreira-Matias et al, 2013)	75
4.3.2. Online prediction models	76
4.3.2.1. The ARIMA model	76
4.3.2.2. IntGARCH model (Matteson et al, 2011)	77
4.3.2.3. Non-parametric Bayesian approach (Adams et al. 2009)	79
4.4. Estimation and validation results	80
4.4.1. Estimation results	81
4.4.1.1. Piecewise linear function	81
4.4.1.2. Temporal seasonality factors	82
4.4.1.3. The ARIMA model	82
4.4.1.4. IntGARCH model	84
4.4.1.5. Non-parametric Bayesian approach	84
4.4.1.6. Comparison of goodness-of-fit	85
4.4.2. Validation of results	88
4.5. Discussion	94
Chapter 5. Non-myopic prepositioning of idle vehicles	95

5.1. Non-myopic prepositioning of idle vehicles problem.....	96
5.2. Motivational example for queueing constraint.....	101
5.3. Computational experiments	103
5.4. Discussion	108
Chapter 6.Conclusions and future research	110
6.1. Future research	112

LIST OF TABLES

Table 2.1. Summary of non-myopic dynamic DARP studies.....	23
Table 2.2. Facility location problem with queueing	36
Table 3.1. Input arrival times in simulation of dynamic DARP policies.....	53
Table 3.2. Output values under non-myopic dynamic DARPP	54
Table 3.3. Output times under non-myopic pricing	54
Table 3.4. Output times under non-myopic marginal pricing.....	55
Table 3.5. Summary of system cost, user cost, and social welfare averaged over 30 runs	57
Table 3.6. Summary of sensitivity tests of the policies to Λ and θ	59
Table 3.7. Comparison of elasticity of social welfare to vehicle capacity	62
Table 4.1. The number of driver and vehicle licenses (TLC Annual Report 2014)	67
Table 4.2. Temporal seasonality parameters.....	83
Table 4.3. Goodness of fitting metrics.....	87
Table 4.4. Validation metrics.....	91
Table 4.5. The ECDF metric and notes on approaches.....	93
Table 5.1. Relocation costs from zone i to zone j	101
Table 5.2. Travel times from zone i to zone j	102
Table 5.3. Customer arrivals and service rates for each zone.....	104
Table 5.4. The value ρ_{ajm} with different confidence levels	106

Table 5.5. The cost of relocation from zone i to zone j	107
Table 5.6. Travel time from zone i to zone j	107
Table 5.7. Summary of non-myopic prepositioning of idle vehicle with $\alpha = 0.95$	108
Table 5.8. Summary of non-myopic prepositioning of idle vehicle with $\alpha = 0.90$	108
Table 5.9. Summary of non-myopic prepositioning of idle vehicle with $\alpha = 0.85$	109

LIST OF FIGURES

Fig1.1. Smart transit and informatics framework.	4
Fig2.1. Static vehicle routing problem.	13
Fig2.2. A dynamic vehicle routing problem with two vehicles.	17
Fig 2.3. Illustration of the non-myopic dynamic DARP as a multi-server queue system.	24
Fig 3.1. Illustration of the inefficiency present assuming inelastic demand and using myopic pricing.	39
Fig3.2. Motivational example for non-myopic pricing (adapted from Figliozzi et al., 2007).	40
Fig 3.3. Simulated tours under proposed non-myopic pricing policy.....	52
Fig 3.4. Simulated tours under the marginal pricing policy from Hyytiä et al. (2012).	53
Fig 3.5. Comparison of the social welfare for each of 30 runs over the four policies under TSPPD operating with (a) Christofides' algorithm vs (b) CPLEX.....	56
Fig 3.6. Comparison of elasticity of social welfare to fleet size.....	60
Fig 3.7. Comparison of elasticity of social welfare to price.	61
Fig 4.1. (a) Pick-up and (b) drop-off of passengers in New York in March 2013 (via QGIS).....	66
Fig 4.2. Format of data.....	67
Fig 4.3. A comparison of the number of trips and passengers in NY and Manhattan.	69
Fig 4.4. (a) Pick-up and (b) drop-off of passengers at the Lincoln Square in March, 2013.	69

Fig4.5 Passenger arrival times per second.	69
Fig 4.6. Original data: The number of arrivals.	70
Fig 4.7. The sample of passenger arrivals-Lincoln Square (Zone Id: 3806)	71
Fig 4.8. Observed average arrival rates at Lincoln Square for different time periods on October 7, 2013.....	72
Fig4.9. Bayesian approach parameters for the NY taxicab data.	85
Fig 4.10. Fitting the off-line approaches for the Lincoln Square (Zone Id: 3806)	89
Fig 4.11. Fitting the on-line approaches for the Lincoln Square (Zone Id: 3806).....	90
Fig 4.12. Predictive mean of Poisson intensity $\lambda(t)$	94
Figure 5.1. The resulting nodes that are assigned to each server with the two methods.	103
Figure5.2.The (X,Y) coordinates of each node.	104

Chapter 1

Introduction and motivations

1.1. Problem definition

With the compounding effects of climate change, population growth, and urbanization, travel in societies is becoming increasingly congested, and the effects of that congestion are increasingly worse. In particular, the rise of mega-regions due to urbanization poses a new set of problems for travelers. Many neighborhoods can benefit from public transit access to other parts of the mega-region, but the cost of maintaining a consistently connected system increases disproportionately with the coverage area – what transportation planners call the “last mile problem” (Chandra and Quadrifoglio, 2013).

To illustrate, Toronto public transportation agencies typically provide bus, taxi, and rail services that may frame the core of such trips, but customers must complete the first and last portion on their own. They must first walk, take a taxi, drive, or take a bus to the nearest subway station. For a transit agency, the first-last mile problem is challenging because it demands high flexibility with under relatively low density, which implies a high cost to implement. Complete solutions for the first-last mile problem, from a public transit agency’s perspective, need to provide cost-efficient routing of customers from their homes to a subway hub. The communication between the drivers of the vehicles, customers and the dispatching center is essential in order to feed the most up-to-date information into the routing system, as those fleets would require “*smart*” dispatch and pricing algorithms that make use of abundant real-time data.

Consider the challenges and opportunities in the City of Toronto as an example. There are 4,849 licensed taxicabs in Toronto and it is estimated that more than 60,000 trips are made per day. City inspection statistics show the average Toronto taxicab operates more than 200km every day, and much of that time is spent in downtown Toronto. There is broad consensus that the current issuing formula for fare pricing is wrong and there have been suggestions that the city should hire an independent consultant to determine the formula for managing the number of taxis and demand-responsive system (Toronto's taxicab report, 2012). These issues suggest a need for effective dynamic optimization and informatics, which is the science of information processing.

A key to effective dynamic optimization and informatics (e.g. routing, scheduling, fare setting, arrival prediction, and prepositioning of idle vehicles using real time information) is the use of the real time traveler/customer arrival data to infer short term future conditions. Real time data may be used by a transportation network company or taxi fleet to set dynamic prices or dispatch vehicles efficiently; for a bus operator to manage their vehicle arrivals dynamically to reduce bus bunching; for a car sharing or bike sharing service (Sayarshad et al. 2012) to effectively rebalance their vehicles; among others. We propose addressing this need using a “smart” transit system framework.

1.2. Research objective

A smart transit system framework takes real time data to optimize various functions to serve passengers, which can increase the effective capacity, accessibility, manage demand, and maximize efficiency and benefit. For example, combining advanced technologies like mobile

devices with optimization models like the dynamic vehicle routing problem (DVRP) can improve the ability to make routing decisions for a fleet of vehicles in real time. The objective of this research is to propose a smart transit system framework under which several components are improved upon.

The framework is shown in **Figure 1.1**, which lists the various technologies and data sources needed to ensure a viable “smart” transit system. For example, communications equipment between the vehicle and the dispatching center is needed. Customers can send travel requests to dispatch centers via the communication equipment. Mobile device communication systems are one example of a technology capable of providing this information.

Some of the terminology in **Fig. 1.1** are explained as follows.

EDI (Electronic Data Interchange): It is the electronic transfer from computer to computer to increase speed of communication and control all aspects using message data. The data format is the same for all modes (rail, road, air, and marine). In addition, it can improve the speed of communication in a dispatch system and will enhance the quality, availability and value of information to manage demand, drivers, vehicles, etc.

GPS (Global Positioning Systems): This is the system for transferring data between a vehicle and the dispatch office using advanced satellite communications and cellular technology. In particular, the position of vehicles and customers becomes much easier to monitor, allowing for better management of the vehicle fleet. Our information includes speed, route planning, the position of passengers and vehicles, and drivers’ work schedules. Telephone, email, fax, automated communications can be used.

Fig1.1. Smart transit and informatics framework.

GIS (Geographical Information Systems): It is a generalized electronic road map with a high resolution graphic screen. This system provides the static (e.g. maps and transport networks) and dynamic (e.g. vehicle positions) geographical data.

IVHS (Intelligent Vehicle-Highways Systems): This system is an application of information and communications technologies (ICT) in order to better control the flow of vehicles. It attempts to create algorithms and advanced technologies to control traffic congestion and hazardous materials.

Traffic Information Center (TIC): This function gets the traffic data from a variety of data sources to share with the operation center. The processed information is then distributed to different applications. The center also updates the traffic maps as well. To summarize, the center provides an opportunity for operation centers to obtain and update transit network information.

Dispatching Center: There are several important tasks in the dispatch centers: (1) the ability to determine cluster of users to be served, (2) the ability to consider scheduling pickup and drop-off for each route, and (3) the ability to sequence the users into a route. In chapter 3, we will discuss on this component and our new approach.

Predictions-Customer Arrival: This informatics function predicts arrival processes using historical user arrival data. In chapter 4, we will be discussing this component and an experiment using New York City (NYC) taxi data to evaluate different approaches in more detail.

Positioning: This function addresses the issue of the location of vehicles so that all passengers can be served within a standard user and system costs. In chapter 5, we will discuss this component and our new proposed dynamic optimization model.

The smart system combines software and hardware technologies so that transit agencies can be made more efficient not only through *automated customer responses*, but the systems also focus on *control* and *delivery* of a large quantity of data. Some studies have introduced an overview of models, communication equipment, and computational methods applied in this context. Fu and Tepley (1999), Fu (2002a) and Fu and Xu (2002) introduced an overview on computational models in the automated dial-a-ride paratransit operation. Dial (1995) proposed an automated version of dial-a-ride transit system, which employs fully-modernized order-entry and routing-and-scheduling system. The usage of real-time optimization models, communication devices and computational equipment enable transit systems to be smarter and more reliable.

The research objectives specific to the three components investigated in this thesis are now presented.

1.3. Dynamic routing and pricing

Dispatch decisions for a fleet of vehicles serving passenger demand in real time can be determined by solving an optimization model called a “dial-a-ride problem” (DARP). The DARP, a subclass of the pickup and delivery problem, is defined as the problem of picking up passengers and delivering them to their destinations under tight time restrictions in a most efficient manner. There are two important decisions in the DARP solutions for single or multiple vehicles. (1) to determine the cluster of users to be served by each vehicle, and (2) to schedule pickup and drop-off for each route. Some of the earliest forms of DARP have been around for decades (Wilson et al., 1976; Jaw et al., 1986; Fu, 2002c; Cordeau and Laporte, 2007). However, a third decision, pricing, is typically ignored in the DARP literature. Dynamic dial-a-ride

problem and other related dynamic vehicle routing problems often ignore the need for non-myopic pricing under the assumption of elastic demand, which leads to an overestimation of the benefits in level of service and resulting inefficiencies.

If the rate of customer arrivals is more than service times, the system will become congested and the static dial-a-ride and pricing problem are bound to provide meaningless results. Thus, dynamic dial-a-ride and pricing problem and queue theory can provide better performance. To correct this problem, in this thesis a new dynamic dial-a-ride policy is introduced, one that features non-myopic pricing based on optimal tolling of queues to fit with the multi-server queueing approximation method for large-scale systems. The dial-a-ride and pricing problem has the potential to manage demand-responses by differentiating prices over users by accounting for their future effects on costs.

1.4. Arrival process prediction

With the availability of customer arrival data via ICTs, a smart transit system can make predictions about the arrivals more effectively. For example, customers can send travel requests to operational centers via the communication equipment provided in order to obtain real time information.

While the literature suggests there have been a number of recent developments in prediction models of arrival processes, no study has reviewed them and compared their performances using a common data set from a transportation system. This thesis serves this purpose by providing a broad review of state of the art methods in two classes of prediction models—online and offline models—and applying them to the same set of taxi demand data so

that comparisons and insights can be gained. An online model, unlike an offline model, is based on real time data inputs. The contribution is to provide empirical insights on which of the latest five state-of-the-art prediction and machine learning algorithms are most appropriate to predicting the arrival patterns of taxi demand as representative transport systems. As a result, we expect this study to serve as a definitive research reference for future studies in “urban informatics”—the use of information and communications technology to better understand cities’ needs—and design of new types of urban mobility systems like carsharing, ridesharing, demand-responsive transit, smart taxis, and autonomous vehicle fleets, all of which require user demand arrival process prediction.

1.5. Prepositioning of idle vehicles

The third component deals with relocation of idle vehicles. Records show that during 2011, 13,586 taxicabs in New York carried over 172 million trips with passengers. From these 172 million trips over 150 million taxi trips originated or ended in Manhattan (Santi, 2013). Thus, a significant portion of the time and fuel is spent by taxis on non-revenue generating activities without passengers, leading to inefficiencies (Li, 2006). When dealing with the design of a transit service, the location of vehicles and customers have a strong influence of the amount of density of demands at each zone and on the quality of service. The vehicles must be reached to customers within a reasonably waiting time. Thus, waiting time is an important determinant of user costs. To decrease the system costs, we must be able to use idle vehicles to serve customers more efficiently. So, a key to improve the system and user cost is to preposition idle vehicles to improve the quality of service within a reasonable waiting time.

The facility location problem is an optimization model used to determine locations of facilities or servers to optimize coverage of demand in a network. While there are a number of studies that have examined dynamic relocation of servers, and numerous applications to taxi prepositioning, there are no studies of non-myopic relocation of idle taxis that look ahead. Location problems have incorporated queueing for services that have physical queue delays (e.g. electric vehicle charging stations that can take 30 minutes or more to recharge (Jung et al., 2014)) or probabilistic availability of a server that may be busy. In this thesis, we propose a dynamic relocation model where queueing delay is used to approximate the look-ahead cost of taxi service.

1.6. Outline of the Thesis

Chapter 2 presents the existing literature dealing with the dynamic vehicle routing and pricing, arrival process prediction models, and prepositioning problem in order to provide an overview of these problems and the role they play in the field of smart transit informatics.

In Chapter 3, a non-myopic dynamic dial a ride and pricing problem is proposed.

Chapter 4 of this thesis provides a broad review of state of the art methods in two classes of prediction models—online and offline models—and presents an original empirical study conducted with NYC taxi data.

In Chapter 5, we propose a new model formulation repositioning idle vehicles to anticipate new requests.

Finally, in Chapter 6 we give our conclusions in a summary of the scientific contributions of this thesis.

Chapter 2

Literature review

In this chapter the existing literature covering the methods associated with smart transit systems is reviewed. This chapter is organized as follows. First, we provide a brief discussion on the traveling salesman problem and solutions in Section 2.1. Having built this foundation, we then give an overview of some of the most important works on the static vehicle routing problem in Section 2.2. In Section 2.3, we introduce the dynamic vehicle routing problem. In Section 2.4, we consider prediction models to predict new information for customer arrival process. Lastly, we give an overview of some of the most important works on location problem in Section 2.5.

2.1. Traveling salesman problem (TSP)

The traveling salesman problem is a classic optimization model that seeks a minimum length tour that visits each node exactly once given a list of nodes and distances between each pair of nodes. This problem is notable for its integer decision variables, which makes it a computationally complex combinatorial problem. In particular, the optimality of some solutions can be evaluated in time that scales as a polynomial of the number of variables in the problem. TSP, however, belongs to the set of problems whose solutions cannot be evaluated for optimality in polynomial time (Sayarshad and Ghoseiri 2009; Sayarshad et al. 2010, Javadian and Sayarshad 2011), which we call NP-hard problems.

There are methods to solve the TSP exactly. For example, the problem can be formulated as a binary programming problem, and then binary programming exact methods like branch and bound can be applied to solve the model. However, the computation time of an exact method cannot be predicted for realistic size instances, and so heuristic methods are needed.

Tour improvement algorithms are one class of heuristics. These algorithms begin with a feasible solution, and then changes are iteratively made. The 2-opt exchange is a simple local search algorithm first proposed by Croes (1958) to find local solutions in the TSP.

Christofides (1976) proposed a heuristic for the TSP that involves the following steps.

Step1. Create a minimal spanning tree (MST) or subgraph from the set of all passengers to connect all nodes together.

Step2. Find the minimum-cost matching on the set of nodes having an odd degree. Then, we add the MST together with the edges.

Step3. Build an Euler cycle from the combined graph, and traverse it taking shortcuts to avoid visited nodes.

The Christofides' algorithm is proven to have a worst case bound of $3/2$ of the optimal value, which means that any solution obtained using the heuristic is guaranteed to never be worse off than 150% of the true optimal value, and in practice they often perform much better.

An overview and classification of exact and approximation algorithms are discussed by Laporte (1992).

2.2. Vehicle routing problem

Since the vehicle routing problem (VRP) was first introduced by Dantzig and Ramser (1959), many different variations have been studied (see Golden and Assad 1988, Fisher 1995). The VRP is one of the most popular problems in combinatorial optimization. The problem is to determine m vehicle routes, where a route is a tour that begins at the origin node (depot) and traverses a subset of customers and the total size of deliveries for customers assigned to each vehicle must not exceed the vehicle capacity. The routes should be chosen to minimize total travel cost under such constraints as 1) each vehicle route begins at the depot and ends at the depot, 2) the total demand of each route cannot exceed the vehicle capacity, subtour elimination constraints, and time windows constraints (Balinski and Quandt; 1964; Fisher and Jaikumar 1981; Fisher et al 1997).

There are two types of approaches for the vehicle routing problem: static and dynamic models. In the static case, a complete route is output. This problem may be deterministic or stochastic, and many algorithms have been developed to deal with it. In the deterministic case, all parameters (travel times or demands) are known. But, in the stochastic case, the complete route serves as an *a priori* plan for re-optimization. Thus, these models must operate under strict temporal restrictions. **Figure 2.1** illustrates the static vehicle routing problem.

Fig2.1. Static vehicle routing problem.

The dial-a-ride problem is a subclass of the VRP, defined as the problem of picking up passengers and delivering them to their destinations under tight time restrictions in a most efficient manner. The dial-a-ride problem was first examined by Wilson et al., (1971), Wilson and Weissberg (1976), and Wilson and Colvin (1976).

Stochastic problems arise whenever some elements of the problems are random (e.g. stochastic demands and/or stochastic travel times), so sometimes the set of customers to be visited is not known with certainty or each customer has a probability of being present. Stewart and Golden (1983), Tillman (1969), Bertsimas et al (1995), Bertsimas (1992) and Allahviranloo et al (2014) consider the classical vehicle routing problem (VRP) in which demands are stochastic. In these models, each vehicle of limited capacity must meet demands at fixed locations, returning periodically to the depot to empty its current load. Gendreau et al. (1995) proposed a two stage stochastic programming for vehicle routing problem with stochastic customers, in which each customer has a deterministic demand and a probability of being present. Some models combine the vehicle routing problem with stochastic demand and stochastic customers (Gendreau et al., 1996a).

Stochastic vehicle routing problems are fundamentally different from deterministic problems. It is not even possible to determine in advance a feasible solution since we do not know what the actual demand will be. It is reasonable to optimize the *expected value* of the system subject to constraints representing the *expected value* of customer demands. We need to fit a distribution for random customer demands or travel times before we solve the model. But, an environment is changing more quickly than before and it requires a repeated re-optimization. Even with a stochastic formulation, as in Gendreau et al. (1996a, b), these static models do not

consider dynamic routing because the stochastic variables in those cases are not time-dependent. Thus, a smarter system that anticipates stochastic elements and dynamically adapts to them is necessary (Powell et al., 1995, Dror & Powell, 1993).

The vehicle routing problem is a NP-hard problem. It is not always possible to solve instances to optimality within reasonable time. Thus, heuristic approaches such as simulated annealing, genetic algorithm, and adaptive neighborhood search have been used to find a solution near to optimum points for the VRP. An overview and classification of exact and approximation algorithms are discussed by Laporte (2009).

Clarke and Wright (1964) proposed an algorithm for solving the VRP. This algorithm is based on a savings method. In short, we first assign a single vehicle to every customer, and then we iteratively merge vehicle routes together if there are savings to doing so.

2.3. Dynamic vehicle routing problem

In the *dynamic vehicle routing problem* (DVRP), new orders dynamically arrive when the vehicles have already started their tours, which consequently requires the vehicles to update their itineraries in order to serve the orders. Dynamic vehicle routing problems involve making decisions, then observing information, after which we make more decisions, and then more information, and so on. Psaraftis (1995) considered attributes of the information forming the input for the vehicle routing problem. The attributes are as following:

- **Evolution of information**

In static VRPs neither does the information change, nor is the information updated. In dynamic models inputs will generally change over time.

- **Quality of information**

Information can be known with certainty (deterministic), and it can be known with uncertainty (forecasts), or it follows a probability distribution (probabilistic). As mentioned in Psaraftis (1995), the quality of information is usually good for near-time events in dynamic vehicle routing models, and becomes poorer for static models.

- **Availability of information**

Information could be either global or local. One example of local information is when customer arrivals are random, and the actual value may be revealed only when customers arrive at the starting node. On the other hand, some inputs may be available globally. For example, customer arrivals may be received by mobile device. ICTs will generally increase the global availability of information.

- **Processing of information**

Processing of information could either be centralized or decentralized. For example, in a centralized system all customer arrivals at nodes are collected and processed by a central unit. On the other hand, in a decentralized system some of the customer arrival information could be processed by each zone.

Figure 2.2 shows a possible solution routing of two vehicles that have to dynamically adjust their tours due to a new request. Imagine that we are routing two vehicles (labeled A and B), and that at time t we get new information about customer orders. The vehicles may stop at any location and receive the benefits or continue to the next customer.

Fig2.2. A dynamic vehicle routing problem with two vehicles.

Below we give a brief summary of the differences between static and dynamic vehicle routing problem:

- **Time dimension**

In dynamic vehicle routing problems, time is always essential, but in static routing problems the time dimension may or may not be important.

