July 2016

Modeling Choice Problems with Heterogeneous User Preferences in the Transportation Network

Mahyar Amirgholy

Modeling Choice Problems with Heterogeneous User Preferences in the Transportation Network

A Dissertation Presented

by

MAHYAR AMIRGHOLY

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2016

Civil and Environmental Engineering Department

MODELING CHOICE PROBLEMS WITH HETEROGENEOUS USER PREFERENCES IN THE TRANSPORTATION NETWORK

A Dissertation Presented

by

MAHYAR AMIRGHOLY

Approved as to style and content by:

Dr. Eric J. Gonzales, Chair

Dr. Song Gao, Member

Dr. Anna Nagurney, Member

Richard N. Palmer, Department Head
Civil and Environmental Engineering Department

DEDICATION

To my parents, for their endless love and support.

ACKNOWLEDGMENTS

My last couple of years in Civil and Environmental Engineering Department of University of Massachusetts, Amherst, have afforded me a desirable opportunity to learn, work and grow in an inspiring environment. During this period, I have had the honor to know and work with remarkable people that completion of this PhD could not be possible without their generous support and guidance.

First of all, I want to thank my advisor, Eric J. Gonzales, for being a terrific teacher and a great advisor for me. He patiently taught me to see through problems from the right angles, and use my logic to attain intuitive insight from solving problems. His valuable suggestions have always been the main source of inspiration and encouragement for creativity in my work. Eric has been a brilliant, knowledgeable, and supportive adviser for me and I am deeply thankful to him for his time, energy, and support during these years.

I also want to acknowledge the help and support of my professors at UMass and Rutgers, and especially my committee members, Song Gao and Anna Nagurney for their valuable guidance and comments. It has been my pleasure to know Song and benefit from her advice regarding the future research and careers. I have also had the privilege of learning and receiving advice from Anna Nagurney, which has provided great inspiration to both my research and career goals. I also want to thank all the Transportation Engineering faculty members for instructing me in developing the skills that I require as an independent researcher. In particular, I want to thank Eleni Christofa whose generous advice and support have always been encouraging in achieving my research objectives and career goals, and also Daiheng Ni whose help has been very valuable to me. I am also very

grateful to John Collura and Michael Knodler for their outstanding leadership of the program. I also want especially thank to Thomas O. Boucher from Industrial Engineering Department of Rutgers University whose lectures have been very inspiring to me and his valuable comments have been very helpful in improving the quality of my work. Finally, I would like to thank the Academic Assistant of CEE department, Jodi Ozdarski, and Administrative Manager of UMass Transportation Center, Kris Stetson, who very frequently granted me valuable help.

During all these years, it was the love and support of my parents that has always kept my heart warm, and helped me through critical corners. I am truly grateful to them for believing in me and supporting me to follow and build my dreams in all stages of my life.

ABSTRACT

MODELING HETEROGENEITY OF USER PREFERENCES IN CHOICE PROBLEMS IN THE TRANSPORTATION NETWORK

MAY 2016

MAHYAR AMIRGHOLY,
B.S., AMIRKABIR UNIVERSITY OF TECHNOLOGY, TEHRAN

M.S., SHARIF UNIVERSITY OF TECHNOLOGY, TEHRAN

M.S., RUTGERS, THE STATE UNIVERISY OF NEW JERSEY, NEW BRUNSWICK

Ph.D., UNIVERSITY OF MASSACHUSETTS, AMHERST

Directed by: Eric J. Gonzales

Users of transportation systems need to make a variety of different decisions for their trips in the network, while their objective is to keep the generalized costs of their own trips minimized. In the transportation network, there is a diversity of different factors that can influence the decisions of the users, while the relative importance of these factors varies among the heterogeneous users with different trip purposes. Nonetheless, the cumulative result of the individual decisions of the users seeking to minimize their costs according to their own preferences leads to the user equilibrium condition in which no one can reduce his/her cost by changing his/her decision. In this research, we adapt the concept of the efficient frontier from portfolio theory (Markowitz, 1952) in finance in order to model the bicriterion choice behavior of users with heterogeneous preferences in transportation networks. We show that the efficient frontier has a set of primary properties that remains general in different problems. Thus, the primary properties of the efficient frontier can be

employed to analytically model and solve different bicriterion choice problems in transportation.

For the first application, we use these properties to propose an analytical model for the morning commute problem when there is a heterogeneity associated with preferences of the users (Vickrey, 1969; Daganzo, 1985). A dynamic pricing strategy is also proposed to optimize the bottleneck by minimizing the total cost for users. In addition to the morning commute problem, Vickrey's congestion theory is also shown to have applications in modeling and optimizing the operation of the demand responsive transit (DRT) system with time-dependent demand and state-dependent capacity as queueing systems. The efficiency of the DRT system can be improved by implementing a dynamic pricing strategy. The analytical solution of the morning commute problem can be also extended for modeling and pricing the DRT system when there is a heterogeneity associated with the preferences of the DRT service users.

For another application of the efficient frontier in modeling choice problems in transportation, we propose a traffic assignment model to account for the heterogeneity in sensitivity of the users to travel time reliability in a network under travel time variability. However, the proposed model can have wide applications in modeling the equilibrium condition of different multicriterion choice problems in transportation.

“People take different roads seeking fulfilment and happiness. Just because they’re not on your road doesn’t mean they’ve gotten lost.”