- **The problem may be open-ended**

The routes start and end at the depot in a static vehicle routing problem, but dynamic vehicle routing may consider paths for vehicle to follow.

- **Myopic and non-myopic**

We assume all information to be known in a static vehicle routing problem, but in dynamic vehicle routing the future is almost never known with certainty. As a result, the solution of a dynamic model is a state-dependent policy (a rule for setting a decision variable dependent on the state of a system). Policies that do not anticipate future

requests are referred to as *myopic policies*, while policies that anticipate unknown future requests are called *non-myopic policies*.

- **Information update mechanisms**

Information update mechanisms are not essential in a static vehicle routing problem, but it is essential in the dynamic vehicle routing problem, because inputs are changing throughout time.

- **Time constraints**

In a real-time vehicle routing problem, time constraints such as latest pickup times tend to be softer than in a static vehicle routing problem. Thus, time constraints may be different.

For many years, such dynamic DARPs have been deemed economically unsustainable and require heavy subsidies. This is because the technologies to enable dynamic operations with seamless transfer of information between passengers, vehicles, and dispatch have been limited. As a result, hybrid solutions that combine flexible routes or schedules with fixed routes have been proposed (e.g. Hadas and Ceder, 2008; Quadrifoglio et al., 2008; Nourbakhsh and Ouyang, 2011; Alshalalfah and Shalaby, 2012). In addition to the studies mentioned above, heuristics to solve DARPs have been proposed by Psaraftis (1983), Jaw et al. (1986), Fu (1999, 2001), Cordeau and Laporte (2003), Diana and Dessouky (2004), Kirchler and Calvo (2013), Paquette et al. (2013), Ma et al. (2013), and Martinez et al. (2014), among others.

There are a couple of non-myopic methods proposed in the literature for the dynamic DARP, as well as for the related dynamic pickup and delivery problems. Both are reviewed here. Berbeglia et al. (2010) classified these problems into three different classes: the dynamic vehicle

routing problem with pickups and deliveries (Dynamic VRPPD), the dynamic stacker crane problem (Dynamic SCP), and in the case where transportation requests consist of passengers the problem is the dynamic dial-a-ride problem (Dynamic DARP). While the static DARP is primarily challenging because it's an NP-hard problem, the dynamic problem with look ahead faces a different set of challenges.

A non-myopic dynamic optimization problem is solved either as an optimal control differential equation (for continuous time and decision problems) or a Markov decision process (for discrete time decision problem) of a form shown in Eq (2.1) (in discrete time form), called the Bellman equation (Powell, 2011).

$$V_t(R_t) = \min_{x_t} (C_t(R_t, x_t) + \gamma E[V_{t+1}(R_{t+1}) | (R_t, x_t)]) \quad (2.1)$$

where V_t is the value of the policy, C_t is the immediate payoff of the decision x_t under state R_t (which is also typically driven by information on exogenous stochastic variables, and varies in size based on the underlying distribution of the variable(s)), and γ is a discount factor. The challenge is in determining an appropriate value for the last term, $E[V_{t+1}(R_{t+1}) | (R_t, x_t)]$. The conditional expectation term depends on the future state R_{t+1} , but R_{t+1} depends on R_t and x_t . As either of these variables increase, and as the number of discrete time steps increases, the problem quickly becomes intractable to solve.

The generalized problem with network effects and explicit timing is shown by Chow and Regan (2011a) to be altogether unsolvable in its complete form because even when R_{t+1} is independent of x_t , there is an additional complexity due to $V_{t,i}$ for network component i being dependent on $V_{t-1,j}$ for network component j . As a result, the Bellman equation becomes one that depends on both past (through network effects) and future (through expected value

function), leading to no exact solution method possible. The intractable problem is shown in Eq (2.2), where K is the number of link components that can be decided upon.

$$V_{t,i}(R_t) = \min_{x_{t,i}} (C_t(R_t, x_{t,i}) + (V_{t,1}(R_t, x_{t,i}) + \dots + V_{t,i-1}(R_t, x_{t,i}) + V_{t,i+1}(R_t, x_{t,i}) + \dots + V_{t,K}(R_t, x_{t,i}) + \rho E[V_{t+1,i}(R_{t+1})|(R_t, x_t)]]) \quad (2.2)$$

Chow and Regan (2011a) considered the network design and timing problem in Eq (2.2), and proposed an intermediate value function as a policy approximation, which can be solved with asymptotic convergence based on a combination of simulation and enumerating multiple sequences $h \in H$ for timing a set of network decisions. Even the intermediate function, shown in Eq (2.3), exhibits a greater curse of dimensionality (due to network dependencies) than the form shown in Eq (2.1).

$$V_{t,i_h}(R_t) = \min_{x_{t,i_h}} (C_t(R_t, x_{t,i_h}) + V_{t,(i+1)_h}(R_t, x_{t,i_h}) + \rho E[V_{t+1,i_h}(R_{t+1})|(R_t, x_t)]) \quad (2.3a)$$

$$\text{where } V_t = \min_h (V_{t,1_h}) \quad (2.3b)$$

where the value of one network decision also depends on the performance of other network decisions at the same time (and approximated using minimum of a series of enumerated sequences).

Due to these challenges, many have proposed approximate dynamic programming methodologies, where either one or more of the following elements are approximated: policy (Secomandi, 2001) (state dependent decision rule), value function (Thomas and White, 2004), contribution function, or state (Figliozzi et al., 2007). Researchers working with non-myopic dynamic DARP have dealt with the approximation using Markov decision processes under different structural variations. Dror et al. (1989) and Dror (1993) proposed a Markov decision

model for the vehicle routing problem. These models were formulated as a multistage stochastic programming problem. Secomandi (2001) presented a version of dynamic programming model for the vehicle routing problem where customers' demands are uncertain. The objective function minimizes the expected distance traveled in order to serve all customer demands. Neuro-dynamic programming was used to provide approximate solutions to this difficult stochastic combinatorial optimization problem. Spivey & Powell (2004) proposed a generalized dynamic assignment problem for fleet management as a Markov decision process, where a resource (container, vehicle, or driver) can serve only one task at a time. Novoa and Storer (2009) presented an approximate dynamic programming algorithm for the single-vehicle routing problem with stochastic demands from a dynamic or re-optimization perspective. They extended Secomandi's one-stage look-ahead to a two-stage look-ahead. Cortés et al. (2009) formulated the problem as a hybrid predictive control problem using state space models, also using a two-stage look-ahead, and proposed a particle swarm optimization (PSO) heuristic to solve it.

Innovative approximations of future states have also been proposed. Mitrović-Minić et al. (2004) introduced a dynamic problem with two rolling horizons, one for short term and one for long term. Spivey and Powell (2004)'s policy includes the use of gradients to approximate the future values. Ichoua et al. (2006) introduced a strategy based on dummy customers (representing forecasted requests) in the vehicle routes to respond to future request arrivals. Chow and Regan (2011a) approximated the future states in adaptive network design problems with a Least Squares Monte Carlo simulation method that is widely used in the real option literature. This method was shown to be applicable to dynamic DARP by Chow and Sayarshad (2015).

A third direction in the literature focused on alternative strategies. In addition to the dispatch strategy considered by many studies, Thomas and White (2004) formulated a dynamic

route selection for a single pickup and delivery as a Markov decision process that anticipates demand requests. Thomas (2007) also looked at the waiting strategy for vehicles to determine a policy for when to wait and when to reposition.

Much of the dynamic programming based approaches cannot solve large scale problems without approximating very short time horizons (e.g. 1 step look ahead) or aggregated states. Hyytiä et al. (2012) proposed approximation in the service process instead, treating it as a multi-server queue to obtain analytical expressions for steady state performance that can be easily computed for large scale systems. A summary of recent contributions to non-myopic dynamic DARPs is provided in **Table 2.1**.

Recent advances in information and communications technologies (ICTs) and “Big Data” helped alleviate the growing problem by making alternative transportation modes more viable (Chow, 2014). Due to the availability of real time vehicle location and customer request data, it is much more feasible to implement dynamic or “on-line” vehicle allocation or routing policies in smart transit systems, e.g. Psaraftis (1980), Psaraftis (1995), Madsen et al. (1995), Cortés and Jayakrishnan (2002), Horn (2002), Jung and Jayakrishnan (2011), Agatz et al., (2011), d'Orey et al. (2012), and Hosni et al. (2014). For example, Thien (2013) and Furuhata et al. (2014) present cost-sharing mechanisms for transport systems where the fare price is a proportion of total cost, and can be implemented due to the maturity of ICTs.

Table 2.1. Summary of non-myopic dynamic DARP studies

Studies	Approximation features(s)
Dror et al. (1989)	Markov decision process framework; No solution provided
Secomandi (2001)	Cyclic rollout policy for value
Thomas and White (2004)	One-stage look-ahead
Mitrović-Minić et al. (2004)	Double horizon insertion heuristic
Spivey and Powell (2004)	Gradient approximations of value
Ichoua et al. (2006)	Dummy customers for state; Tabu search heuristic
Thomas (2007)	Five insertion heuristic policies
Novoa and Storer (2009)	Cyclic rollout policy for value; Monte Carlo simulation
Cortés et al. (2009)	Two-stage look-ahead; Particle swarm optimization
Hyttiä et al. (2012)	Multi-server queue for value
Chow and Sayarshad (2015)	Least Squares Monte Carlo method

2.3.1. Hyttiä et al.'s model

One recent methodology was to approximate the value function with a multi-server queue system (Hyttiä et al., 2010a, b, 2012). Their argument is that, instead of approximating short time horizons, value functions, or states, why not approximate the overall system as a queueing system? As a stochastic queueing system, it is possible to derive analytical expressions for the steady state average delay costs, which is computationally cheap. It is then possible to approximate steady state costs of an infinite horizon for a large scale problem setting, which is not even possible with the current state of the art non-myopic DARP methods. An illustration of

the approximation is shown in **Figure 2.3**. By assuming that request arrivals with rate λ are time-invariant Poisson processes and approximating the server time with rate μ as Poisson processes, the steady state (assuming infinite horizon) cost to allocate a new customer to a vehicle can be determined as the change to the steady state queue delay. Whereas a non-myopic approach under a queueing system uses the steady state variables, the myopic approach simply uses the current delays observed in the queues.

Fig 2.3. Illustration of the non-myopic dynamic DARP as a multi-server queue system.

Under this approximation, both the network effects and the timing effects (as illustrated with an intermediate value function from Eq (2.2)) are lumped together into one effect as shown in Eq (2.4).

$$V_{t,i}(R_t) = \min_{x_{t,i}} \left(C_t(R_t, x_{t,i}) + \Omega(V_{t,1}, \dots, V_{t,i-1}, V_{t,i+1}, \dots, V_{t,K}, E[V_{t+1}|(R_t, x_t)]) \right) \quad (2.4)$$

where Ω is approximated by an expression derived from M/M/s queue and the whole expression is split into user costs and system costs, shown in Eq (2.5).

$$\operatorname{argmin}_{v,\xi} [c(v, \xi) - c(v, \xi')] \quad (2.5a)$$

where

$$c(v, \xi) = \theta T(v, \xi) + (1 - \theta) \left(\beta T(v, \xi)^2 + \sum_i S_i(v, \xi) \right) \quad (2.5b)$$

where v is a vehicle, ξ is a tour obtained for a traveling salesman problem with pickup and delivery (TSPPD), ξ' is the previous tour updated to the time of the current customer arrival, c is the value function, T is the tour length, S_i is the total delay for customer i (service plus wait time, i.e. time from call in to time they are delivered). θ and β are parameters to adjust the degree of system cost versus user cost (θ) and degree of look ahead (β).

The method is shown to be easy to implement even for large systems (up to fleets of 140 vehicles) and improves upon the myopic solution for a range of parameter values. While an M/M/s approximation is not how the system truly behaves, it is empirically shown using simulation to outperform a myopic policy. We know that some form of approximation is required for such systems, whether it is a finite rolling horizon (as done by Figliozzi et al., 2007), an approximation of the states, an approximation of the value functions, or otherwise, and the approximation method from Hyytiä et al. (2012) has the added benefit of computational efficiency. Fare prices are not discussed by Hyytiä et al., and as such are not included in the optimal policy.

2.3.2. Non-myopic pricing in queues

Since the approximation in Hyytiä et al. (2012) is based on an M/M/1 queue, we examine the literature on tolling in queues as a means of incorporating pricing into the non-myopic policy. It has long been known (Naor, 1969) that a queue for a public good that does not impose any toll would lead to a suboptimal social value. Tolling and other strategies to assign priority to customers based on their willingness to pay are needed to attain a social optimum. While Naor (1969) made this argument for an M/M/1 queue, Knudsen (1972) also showed this for an M/M/s queue (this is also obvious from the marginal cost pricing literature for roads, as discussed in Yang and Huang, 1998).

Mendelson and Whang (1990) distinguished the aggregate arrival rate from an individual decision structure of whether to join the system. In the individual decision system, there exists an actual rate of pre-arrivals Λ where individuals each have an effective arrival rate λ , $0 < \lambda \leq \Lambda$, representing their expected arrival (no balking or reneging) of entering the system at the level of service that they are willing to pay. The authors derive a demand function at the individual level by first defining the probability that an arrival values the service at x or higher as shown in Eq (2.6).

$$\bar{\Phi}(x) = 1 - \Phi(x) = \int_x^{\infty} \phi(z) dz \quad (2.6)$$

where $\Phi(x)$ is the cumulative probability of willingness to pay for service x . Since the arrival rate is then $\lambda = \Lambda \bar{\Phi}(x)$, the marginal value (willingness to pay), $V'(\lambda)$, is the value x corresponding to the expression, shown in Eq (2.7).

$$V'(\lambda) = \bar{\Phi}^{-1}\left(\frac{\lambda}{\Lambda}\right) \quad (2.7)$$

Let $W(\lambda)$ be the time spent in the system. At the equilibrium, the user would be indifferent between joining and not joining the system, resulting in Eq (2.8).

$$V'(\lambda) = p + \psi W(\lambda) \quad (2.8)$$

where p is the cost to the user, and ψ is value of time faced by an incoming arrival. At the social optimum, the first order condition results in Eq (2.9), and combining that with Eq (2.8) leads to the social optimal toll p^* shown in Eq (2.10).

$$V'(\lambda^*) = \psi \lambda^* \frac{dW}{d\lambda} \Big|_{\lambda=\lambda^*} + \psi W(\lambda^*) \quad (2.9)$$

$$p^* = \psi \lambda^* \frac{dW}{d\lambda} \Big|_{\lambda=\lambda^*} \quad (2.10)$$

2.4. Prediction models: Online-offline information

There have been many recent studies examining short term predictions of travel demand or general arrival processes. Contrary to models that focus on spatial aspects (e.g. spatial queueing models or spatial Poisson arrival processes), our focus is on models driven by real time arrival data that capture *temporal dependencies*. For example, Wang (2004), Chang et al. (2010), Yuan et al. (2011), and Li et al. (2012) studied prediction of customer demand for taxi services to inform the dispatcher or the prepositioning of idle taxis. Kaltenbrunner et al. (2010) and Regue and Recker (2014) developed prediction models for shared bike systems. More generally, Ihler et

al. (2006), Weinberg et al. (2007), and Matteson et al. (2011) examined more sophisticated Poisson process models of arrival prediction applicable to any call center or service fleet.

2.4.1. The arrival process

Customer arrival processes have long been modeled in urban transportation service operations and planning, with notable early examples discussed in Larson and Odoni (1981). The most common probabilistic model for customer arrivals is the Poisson distribution, whose probability mass function is given by Eq (2.11).

$$P(k; \lambda) = e^{-\lambda} \lambda^k / k!, \quad k = 0, 1, \dots \quad (2.11)$$

where the intensity parameter λ represents the arrival rate, or average number of occurrences in a unit time interval. When the intensity λ is a function of time, i.e. $\lambda \equiv \lambda(t)$, it becomes a *non-homogeneous* Poisson process. Travel demands vary significantly by time and location, so it makes most sense to model such demand as a non-homogeneous (also called inhomogeneous) Poisson process (NHPP). An NHPP is obtained by integrating the intensity function from time s to time t , resulting in Eq. (2.12).

$$P(N_t - N_s = k) = \exp\left(-\int_s^t \lambda(u) du\right) \frac{\left(\int_s^t \lambda(u) du\right)^k}{k!} \quad (2.12)$$

where $\Lambda(t) = \int_0^t \lambda(u) du$ and $\int_s^t \lambda(u) du = \Lambda(t) - \Lambda(s)$.

We categorize the models in the literature into two types. Offline models are those estimated from historical data (prior day $d - 1$ and before) to predict arrivals during throughout

a day d . Prior and posterior distributions in the offline models typically rely on household surveys and roadside surveys. The survey data cannot provide up- to-date information to reflect the rapid changes in customer arrival patterns. The second type are online models, which are run on a real time or near-real-time basis, and tend to rely on autocorrelation between time intervals to predict the next time step within a day from earlier time steps in that day.

2.4.2. Offline models

Offline models consist of an NHPP that is estimated from historical data from prior days. The simplest way to employ an NHPP is to define time intervals exogenously (e.g. every hour) at which average historical arrival rates are used. Ihler et al. (2006) present one such method that estimates the arrival rate as such, breaking down the value into seasonal factors to account for time of day and day of the week factors. Moreira-Matias et al. (2013) apply this method to predicting taxi passenger arrivals and compare it to an online method—an ARIMA (Auto Regressive Integrated Moving Average) time series model (see Chapter 5)—as well as an ensemble method that combines the two by learning the threshold for selecting one model over another.

The problem with defining exogenous time intervals is that the model is not truly time dependent; it is just essentially a number of homogeneous Poisson processes superimposed on different time intervals as the arrival rate is assumed constant or piecewise linear within those intervals. For example, Kaltenbrunner et al. (2010) showed using shared bike data that prediction errors can vary depending on the specification of the time intervals and ended up using 30 minute intervals. To avoid this, one method is to assume a parametric function for the intensity

function, such as a Weibull process (Leemis, 1991). Other examples of parametric techniques are shown in Kuhl et al. (1997) and Massey et al. (1996). However, parametric methods impose an assumption about the form of the intensity function. This can be a problem when estimating the parameters if the true intensity function does not fit the assumed parametric class.

Alternatively, Leemis (1991) and Arkin and Leemis (2000) proposed a nonparametric technique for estimating the cumulative intensity function of a NHPP on the time interval from historical data. This method does not require specifying any parameters or weights (as done in Ihler et al., 2006) for the intensity functions. By avoiding parametrization, it makes estimation via maximum likelihood a simpler computational task.

When the true intensity function is known to be continuous, piecewise-linear functions are convenient forms for fitting intensity functions from point processes (Leemis, 2004). Typically, observed counts can be averaged to find the constant value for each time period. Chen and Schmeiser (2011, 2013) introduce the algorithm I-SMOOTH, which cyclically smooths any piecewise-constant function by bisecting time intervals as needed and fitting new piecewise linear functions based on gradients. Nicol and Leemis (2014) consider a class of continuous piecewise-linear algorithms.

No studies have used the method from Nicol and Leemis (2014) to estimate an NHPP for predicting passenger arrival patterns in transportation systems nor compared it to earlier methods. In this study, we conduct an empirical experiment to compare the performance of the NHPP estimation method from Nicol and Leemis (2014) with that of the NHPP portion from Ihler et al. (2006).

2.4.3. Online models

Moreira-Matias et al. (2013) argue that the true potential benefit of real time arrival data lies in online prediction models, and that this potential has largely been unrealized. Whereas offline models can capture recurrent trends from historical data, they are not well designed for scenarios involving non-recurrent events (Ihler et al., 2006).

There are only a few online arrival process models applied to transportation systems in recent years. A majority of these models are based on time series methods, such as the ARIMA model (e.g. Kaltenbrunner et al., 2010; Li et al., 2012). In the latter study, Li et al. compare four methods: a naïve method based on time of day factors, a Bayesian network approach, and two ARIMA models. They conclude that the ARIMA model fit best among those methods. ARIMA models can incorporate autocorrelation between time steps and lag effects. However, they are based on continuous variables whereas arrival counts are discrete in nature.

The discrepancy between continuous and discrete variables is noted by Matteson et al. (2011) and Moreira-Matias et al. (2013). To test their argument, Moreira-Matias et al. applied the NHPP model from Ihler et al. (2006) and compared the model to an ARIMA model on taxi arrival data from the city of Porto, Portugal. However, their results were inconclusive with one model outperforming the other in different periods. On the other hand, Matteson et al. (2011) propose a class of discrete variable time series models called IntGARCH. Using Toronto emergency medical service data, they empirically tested different forms of the IntGARCH model and showed significant reduction in forecast error compared to offline factor models.

A third group of online models are Bayesian methods. Examples of Bayesian methods for NHPP include Weinberg et al. (2007) and Li et al. (2012). There are two types of Bayesian

models applied to count data: parametric and non-parametric. The non-parametric class refers to a population in which their distribution to different parameter values, i.e. the mixture, is unknown (Krnjajić et al., 2008). In other words, non-parametric Bayesian modeling assumes a *doubly-stochastic* Poisson process, or a Gaussian Cox process (Adams et al., 2009), in terms of random arrivals with random intensities. Krnjajić et al. (2008) found that nonparametric models are more accurate and flexible than parametric models.

Arjas and Gasbarra (1994) and Arjas and Heikkinen (1997) introduce Markov chain Monte Carlo (MCMC) for non-parametric Bayesian estimation of an intensity function from real time data. Weinberg et al. (2007) propose a hybrid MCMC method to estimate both latent states and model parameters. Because MCMC can be computationally burdensome, Adams et al. (2009) propose approximations in the form of a Sigmoidal Gaussian Cox Process to be more computationally efficient. Compared to frequentist or non-Bayesian method, it is shown to have better results.

2.5. Relocation of idle vehicles

Positioning of idle vehicles is a type of resource allocation problem categorized within the field of location science, where the location decision impacts the level of service that a vehicle provides to neighboring demand. Optimization problems in this area are called facility location, or set covering, problems, and like in routing, can fall into different variations and subclasses depending on the objective and constraints desired.

Facility location models have been developed for a variety of application areas. Some common applications are locating gas stations, refueling stations, convenience stores, and

billboards (Berman et al. 1995; Jung et al. 2013) and emergency services as ambulances and air tankers for wild fires (Chow and Regan 2011; Berman et al. 1985; Galvao and Morabito 2008; Marianov and Revelle 1996). Furthermore, facility location problems can be related to communication networks (Li et al. 1999). Facility location is also important in manufacturing environments (De Alvarenga et al. 2000). Jia et al., (2011) proposed a model for the prepositioning of supplies in emergency services.

One of the more well known facility location problems is the p-median problem. In that problem, p facilities have to be located on a network in such a way that it minimizes total distances of each customer to their closest facility. Hakimi (1964) showed that the optimal location always lies on a node (as opposed to any continuous point on a link between nodes), so finite sets of candidate locations can be used without degrading the optimality of the solution. When more than a facility is to be located, the problem becomes known as the “p-median” problem, term that was first used by Hakimi (1965) and Hakimi and Naheshwari(1977). The first integer programming formulation for finding the solution to a reasonably sized instance of p-median problem was proposed by ReVelle and Swain (1970) and Swain (1974). A standard formulation of the p-median problem is shown in Eq (2.13).

$$\min \sum_i \sum_j d_{ij} X_{ij} \quad (2.13a)$$

Subject to

$$X_{ij} - Y_j \leq 0 \quad \forall i, j \quad (2.13b)$$

$$\sum_j X_{ij} = 1 \quad \forall i \quad (2.13c)$$

$$\sum_j Y_j = B \quad (2.13d)$$

$$X_{ij}, Y_j \in \{0,1\} \quad (2.13e)$$

where

X_{ij} = whether node j covers node i

d_{ij} = matrix of distances from i to j

Y_j = whether a server is based at node j

B = the number of servers (vehicles) in the network

The objective function is to minimize the distances of the closest servers in terms of deployment time. Constraint (2-13b) guarantees that coverage is only possible if a server is located at that node. Constraint (2-13c) forces that the demand is met by only one server. (2-13d) and (2-13e) are the facility budget constraint and binary constraints.

Another facility location problem was proposed by Church and ReVelle (1974) (maximum covering location problem). Chow and Regan (2011) proposed a k server p -median problem with a rolling horizon to assign co-located servers for air tanker initial attack basing in California for regional wildland fires. Brotcorne et al. (2003) provide an overview on location/relocation in emergency problem. Some work has been completed with location/relocation problems for vehicles in a transit system. Yuan et al. (2011) proposed a probabilistic model based on parking place detection algorithm to recommend locations to taxi drivers and people where they can easily find vacant taxis. The approach was tested on 1200

taxis in 110 days. Phithakkitnukoon et al. (2011) developed a predictive model for the number of vacant taxis in a given area based on time and weather condition. The approach tested on 150 taxis in Lisbon, Portugal. Powell et al. (2011) introduced an approach based on historical data (GPS dataset) to direct taxi drivers in order to reduce the number of cruising miles (taxi without passenger) while increasing the number of live miles (taxi with passenger). This approach was tested on 600 taxis in Shanghai. Li et al. (2011) proposed a time-location model with GPS data for passenger-finding strategies. This approach was tested on 5350 taxis in China. Yang et al. (2010) modeled urban taxi services in a network context which can reduce the number of cruising miles while increasing the number of live miles by defining the relationship between customer and taxi waiting times.

Some models explicitly consider the probabilistic nature of requests and availability of servers. In addition, many authors consider queue theory to design a model to determine optimal location. Berman and Larson (1985) developed two server networks, and Berman et al. (1987) developed a two heuristics for locating n servers in a congested network. Then, Berman and Drezner (2007) and Berman et al. (1987) presented a generalization of the p -median queue, concerned with waiting cost and service time. Their algorithms start with guessing a location, taking a solution that minimizes total cost in queue, and continuing until no improvement can be obtained. Then, Batta (1989, 1988) proposed a one-server model for the optimal location. These problems were solved by greedy heuristic algorithms to find the location of the server. Pasandideh and Niaki (2012) solved a multi-objective facility location problem based on queue theory by genetic algorithm. They used desirability function by defining a geometric mean to minimize waiting time, and idle probability at open facility. All these models are nonlinear and their solution procedures are heuristic.

Marianov and ReVelle (1994, 1996) proposed a queueing maximal availability location problem. Marianov and Serra (1998) extended the maximum covering by taking the fact that the number of request services behaves following some distribution function. Marianov and Serra (2002) modified these models, which seeks a probabilistic location-allocation to cover population to cover all population and includes server allocation to the facilities. Jung et al. (2014) proposed a simulation-based optimization approach to locate electric charging stations that accounted for queueing delay and allocating to taxi tours instead of demand nodes. A summary of recent contributions in location theory with queueing is provided in **Table 2.2**.

Table2.2. Facility location problem with queueing

Studies	model features(s)
Hakimi(1964)	p-median policy problem under myopic policy
Revelle and Swain(1970)	p-median policy problem under myopic policy
Church and ReVelle(1974)	Maximum covering location problem
Batta (1989&1988)	Geedy algorithm for single server(M/M/1)
Berman and Larson (1985)	Myopic policy for M/M/2 system
Berman et al. (1985)	Myopic policy for M/G/1 system
Berman et al. (1987)	p-median problem for multi-server(M/M/S)
Marianov and Serra(1998)	Maximum covering problem with myopic policy(M/M/1 and M/M/S)
Marianov & Serra's (2002)	Myopic policy for multi-server queue(M/M/S)
Jung et al. (2014)	Simulation-based optimization for queueing delay with tour allocation

Chapter 3

Routing and pricing

In this chapter, we discuss some important issues in dynamic vehicle routing and pricing problems that should be addressed with our new methodology. We propose to incorporate non-myopic pricing into the non-myopic DARP as a solution for last mile problem and smart cities. The approach introduces a social optimal toll for queues under multi-server queue-based DARP. We also show improvement in social welfare over prior policies that ignored price. Then, we evaluate the elasticity of the social welfare with respect to fleet size, capacity and price. Finally, we prove that our approach can provide the guidelines for fleet design and shared-use policies under non-myopic pricing.

3.1. Introduction

The benefit of anticipating future states is that the true cost is better estimated at a specific time so as to make more informed decisions. For example, delivering a user from point A to point B does not just imply there is a marginal cost of adding the trip A-B into the vehicle's tour, but there is an opportunity cost due to having a resource (the vehicle) tied up during that time and being repositioned in a location that augments the vehicle's level of service for future users. If each vehicle is viewed as a server, the dynamic DARP is in turn a multi-server queueing system. In fact, Hyytiä et al. (2010a,b, 2012) exploit this characteristic to approximate these opportunity costs made in vehicle allocation and routing decisions.

However, a fundamental assumption present in the non-myopic dynamic DARP models causes an inefficiency that has not yet been addressed in the literature: the models assume user demand is not elastic even when it should be. This flawed assumption is illustrated in **Figure 3.1** with a simplified supply-demand graph. The essential argument made by non-myopic models is that systems operating under a myopic approach assume a less efficient cost curve than the true cost curve that could be achieved with a non-myopic approach. Their conclusion is that moving to a more efficient cost curve that properly accounts for opportunity costs should lead to a cost reduction represented by $B - C$, resulting in the point $\{A, C\}$. However, if demand is elastic, then the optimal point is not at $\{A, C\}$, but at the point where demand meets the marginal cost curve, i.e. $\{F, G\}$. Because of presuming the demand is inelastic, the system would operate at cost C , leading to excess demand of D that requires level of service at cost of E to sufficiently meet.

The inefficiency in non-myopic dynamic routing models can be corrected by considering non-myopic dynamic pricing simultaneously, which can allow a system to move users toward $\{F, G\}$. However, no studies have jointly considered vehicle assignment, routing, and pricing in the non-myopic dynamic DARP. Figliozzi et al. (2007) conducted one of the few studies on non-myopic pricing in dynamic vehicle routing problems for truckload routing, examining the benefits of pricing when looking ahead even one time step.

A – Myopic quantity

B – Myopic level of service

C – Cost assumed using the non-myopic policy under inelastic demand

D – Unrealized demand due to assuming full non-myopic cost at myopic quantity

E – Actual cost (and resulting inefficiency) that a non-myopic policy with fixed demand assumption would induce

Fig 3.1. Illustration of the inefficiency present assuming inelastic demand and using myopic pricing.

3.2. Motivational example for non-myopic pricing

Figliozi et al. (2007) illustrate the need for non-myopic pricing using a simple 4-node example, which we reproduce here as a DARP and expand upon with respect to the motivation.

Consider a network illustrated in **Figure 3.2**, where a single vehicle fleet serves 4 nodes located around a square block. The vehicle would have to traverse to node C or node A from node B in order to reach node D. The lengths of each link is 1 unit time step, and one of two possible requests occurs each step: a customer either (a) desires to be picked up at node A to be

dropped off at node B (50% probability), or (b) they wish to be picked up at node D to be dropped off at node A (50% probability). We assume no dwell times and no repositioning costs. The willingness to pay of each customer is distributed as $p(1) = 0.25, p(2) = 0.50, p(3) = 0.25$.

Fig3.2. Motivational example for non-myopic pricing (adapted from Figliozzi et al., 2007).

Static price: Both AB and DA have the same cost of 1 unit, so under a static pricing scheme the vehicle would charge the passengers the same cost of 1 unit. This is the pricing policy used in many taxis, for example.

Myopic dynamic price: In this case, the price depends on where the vehicle is currently positioned. Suppose the vehicle is located at node A. In order to serve customer AB, the vehicle would have a cost of 1 unit. On the other hand, serving customer DA would have a cost of 2 units (A-D-A).

Non-myopic dynamic price: Now let's consider the consequence of serving customer AB from node A. After serving that customer, the vehicle would be located at node B. If a new customer arrives at AB ($S = AB$) and the vehicle status is at node B ($Z = \{B\}$), the cost would be 2 (B-A-B), $C(S = AB | Z = \{B\}) = 2$. With the distribution of the willingness to pay, the conditional

expected user benefit is $(3 - 2)(0.25) = 0.25$. If the new customer instead requests DA, the cost is 3 (B-C-D-A), $C(S = DA | Z = \{B\}) = 3$. With the distribution of the willingness to pay, the conditional expected user benefit is $(3 - 3)(0.25) = 0$. The expected user benefit from this scenario is $0.5(0.25) + 0.5(0) = 0.125$.

If the vehicle is at node A because the initial AB customer turns down the service due to price, then $C(S = AB | Z = \{A\}) = 1$ and $C(S = DA | Z = \{A\}) = 2$. The conditional expected user benefit of a subsequent AB is $0.5(2 - 1) + 0.25(3 - 1) = 1$, while that for DA is $0.25(3 - 2) = 0.25$. The expected benefit is $0.5(1) + 0.5(0.25) = 0.625$.

The price for the initial AB is set equal to the cost of the accepted service (1) minus expected user benefit from accepted service (0.125) plus the opportunity cost due to expected benefit from rejected service (0.625). The result is an optimal price of $1 - 0.125 + 0.625 = 1.5$, which differs from the myopic dynamic price of 1 unit for AB at node A.

As this example shows, not properly anticipating/looking ahead for the opportunity costs or benefits can result in suboptimal pricing. Figliozzi et al. (2007) modeled the future costs by approximating them with a single time step finite horizon. These existing approximation methods have not included the pricing strategy as a non-myopic policy. The challenge with incorporating pricing in many of these problems, which are typically formulated as discrete state spaces with discrete decisions, is the continuous nature of the pricing variable.

We propose to incorporate non-myopic pricing into the non-myopic dynamic DARP, as Figliozzi et al. (2007) had done for truckload routing. Contrary to that earlier study, our proposed contribution is not limited to only a single time step look-ahead, but instead approximates the expected average future value from an infinite horizon. This is accomplished by adopting the

multi-server queue approach from Hyytiä et al. (2012) to approximate the average future value. We introduce a non-myopic pricing component from the queueing literature into the dispatch and routing policy function. The methodology is numerically illustrated for a replicable example, and then compared over a range of test instances against the myopic pricing assumption to demonstrate its effectiveness. Our approach to integrating non-myopic pricing is as different from Figliozzi et al.'s (2007) approach as Hyytiä et al.'s (2012) approach is different from the existing non-myopic dynamic DARPs that only provide one- to two-step look ahead.

3.3. Proposed methodology

Our approach to integrating non-myopic pricing is as different from Figliozzi et al.'s (2007) approach as Hyytiä et al.'s (2012) approach is different from the existing non-myopic dynamic DARPs that only provide one- to two-step look ahead. The precise dynamic DARP that we are solving is described below.

- A fleet of vehicles receives dispatch, routing, and pricing instructions from a central authority.
- Customers call in requesting service from one location to another; as a result, dispatch, routing, and pricing decisions are made in real time by the central authority. The subgraph during each dispatch assignment consists of $G = \{O, D, P^+, P^-\}$, where O is the current vehicle location, D is a pre-specified idle vehicle destination, P^+ is the set of pickup nodes, and P^- is the set of drop-off nodes, where $|P^+| \leq |P^-|$ since a vehicle may be en-route to drop off a passenger when a new passenger arrives. In the numerical experiments in this study, D is set to be the fleet depot, but in future studies we intend to call an assignment

routine to preposition the idle vehicle. All vehicles at the start of the day are located at the depot, but future studies will examine the initial prepositioning problem as well. Pre-booking, reservations, etc., are not explicitly handled. As noted in de Borger and Fosgerau (2012), the cost of waiting can be transferred to the cost of inconvenience in trip planning, depending on information provision.

- Customers' requests in typical DARPs include time windows for pickup and drop-off. Hyttiä et al. (2012) studied the DARP without time windows. We follow this assumption of no time windows, but justify it as being built into the elastic demand function – since the objective has user cost, the pricing would reflect late deliveries, which in turn feeds back to the demand for service.
- Requests are assumed to be from single individuals arriving in a Poisson process. Group arrivals and more complex arrival processes are not modeled here; these properties can be explored in future research. Vehicles have a capacity for the case of shared use service.
- Based on the price quoted by the system, a customer may choose to reject the service. As a result, user demand is elastic.
- Once a vehicle is assigned to a customer, the dispatch decision and price do not change (the route sequence may change if the vehicle is assigned to subsequent arrivals). As discussed in Furuhata et al. (2014), different cost-sharing mechanisms may be possible and user behavior may be sensitive to them, but price updating as well as cost-sharing mechanisms will be considered in future research.

By assuming no elasticity in the demand (i.e. fixed λ that is not based on cost of service), the non-myopic policy will over-discount the cost of service, leading to inefficiency. To correct the

problem, we need to introduce the fare pricing decision variable that achieves social optimum for the steady state queue. $V'(\lambda)$ in Eq (2.7) to Eq (2.9) is an inverse demand function. The $V(\lambda)$ is required to be strictly concave. We assume a linear demand function, $V' = \alpha - \frac{\alpha}{\Lambda}\lambda$, such that the value function is shown in Eq (3.1), although other concave functions can also be used.

$$V = \alpha\lambda - \frac{\alpha}{2\Lambda}\lambda^2, 0 < \lambda \leq \Lambda \quad (3.1)$$

If we make the same M/M/1 approximation as Hyytiä et al. (2012), the average delay for a user can be defined as shown in Eq (3.2).

$$W = \frac{1}{\mu - \lambda} \quad (3.2)$$

The parameter μ is exogenous and approximated with $\hat{\mu}$ obtained from historical data, as is the parameter α . For the non-myopic DARP, the service time is the time a user is picked up until they are dropped off. If we call this time for user i (from a finite population N) to be x_i , then the value of $\hat{\mu}$ is determined by Eq (3.3).

$$\hat{\mu} = \frac{1}{N} \sum_i \frac{1}{x_i} \quad (3.3)$$

Taking the derivative of W with respect to λ , we get:

$$\frac{dW}{d\lambda} = \frac{1}{(\mu - \lambda)^2}$$

Then the social optimal fare price to charge incoming user n , considering steady state queue performance, is dependent on the social optimal λ_n^* as shown in Eq (3.4). We assume that the

prices are set marginally, based on the additional cost to the tour and the user cost. This assumption will be used when computing the social welfare of this policy in the experimental study.

$$p_n^* = \psi \lambda_n^* \frac{1}{(\hat{\mu} - \lambda_n^*)^2} \quad (3.4)$$

where p_n^* is the social optimal fare price to charge incoming user n , and ψ is value of time faced by an incoming arrival. In the individual decision system, there exists an actual rate of pre-arrivals Λ where individuals each have an effective arrival rate λ , $0 < \lambda \leq \Lambda$.

When implementing this as a dynamic DARP, the prices of all passengers that already accepted service are fixed (per our assumptions described earlier), and we only need to solve ξ, λ, p for each vehicle v for the new passenger under consideration, and select the maximum value increase. If cost-sharing mechanisms are to be considered, it may be implemented here in a future study as a price updating scheme for all passengers being served by the vehicle. For each new passenger n , the optimal policy when including the price from Eq (3.4) is shown in Eq (3.5) derived from Eq (2.5). In Eq (3.5), the objective is to maximize value instead of minimizing cost, and the expression includes the added benefit of a single new customer $\left(\alpha \lambda_n - \frac{\alpha}{2\Lambda} \lambda_n^2\right)$ as well as the price transfer $\left(\frac{\psi \lambda_n}{(\hat{\mu} - \lambda_n)^2}\right)$ in the user and system costs. We also divide the costs by $T(v, \xi_n)$ to ensure the units are the same as the benefits.

$$\max Z_n(\lambda_n; v, \xi_n) - Z'(v, \xi_{n-1}) \quad (3.5a)$$

where

$$\begin{aligned}
Z_n(\lambda_n; v, \xi_n) = & \alpha\lambda_n - \frac{\alpha}{2\Lambda}\lambda_n^2 - \frac{\theta \left(T(v, \xi_n) - \frac{\psi\lambda_n}{(\hat{\mu} - \lambda_n)^2} \right)}{T(v, \xi_n)} \\
& - \frac{(1 - \theta) \left(\frac{\hat{\mu}\lambda_n}{2(\hat{\mu} - \lambda_n)} T(v, \xi_n)^2 + \sum_{i=1}^n S_i(v, \xi_n) + \frac{\psi\lambda_n}{(\hat{\mu} - \lambda_n)^2} \right)}{T(v, \xi_n)}
\end{aligned} \tag{3.5b}$$

The parameter θ , $0 \leq \theta \leq 1$, is a weight to differentiate user cost from system cost. To solve for λ_n^* , we take the derivative of Z_n with respect to λ_n and set it to zero, $\partial Z_n / \partial \lambda_n = 0$, shown in Eq (3.6) which can be solved with a simple one-dimensional line search. A price updating mechanism would require more sophisticated solution method over all passengers being served by the vehicle, which is avoided here. T is the tour length, S_i is the total delay for customer i (service plus wait time, i.e. time from call in to time they are delivered). θ and β are parameters to adjust the degree of system cost versus user cost (θ) and degree of look ahead (β).

$$\begin{aligned}
\alpha - \frac{\alpha}{\Lambda}\lambda_n^* + \theta \left(\frac{\psi}{T(v, \xi_n)(\hat{\mu} - \lambda_n^*)^2} + \frac{2\psi\lambda_n^*}{T(v, \xi_n)(\hat{\mu} - \lambda_n^*)^3} \right) \\
- (1 - \theta) \left(\frac{\hat{\mu}^2 T(v, \xi_n)}{2(\hat{\mu} - \lambda_n^*)^2} + \frac{\psi}{T(v, \xi_n)(\hat{\mu} - \lambda_n^*)^2} + \frac{2\psi\lambda_n^*}{T(v, \xi_n)(\hat{\mu} - \lambda_n^*)^3} \right) = 0, \tag{3.6}
\end{aligned}$$

$$0 < \lambda_n^* \leq \Lambda$$

In real-time operations (and for simulations of such operations), Algorithm 1 is used.

Algorithm 1

Upon arrival of a new user n ,

1. Update the positions and service statuses of the ξ_{n-1} tours of every vehicle from the time of previous user $n - 1$.
2. For each vehicle v ,
 - a. Solve a TSPPD algorithm (Algorithm 2) to obtain potential tour ξ_n
 - b. Determine λ_n^* and $Z_n(\lambda_n^*)$ (the prices p_n^* can also be computed using Eq (3.4))
3. Pick the vehicle that maximizes Eq (3.5a). Update ξ_n as the new tour for that vehicle, while keeping the other vehicles' tours the same as before.

In this optimization model, the price assigned to the new customer is a function also of (v, ξ) because the assignment and route determine the resulting x_i (as a DARP, some existing customers might get reshuffled). For example, user 1 may be assigned to vehicle 1, but when user 2 arrives it might get picked up and dropped off by vehicle 1 if it's along the way and the solution maximizes the value. As such, we acknowledge (just like Hyytiä et al.) that the M/M/1 assumption is not reflective of operating conditions since FIFO conditions are violated. The approximation is assumed to build in the steady state considerations into the policy. All prices are set at an equilibrium that assumes users have just enough incentive to choose the service, so no rejections or balking actually occurs during a simulation of the dynamic policy.

For the TSPPD, a heuristic is introduced based on the one discussed in Mosheiov (1994) and shown in Algorithm 2. The difference is that we are evaluating a reoptimization-based

approach, where some users might have already been picked up earlier. In those cases, those users would only have a drop-off node while counting toward the initial capacity.

Algorithm 2 – reoptimization-based TSPPD insertion algorithm

For a given current vehicle location O , a set of pickup nodes P^+ and drop-off nodes P^- , a vehicle capacity K , and distances between every node pair c_{uw} ,

1. Create a TSP tour S for $\{P^-, O\}$
2. For each $u \in P^+$, insert u into S such that
 - a. capacity constraint at that insertion point and all subsequent nodes is feasible,
 - b. u is visited before its corresponding drop-off, and
 - c. increased tour length is minimized
3. Update capacities at each node in the tour from insertion point, and go to step 2

For the experimental study in Section 3.4, the TSP tour construction of the drop off nodes in step 1 of Algorithm 2 is solved using Christofides' (1976) algorithm. Alternative routing and insertion algorithms can certainly be substituted in (and encouraged for practitioners). While the contribution of the study is focused on evaluating the improvement of the non-myopic pricing, the Christofides' algorithm is known to have a worst case bound of 50% of the optimal value. To ensure that the algorithm works well enough for the experimental study, the results are compared against an exact solution obtained using CPLEX on an integer programming (IP) formulation of the TSP for step 1. The IP formulation that can be used on the drop-off node tour construction is shown in Eq (3.7).

$$\min \sum_{u \in \{P^-, D\}} \sum_{\substack{w \neq u, \\ w \in \{P^-, D\}}} c_{uw} X_{uw} \quad (3.7a)$$

Subject to

$$\sum_{w \in P^-} X_{ow} = 1 \quad (3.7b)$$

$$\sum_{w \in P^-} X_{wD} = 1 \quad (3.7c)$$

$$T_u + c_{uw} - T_w \leq (1 - X_{uw})\mathcal{M}, \quad u, w \in P^- \quad (3.7d)$$

$$c_{ow} - T_w \leq (1 - X_{ow})\mathcal{M}, \quad w \in P^- \quad (3.7e)$$

$$X_{uw} \in \{0, 1\}, \quad T_u \geq 0 \quad (3.7f)$$

In terms of performance measures, let τ_{1n} be the realized clock arrival time of user n , τ_{2n} be the realized pickup time of user n , and τ_{3n} be the drop-off time of user n . The marginal pricing for user n (for determining social welfare for the policy from Hyytiä et al. (2012)) is then set equal to the value shown in Eq (3.8).

$$p_{\text{marginal},n} = \psi[T(v, \xi_n) - T(v, \xi'_n)] \quad (3.8)$$

The social welfare, computed as the total value to the users minus the total costs to the users, is computed as shown in Eq (3.9). More specifically, for a given vehicle v and (feasible) route ξ , let $S_n(v, \xi_n)$ denote the total delay of a customer n in the system, and $T(v, \xi_n)$ the tour length of vehicle v for passenger n . Thus, total costs to the users are computed by the total sojourn time and fare price of all customers divided by the current work backlog of vehicle v .

$$SW = \sum_n \left(\alpha \lambda_n - \frac{\alpha}{2\Lambda} \lambda_n^2 \right) - \sum_n \frac{\left(S_n(v, \xi_n) + \frac{p_n}{\psi} \right)}{T(v, \xi_n)} \quad (3.9)$$

3.4. Experimental study

The experimental study is designed to address five questions:

- 1) Does the proposed methodology work? This question is resolved in Section 3.4.1 with a replicable example (input parameters and outputs provided in **Table 3.1** to **Table 3.4**) and comparison between the policy from Hyytiä et al. (2012) (which we will call the *marginal pricing policy*) and our proposed policy that accounts for non-myopic pricing.
- 2) Does the assumption of TSPPD heuristic matter to the relative comparison between myopic and non-myopic policies? To answer that question, we run the same base scenario instances using Algorithm 2 with step 1 solved separately by Christofides' algorithm and using CPLEX to solve the IP formulation.
- 3) How much better does our policy work in terms of social welfare of our users? Assuming that a marginal cost policy is used when not considering non-myopic pricing, we can compute the social welfare for simulated runs. This comparison over a range of test instances is shown in Section 3.4.2.
- 4) What is the capability of this model for sensitivity analysis? We vary the arrival rate Λ as well as the parameters θ, β in Section 3.4.3.
- 5) Can we measure the elasticity of social welfare of the system to different variables? We test this by varying the fleet size, price, and vehicle capacity in Section 3.4.4 to quantify the elasticity of the social welfare to those variable so trade-offs can be made. The vehicle

capacity test also illuminates the trade-offs for a shared-taxi service (capacity one versus higher).

All the numerical tests are conducted on the same square Euclidean plane defined by Hyytiä et al. (2012). The space is bounded by $x = [-5,5]$ and $y = [-5,5]$, and pickup and drop-off demand is assumed to be independently and uniformly distributed over that region. We assume there is no dwell time at the nodes, and that idle vehicles are sent back to the depot at the origin since demand is uniformly distributed. The simulation of the policies and arrivals were run in Matlab R2012 on an Intel Core i5-2450 CPU with 2.5 GHz and 8 GB RAM, running on a 64-bit Windows 7 operating system.

3.4.1. Illustrative example

For the illustration, the following inputs are used for the non-myopic pricing policy: $\Lambda = 0.4$, $\alpha = \$30$, $\hat{\mu} = 1$, $\psi = 0.33$, $\theta = 0.4$, $\beta = 0.5$, speed = 4/6 km/min, vehicle capacity = 4, fleet size = 2. The input inter-arrival times for the simulated scenario of 13 arrivals is shown in **Table 3.1**. The effective λ_n 's and fare prices p_n 's are provided in **Table 3.2**. The arrival times, pickup times, and drop-off times of each of the 13 arrivals under the non-myopic pricing policy are shown in **Table 3.3**. **Figure 3.3** shows the resulting tour of the two vehicles under the non-myopic policy, where vehicle 1 serves passengers 1, 2, 4, 6, 7, 9, 10, and 13. Vehicle 2 serves passengers 3, 5, 8, 11 and 12. The tour length of vehicle 1 is 83.875 min; for vehicle 2 it is 78.379 min.

Fig 3.3. Simulated tours under proposed non-myopic pricing policy.

The arrival times, pickup times, and drop-off times under the marginal pricing policy are shown in **Table 3.4**, and the realized tour is illustrated in **Figure 3.4**. Under this policy, the vehicle 1 serves passengers 1, 4, 5, 7,8, 9, and 11. Vehicle 2 serves passengers 2, 3, 6, 10, 12, and 13. The tour length of vehicle 1 is 81.463; for vehicle 2 it is 91.499. This example illustrates the potential differences between the proposed policy and the policy from Hyytiä et al.

Fig 3.4. Simulated tours under the marginal pricing policy from Hyytiä et al. (2012).

Table 3.1. Input arrival times in simulation of dynamic DARP policies

Arrival	Inter-arrival times (min)	Arrival	Inter-arrival times (min)
1	9.7323	8	14.7701
2	2.1112	9	3.6244
3	0.8318	10	6.0634
4	0.2276	11	0.7677
5	6.8031	12	3.8909
6	1.8806	13	2.1255
7	2.5211		

Table 3.2. Output values under non-myopic dynamic DARPP

n	λ_n	p_n (\$)	n	λ_n	p_n (\$)
1	0.0158	2.8595	8	0.0213	1.3440
2	0.0250	1.2866	9	0.0005	0.0158
3	0.0609	3.9841	10	0.0098	0.3666
4	0.0112	0.3633	11	0.0280	1.7460
5	0.0404	3.0773	12	0.0058	0.2409
6	0.0065	0.2877	13	0.0103	0.4438
7	0.0032	0.1065			

$\hat{\mu} = 0.1$

Table 3.3. Output times under non-myopic pricing

User	Arrival time (min)	Pickup time (min)	Drop-off time (min)
1	9.7323	15.2436	25.2057
2	11.8436	32.3054	59.7914
3	12.6753	14.9985	16.9827
4	12.9029	20.1436	38.4632
5	19.7060	25.7877	28.4892
6	21.5866	47.7197	68.2632
7	24.1077	41.9567	60.5426
8	38.8779	45.9737	62.7199
9	42.5022	51.1024	52.7237
10	48.5657	70.5082	89.6107
11	49.3334	54.1612	78.2368
12	53.2243	71.0214	89.3975
13	55.3498	66.5899	78.5801

3.4.2. Social welfare comparison and algorithm evaluation

In this section, we evaluate the average performance of the proposed policy against the marginal pricing policy over 30 randomly simulated days. The following parameters are used: $\Lambda = 0.3$, $\alpha = \$30$, $\hat{\mu} = 1$, $\psi = 0.33$, speed = 4/6 km/min, vehicle capacity = 4, fleet size = 2, and 13 passengers in each of the 30 runs. For each simulated day, we determine the average system cost, the average user cost, and the average social welfare based on Eq (3.7) to Eq (3.8). We run four different policies on these 30 sample runs: marginal pricing under $\theta = 0.4$, $\beta = 0$ (myopic), and $\theta = 0.4$, $\beta = 0.5$ (non-myopic); the proposed non-myopic pricing under the same myopic allocation policy $\theta = 0.4$, $\beta = 0$ and non-myopic allocation policy $\theta = 0.4$, $\beta = 0.5$. The summary of the costs and social welfare are shown in **Table 3.5** and the comparison of the social welfare over the 30 runs is shown in **Figure 3.5**. Results in **Table 3.5** are shown for both the heuristic and using CPLEX to solve Eq (3.7) to obtain the initial tour in the TSPPD subroutine.

Table 3.4. Output times under non-myopic marginal pricing

User	Arrival time (min)	Pickup time (min)	Drop-off time (min)	User	Arrival time (min)	Pickup time (min)	Drop-off time (min)
1	9.7323	15.2436	86.5867	8	38.8779	66.4502	82.0117
2	11.8436	15.3698	36.0501	9	42.5022	53.7128	70.5795
3	12.6753	21.2192	42.8890	10	48.5657	54.6441	64.0175
4	12.9029	20.1436	30.1711	11	49.3334	73.4530	85.2951
5	19.7060	39.3330	42.0346	12	53.2243	60.1662	84.3470
6	21.5866	30.1901	41.7473	13	55.3498	72.8857	98.0073
7	24.1077	26.6777	61.0581

(a)

(b)

Fig 3.5. Comparison of the social welfare for each of 30 runs over the four policies under TSPPD operating with (a) Christofides' algorithm vs (b) CPLEX.

As **Figure 3.5** and **Table 3.5** illustrate, the non-myopic allocation ($\beta = 0.5$) can improve over the myopic allocation policy by 15% in social welfare, which reaffirms the results from Hyytiä et al. (2012). Furthermore, the non-myopic pricing can have an even greater effect. The non-myopic pricing with myopic allocation performs better than myopic case by 31%, while the non-myopic pricing and allocation outperforms the purely myopic policy on average by 35% (4% better than the non-myopic pricing with myopic allocation).

Table 3.5. Summary of system cost, user cost, and social welfare averaged over 30 runs

	Marginal pricing		Non-myopic pricing	
	$\theta = 0.4,$ $\beta = 0$	$\theta = 0.4,$ $\beta = 0.5$	$\theta = 0.4,$ $\beta = 0$	$\theta = 0.4,$ $\beta = 0.5$
Avg. System Cost per User	30.838	26.289	24.330	22.995
Avg. User Cost per User	30.288	27.960	25.332	23.616
Avg. Social welfare	14.2170	16.2906	18.6515	19.2391
(using Christofides' algorithm)		(+15%)	(+31%)	(+35%)
Avg. Social welfare	15.4914	18.0210	20.0877	20.7892
(using CPLEX)		(+16%)	(+30%)	(+34%)
Avg. % difference between				
Christofides' algorithm and	8.2%	9.6%	7.1%	7.5%
CPLEX				

The last row of **Table 3.5** also points out the relative gap between running Algorithm 2 with Christofides' algorithm as opposed to solving with CPLEX. The difference of only 7.1% to 9.6% is acceptable noise when compared to the 15% to 35% improvements in the non-myopic policies.

Moreover, we find that whether or not we use CPLEX or Christofides algorithm in the TSPPD tour construction, the average relative comparison of the myopic policies with the non-myopic policies are approximately stable (+15% vs +16%, +31% vs +30%, or +35% vs +34%). This suggests that policy performances are independent of the noise due to the heuristic suboptimality. In other words, while substituting the subroutine with a better performing heuristic can improve the performance of the overall policy, it appears for the purpose of this study that we need not worry about changes in the relative comparison between policies. This is expected, as any noise caused by suboptimality gaps algorithms applied consistently within both the myopic and non-myopic policies should have its noise cancelled out in a relative comparison.

The remainder of the computational tests are conducted using only Christofides' algorithm for drop off tour construction.

3.4.3. Sensitivity evaluation

We test the sensitivity of the policies under a range of different arrival rates Λ 's, and also over two different θ values. Once again, 30 runs are conducted for each value of Λ to obtain average social welfare values, using the same input parameters as earlier. The results are shown in **Table 3.6**. The relative improvements are independent of both Λ and θ , as these results show.

3.4.4. Elasticity evaluation

We evaluate the elasticity of the social welfare with respect to fleet size, capacity and price, assuming all else remain the same. In this scenario, we use 13 customers, with $\beta = 0.5, \theta = 0.4$.

All the other inputs remain the same as earlier, with 30 simulated runs used to obtain the average results. The results between the marginal pricing policy and non-myopic pricing policy are summarized graphically in **Figure 3.6**.

Based on **Figure 3.6**, the system appears to be perform best for the 13 customers when the fleet size is 3, as the social welfare does not increase after that. This makes sense as there are more than enough vehicles after a certain threshold to handle a given population demand function.

Table 3.6. Summary of sensitivity tests of the policies to Λ and θ

		social welfare ($\beta = 0.5, \theta = 0.40$)		social welfare ($\beta = 0.5, \theta = 0.60$)	
	Λ	Marginal pricing	Non-myopic pricing	Marginal pricing	Non-myopic pricing
1	0.1	12.352	13.112(+6%)	12.943	13.731(+6%)
2	0.11	12.846	13.956(+9%)	14.005	14.951(+7%)
3	0.12	13.651	14.637(+7%)	15.778	16.826(+7%)
4	0.13	13.745	15.010(+9%)	16.347	17.707(+8%)
5	0.14	13.854	15.582(+12%)	16.426	18.828(+15%)
6	0.15	14.124	15.839(+12%)	17.034	19.338(+14%)
7	0.16	14.652	16.147(+10%)	17.576	20.251(+15%)
8	0.17	14.983	16.728(+12%)	19.324	21.853(+13%)
9	0.18	15.156	17.027(+12%)	19.772	22.496(+14%)
10	0.19	15.298	17.036(+11%)	20.447	23.457(+15%)
11	0.2	15.334	17.040(+11%)	22.134	25.554(+15%)
Avg:			+10%		+12%

Fig 3.6. Comparison of elasticity of social welfare to fleet size.

Figure 3.7 evaluates the elasticity of the social welfare under a policy between 0 to 3.5 times the optimal price (represented as “0p” to “3.5p” in the figure). Once again, 30 runs are conducted for each value of price to obtain average social welfare values. Since the price is endogenous to the model, this evaluation is based on changing the prices with respect to the optimal values within the policies. The users’ social welfare appears to be best when it is at the equilibrium point, whereas the optimal price for the myopic marginal pricing policy appears to be ~ 1.5 times higher on average. It seems that the system has major deadweight loss after increasing price to 3.5 times the optimal price. Note that **Figure 3.7** is intentionally zoomed in to better visualize these peaks.

Fig 3.7. Comparison of elasticity of social welfare to price.

Lastly, we compare the elasticity of the social welfare under the two policies for changes in vehicle capacity. For this test, we use the same parameters as in the fleet size sensitivity test, with a fleet of 2 vehicles. The results are summarized in **Table 3.7**. As capacity goes up to 4 – 6 range, the social welfare appears fixed for this setting. Otherwise, the system cost goes down as vehicle capacity goes up. Under elastic demand with maximum arrival rate Λ , there appears to be a maximum capacity from which increasing it further would not improve the welfare.

One interesting finding from this test is that as capacity goes up from 1 to being a shared use service with increasing capacity, the social welfare declines down to a certain point while the system costs also go down. These findings match that of the literature (e.g. Jung et al., 2013), where the decrease in social welfare represents the increased inconvenience to passengers.

Table 3.7. Comparison of elasticity of social welfare to vehicle capacity

Social welfare ($\beta = 0.5, \theta = 0.4$)			System costs	
Capacity	Marginal pricing	Non-myopic pricing	Marginal pricing	Non-myopic pricing
1*	21.589	24.606(+14%)	26.7130	25.1625
2	17.987	20.262(+13%)	24.7304	23.9502
3	17.236	19.312(+12%)	23.9931	23.7672
4	17.184	19.239(+12%)	23.9010	23.7672
5	17.184	19.239(+12%)	23.9010	23.7672
6	17.184	19.239(+12%)	23.9010	23.7672

3.5. Discussion

Service fleets in public transit have evolved over the years because of opportunities enabled by ICT and Big Data. Design of such systems can now consider flexible, dynamic policies. Earlier studies in this area of designing flexible transit systems have considered user data to anticipate future states and make more informed decisions in vehicle allocation, routing, and scheduling. However, when ignoring the elasticity of demand, such policies can result in inefficiencies due to overestimating the improvement in level of service that non-myopic considerations can have.

Our main contributions to this literature are summarized as follows:

- We embed non-myopic pricing into the non-myopic dynamic dial-a-ride problem. This is achieved by replicating and extending the multi-server queue approach proposed by Hyytiä et al. (2012) to include social optimal tolling for queues.

- We experimentally prove that our approach can perform better in terms of social welfare over a range of different scenarios and tests. As a result, we have a policy that can increase social welfare of users and attract the most ridership for the same routing policy.
- We demonstrate that relative performance of the pricing policies are independent of the suboptimality gap caused by using a simple tour construction heuristic. Even as switching from the heuristic to CPLEX resulted in social welfare improvement of 7 – 10%, the relative comparison between policies did not change by more than 1%.
- The price obtained from the pricing policy is experimentally proven to be the optimal price point with respect to the realized social welfare.
- By varying the fleet size, we demonstrate how to design for the optimal fleet size corresponding to the pricing policy in a given scenario.
- By varying vehicle capacity, we evaluate the trade-offs for shared-use service with and without non-myopic policy. We found that the myopic policy agrees with the literature (e.g. Jung et al., 2013), and our study provides the first simulation evaluation of trade-offs for shared-taxis under non-myopic allocation and pricing.

Chapter 4

Arrival process prediction

While the literature suggests there have been a number of recent developments in prediction models of arrival processes, no study has reviewed them and compared their performances using a common data set from a transportation system. Our study serves this purpose by providing a broad review of state of the art methods in two classes of prediction models—online and offline models—and applying them to the same set of taxi demand data so that comparisons and insights can be gained. In this chapter, an empirical experiment will be conducted with five methods for the first time using taxi pickup New York data. The experiment is conducted using the same data set to draw comparisons for arrival process modeling and to provide future urban informatics researchers and mobility service operators with new insights.

4.1. Introduction

The literature review highlights several types of state of the art models for arrival processes. Among the offline models, we have a nonparametric piecewise linear function from Nicol and Leemis (2014) and the time varying Poisson process from Ihler et al. (2006) that was used by Moreira-Matias et al. (2013). Among the online models, we have the IntGARCH model from Matteson et al. (2011), an ARIMA model as illustrated by Li et al. (2012), and a non-parametric Bayesian approach from Adams et al. (2009).

An empirical experiment is conducted with these five methods for the first time using taxi pickup data. The experiment is conducted using the same data set to draw comparisons for arrival process modeling and to provide future urban informatics researchers and mobility service operators with new insights.

4.2. New York City Taxi Data

The empirical experiment is set out to compare the five prediction models using a common data set. Taxi pickup data is used to represent arrival processes of “picked up passengers”. The data is collected by the NY Taxi and Limousine Commission (TLC) (see Gonzales et al., 2014), who logs the GPS data for every taxi trip in the city, and pickup/drop-off locations and times are made available to researchers as an open data source. This open data set is chosen for ease of replication by other researchers, and because existing studies using the data (e.g. Yang and Gonzales, 2014) have not considered online prediction models.

The records show that during 2011, 13,586 taxicabs in New York carried over 172 million trips with passengers. From these 172 million trips over 150 million taxi trips originating or ending in Manhattan (Santi, 2013). Thus, a significant portion of the time and fuel is spent by taxis on non-revenue generating activities without passengers, leading to inefficiencies (Li, 2006). **Figure 4.1** shows the pickup and drop-off of taxi passengers in March, 2013.

a. Pickup passengers

b. Drop-off passengers

Fig 4.1. (a) Pick-up and (b) drop-off of passengers in New York in March 2013 (via QGIS).

The New York City Taxi and Limousine Commission (TLC) was created in 1971, and is the agency responsible for licensing and regulating New York City's medallion (yellow) taxicabs, for-hire vehicles (black cars). TLC also licenses and regulates paratransit vehicles (ambulettes) and commuter vans that are authorized to transport passengers within specific geographic zones. **Table 4.1** shows the number of driver and vehicle licenses in New York City.

As shown in **Table 4.1**, the TLC licenses and regulates over 60,000 vehicles and approximately 132,000 drivers. The data that were obtained from Taxi and Limousine Commission (TLC) were used to evaluate the framework.

Table 4.1. The number of driver and vehicle licenses (TLC Annual Report 2014)

Drivers		Vehicles	
Medallion Drivers	53,801	For-Hire Vehicles	58,295
For-Hire Vehicle Drivers	75,517	Paratransit	1,716
Paratransit Drivers	2,612	Commuter Van	534
Commuter Van Drivers	289	Total Vehicle Licenses	60,545
Total Driver Licenses	132,219		

Each file has about 13 million rows, and each row contains household id, pickup date/time drop-off date/time, passenger count, trip time in seconds, trip distance, and latitude/longitude coordinates for the pickup and drop-off locations.

A	B	C	D	F	G	I	J	K	L	M	N
medallion	hack_license	vendor_id	rate_code	pickup_datetime	dropoff_datetime	trip_time_in_sec	trip_distance	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
007302E11B2860D1571777148E0	VT	1	1	2/1/2013 0:02	2/1/2013 0:17	900	4.55	-73.990334	40.76746	-73.93808	40.796532
4DAA988E50E317911E94E3D581	VT	1	1	2/1/2013 0:04	2/1/2013 0:08	240	0.96	-73.982269	40.727707	-73.983948	40.737598
ID31891D29C4AID5E84EF2D03F0	VT	1	1	2/1/2013 6:43	2/1/2013 6:46	180	1.1	-73.972351	40.782539	-73.95813	40.778664
ID31891D29C4AID5E84EF2D03F0	VT	1	1	2/1/2013 6:49	2/1/2013 6:55	360	0.9	-73.960785	40.772846	-73.969872	40.763245
ID31891D29C4AID5E84EF2D03F0	VT	1	1	2/1/2013 7:02	2/1/2013 7:18	960	7.05	-73.956451	40.775639	-74.006706	40.705547
ID31891D29C4AID5E84EF2D03F0	VT	1	1	2/1/2013 7:25	2/1/2013 7:32	420	1.77	-74.017067	40.708683	-74.009186	40.731129
ID31891D29C4AID5E84EF2D03F0	VT	1	1	2/1/2013 7:35	2/1/2013 7:43	480	1.57	-74.006149	40.735073	-73.987885	40.744907
IE7FD7C99348919FD6D8D926758	VT	1	1	2/1/2013 7:43	2/1/2013 8:07	1440	5.3	-73.993027	40.737007	-73.993027	40.737007
ID31891D29C4AID5E84EF2D03F0	VT	1	1	2/1/2013 7:48	2/1/2013 7:56	480	1.58	-73.990929	40.740021	-73.975937	40.754993
AC5A467F24A5D484D2DA847E1	VT	1	1	2/1/2013 7:57	2/1/2013 8:15	1080	4.73	-73.993179	40.757641	-73.965088	40.806206

Fig 4.2. Format of data.

Figure 4.2 show the format of real data. The data file has about 13 million rows with complete information of taxi pick-up and drop off dates, times, locations (latitude/longitude), fares (including tolls, tips, total fare amounts), and distances travelled. **Figure 4.3** shows the time of day distribution of trips and pickups on October 11 of 2013, categorized by all trips in NYC and all trips originating from the primary borough, Manhattan.

Fig 4.3. A comparison of the number of trips and passengers in NY and Manhattan.

Since the purpose of the study is on the temporal dependency as opposed to the spatial dependencies, all the experiments are conducted using data from a single origin zone—the Lincoln Square (zone ID 3806) in Manhattan. Pick-up and drop-off data are shown in **Figure 4.4**. The Lincoln Square is located on the intersection of Broadway and Columbus Avenue, between West 65th and West 66th streets.

a. Pick-up passengers

b. Drop-off passengers

Fig 4.4. (a) Pick-up and (b) drop-off of passengers at the Lincoln Square in March, 2013.

Now, we can find arrival times and inter arrivals for all passengers. Using excel functions, we could separate event times base on hour, minute and second. **Figure 4.5** shows passenger arrival times.

A	B	C	D	E	F
pickup_datetim	event time	H	M	S	
3/25/2013 0:00	0:01:00	0	1	0	60
3/25/2013 0:00	0:02:00	0	2	0	120
3/25/2013 0:00	0:03:00	0	3	0	180
3/25/2013 0:00	0:03:00	0	3	0	180
3/25/2013 0:00	0:04:00	0	4	0	240
3/25/2013 0:00	0:04:00	0	4	0	240
3/25/2013 0:00	0:04:00	0	4	0	240
3/25/2013 0:00	0:04:00	0	4	0	240
3/25/2013 0:00	0:05:00	0	5	0	300
3/25/2013 0:00	0:05:00	0	5	0	300
3/25/2013 0:00	0:06:00	0	6	0	360
3/25/2013 0:00	0:07:00	0	7	0	420
3/25/2013 0:00	0:10:00	0	10	0	600
3/25/2013 0:00	0:10:00	0	10	0	600
3/25/2013 0:00	0:11:00	0	11	0	660
3/25/2013 0:00	0:11:00	0	11	0	660

Fig4.5 Passenger arrival times per second.

When arrival times can find from data, we can find intensity values for hours and days. **Figure 4.6** illustrates the intensity values for our sample.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
day 1	day 2	day 3	day 4	day 5	day 6	day 7	day 8	day 9	day 10	day 11	day 12	day 13	day 14	day 15	day 16	day 17	day 18	day 19	day 20
70	100	135	133	155	32	17	38	19	23	60	18	156	260	253	104	80	112	148	146
24	28	51	68	92	19	4	19	10	11	36	10	75	82	95	62	36	58	55	75
18	14	22	33	46	14	1	5	3	5	13	14	28	33	48	29	25	16	30	81
11	7	15	19	18	5	0	3	3	1	8	3	15	22	25	9	14	15	15	26
27	16	18	24	24	6	0	5	2	4	5	10	22	32	20	25	12	22	16	22
50	43	40	39	27	10	1	15	1	5	22	19	79	73	83	49	36	30	26	38
139	119	129	110	72	44	9	51	13	45	55	60	217	244	247	121	133	115	127	127
273	257	259	302	150	92	24	82	19	69	144	154	439	503	450	243	270	255	261	238
304	309	334	341	282	99	27	61	23	63	193	126	553	571	448	294	296	269	292	285
318	320	348	335	310	56	17	28	21	71	173	151	523	511	517	274	271	270	326	279

Fig 4.6. Original data: The number of arrivals

We use two sets of samples, one for calibration, and one for validation. For the online models, the validation data is also used to continually train the models within each day. For the experiment, 28 full week days of data are extracted from the file to use as a training set. **Figure 4.7** shows the variation in the number of passengers throughout those 28 days originating from Zone 3806.

Figure 4.8 shows passenger arrivals originating from Lincoln Square over different time periods of a single day, October 7, 2013.

Two hypothesis tests are conducted. In the first one, the chi-square distribution is used to test the goodness of fit to show that our data is following a Poisson distribution. A null hypothesis in this test is set to be $H_0: N \sim \text{Poisson}$. We obtain a p-value of 0.970, which means we cannot reject the Poisson distribution hypothesis and are justified to assume it follows such a distribution.

Fig 4.7. The sample of passenger arrivals-Lincoln Square (Zone Id: 3806)

In the second test, we perform another chi-squared test to see whether the data is nonhomogeneous. In this case, the null hypothesis is set to be $H_0: \lambda(t) = c$. Contrary to the previous test, a low p-value affirms that this data cannot be constant over time. We obtained a p-value of 0.0009, which strongly supports the hypothesis that the data is nonhomogeneous throughout the day. This indicates that the more general NHPP model is required for the NY taxicab data.

Let $x = \{x_{i,1}, x_{i,2}, \dots, x_{i,T}\}$ denote the $n \times T$ matrix of observed counts for each sample day $i = 1, 2, \dots, n$ over each time of day period t . We suppose aggregated customer arrivals, $X_{i,t}$, follow a non-homogeneous Poisson process over each day with $\lambda(t)$ estimated from observed x .

The estimation and application of the models were run with Matlab R2014 and R Software on an Intel Core i5-2450 CPU with 2.5 GHz and 8 GB RAM, running on a 64-bit

Windows 7 operating system. **Section 4.3** provides an overview of the model specifications used for this experiment, and **Section 4.4** presents the results.

a. 12:01:28 AM to 5:59:54 AM
 $\lambda = 38.833$

b. 6:02:39 AM to 11:59:54 AM
 $\lambda = 292.5$

c. 12:00:08 PM to 6:59:41 PM
 $\lambda = 334.6666$

d. 7:00:12 PM to 11:59:16 PM
 $\lambda = 362.666$

Fig 4.8. Observed average arrival rates at Lincoln Square for different time periods on October 7, 2013.

4.3. Model specifications

4.3.1. Offline prediction models

4.3.1.1. Piecewise linear function (Nicol and Leemis, 2014)

Nicol and Leemis (2014) propose a piecewise-linear intensity function (which we label $\lambda_{NL}(z)$) using historical data $x = \{x_{i,1}, x_{i,2}, \dots, x_{i,T}\}$. Daily averages of the historical data are obtained for each time period ($\bar{x} = \{\bar{x}_1, \dots, \bar{x}_T\}$), which are used to fit continuous intensity curves $\lambda_{NL}(z)$ such that $\bar{x}_t = \Lambda_{NL}(t) - \Lambda_{NL}(t-1) = \int_{t-1}^t \lambda_{NL}(x) dx$ for discrete time interval $0 \leq t \leq T$. The precision of the intensity estimation model can be increased by further dividing each time interval into w subintervals. For example, having $w = 2$ with hourly intervals means fitting 48 piecewise linear intensity functions and points where adjacent lines intersect, which we call “knots”. Nicol and Leemis (2014) provide two variations of their method consisting of a cyclic version (where the starting intensity value has to equal the ending intensity value) and a non-cyclic version; we employ the non-cyclic one since we update the model each day in the validation test.

The estimation method is based on the premise of maximizing smoothness or minimizing sharp angles of the intersection of lines at each knot, which turns out to be a quadratic expression, as shown in Eq. (4.1).

$$\min f(Z) = (1/2)Z^T A Z \quad (4.1)$$

where

$$A_{t,j} = \begin{cases} 12w^2, & \text{if } j = t \\ -8w^2, & \text{if } j = (t+1)\text{mod}(Tw), \text{ or } j = (t-1)\text{mod}(Tw) \\ 2w^2, & \text{if } j = (t+2)\text{mod}(Tw), \text{ or } j = (t-2)\text{mod}(Tw) \end{cases}$$

Z is an $(wT + 1)$ vector of intensity values z_j at each knot j in the day and $A_{t,j}$ is a symmetric coefficients for expression in quadratic matrix form. The objective is clearly convex, and subject to the following linear constraints in Eq. (4.2).

Subject to:

$$\bar{x}_t = \frac{\sum_{j=1}^{w-1} (z_{tw+j} + z_{tw+j+1})}{2w}, \quad \forall t \quad (4.2)$$

$$z_j \geq 0 \quad \forall j$$

Eq. (4.1) – (4.2) can be solved using any standard quadratic programming solution method. Due to the structure of the problem, the Karush-Kuhn-Tucker conditions lead to a system of equations that can be solved using any numeric solver. For this study we used the function “*solve*” in *R*.

Once Z is estimated from taxi data x , the intensity function for any future time of day z , $\frac{(j-1)}{w} \leq z \leq \frac{j}{w}$, can forecasted as an NHPP in Eq. (2.12) using the arrival rate from Eq. (4.3).

$$\lambda_{NL}(z) = z_j + w(z_j - z_{j-1})z - j(z_j - z_{j-1}) \quad (4.3)$$

The cumulative intensity function $\Lambda_{NL}(z)$ is shown in expanded form in Eq. (4.4).

$$\begin{aligned} \Lambda_{NL}(z) = \Lambda_{NL}(j-1) + \frac{w}{2} * (z_j - z_{j-1})z^2 + (z_j - j(z_j - z_{j-1}))z \\ - \frac{(j-1)}{w} \left[\frac{(j-1)(z_j - z_{j-1})}{w} + z_j - j(z_j - z_{j-1}) \right] \end{aligned} \quad (4.4)$$

4.3.1.2. Temporal seasonality factors (Ihler et al, 2006; Moreira-Matias et al, 2013)

An alternative approach is to estimate a set of randomly sampled seasonality factors. Moreira-Matias et al. (2013) compared such an approach to a time series method for taxi data, and Ihler et al. (2006) used the factors to estimate the baseline trend portion of a “Markov Poisson” model. The estimated intensity rate (which we’ll call $\lambda_I(t)$) for a given time of day t is obtained by the product of a set of seasonality factors as shown in Eq. (4.5).

$$\lambda_I(t) = \lambda_0 \delta_{d(t)} \eta_{d(t),h(t)} \quad (4.5)$$

where λ_0 is the baseline rate, $\delta_{d(t)}$ indicates the relative change for day of week $d(t)$, and $\eta_{d(t),h(t)}$ indicates the relative change for time period $h(t)$ on day $d(t)$. Since the factors vary over multiple observations of x and may be updated periodically, the sampled λ_0 is assumed to belong to a Gamma distribution and the sampled seasonality factors from Dirichlet distributions, as shown in Eq. (4.6) per Ihler et al. (2006).

$$\lambda_0 \sim \Gamma(\lambda; a, b), \quad \Gamma(\lambda; a, b) \propto \lambda^{a-1} e^{-b\lambda} \quad (4.6a)$$

$$\frac{1}{5} [\delta_1, \dots, \delta_5] \sim \text{Dir}(\alpha_1^d, \dots, \alpha_5^d) \quad (4.6b)$$

$$\frac{1}{T} [\eta_{j,1}, \dots, \eta_{j,T}] \sim \text{Dir}(\alpha_1^h, \dots, \alpha_T^h) \quad (4.6c)$$

The factors are estimated using maximum likelihood method via the *Dirichlet.MLE* and *gamfit.MLE* functions in *R*. Then, Eq. (2.12) is applied with the estimated $\Lambda_I(t) - \Lambda_I(s)$ from Eq. (4.5) to predict the number of arrivals for a time interval $[s, t]$.

4.3.2. Online prediction models

In online models, the data is no longer a static data set like $x = \{x_{i,1}, x_{i,2}, \dots, x_{i,T}\}$. Instead, it is assumed to be for a previous set of time periods from time period t , i.e. $\{x_{i,t-1}, x_{i,t-2}, \dots, x_{i,t-T}\}$. Based on this information, a prediction is made for the upcoming time period(s).

4.3.2.1. The ARIMA model

In an ARIMA model, the future value of a variable is assumed to be a linear function of several past observations and random unobservable errors. This model is fundamentally different from the earlier models not only because it's an online model, but also because the prediction model is

not an NHPP, but a continuous linear regression model. Using historical data $x(t)$, the estimated number of arrivals (which we'll call $X(t)$) for a given time of day t is obtained by Eq. (4.7).

$$X_t = k_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t - k_1 \varepsilon_{t-1} - k_2 \varepsilon_{t-2} - \dots - k_q \varepsilon_{t-q} \quad (4.7)$$

where X_t is the number of arrivals at time period t ; $\varepsilon_t \sim N(0, \sigma_t)$ is the unobservable error assumed to have a Gaussian distribution; ϕ_l ($l = 1, 2, \dots, p$) and k_m ($m = 0, 1, 2, \dots, q$) are the model parameters/weights; and p and q are positive integers that are often referred to as the order of the model. R was employed to estimate the model.

Moreira-Matias et al. (2013) apply this method to predicting taxi passenger arrivals and compare it to offline methods. The results showed the ARIMA did not have favorable results compared to discrete NHPP methods, which concurs with Matteson et al. (2011).

4.3.2.2. IntGARCH model (Matteson et al, 2011)

The integer-valued time series model from Matteson et al. (2011) is designed to overcome the continuous variable flaw from the ARIMA model, and integrate temporal auto-correlation into the Poisson arrival process. Matteson et al. tested several variations of the model and recommended one model as the best performing, which is described below.

Suppose aggregated arrivals follow a non-homogeneous counting process $X_{i,t}$ in discrete time interval t with a latent nonnegative intensity function which we denote $\lambda_M(t)$. The intensity

function is now conditional on the time of day average intensity θ_t and a multiplicative factor R_t to capture non-recurrent effects, as shown in Eq. (4.8).

$$\lambda_M(t) = \theta_t R_t \quad (4.8)$$

Let $\hat{e}_t = X_t / \hat{\theta}_t$ denote the multiplicative residual in period t implied by $\hat{\theta}_t$ fitted to data, which may be obtained using a factor method such as Eq. (4.5) from Ihler et al. (2006). The k-factor model provides a parsimonious representation of nonstationary pattern in high dimensional passenger arrivals, but time series capture the remaining serial dependence in the intensity values. Thus, the k-factor method is combined with time series models to calculate $\hat{\theta}_t$.

A generalized autoregressive time series model is used to model $R(t)$, as shown in Eq. (4.9).

$$R_t = (A_1 + B_1 \hat{e}_{t-1} + C_1 R_{t-1}) I_{\{t \in (t_1, t_2)\}} + (A_2 + B_2 \hat{e}_{t-1} + C_2 R_{t-1}) I_{\{t \notin (t_1, t_2)\}} \quad (4.9)$$

where $A_1, B_1, C_1, A_2, B_2, C_2$ are parameters to be estimated such that $A_1 = 1 - B_1 - C_1$, $B_1 + C_1 < 1$, $A_2 = 1 - B_2 - C_2$, and $B_2 + C_2 < 1$. I is an indicator variable where the threshold boundaries satisfy $0 < t_1 < t_2$. The model features regime switching at deterministic times t_1 and t_2 to allow to more accurately capture peak and off-peak patterns in a day. In the empirical experiment in Section 4.5, a regime is defined over the period $(t_1, t_2) = (10 \text{ h}, 16 \text{ h})$. A maximum likelihood algorithm written a *gam* function (generalized additive models) in a *mgcv* library in *R* is used to estimate the parameters associated with Eq. (4.9), which can be referenced in Matteson et al. (2011). The forecasted R_t is then applied to Eq. (4.8) to obtain the conditional intensity $\lambda_M(t)$ as part of the NHPP. For this study, we used an *R* script provided by Matteson et al. (2011) for inferring and estimating the NHPP.

4.3.2.3. Non-parametric Bayesian approach (Adams et al. 2009)

The last model we consider is a non-parametric Bayesian method from Adams (2009) and Adams et al. (2009). The authors proposed a Sigmoidal Gaussian Cox Process (SGCP) which transforms a Gaussian process into a nonparametric prior distribution of an intensity function for an NHPP. The SGCP is able to find prior and posterior distributions without performing intractable integrals, thus making it a faster algorithm than parametric Bayesian models.

The SGCP transforms a random Gaussian scalar function $g(x)$ into a random intensity function which we denote as $\lambda_A(x)$ as shown in Eq. (4.10).

$$\lambda_A(x) = \lambda^* (1 + e^{-g(x)})^{-1} \quad (4.10)$$

where λ^* is an upper bound on $\lambda_A(x)$ and the logistic function forces the $g(x)$ to be between 0 and 1. Because the NHPP in this case is doubly stochastic, values of $g(x)$ may be sampled from a Gaussian process (GP) to simulate the value of $\lambda_A(x)$ and consequently the distribution of the NHPP.

A Gaussian process is a collection of random variables, any finite number of which has a joint Gaussian distribution (Wilson and Adams; 2013). Using a Gaussian process, we can define a distribution over functions $f(x) \sim gp(m(x), k(x, x'))$, where $x \in R^+$ is an arbitrary input variable, and the mean function $m(x)$ and covariance kernel $k(x, x') = cov(f(x), f(x'))$. As a Bayesian approach, the posterior Gaussian function g is updated from a prior SGCP-based $\lambda_A(x)$ and observed arrivals $\{x_{i,t-1}, x_{i,t-2}, \dots, x_{i,t-T}\}$. The posterior estimate is shown in Eq. (4.11).

$$p(g|\{x_i\}_{i=t-1}^{t-T}) = \frac{\mathcal{GP}(g) \exp\left\{-\int \lambda^*(1 + e^{-g(x)})^{-1} ds\right\} \prod_{i=t-1}^{t-T} \lambda^*(1 + e^{-g(x_i)})^{-1}}{\int \mathcal{GP}(g) dg \exp\left\{-\int \lambda^*(1 + e^{-g(x)})^{-1} ds\right\} \prod_{i=t-1}^{t-T} \lambda^*(1 + e^{-g(x_i)})^{-1}} \quad (4.11)$$

where $\mathcal{GP}(\cdot)$ is a sampling from the GP. Because there are intractable integrals in both the numerator and denominator, Adams et al. (2009) propose using thinning to modify the set of events from $g_T = \{g(s_i)\}_{i=t-1}^{t-T}$ to $g_M = \{g(\tilde{s}_m)\}_{m=1}^M$. The thinning procedure is used to choose which $M \leq T$ points are kept via the Metropolis-Hastings sampling algorithm. The resulting thinned events are used to substitute the events and Hamiltonian Monte Carlo is applied to infer the posterior distribution. The procedure is a gradient search method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. The system is defined by a function of x and q known as Hamiltonian. $H(x, q) = PE(x) + KE(q)$, x is position or state vector and q is momentum vector, $PE(x)$ is called potential energy, and is defined from sampling of the posterior distribution and $KE(q)$ is called the kinetic energy, and defined from $KE(q) = p'p/2$. For additional details of the inference methodology, readers are referred to Adams et al. (2009). For the experiment, we used Matlab scripts provided by Adams et al. (2009) for inferring and simulating the NHPP, the Lightspeed Matlab toolbox version 2.7 by Tom Minka, and the Gaussian Processes for Machine Learning (GPML) Matlab toolbox by Rasmussen and Williams.

4.4. Estimation and validation results

With the five candidate models specified, their parameters are estimated from the same 28-day data set for taxi pickups in zone 3806 in New York. The models are compared in terms of

goodness of fit in **Section 4.4.1** In **Section 4.4.2**, a separate four days of arrivals are used to validate the models to compare their predictive capabilities.

4.4.1. Estimation results

4.4.1.1. Piecewise linear function

Based on 28 days of historical hourly data, the average $\bar{x}_t$ for each t hour is shown below.

104.14, 47.50, 24.18, 12.54, 17.11, 34.71, 108.54, 217.18, 249.71, 248.04, 223.68, 216.79,

239.29, 217.07, 227.36, 222.89, 197.86, 248.11, 331.04, 313.61, 283.61, 296.32, 307.54, 234.64

We assume a subinterval number of $w = 2$ and using Eq. (4.1) – (4.2) we estimate each of the 48 knots z_j as follows.

$z_1 \dots z_{48} =$	[1]	164.47	95.89	60.32	47.00	35.68
	[6]	23.32	14.40	11.21	13.32	17.08
	[11]	20.95	30.92	56.07	104.04	169.99
	[16]	225.45	247.82	249.86	251.31	250.44
	[21]	239.95	222.84	209.09	212.24	233.58
	[26]	246.28	231.00	211.44	214.40	229.66
	[31]	235.71	225.47	204.91	191.70	203.11
	[36]	243.89	301.55	341.58	339.44	312.59
	[41]	289.81	280.53	283.56	295.79	310.16
	[46]	313.57	292.85	240.63		

Since this is a non-parametric approach, there are no parameters estimated. Eq. (4.3) is used with these values of z_j to construct the piecewise linear intensity function $\lambda_{NL}(x)$. In section 4.4.2, we compare the validation of intensity $\lambda_{NL}(x)$ by new samples.

4.4.1.2. Temporal seasonality factors

Using the same sample, we can estimate $\lambda_l(t)$ using temporal seasonality factors. To find parameters δ , η , and λ_0 , we need to calculate DP parameters (α^d, α^h) using the MLE method. **Table 4.2** shows the parameters. The base λ_0 is sampled from gamma distribution with parameters $a = 0.9674$ and $b = 199.1308$. The standard error for a and b is 0.031 and 8.45 with 0.95% confidence intervals (CI) $0.8808 \leq \hat{a} \leq 1.0626$ and $176.3788 \leq \hat{b} \leq 224.8176$.

4.4.1.3. The ARIMA model

The ARIMA model is estimated from the 28 days of data and the best fit model is automatically determined by R with $p = 2$, $d = 1$, and $q = 5$. The parameters lead to the following coefficients and standard error for the error term $\varepsilon_t \sim N(0, \sigma_t)$: $\{\phi_1 = 1.7085, s.e = 0.0129\}$ $\{\phi_2 = -0.9795, s.e = 0.0122\}$ $\{k_1 = -1.5288, s.e = 0.041\}$ $\{k_2 = 0.5216, s.e = 0.072\}$ $k_3 = 0.4187, s.e = 0.0724$ $\{k_4 = -0.0996, s.e = 0.0804\}$ $\{k_5 = -0.0845, s.e = 0.0515\}$ $\{k_0 = 0.0181, s.e = 2.01\}$, and $\sigma_t = 62.2$.

Table 4.2. Temporal seasonality parameters

d		1	2	3	4	5
		$\alpha_1^d = 16.1341$	$\alpha_2^d = 16.8592$	$\alpha_3^d = 17.7187$	$\alpha_4^d = 19.3569$	$\alpha_5^d = 17.8231$
		$\delta_1 =$ 1.100987	$\delta_2 =$ 0.816887	$\delta_3 =$ 0.759256	$\delta_4 =$ 1.192905	$\delta_5 =$ 1.129964
h		$\eta_{1,h}$	$\eta_{2,h}$	$\eta_{3,h}$	$\eta_{4,h}$	$\eta_{5,h}$
1	$\alpha_1^h = 5.4310$	0.5081	0.3999	0.3598	0.5974	0.7429
2	$\alpha_2^h = 2.6943$	0.3802	0.4856	0.4292	0.5135	0.0856
3	$\alpha_3^h = 1.5239$	0.1900	0.4838	0.2166	0.0172	0.1492
4	$\alpha_4^h = 0.4719$	0.0391	0.1979	0.0145	0.0050	0.0677
5	$\alpha_5^h = 0.5408$	0.0990	0.0034	0.0191	0.0031	0.0468
6	$\alpha_6^h = 1.9167$	0.5875	0.1023	0.0789	0.2216	0.0619
7	$\alpha_7^h = 5.8516$	0.4253	0.6991	1.0928	0.7401	0.6091
8	$\alpha_8^h = 11.2327$	0.9026	1.2381	0.9059	0.8109	1.1289
9	$\alpha_9^h = 12.6404$	1.4965	1.3405	1.5470	1.4340	1.5177
10	$\alpha_{10}^h = 11.8221$	1.3482	1.0906	1.4191	0.8903	1.5101
11	$\alpha_{11}^h = 10.8528$	0.7036	1.0301	1.0426	1.1810	1.0974
12	$\alpha_{12}^h = 10.4685$	0.5365	0.8505	1.2323	1.3278	0.7815
13	$\alpha_{13}^h = 11.2838$	1.9921	1.2146	0.9113	1.0400	0.8433
14	$\alpha_{14}^h = 10.3622$	1.0682	1.1944	1.0893	1.7992	0.9834
15	$\alpha_{15}^h = 11.0132$	1.3344	1.4457	1.1463	1.1187	1.2399
16	$\alpha_{16}^h = 10.6479$	1.1061	0.7222	1.0264	1.2254	1.4203
17	$\alpha_{17}^h = 9.3406$	0.7231	1.1572	0.9180	0.9945	0.8425
18	$\alpha_{18}^h = 12.5525$	1.0446	1.0066	1.6069	1.6071	1.3855
19	$\alpha_{19}^h = 16.7169$	1.5927	1.8030	1.3634	2.3089	1.6924
20	$\alpha_{20}^h = 15.9101$	2.2922	1.8211	2.1423	0.9416	1.8184
21	$\alpha_{21}^h = 14.5582$	1.3289	2.2687	1.2265	1.1251	1.5065
22	$\alpha_{22}^h = 15.3918$	1.4189	1.0142	1.5952	1.4000	1.7024
23	$\alpha_{23}^h = 16.2899$	1.9155	1.3677	1.3608	1.7332	1.5869
24	$\alpha_{24}^h = 11.9686$	0.9665	1.0630	1.2558	0.9645	1.1798

4.4.1.4. IntGARCH model

Using the IntGARCH model, we estimated the following parameters, standard error and confidence intervals (CI) 95%: $\{B_1 = 0.074, s.e = 0.003, 0.068 \leq \widehat{B}_1 \leq 0.079\}$, $\{C_1 = 0.7901, s.e = 0.014, 0.76 \leq \widehat{C}_1 \leq 0.82\}$, $\{B_2 = 0.0366, s.e = 0.003, 0.031 \leq \widehat{B}_2 \leq 0.042\}$ and $\{C_2 = 0.906, s.e = 0.015, 0.87 \leq \widehat{C}_2 \leq 0.93\}$. As mentioned earlier, the regime thresholds are $(t_1 = 10 h, t_2 = 16 h)$. We found the factor model fit with $K = 4$ is the most appropriate (with less residual) of the factor models with $K = 1, 2, 3$.

4.4.1.5. Non-parametric Bayesian approach

In the Bayesian approach, an upper bounding parameter λ^* is inferred as part of the MCMC procedure. For example, λ^* is a part of the prior and can be sampled from the conditional posterior distribution which is Gamma distribution. The Gamma distribution with shape parameter α and inverse-scale parameter β provides a conditionally-conjugate prior for λ^* . Online parameters are updated every hour. Thus, we input customer arrival accounts for our sample over a 1-hour. Then, our prediction result was verified with our sample over a 1-hour under the 28 days of samples as shown in **Figure 4.9(a)** as a histogram of the inferred upper bounding intensity λ^* with a mean 84.55. **Figure 4.9(b)** is a histogram of the number of latent thinned events with a mean 306.06. **Figure 4.9(c)** is a histogram of the locations of rejections during the latent history inference with a mean 35.5. **Figure 4.9(d)** is a histogram of the α and β parameters for the beta base density with a mean of 2.84 and 12.26 on α and β parameters.

a. Histogram of sampled λ^* values

b. Histogram of sampled number of thinned events

c. Histogram of the locations of rejections

d. Gamma parameters for resampling λ^*

Fig4.9. Bayesian approach parameters for the NY taxicab data.

4.4.1.6. Comparison of goodness-of-fit

Now, we report the goodness-of-fit measures to compare predictions based on our sample. Matteson et al. (2011) used the following measures to evaluate the fitness values, which we adopt for our comparison. Three types of residuals are computed: multiplicative, Pearson, and Anscombe. Their respective formulas for the Poisson distribution are given by Eq. (4.12).

$$R_{M,t} = \frac{X(t)}{\hat{\lambda}(t)} - 1, R_{p,t} = \frac{(X(t) - \hat{\lambda}(t))}{\sqrt{\hat{\lambda}(t)}}, R_{A,t} = \frac{\frac{3}{2}(X(t)^{2/3} - \hat{\lambda}(t)^{2/3})}{\hat{\lambda}(t)^{1/6}} \quad (4.12)$$

We refer to the root mean square error (RMSE) corresponding to each metric as RMSME (for $R_{M,t}$), RMSPE (for $R_{p,t}$), and RMSAE (for $R_{A,t}$). Our result is shown in **Table 4.3**. The online approaches show that the IntGARCH has an error performance with 25% less than other approaches. The Bayesian and ARIMA approaches have an error performance with 10% less than offline approaches.

Several efficiency metrics shown in **Table 4.3** were used to evaluate the fitted values of the proposed models. Nash-Sutcliffe efficiency (**NSE**), and Modified Nash- Sutcliffe efficiency (**mNSE**) are normalized statistics that determine the relative magnitude of the residual variance ("noise") compared to the measured data variance. The value range is from 0 to 1, where 1 represents a perfect match between modelled to observed data (Nash and Sutcliffe; 1970). These results show that the online approaches have better performance than offline approaches. The Index of Agreement (**d**) and Modified Index of Agreement (**md**) were developed by Willmott (1981) as a standardized measure of the degree of model prediction error and varies between 0 and 1. A value of 1 indicates a perfect match, and 0 indicates no agreement at all (Willmott, 1981). The error performance metrics shows the IntGARCH approach is preferred for the online approaches and the temporal seasonality factors method is preferred for the offline approaches.

Table 4.3. Goodness of fitting metrics

Performance metrics	$\lambda_M(t)$	X_t	$\lambda_A(t)$	$\lambda_{NL}(t)$	$\lambda_I(t)$
RMSME	0.214	0.341	0.332	0.347	0.444
RMSPE	2.701	3.518	3.468	3.598	3.526
RMSAE	2.725	3.391	3.386	3.523	3.434
NSE	0.81	0.86	0.82	0.81	0.8
mNSE	0.64	0.67	0.62	0.55	0.57
d	0.96	0.96	0.95	0.94	0.95
md	0.83	0.82	0.78	0.75	0.78
r	0.92	0.95	0.95	0.92	0.9
R2	0.85	0.9	0.91	0.84	0.81
bR2	0.82	0.8	0.77	0.74	0.76
KGE	0.87	0.86	0.78	0.83	0.89

The Pearson product-moment correlation coefficient (**r**) is a linear correlation between the observed and predicted time series. The Coefficient of Determination (**R2**) is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from observation. The coefficient of determination multiplied by the slope of the linear regression between prediction and observation (**bR2**) allows accounting for the discrepancy in the magnitude of pair data. The results show the IntGARCH approach is a perfect model with a correlation coefficient equal to 0.82.

The Kling-Gupta efficiency (**KGE**) was developed by Gupta et al. (2009) to provide the analysis of the relative importance of *correlation*, *bias* and *variability*. Kling-Gupta efficiencies range from $-\infty$ to 1. The value range is from 0 to 1, where 1 represents a perfect match of the model. The results show IntGARCH and ARIMA approaches are the best matching models with correlation coefficients of 0.87 and 0.86.

4.4.2. Validation of results

In this section, a separate four days ($4 \times 24 = 96$ sample points) of arrivals are used to validate the models to compare their predictive capabilities. **Figure 4.10** shows our results by fitting offline approaches to the data, where the models are re-estimated at the end of each day with the rest of the 28 days of historical day for the following day. **Figure 4.10(a)** shows a comparison on intensity values between true and prediction values. There are more gaps in the first day, and then these approaches are predicted better by updating data. **Figure 4.10(b)** shows the empirical cumulative distribution function (ECDF) to evaluate two approaches using data. The results show that the Temporal seasonality factors is better than the piecewise linear function.

a. A comparison of the predictive capability by the four days

b. The ECDF of offline approaches

Fig 4.10. Fitting the off-line approaches for the Lincoln Square (Zone Id: 3806)

a. A comparison of the predictive capability by the four days

b. The ECDF of online approaches

Fig 4.11. Fitting the on-line approaches for the Lincoln Square (Zone Id: 3806)

Table 4.4. Validation metrics

Performance metrics	$\lambda_M(t)$	X_t	$\lambda_A(t)$	$\lambda_{NL}(t)$	$\lambda_I(t)$
RMSME	0.253	0.301	0.298	0.339	0.305
RMSPE	3.187	3.907	3.317	4.422	3.926
RMSAE	3.115	3.769	3.172	4.151	3.815
NSE	0.86	0.8	0.87	0.77	0.8
mNSE	0.66	0.61	0.71	0.57	0.59
d	0.97	0.94	0.97	0.93	0.95
md	0.83	0.8	0.86	0.76	0.79
r	0.94	0.9	0.94	0.91	0.9
R2	0.88	0.8	0.88	0.84	0.81
bR2	0.83	0.76	0.84	0.71	0.75
KGE	0.92	0.88	0.92	0.8	0.88

Using the data, we consider three online approaches to validate the models to compare intensity values. In this case, the models are re-estimated every hour and applied to forecast the next hour. The ARIMA model (p, d, and q values, and seasonality) was first set by learning and detecting the coefficients and correlations on the 28 days curve of each pick-up node and then updating the model each hour over the 4 days. **Figure 4.11** shows the resulting prediction of the three online approaches applied hourly. **Figure 4.11(a)** shows a comparison on intensity values between true and prediction values. The results show that the online approaches can be updated are more flexible than offline approaches, which makes sense. **Figure 4.11(b)** shows the empirical

cumulative distribution function (ECDF) to evaluate three approaches using data. These result shows that the Bayesian approach is straightforward to implement and demonstrates improvements in forecast performance.

Several efficiency metrics are shown in **Table 4.4** and were used to evaluate the fitted values of the proposed models for the four days. The error and efficiency metrics show the IntGARCH and Bayesian approaches have better performance. **Table 4.5** presents the Kolmogorov-Smirnov values, and our experiences on these approaches.

Using the data, we consider the Bayesian approach to fit a model for six hours. **Figure 4.12** shows the convergence of Bayesian approach to estimate non-homogeneous Poisson process by simulating a set of events from a homogeneous Poisson process with intensity λ_* . The green and blue lines show the mean and simulated runs during the expected number of events, respectively. The red circles show the actual arrival events.

Table 4.5. The ECDF metric and notes on approaches

Model	Ks-test	notes
Temporal seasonality factors $\lambda_1(t)$	k = 0.1875 p-value =0.06 H0=accepted	No guarantee to find the DP parameters. It is a simple and fast method. It is able to generate new points and gradients by DP. It is a discrete method to find intensities.
Piecewise linear function $\lambda_{NL}(t)$	k = 0.2813 p-value =0.0007 H0=rejected	It is a simple and fast method. It requires that the intensity function be continuous and piecewise linear. It requires that the estimator's mean value over an observation interval be equal to the mean of events in that interval. Using cyclic and non-cyclic period such as a day or a week. We have the biggest error by this method.
ARIMA A(t)	k=0.1042 p-value =0.06 H0=accepted	It is a simple and fast estimator without performing intractable computation. It is not flexible by updating new data. It is a short term time series forecasting. No thinning process.
Bayesian approach $\lambda_A(t)$	k= 0.0521 p-value = 0.9992 H0=accepted	It is continuous and infinite-dimensional intensity function- Using thinning process to find exact generated points. It can find prior and posterior distribution without performing intractable integrals. It makes the method faster by Hamiltonian Monte Carlo. Do not need to store more information. Sampling λ^* from conditional posterior distribution which is Gamma (α, β).
IntGARCH $\lambda_M(t)$	k=0.1875 p-value =0.06 H0=accepted	The k-factor model provides a parsimonious representation of high dimensional data. It is a discrete method to find intensities. Do not need to take intractable integrals. No missing observations within a day.

Fig 4.12. Predictive mean of Poisson intensity $\lambda(t)$.

4.5. Discussion

This research presents a comparison of forecasting techniques (online and offline) to predict passenger arrival patterns in transport systems, using taxis as representative of such. To summarize, among the offline approaches, the temporal seasonality model appears to fit the NYC taxi demand data better than the piecewise linear model. Among the online approaches, the IntGARCH model appears to outperform the ARIMA continuous time series model and non-parametric Bayesian model.

The contribution of this study is to serve as a definitive research reference for future studies in urban informatics such as prepositioning of idle vehicles, emergency services, dynamic pricing or dispatching vehicles efficiently, which all require effective arrival process models that depend on real time information.

Chapter 5

Non-myopic prepositioning of idle vehicles

In a demand-responsive smart transit system, user arrivals can be very irregular and only a few users are known in advance to the operator at the start of operations. When vehicles finish dropping off the last user, an idle vehicle can be sent to another location in anticipation of new user arrivals. Proper prepositioning of idle vehicles in this context can reduce idle time to save costs.

An idle vehicle prepositioning policy can be run at the end of every time interval (e.g. 10 minutes) to relocate idle vehicles in a centralized manner. The relocation problem should account for transportation costs of relocation (see Chow and Regan, 2011b). In order to account for the uncertainty in passenger arrivals, look ahead is approximated with steady state queue delay. This leads to a dynamic relocation problem with queueing delay to estimate the future cost of service at a certain location. The bulk of the queueing-based location methods in the literature in Chapter 2 either feature physical queues (e.g. Jung et al., 2014) or they are static location models for maximizing availability of servers. In our case, we are maximizing availability of idle vehicles in a dynamic relocation problem. In short, we extend the p-median relocation problem from Chow and Regan (2011) to incorporate the queueing formulation from Marianov and Serra (2002).

5.1. Non-myopic prepositioning of idle vehicles problem

The following notation is introduced.

Notation:

(i, j) : the number of zones $i = 1 \dots n, j = 1 \dots n$.

m : the number of idle vehicles $m = 1, 2, \dots, C_j$.

Inputs:

λ_i : predicted intensity values of customer demand

μ_j : service rate per vehicle in zone j

α : confidence level.

$\rho_{\alpha jm}$: utilization ratio that corresponds to confidence level α for zone j with m servers

t_{ij} : travel time from zone i to zone j .

r_{ij} : the cost of relocation from zone r to zone s .

θ : is a scalar conversion of the relocation cost to the value of improved deployment time.

C_j : maximum number of taxis allowed at a zone j .

B : fleet of idle taxis.

y_j^0 : total number of idle taxis currently located at zone j .

Outputs:

X_{ij} : a binary variable that indicates customers in zone i are assigned to zone j if equal to 1, 0 otherwise.

Y_{jm} : a binary variable set to 1 if at least m vehicles are serving from zone j , 0 otherwise.

W_{ij} : relocation of vehicles from i to j .

S_i : dummy variable for the surplus.

D_j : dummy variable for the demand.

Problem 1: Non-myopic p-median prepositioning of idle vehicle problem

$$Z = \min \sum_{i \in N} \sum_{j \in N} \lambda_i t_{ij} X_{ij} + \theta \sum_{i \in N} \sum_{j \in N} r_{ij} W_{ij} + \sum_{j \in N} \frac{\sum_i \lambda_i X_{ij}}{\mu_j \sum_m Y_{jm} - \sum_i \lambda_i X_{ij}} \quad (5.1)$$

Subject to:

$$\sum_{j \in N} X_{ij} = 1, \quad \forall i \in N \quad (5.2)$$

$$Y_{jm} \leq Y_{j,m-1}, \quad \forall j, m = 2, 3, \dots, C_j \quad (5.3)$$

$$X_{ij} \leq Y_{j1}, \quad \forall i, j \in N \quad (5.4)$$

$$\sum_{j \in N} \sum_{m=1}^{C_j} Y_{jm} = B \quad (5.5)$$

$$\sum_{j \in N} W_{ij} = S_i, \quad \forall i \in N \quad (5.6)$$

$$\sum_{i \in N} W_{ij} = D_j, \quad \forall j \in N \quad (5.7)$$

$$-D_j - y_j^0 + \sum_{m=1}^{c_j} Y_{jm} \leq 0, \quad \forall j \in N \quad (5.8)$$

$$-S_j + y_j^0 - \sum_{m=1}^{c_j} Y_{jm} \leq 0, \quad \forall j \in N \quad (5.9)$$

$$X_{ij}, Y_{jm} \in \{0,1\} \quad (5.10)$$

$$D_j, S_j, W_{ij} \geq 0 \quad (5.11)$$

Objective function (5.1) is to minimize the travel time between demand points and median points, the sum of relocation costs and the waiting time. Constraint (5.2) allocates each customer demand node to one and only one idle vehicle. Constraint (5.3) ensures the $m-1$ idle vehicles are allocated before allocating the m th idle vehicle. Constraint (5.4) forces the allocation of a demand node only to an idle vehicle. Constraint (5.5) sets the number of available idle vehicles. Constraints (5.6) – (5.9) are used to determine the relocation flows. Constraints (5.6) and (5.7) assign the differences in vehicle locations to supply and demand at each node using dummy variables. Constraints (5.8) and (5.9) are flow conservation constraints. The remaining constraints are non-negativity and binary constraints.

The average waiting time is calculated by $\sum_{j \in N} \frac{\sum_i \lambda_i X_{ij}}{\mu_j \sum_m Y_{jm} - \sum_i \lambda_i X_{ij}}$ and this term is a nonlinear function. Thus, in **Problem 1**, the objective function shows the formulation is a nonlinear optimization problem. Following Marianov and Serra (2002), we convert it to **Problem 2** which is an equivalent linear optimization. Customer demands at each zone i appear according to a Poisson process with intensity λ_i . Marianov & Serra's (2002) model showed the waiting time constraint using the relationship between the arrivals assigned to a server node and the servers as shown in Eq (5.12).

$$\sum_{i \in N} \lambda_i X_{ij} \leq \mu_j \rho_\alpha, \quad \forall j \in N \quad (5.12)$$

ρ_α is determined from Eq (5.13).

$$\sum_{k=0}^{m-1} ((m-k)m! m^b / k!) (1/\rho^{m+b+1-k}) \geq 1/(1-\alpha) \quad (5.13)$$

where b is the number of passenger in line upon a random arrival with probability of at least α .

Problem 1 is replaced with a linear objective Eq (5.14) and constraint (5.15). The constraint is the probabilistic constraint to determine the maximum queue delay with a confidence level α . The left-hand side is the sum of the intensities of the process at the zones served by idle vehicles with a mean distance. The right hand side is limited to the number of servers and at each node and to indicate that the m th server cannot be located to zone j without first locating $(m-1)$ th server. The parameter $\rho_{\alpha jm}$ is the value of ρ_α that is computed when Eq (5.13) is a binding equality.

Problem 2:

$$Z = \min \sum_{i \in N} \sum_{j \in N} \lambda_i t_{ij} X_{ij} + \theta \sum_{i \in N} \sum_{j \in N} r_{ij} W_{ij} \quad (5.14)$$

Subject to:

$$\sum_{i \in N} \lambda_i X_{ij} \leq \mu_j \left[Y_{j1} \rho_{\alpha j 1} + \sum_{m=2}^{C_j} Y_{jm} (\rho_{\alpha j m} - \rho_{\alpha j, m-1}) \right], \quad \forall j \in N \quad (5.15)$$

&

Constraints (5.2 to 5.11)

We provide online recommendations to taxi drivers, which help them to serve customer with a reasonable delay time and reduces the disequilibrium of the demand and supply. The model focuses on the on the off-peak hours to help the taxi drivers make the decision whenever and wherever they want to decide to go.

If we let N_1 be the set of demand nodes, N_2 be the set of server nodes, and C be the uniform server capacity for each node, then the number of design variables is: $(2N_1N_2 + N_2C + N_1 + N_1)$. The total number of constraints is equal to $(3(N_1 + N_2) + N_2C)$. The total number of binary variables is equal to $(N_1N_2 + N_2C)$. Note that the input parameters, $i \in N_1$, $j \in N_2$, and $m \in C_j$, govern the size of the problem with respect to both the design variables and the constraints. The increase in the number of variables and constraints is especially inconvenient due to the fact that the probabilistic constraint has the form of a capacity constraint and capacitated problems are especially difficult to solve. Lagrangian Relaxation (LR) is popular method to solve large-scale optimization problems (Marianov and Serra; 2002). Neebe (1978) proposed the Lagrangian decomposition method with subgradient optimization for the p-median transportation problem. Improved heuristics to solve the problem, such as the Lagrangian

decomposition, may be tested on larger instances of this model. Note that our model is separable into two sub-problems: 1) a p-median problem in the space of the X_{ij} and Y_{jm} , and 2) a transportation problem in the space W_{ij} that is the relocation cost.

5.2. Motivational example for queueing constraint

For a better understanding, we solve our model with and without the queueing constraint. We compute the change in the probability of demand being served at the location and the expected delay of a random customer arriving at each of the nodes under the two solution methods. It should illustrate the model difference from the one without queueing constraint.

For this illustration, we use the following parameters with four zones: $\alpha = 0.95, B = 2, \theta = 0.2, b = 0, C_j = 2 \quad \forall m = 1, 2, y^0 = (2, 0, 0, 0)$, arrival times $\lambda_1 = 4.14, \lambda_2 = 3.30, \lambda_3 = 4.78, \lambda_4 = 5.69$, service times $\mu_1 = 28.60, \mu_2 = 25.25, \mu_3 = 25.97, \mu_4 = 25.87$. **Table 5.1** shows the cost of relocation between the four zones. **Table 5.2** shows travel times from zone i to zone j .

Table 5.1. Relocation costs from zone i to zone j

r_{ij}	1	2	3	4
1	-	0.043	0.032	0.049
2	0.043	-	0.030	0.073
3	0.032	0.030	-	0.077
4	0.049	0.073	0.077	-

Table 5.2. Travel times from zone i to zone j

t_{ij}	1	2	3	4
1	0	0.949568	1.264733	0.637883
2	0.949568	0	0.772835	1.473296
3	1.264733	0.772835	0	0.950179
4	0.637883	1.473296	0.950179	0

Figure 5.1 shows the resulting nodes that are assigned to each server with and without the queueing constraint. We compare the average waiting time with queueing constraint and without queueing constraint. The average waiting time is calculated by $\sum_{j \in N} \frac{\sum_i \lambda_i X_{ij}}{\mu_j \sum_m Y_{jm} - \sum_i \lambda_i X_{ij}}$. One interesting finding from this test is the average delay with queueing constraint is 0.45 and the average delay without queueing constraint is 0.52. This example illustrates the potential differences between two methods. This is expected, the model has better performance by including queueing constraint.

a. The allocation of vehicles to zones without queueing constraint

b. The allocation of vehicles to zones with queueing constraint

Figure 5.1. The resulting nodes that are assigned to each server with the two methods.

5.3. Computational experiments

In this section, we illustrate the capability of the model to capture queue effects under different significance levels ($\alpha = 0.95$, $\alpha = 0.9$, and $\alpha = 0.85$). A network instance with 15 zones is solved using commercial software. The (X, Y) coordinates of 15 zones is shown in **Figure 5.2**.

In the instance, the following parameters are used: $B = 7$, $\theta = 0.2$, $b = 0$, $C_j = 10$ $\forall m = 1, 2, \dots, C_j$, and $y^0 = (0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0)$. Customer arrivals and service rates are

shown in **Table 5.3**. In constraint (5.15), the parameter ρ_{ajm} represent of ρ_α that computes by Eq (6.13) as shown in **Table 5.4**. The distances and relocation are shown in **Table 5.5** and **Table 5.6**.

Figure 5.2.The (X, Y) coordinates of each node.

Table 5.3. Customer arrivals and service rates for each zone

i	λ_i	μ_i	i	λ_i	μ_i
1	14.65	53.82	9	13.85	58.63
2	15.89	51.01	10	11.45	51.39
3	10.05	55.96	11	13.55	52.45
4	11.67	58.99	12	13.74	50.45
5	10.40	58.85	13	11.28	50.32
6	12.89	59.58	14	10.41	51.64
7	13.91	50.14	15	10.38	52.20
8	10.71	54.07			

These problems are solved using LINDO 8, which utilizes the branch and bound approach to integer problem. The algorithm was run on an Intel Core i5-2450 CPU with 2.5 GHz and 8 GB RAM, running on a 64-bit Windows 7 operating system. For each run, we determine optimal zones to allocate vehicles and the total cost. We run 5 runs by three different confidence levels. The summary of the cost and optimal vehicle-zone allocations are shown in **Table 5.7 to Table 5.9**.

Table 5.7 evaluates the optimal vehicle allocation to zones with travel and waiting costs under confidence levels 0.95. It shows that the travel cost goes down by dropping the confidence levels to 0.9 and 0.85 as showing in **Table 5.8** and **Table 5.9**, respectively. On the other hand, the waiting cost goes up by decreasing of confidence levels as well. Our results show that by customer arrivals, how transit agencies can find the optimal service level.

This research addresses two problems. What if we have a function value that told us the value of waiting time at the next node? He may choose a load to move if one is available; alternatively, he has the option of doing nothing or moving empty to another state (even if a load is available). It is very easy to show the relationship between Bellman's equation in Eq (2.1) and the original objective function that we stated in Eq (5.1). We may know the contribution of this decision in current state where transportation cost is current payoff, but we do not know the wait cost is the future expected payoff in the next state (although we may estimate it by customer arrival predictions). Thus, we can evaluate each possible decision x_t and simply choose the decision x_t that has the lowest transportation cost $C_t(R_t, x_t)$, plus the value of delay in next state $R_{t+1} = R(R_t, x_t)$, which we represent using $V_{t+1}(R_{t+1})$.

Table 5.4. The value ρ_{ajm} with different confidence levels

CI	$\alpha = 0.95$	$\alpha = 0.90$	$\alpha = 0.85$
m	$j = 1, 2 \dots 30$	$j = 1, 2 \dots 30$	$j = 1, 2 \dots 30$
1	0.32	0.44	0.54
2	0.64	0.82	0.96
3	1.16	1.43	1.61
4	1.73	2.52	2.20
5	2.66	3.66	3.24
6	3.61	4.85	4.34
7	4.52	5.85	5.31
8	5.35	6.69	6.15
9	6.10	7.43	6.90
10	6.81	8.11	7.60
11	7.48	8.78	8.26
12	8.13	9.43	8.91
13	8.77	10.10	9.57
14	9.42	10.78	10.24
15	10.08	11.49	10.93
16	10.76	12.46	11.79
17	11.45	13.58	12.78
18	12.17	14.76	13.85
19	12.92	15.91	14.95
20	14.55	16.96	15.99

Table 5.5. The cost of relocation from zone i to zone j

r_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	0.04	0.03	0.05	0.03	0.07	0.06	0.08	0.03	0.06	0.06	0.08	0.04	0.09	0.09
2	0.04	-	0.08	0.08	0.05	0.03	0.06	0.08	0.03	0.05	0.06	0.07	0.05	0.05	0.08
3	0.03	0.08	-	0.03	0.07	0.03	0.08	0.07	0.05	0.08	0.08	0.03	0.07	0.04	0.07
4	0.05	0.08	0.03	-	0.02	0.03	0.07	0.04	0.05	0.03	0.04	0.09	0.03	0.05	0.08
5	0.03	0.05	0.07	0.02	-	0.07	0.03	0.03	0.05	0.06	0.07	0.08	0.04	0.07	0.08
6	0.07	0.03	0.03	0.03	0.07	-	0.04	0.06	0.02	0.03	0.06	0.06	0.06	0.07	0.05
7	0.06	0.06	0.08	0.07	0.03	0.04	-	0.09	0.08	0.06	0.05	0.05	0.09	0.03	0.09
8	0.08	0.08	0.07	0.04	0.03	0.06	0.09	-	0.08	0.08	0.02	0.07	0.06	0.07	0.07
9	0.03	0.03	0.05	0.05	0.05	0.02	0.08	0.08	-	0.06	0.08	0.03	0.04	0.04	0.05
10	0.06	0.05	0.08	0.03	0.06	0.03	0.06	0.08	0.06	-	0.05	0.09	0.04	0.08	0.06
11	0.06	0.06	0.08	0.04	0.07	0.06	0.05	0.02	0.08	0.05	-	0.07	0.05	0.03	0.08
12	0.08	0.07	0.03	0.09	0.08	0.06	0.05	0.07	0.03	0.09	0.07	-	0.03	0.04	0.04
13	0.04	0.05	0.07	0.03	0.04	0.06	0.09	0.06	0.04	0.04	0.05	0.03	-	0.03	0.05
14	0.09	0.05	0.04	0.05	0.07	0.07	0.03	0.07	0.04	0.08	0.03	0.04	0.03	-	0.06
15	0.09	0.08	0.07	0.08	0.08	0.05	0.09	0.07	0.05	0.06	0.08	0.04	0.05	0.06	-

Table 5.6. Travel time from zone i to zone j

t_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	0.66	0.76	0.50	0.39	0.88	0.71	0.86	0.85	0.30	0.93	0.53	0.28	0.32	0.53
2	0.66	-	0.66	0.43	0.66	1.38	0.06	0.87	0.73	0.37	1.04	1.14	0.39	0.98	0.65
3	0.76	0.66	-	0.30	0.40	0.97	0.71	0.22	0.09	0.61	0.40	0.92	0.70	1.01	0.25
4	0.50	0.43	0.30	-	0.27	0.99	0.48	0.48	0.39	0.31	0.63	0.82	0.40	0.79	0.22
5	0.39	0.66	0.40	0.27	-	0.72	0.72	0.47	0.48	0.39	0.55	0.55	0.46	0.61	0.15
6	0.88	1.38	0.97	0.99	0.72	-	1.44	0.85	1.00	1.06	0.73	0.40	1.10	0.77	0.78
7	0.71	0.06	0.71	0.48	0.72	1.44	-	0.92	0.77	0.42	1.09	1.19	0.43	1.03	0.71
8	0.86	0.87	0.22	0.48	0.47	0.85	0.92	-	0.19	0.77	0.19	0.89	0.86	1.05	0.33
9	0.85	0.73	0.09	0.39	0.48	1.00	0.77	0.19	-	0.69	0.38	0.98	0.78	1.09	0.33
10	0.30	0.37	0.61	0.31	0.39	1.06	0.42	0.77	0.69	-	0.90	0.78	0.09	0.62	0.46
11	0.93	1.04	0.40	0.63	0.55	0.73	1.09	0.19	0.38	0.90	-	0.85	0.98	1.08	0.44
12	0.53	1.14	0.92	0.82	0.55	0.40	1.19	0.89	0.98	0.78	0.85	-	0.80	0.37	0.68
13	0.28	0.39	0.70	0.40	0.46	1.10	0.43	0.86	0.78	0.09	0.98	0.80	-	0.60	0.54
14	0.32	0.98	1.01	0.79	0.61	0.77	1.03	1.05	1.09	0.62	1.08	0.37	0.60	-	0.75
15	0.53	0.65	0.25	0.22	0.15	0.78	0.71	0.33	0.33	0.46	0.44	0.68	0.54	0.75	-

Table 5.7. Summary of non-myopic prepositioning of idle vehicle with $\alpha = 0.95$

Customer arrivals (λ)			Customer arrivals ($2 * \lambda$)	
	(Zones, vehicle)	Cost(T,W)	(Zones, vehicle)	Cost(T,W)
1	(1,1); (3,1);(8,2);(10,2);(15,1)	(36.992,0.341)	(6,1); (10,6)	(127.849,1.060)
2	(3,2); (4,1);(8,2);(9,1);(14,1)	(38.75,0.345)	(3,6); (4,1)	(137.306,1.060)
3	(3,2); (4,1);(5,1);(8,2);(14,1)	(37.142,0.33)	(3,6); (9,1)	(133.706,0.97)
4	(3,1); (4,1);(5,1);(6,1);(9,1);(14,2)	(30.645,0.293)	(3,5); (4,2)	(108.565,0.822)
5	(3,2); (4,1);(8,2);(9,1);(10,1)	(38.351,0.337)	(6,1); (10,6)	(131.351,1.1)
Avg:		(36.376,0.329)		(127.755,0.995)

Table 5.8. Summary of non-myopic prepositioning of idle vehicle with $\alpha = 0.90$

Customer arrivals (λ)			Customer arrivals ($2 * \lambda$)	
	(Zones, vehicle)	Cost(T,W)	(Zones, vehicle)	Cost(T,W)
1	(2,1); (3,2);(4,1);(8,1);(9,1);(11,1);(13,1)	(28.152,0.340)	(3,4); (4,3)	(115.664,1.04)
2	(5,1); (7,2);(8,1);(10,1);(12,1);(13,1);(15,1)	(27.271,0.340)	(3,4); (8,1); (12,2)	(105.270,1.05)
3	(1,1); (5,2);(7,1);(8,1);(11,1);(13,1);(15,1)	(26.635,0.350)	(1,1); (3,4);(4,1);(5,1)	(100.545,0.954)
4	(1,2);(3,1);(4,1);(6,1);(14,1);(15,1)	(26.056,0.330)	(3,4); (4,2);(9,1)	(98.365,0.928)
5	(1,2);(3,1);(5,1);(9,1);(10,1);(12,1);(14,1)	(27.290,0.340)	(3,4); (4,2);(9,1)	(111.331,1.05)
Avg:		(27.290,0.340)		(106.235,1.004)

5.4. Discussion

We proposed a new formulation for the problem of repositioning idle vehicles in a demand responsive service system. Its innovation lies in incorporating a non-myopic optimization concept with the demand-forecasting component into the formulation of the combined facility

location and transportation problem. Our methodology was a linear optimization formulation under non-myopic policy using online and offline data, explicitly including a constraint on the waiting time and queue length, and thus, on the service quality. We experimentally show that our approach can perform better than a problem ignored queueing constraint. As a result, we have a policy that can decrease delay of users for the same allocation/reallocation policy. At last, we demonstrated optimal location of idle vehicles and total costs by three confidence levels (95%, 90%, and 85%).

Table 5.9. Summary of non-myopic prepositioning of idle vehicle with $\alpha = 0.85$

Customer arrivals (λ)		Customer arrivals ($2 * \lambda$)	
(Zones, vehicle)	Cost(T,W)	(Zones, vehicle)	Cost(T,W)
1 (4,1);(5,1);(6,1);(7,1);(9,1);(10,1); (13,1)	(29.060,0.353)	(1,1);(3,5);(9,1)	(104.902,1.0563)
2 (2,1);(3,1);(4,1);(6,1);(8,1);(9,1); (13,1)	(25.942,0.340)	(3,5);(8,1);(9,1)	(109.684,1.009)
3 (3,1);(4,1);(5,1);(6,1);(9,1);(10,1); (14,1)	(27.705,0.337)	(5,1);(6,1);(10,5)	(104.872,1.028)
4 (3,1);(6,1);(8,1);(9,1);(11,1);(12,1); (14,1)	(27.090,0.338)	(6,1);(8,1);(10,5)	(104.739,1.045)
5 (2,1);(7,1);(8,1);(10,1);(10,1);(14,1); (15,1)	(25.872,0.345)	(3,4);(8,3)	(106.794,0.934)
Avg:	(27.134,0.343)		(106.198,1.014)

Chapter 6

Conclusions and future research

In this final section we give a short summary of findings and list the contributions that have been made in this thesis. A new framework for smart transit systems is proposed, and within this framework we identified three areas of functionality that have research gaps. These three areas are the dispatch function, the arrival prediction, and the idle vehicle prepositioning.

First, during our discussion of the existing dynamic dial-a-ride problem research, they often ignore the need for non-myopic pricing under the assumption of elastic demand, which leads to an overestimation of the benefits in level of service and resulting inefficiencies. To correct this problem, in this thesis a new dynamic dial-a-ride policy is introduced, one that features non-myopic pricing based on optimal tolling of queues to fit with the multi-server queueing approximation method for large-scale systems. We experimentally proved that our approach can perform better in terms of social welfare over a range of different scenarios and tests. As a result, we have a policy that can increase social welfare of users and attract the most ridership for the same routing policy. We demonstrated that relative performance of the pricing policies are independent of the suboptimality gap caused by using a simple tour construction heuristic. Even as switching from the heuristic to exact solution resulted in social welfare improvement of 7 – 10% in the instances tested, the relative comparison between policies did not change by more than 1%. In addition, we illustrated how to design for the optimal fleet size corresponding to the pricing policy in a given scenario. A comparison of single passenger taxis to shared-taxis shows that system cost may reduce at the expense of decreased social welfare,

which agrees with the results of Jung et al. (2013). We provided the first simulation evaluation of trade-offs for shared-taxis under non-myopic allocation and pricing.

Design of such systems can now consider flexible, dynamic policies. Earlier studies in this area of designing flexible transit systems have considered user data to anticipate future states and make more informed decisions in vehicle allocation, routing, and scheduling. However, when ignoring the elasticity of demand, such policies can result in inefficiencies due to overestimating the improvement in level of service that non-myopic considerations can have.

User or passenger arrival predictions models can make use of historical information to make short term forecasts for informed operations. While a number of state-of-the-art methods exist, no study has reviewed their performances using a common data set from a transport service. We accomplish this objective by conducting an empirical experiment with NYC taxi pickup data and five state of the art methods in the literature. Our experimental results provide conclusive evidence that certain methods like IntGARCH are better fitting to these types of arrival processes.

The idle vehicle prepositioning problem is a server relocation problem that can be run periodically to better position vehicles and reduce idle time by anticipating arrivals. Relocation problems, particularly those related to taxis and demand responsive services, have ignored look ahead anticipation. We proposed a new mixed integer linear programming model that uses queueing to approximate the future value function. We demonstrated that the model works by solving it for replicable instances over a range of confidence levels (95%, 90%, and 85%). Our experimental results illustrated how the waiting time goes up by decreasing the confidence levels.

6.1. Future research

There are a number of different directions to take this research:

- **Dynamic dial-a-ride and pricing:** The multi-server queueing approach for now assumes time-invariant arrival rates; more realistic scenarios should observe peaks and demand surges. Furthermore, an infinite horizon (using steady state queueing) might not be appropriate in real world operations, and transient queueing measures (see Chow, 2013) might be more appropriate. Implementation of such a system in real field settings (perhaps as a field test of a last mile solution for a transit hub in a mega-region) and measuring the realized performance would be an important validation of the methodology for market adoption. This methodology could be of great use to private companies like Uber or taxi industry as well.
- **Prediction models:** The real-time information problem introduces a new dimension to make a decision in dynamic models. The online and offline prediction models are an important problem in many transport agencies where we face uncertainty about customer arrival and the value of a decision, but the only way to obtain better estimates of the value is to take the decision. Implementation of these models to anticipate future information in dispatch system, bike sharing and emergency problem would be great use in real world application.
- **Preposition problem:** A dynamic optimization model based on the Bellman equation may be a new version of this model. Implementation of such a system in real field settings (perhaps as a field test of a taxicab service in a mega-region) and measuring the realized performance (myopic and non-myopic) would be an important validation of the

methodology for last mile solution. Improved heuristics to solve the problem such as the Lagrangean decomposition may be tested on larger instances of this model. This methodology could be of great use to emergency services, gas stations, and bike sharing as well.

References

- 1) Adams, R., 2009. Kernel methods for nonparametric Bayesian inference of probability densities and point processes. PhD Dissertation, University of Cambridge.
- 2) Adams, R., Murray, I., MacKay, D., 2009. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. 26th Annual International Conference on Machine Learning. ACM, 9–16.
- 3) Agatz, N.A.H., Erera, L.A., Savelsbergh, M.W.P., Wang, X., 2011. Dynamic ride-sharing: A simulation study in metro Atlanta. *Transportation Research Part B* 45(9), 1450-1464.
- 4) Allahviranloo, M., Chow, J.Y.J., Recker, W.W., 2014. Selective vehicle routing problems under uncertainty without recourse. *Transportation Research Part E* 62 (2014) 68–88.
- 5) Alshalalfah, B., Shalaby, A., 2012. Feasibility of flex-route as a feeder transit service to rail stations in the suburbs: a case study in Toronto. *Journal of Urban Planning and Development* 138(1), 90-100.
- 6) Arjas, E., Gasbarra, D., 1994. Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica* 4, 505–524.
- 7) Arjas, E., Heikkinen, J., 1997. An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Computational Statistics* 12(3), 385–402.
- 8) Arkin, B.L., Leemis, L.M., 2000. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations. *Management Science* 46(7), 989–998.
- 9) Balinski, M., Quandt, R., 1964. On an integer program for a delivery problem. *Operations Research* 12, 300–304.

- 10) Batta, R., 1989. A queueing–location model with expected service time dependent queueing disciplines, *European Journal of Operational Research* 39, 192–205.
- 11) Batta, R., Larson, R., Odoni, A., 1988. A single-server priority queueing–location model, *Networks* 8, 87–103.
- 12) Berbeglia, G., Cordeau, J.F., Laporte, G., 2010. Dynamic pickup and delivery problems. *European Journal of Operational Research* 202(1), 8-15.
- 13) Berman O., Drezner, Z., 2007. The Multiple Server Location Problem, *Journal of the Operational Research Society*, 58, 91-99.
- 14) Berman, O., Hodgson, M.J., Krass, D., 1995. Flow-interception problems. In Z. Drezner (Ed.), *Facility location: A survey of applications and methods*. Springer Series in Operations Research. New York, NY: Springer.
- 15) Berman, O., Larson R., Parkan, C., 1987. The stochastic queue p-median location problem, *Transportation Science* 21, 207–216.
- 16) Berman, O., Larson, R., Chiu, S., 1985. Optimal server location on a network operating as a M/G/1 Queue, *Operations Research* 12, 746–771.
- 17) Berman, O., Larson, R. 1985. Optimal 2-facility network districting in the presence of queueing, *Transportation Science* 19, 261–277.
- 18) Bertsimas, D. J. 1992. A vehicle routing problem with stochastic demands. *Operations Research*; 40:574-585.
- 19) Bertsimas, D.J, Chervi, P., Peterson, M., 1995. Computational approaches to stochastic vehicle routing problems. *Transportation Science*; 29: 342-352
- 20) Brotcorne, L, G. Laporte, and F. Semet, 2003, Ambulance location and relocation models, *European Journal of Operational Research*, 147, 451–463.

- 21) Chandra, S., Quadrifoglio, L., 2013. A model for estimating the optimal cycle length of demand responsive feeder transit services. *Transportation Research Part B* 51, 1-16.
- 22) Chang, H., Tai, Y., Hsu, J.Y., 2010. Context-aware taxi demand hotspots prediction, *International Journal of Business Intelligence and Data Mining*, 5(1), 3-18.
- 23) Chen, H., Schmeiser. B., 2013. I-SMOOTH: Iteratively smoothing mean-constrained and nonnegative piecewise-constant functions. *INFORMS Journal on Computing* 25(3), 432–445.
- 24) Chen, H., Schmeiser., B. W., 2011. I-SMOOTH: Iteratively smoothing piecewise-constant Poisson- process rate functions. In Proceedings of the Winter Simulation Conference, 469–480.
- 25) Chow, J.Y.J., 2013. On observable chaotic maps for queueing analysis. *Transportation Research Record* 2390, 138-147.
- 26) Chow, J.Y.J., 2014. Policy analysis of third party electronic coupons for public transit fares. *Transportation Research Part A* 66, 238-250.
- 27) Chow, J.Y.J., 2014. Trading public transport travel demand for electronic coupons through mobile device fare collection. Proc. 93rd Annual Meeting of the TRB, Washington, DC.
- 28) Chow, J.Y.J., Regan, A.C., 2011a. Network-based real option models. *Transportation Research Part B* 45(4), 682-695.
- 29) Chow, J.Y.J., Regan, A.C., 2011b. Resource location and relocation models with rolling horizon forecasting for wildland fire planning. *INFOR* 49(1), 31-43.
- 30) Chow, J.Y.J., Regan, A.C., Ranaiefar, F., Arkhipov, D.I., 2011. A network option portfolio management framework for adaptive transportation planning. *Transportation Research Part A* 45(8), 765-778.

- 31) Chow, J.Y.J., Sayarshad, H.R., 2015. A network-sensitive reference policy for non-myopic sequential network design and timing problems. Proc. 94th Annual Meeting of the Transportation Research Board, Washington, DC.
- 32) Christofides, N., 1976. Worst-case analysis of a new heuristic for the traveling salesman problem. Management Sciences Research Report No. 388, Carnegie-Mellon University, Pittsburgh, PA.
- 33) Christofides, N., Worst-case analysis of a new heuristic for the travelling salesman problem, Report 388, Graduate School of Industrial Administration, CMU, 1976.
- 34) Church, R., ReVelle, C., 1974. The maximal covering location problem, *Papers of the Regional Science Association* 32, 101–118.
- 35) Clarke, G. & Wright, J.W., 1964. Scheduling of vehicles from a central depot to a number of delivery points, *Operations Research*, 12, 568-581.
- 36) Cordeau, J. F., Laporte, G., 2003. A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transportation Research Part B* 37(6), 579–594.
- 37) Cordeau, J.F., Laporte, G., 2007. The dial-a-ride problem: models and algorithms. *Annals of Operations Research* 153(1), 29-46.
- 38) Cortés, C.E., Jayakrishnan, R., 2002. Design and operational concepts of high-coverage point-to-point transit system. *Transportation Research Record* 1783, 178-187.
- 39) Cortés, C.E., Sáez, D., Núñez, A., Muñoz-Carpintero, D., 2009. Hybrid adaptive predictive control for a dynamic pickup and delivery problem. *Transportation Science* 43(1), 27-42.
- 40) Croes, G. A. (1958). A method for solving traveling salesman problems. *Operations Research*, 6, 791-812.

- 41) d'Orey, P. M., Fernandes, R., 2012. Ferreira, M., Empirical Evaluation of a Dynamic and Distributed Taxi-Sharing System , 15th IEEE Intelligent Transportation Systems Conference, Anchorage, USA, 16-19.
- 42) Dantzig, G., Ramser, J., 1959. The truck dispatching problem. *Management Science* 6, 80–91.
- 43) De Alvarenga, A. G., Negreiros-Gomes, F. J., Mestria, M., 2000. Metaheuristic methods for a class of the facility layout problem. *Journal of Intelligent Manufacturing* 11, 421–430.
- 44) De Borger, B., Fosgerau, M., 2012. Information provision by regulated public transport companies. *Transportation Research Part B* 46(4), 492-510.
- 45) Dial, R.B., 1995. Autonomous dial-a-ride transit introductory overview, *Transportation Research Part C*, 3(5), 261-275.
- 46) Diana, M., Dessouky, M. M., 2004. A new regret insertion heuristic for solving large-scale dial-a-ride problems with time windows. *Transportation Research Part B* 38(6), 539-557.
- 47) Dror, M., 1993. Modeling vehicle routing with uncertain demands as a stochastic program: properties of the corresponding solution. *European Journal of Operational Research* 64(3), 432-441.
- 48) Dror, M., Laporte, G., Trudeau, P., 1989. Vehicle routing with stochastic demands: properties and solution frameworks. *Transportation Science*, 23(3), 166-176.
- 49) Dror, M., Powell, W., 1993. Stochastic and dynamic models in transportation. *Operations Research* 41(1), 11-14.
- 50) Figliozzi, M.A., Mahmassani, H.S., Jaillet, P., 2007. Pricing in dynamic vehicle routing problems. *Transportation Science* 41(3), 302-318.

- 51) Fisher, M. 1995. Vehicle routing. Network routing, Volume 8 of Handbooks in Operations Research and Management Science. M. O. Ball, T. L. Magnanti, C. L. Monma, G. L. Nemhauser (eds.). Elsevier, Amsterdam, The Netherlands.
- 52) Fisher, M.L., Jaikumar, R. 1981. A generalized assignment heuristic for the vehicle routing problem. *Networks* 11, 109–124.
- 53) Fisher, M.L., Jornsten, K.O., Madsen, O.B.G., 1997. Vehicle routing with time windows: two optimization algorithms. *Operations Research* 45 (3), 488–492.
- 54) Fu, L. 1999. Improving paratransit scheduling by accounting for dynamic and stochastic variations in travel time”, *Transportation Research Record* 1666, Transportation Research Board, National Research Council, Washington, D.C., 1999, pp. 74-81.
- 55) Fu, L. 2001. An adaptive routing algorithm for in-vehicle route guidance systems with real-time information, *Transportation Research Part B: Methodological*, Vol. 35(8), 749-765.
- 56) Fu, L., 2002a. A simulation model for evaluating advanced dial-a-ride paratransit systems, *Transportation Research Part A*, 36,291-307.
- 57) Fu, L., 2002b. Scheduling dial-a-ride paratransit under time-varying, stochastic congestion, *Transportation Research Part B*, 36 (6)2002, 485-506.
- 58) Fu, L., 2002c. Planning and Design of Flexible Transit Services, Transportation Research Record 1791, *Transportation Research Board*, National Research Council, Washington, D.C., 59-66.
- 59) Fu, L., Teply, S.,1999. On-line and off-line routing and scheduling of dial-a-ride paratransit vehicles, *Computer-Aided Civil and Infrastructure Engineering* 14 (1999) 309–319
- 60) Fu, L., Xu, Y., 2002. Potential effects of automatic vehicle location and computer-aided dispatch technology on paratransit performance: A simulation study”, *Transportation*

Research Record 1760, *Transportation Research Board*, National Research Council, Washington, D.C., 107-113.

- 61) Furuhata, M., Cohen, L., Koenig, S., Dessouky, M., Ordoñez, F., 2014. Characterizing online cost-sharing mechanisms for demand responsive transport systems. In: *Proc. 13th International Conference on Autonomous Agents and Multiagent Systems*, Paris, France.
- 62) Gendreau, M., Laporte, G., Seguin, R., 1996a. Stochastic vehicle routing. *European Journal of Operational Research* 88 (1), 3–12.
- 63) Gendreau, M., Laporte, G., SeHguin, R., 1995. An exact algorithm for the vehicle routing problem with stochastic demands and customers. *Transportation Science*;29:143-155.
- 64) Gendreau, M., Laporte, G., SeHguin, R., 1996b. A tabu search heuristic for the vehicle routing problem with stochastic demands and customers. *Operations Research*;44:469-477.
- 65) Golden, B. L., A. A. Assad. 1988. *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, The Netherlands.
- 66) Gonzales, E.J., Yang, J., Morgul, E.F., Ozbay, K., 2014. *Modeling Taxi Demand with GPS Data from Taxis and Transit*. MNTRC Report 12-16, Mineta National Transit Research Consortium.
- 67) Gupta, Hoshin V., Harald Kling, Koray K. Yilmaz, Guillermo F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, Volume 377, Issues 1-2, 20 October 2009, Pages 80-91. DOI: 10.1016/j.jhydrol.2009.08.003. ISSN 0022-1694.
- 68) Hadas, Y., Ceder, A., 2008. Multiagent approach for public transit system based on flexible routes. *Transportation Research Record* 2063, 89-96.

- 69) Hakimi, S. L. 1965. Optimum distribution of switching centers on a communications network and some related graph theoretic problems. *Operations Research*. 13,462-475.
- 70) Hakimi, S. L., 1964. Optimal locations of switching centers and absolute centres and medians of a graph. *Operations Research*. 12, 450-459.
- 71) Hakimi, S. L., Hasheshwari, S. H., 1977. Optimum locations on centers in networks. *Operations Research*. 20,967-977.
- 72) Horn, M.E.T., Fleet scheduling and dispatching for demand-responsive passenger services. *Transportation Research Part C* 10(1), 35-63.
- 73) Hosni, H., Naoume-Sawaya, J., Artail, H., 2014. The shared-taxi problem: Formulation and solution methods. *Transportation Research Part B* 70, 303–318.
- 74) Hyttiä, E., Aalto, S., Penttinen, A., Sulonen, R., 2010a. A stochastic model for a vehicle in a dial-a-ride system. *Operations Research Letters* 38(5), 432-435.
- 75) Hyttiä, E., Häme, L., Penttinen, A., Sulonen, R., 2010b. Simulation of a large scale dynamic pickup and delivery problem. In: *Proc. 3rd International ICST Conference on Simulation Tools and Techniques*, Brussels, Belgium.
- 76) Hyttiä, E., Penttinen, A., Sulonen, R., 2012. Non-myopic vehicle and route selection in dynamic DARP with travel time and workload objectives. *Computers & Operations Research* 39(12), 3021-3030.
- 77) Ichoua, S., Gendreau, M., Potvin, J.Y., 2006. Exploiting knowledge about future demands for real-time vehicle dispatching. *Transportation Science* 40(2), 211-225.
- 78) Ihler, A., Hutchins, J., Smyth, P., 2006. Adaptive event detection with time varying Poisson processes. 12th Int. Conf. ACM SIGKDD, 207–216.

- 79) Jaillet, P., 1988. A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research* 36(6), 929–936.
- 80) Javadian, N., Sayarshad H.R., 2011 Using simulated annealing for determination of the capacity of yard stations in a railway industry, *Applied Soft Computing*, 11(2), 1899-1907.
- 81) Jaw, J.J., Odoni, A.R., Psaraftis, H.N., Wilson, N.H.M., 1986. A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. *Transportation Research Part B* 20(3), 243-257.
- 82) Jia, H., Ordonez, F., Dessouky, M., 2007. A modeling framework for facility location of medical services for large-scale emergencies, *IIE Transactions* 39 (1), 41-55.
- 83) Jung, J., Chow, J.Y.J., Jayakrishnan, R., Park, J.Y., 2013. Stochastic dynamic itinerary interception refueling location with queue delay for electric taxi charging stations. *Transportation Research Part C: Emerging Technologies* 40, 123-142.
- 84) Jung, J., Jayakrishnan, R., 2011. High-coverage point-to-point transit: study of path-based vehicle routing through multiple hubs. *Transportation Research Record* 2218, 78-87.
- 85) Jung, J., Jayakrishnan, R., Park, J.Y., 2013. Design and modeling of real-time shared-taxi dispatch algorithms. *Proc. 92nd Transportation Research Record Annual Meeting*, 13, 1798.
- 86) Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R., 2010. Urban cycles and mobility patterns: exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing* 6, 455-466.
- 87) Kirchler, D., Calvo, R.W., 2013. A granular tabu search algorithm for the dial-a-ride problem, *Transportation Research Part B* 56, 120–135.
- 88) Knudsen, N.C., 1972. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* 40(3), 515-528.

- 89) Krnjajić, M., Kottas, A., Draper, D., 2008. Parametric and nonparametric Bayesian model specification: a case study involving models for count data. *Computational Statistics & Data Analysis* 52(4), 2110–2128.
- 90) Kuhl, M.E., Wilson, J.R., Johnson, M.A., 1997. Estimating and simulating Poisson processes having trends and multiple periodicities. *IIE Transactions* 29(3), 201–211.
- 91) Laporte, G., 1992. The traveling salesman problem an overview of exact and approximate algorithms,, *European Journal of Operational Research*, 59, 231–247.
- 92) Laporte, G., 2009. Fifty years of vehicle routing, *Transportation Science*, 43, 408–416.
- 93) Larson, R.C., Odoni, A.R., 1981. *Urban Operations Research*, Prentice-Hall, Inc.
- 94) Leemis, L.M., 1991. Nonparametric estimation of the intensity functions for a non-homogeneous Poisson process. *Management Science* 37(7), 886–900.
- 95) Leemis, L.M., 2004. Nonparametric Estimation and variate generation for a non-homogeneous Poisson process from event count data. *IIE Transactions* 36(12), 1155–1160.
- 96) Li, B., Golin, M. J., Italiano, G. F., Deng, X., & Sohraby, K. 1999. On the optimal placement of web proxies in the Internet. In Proceeding of the INFOCOM ‘99 conference 1282–1290.
- 97) Li, B., Zhang, D., Sun, L., Chen, C., Li, S., Qi, G., Yang. Q. Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference, 63 –68.
- 98) Li, X., Pan, G., Wu, Z., Qi, G., Li, S., Zhang, D., Zhang, W., Wang, Z., 2012. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science in China* 6(1), 111–121.

- 99) Ma, S., Zheng, Y., Wolfson, O., 2013. Ferreira, M., T-Share: A large-scale dynamic taxi ridesharing service, *IEEE 29th International Conference on Data Engineering (ICDE)*, 410 – 421.
- 100) Madsen, O.B.G., Ravn, H.F., Rygaard, J.M., 1995. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research* 60(1), 193-208.
- 101) Marianov, V., ReVelle, C., 1994. The queueing probabilistic location set covering problem and some extensions, *Socio-Economic Planning Sciences* 28, 167–178.
- 102) Marianov, V., ReVelle, C., 1996. The queueing maximum availability location problem, *European Journal of Operational Research* 93, 110–120.
- 103) Marianov, V., Serra, D., 1998. Probabilistic, maximal covering location–allocation models for congested systems, *Journal of Regional Science* 38, 401–424.
- 104) Marianov, V., Serra, D., 2002. Location–allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research* 111, 35–50.
- 105) Martinez, L.M., Correia, G.H.A., Viegas, J.M., 2014. An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal). *Journal of Advanced Transportation*, 49(3), 475–495
- 106) Massey, W.A., Parker, G.A., Whitt, W., 1996. Estimating the parameters of a non-homogeneous Poisson process with linear rate. *Telecommunication systems* 5(2), 361-388.
- 107) Matteson, D.S., McLean, M.W., Woodard, D.B., Henderson, S.G., 2011. Forecasting emergency medical service call arrival rates, *Annals of Applied Statistics* 5(2B), 1379-1406.
- 108) Mendelson, H., Whang, S., 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research* 38(5), 870-883.

- 109) Mitrović-Minić, S., Krishnamurti, R., Laporte, G., 2004. Double-horizon based heuristics for the dynamic pickup and delivery problem with time windows. *Transportation Research Part B* 38(8), 669-685.
- 110) Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., Damas, L., 2013. Predicting taxi-passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems* 14(3), 1393–1402.
- 111) Mosheiov, G., 1994. The traveling salesman problem with pickup and delivery, *European Journal of Operational Research*, 79(2).
- 112) Naor, P., 1969. The regulation of queue size by levying tolls. *Econometrica* 37(1), 15-24.
- 113) Nash, J.E., Sutcliffe, J.V. 1970. River flow forecasting through conceptual models: Part 1- A discussion of principles. *Journal of Hydrology* 10: 282–290.
- 114) Neebe, W.A., 1978. A Branch and Bound Algorithm for the p-Median Transportation Problem. *Journal Operational Research*, 10 (29), 995-989.
- 115) Nicol, D.M., Leemis, L.M., 2014. Continuous piecewise-linear intensity function estimation for non-homogeneous Poisson process count data. Technical Report, Department of Mathematics, the College of William & Mary.
- 116) Nourbakhsh, S.M., Ouyang, Y., 2011. A structured flexible transit system for low demand areas. *Transportation Research Part B* 46(1), 204-216.
- 117) Novoa, C., Storer, R., 2009. An approximate dynamic programming approach for the vehicle routing problem with stochastic demands. *European Journal of Operational Research* 196(2), 509-515.
- 118) Paquette, J., Cordeau, J.F., Laporte, G., Pascoal, M.M.B., 2013. Combining multi criteria analysis and tabu search for dial-a-ride problems. *Transportation Research Part B* 52, 1–16.

- 119) Pasandideh, S.H.R., Niaki, S.T.A., 2012. Genetic application in a facility location problem with random demand within queuing framework. *Journal of Intelligent Manufacturing* 23, 651–659.
- 120) Phithakkitnukoon, S., Veloso, M., Bento, C. Biderman, A., Ratti, C., 2010. Taxi-aware map: Identifying and predicting vacant taxis in the city. In Proc. AMI, 86.
- 121) Powell, J., Huang, Y., Bastani, F., Ji. M., 2011. Towards reducing taxicab cruising time using spatio-temporal profitability maps. In Proceedings of the 12th International Symposium on Advances in Spatial and Temporal Databases, SSTD '11.
- 122) Powell, W. B., 2011. Approximate Dynamic Programming: Solving the curses of dimensionality (2nd ed.), John Wiley and Sons, New York.
- 123) Psaraftis, H.N., 1980. A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. *Transportation Science* 14(2), 130-154.
- 124) Psaraftis, H.N., 1983. Analysis of an $O(N^2)$ heuristic for the single vehicle many-to-many euclidean dial-a-ride problem. *Transportation Research Part B* 17(2), 133–145.
- 125) Psaraftis, H.N., 1995. Dynamic vehicle routing: status and prospects. *Annals of Operations Research* 61(1), 143-164.
- 126) Quadrifoglio, L., Dessouky, M.M., Ordóñez, F., 2008. Mobility allowance shuttle transit services: MIP formulation and strengthening with logic constraints. *European Journal of Operational Research* 185(2), 481-494.
- 127) Regue, R., Recker, W.W., 2014. Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem. *Transportation Research Part E* 72, 192-209.
- 128) ReVelle, C., Swain, S., 1970. Central facilities location, *Geographical Analysis* 2, 30–42.

- 129) Santi, P. Resta, G. Szell, M., 2013. Sobolevsky, S., Strogatz S., Ratti. C. Taxi pooling in New York city: a network-based approach to social sharing problems, working paper.
- 130) Sayarshad, H.R, Ghoseiri K., 2009. A simulated annealing approach for the multi-periodic rail-car fleet sizing problem, *Computers and Operations Research*, 36(6), 1789-1799.
- 131) Sayarshad, H.R., 2010. Using bees algorithm for material handling equipment planning in manufacturing systems, *International Journal of Advanced Manufacturing Technology*, 48(9), 1009-1018.
- 132) Sayarshad, H.R., Chow, J.Y.J., 2015. A scalable non-myopic dynamic dial-a-ride and pricing problem. *Transportation Research Part B*, DOI: 10.1016/j.trb.2015.06.008.
- 133) Sayarshad, H.R., Tavassoli, S., Zhao, F., 2012. A Multi-Periodic Optimization Formulation for Bike Planning and Bike Utilization, *Applied Mathematical Modeling* 36(10), 4944-4951.
- 134) Secomandi, N., 2001. A rollout policy for the vehicle routing problem with stochastic demands. *Operations Research* 49(5), 796-802.
- 135) Spivey, M.Z., Powell, W.B., 2004. The dynamic assignment problem. *Transportation Science* 38(4), 399-419.
- 136) Stewart, J., Golden, B.L., 1983. Stochastic vehicle routing: a comprehensive approach. *European Journal of Operational Research* 14 (4), 371–385.
- 137) Swain, R., 1974. A parametric decomposition algorithm for the solution of uncapacitated location problems, *Management Science* 21, 189-198.
- 138) Thien, N.D., 2013. Fair cost sharing auction mechanisms in last mile ridesharing. Master Thesis, Singapore Management University.

- 139) Thomas, B.W., 2007. Waiting strategies for anticipating service requests from known customer locations. *Transportation Science* 41(3), 319-331.
- 140) Thomas, B.W., White III, C.C., 2004. Anticipatory route selection. *Transportation Science* 38(4), 473-487.
- 141) Tillman, F., 1969. The multiple terminal delivery problem with probabilistic demands. *Transportation Science* 3, 192-204.
- 142) Toronto's taxicab industry. Taxicab industry review, <http://www.toronto.ca/legdocs/mmis/2012/ls/bgrd/backgroundfile-50094.pdf>, September 2012.
- 143) Wang, H. PhD Thesis, Improving taxi dispatch services with real-time traffic and customer information, National University of Singapore, 2004.
- 144) Weinberg, J., Brown, L. D. and Stroud, J. R. 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association* 102(480), 1185-1198.
- 145) Willmott, C. J. 1981. On the validation of models. *Physical Geography*, 2, 184-194.
- 146) Wilson, A.G., Adams, R.P., 2013. Gaussian process kernels for pattern discovery and extrapolation. *JMLR W&CP* 28(3), 1067-1075.
- 147) Wilson, N.H.M., Colvin, N., 1976. Computer control of the Rochester dial-a-ride system. Technical Report R77-31, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- 148) Wilson, N.H.M., Sussman, J., Wong, H., Higonnet, B., 1971. Scheduling Algorithms for dial-a-ride systems. Technical Report USL TR-70-13, Urban Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

- 149) Wilson, N.H.M., Weissberg, H., 1976. Advanced dial-a-ride algorithms research project: Final report. Technical Report R76-20, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- 150) Wilson, N.H.M., Weissberg, R.W., Hauser, J., 1976. Advanced dial-a-ride algorithms research project: final report, UMTA-MA-11-0024, Massachusetts Institute of Technology, Cambridge, MA.
- 151) Yang, H., Fung, C.S., Wong, K.I., Wong, S.C., 2010. Nonlinear pricing of taxi services. *Transportation Research Part A: Policy and Practice*, 44(5), 337-348.
- 152) Yang, H., Huang, H.J., 1998. Principle of marginal-cost pricing: how does it work in a general road network? *Transportation Research Part A* 32(1), 45-54.
- 153) Yang, M.S., Gonzales, E.J., 2014. Modeling taxi trip demand by time of day in New York City. *Transportation Research Record* 2429, 110-120.
- 154) Yuan, J., Zheng, Y., Zhang, L., Xie, X., Sun, G., 2011. Where to find my next passenger? *Proc. UbiComp 2011*, New York, NY.