- Dalai Lama

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Motivation and Research Problem

Rapid growth in the demand for transportation systems in the urban networks has given rise to the critical need for the conscious plans and policies that can make efficient use of available resources to maintain and improve the quality of service in the transportation networks. In this respect, designing successful network plans and management policies crucially depends on deep understanding of travel behavior of users in transportation networks. To demonstrate the behaviors of the users of the transportation system, it can be generally considered that rational users tend to follow their own interests by making the decisions that maximize their utilities in the network. Nevertheless, the discrepancy in the decisions that users make in comparable situations in the network can be attributed to the heterogeneity in their preferences due to the dissimilarities in the socioeconomic backgrounds and conditions in which they make these decisions. As a result, it is of great importance to recognize the heterogeneity in preferences of the users as an aspect of their decision-making procedure that should be taken into account in modeling and optimizing the transportation systems.

Transportation system users need to make different decisions regarding the starting time, destination, mode and route of their trip in a network. For each of these decisions, users consider a variety of different factors in their decision-making procedures, while the

relative importance of these factors varies among the heterogeneous users with different trip purposes. Rational users tend to make choices that minimize the cost of their trips according to their own preferences. The cumulative result of the individual decisions of the heterogeneous users is the user equilibrium condition in which no one can reduce his/her cost by changing his/her decisions. In this respect, the focus of this research is to account for heterogeneity in the preferences of users on their cumulative choice behavior in the equilibrium condition of the network. For this purpose, we adapt the concept of the *efficient frontier* from *portfolio theory* (Markowitz, 1952) in finance to represent the equilibrium condition when there is heterogeneity associated with the preferences of users. The efficient frontier can be shown to have primary properties that remain general in different problems. Thus, we make uses of the primary properties of the efficient frontier to model the equilibrium condition of different bicriterion choice problems with heterogeneous user preferences in the transportation network.

The concept of the efficient frontier was introduced for the first time by Markowitz (1952) as a part of portfolio theory in order to formulate the modern investment theory that accounts for the heterogeneity in risk sensitivity of the investors in maximizing the expected return and minimizing the associated risk of their investments in a free market. The investment problem in finance can be shown to have many parallels with bicriterion choice problems in transportation. On this basis, the concept of the efficient frontier can be adapted to develop a bicriterion equilibrium model for different choice problems in transportation.

As the first step, we show that when the generalized cost of the choices can be approximated as a linear combination of the components weighted by their relative importance factors among the heterogeneous users, the efficient frontier of equilibrium choices has primary properties that remain general in different problems. It can be shown that the efficient frontier is the convex hull of the equilibrium choices of the heterogeneous users that remains non-increasing all the time. Moreover, the efficient frontier is shown to have a specific geometric property that is determined by the probability distribution of the preferences of the heterogeneous users. Thus, the primary properties of the efficient frontier are used to model the equilibrium condition of different bicriterion choice problems in a transportation network.

One important decision that users make in the network is what time to start traveling. Vickrey (1969) introduced a model of congestion dynamics based on a first-in, first-out single bottleneck with time-dependent demand and a fixed capacity. Insufficient capacity of bottleneck to meet the demand results in the formation of a queue, which causes users to experience a combination of delay and schedule deviation in their commutes. Rational users tend to minimize the combination of these costs in their own trip by adjusting their arrival times to the bottleneck. However, the relative importance of the components of the cost may vary among the heterogeneous commuters with different trip purposes. Such heterogeneity in preferences of the commuters can be represented by a set of independent probability distributions over the population of the users. Competition between the users in

minimizing their own cost eventually leads to the user equilibrium condition in which no one can reduce his/her cost by changing his/her own arrival time to the bottleneck.

In the first part of this research, we use the concept of the efficient frontier to propose an analytical solution to this bicriterion choice problem when there is heterogeneity associated with preferences of the commuters. On this basis, we analytically approximate the equilibrium arrivals of the heterogeneous commuters to the bottleneck. We also use the results to propose a dynamic pricing pattern to optimize the system by avoiding the formation of a queue, which can also be adapted for designing dynamic pricing for an urban network. The proposed model is also employed to derive a closed form solution for the morning commute problem when the probability distributions of the preferences are uniform. We also use the proposed model to solve a numerical example for different distributions of the schedule penalty preferences. In addition, we provide an explanation for retrieving independent probability distributions of the schedule penalty factors from a joint distribution. We also demonstrate the approximation procedure of the schedule penalty factors of the heterogeneous commuters using empirical data from observing arrival time of the users at the bottleneck.

The congestion theory proposed by Vickrey (1969) can also have applications in analyzing other transportation systems with fixed capacity and time-dependent demand that can be modeled as a queueing system. In this respect, we show that a demand responsive transit (DRT) service can be modeled as a queueing system with limited capacity and time-dependent demand. Accordingly, Vickrey's congestion model can be employed to analyze

and optimize the operation of the DRT system, while the analytical solution derived for the morning commute problem can be extended to account for the heterogeneity in preferences of the DRT service users. The operating cost of a DRT system strictly depends on the quality of service that it offers to its users. The agency seeks to minimize the operating costs, meanwhile keeping the quality of service high for the users. In this research, an analytical model is employed to approximate the agency's operating cost for running a DRT system with dynamic demand and the total generalized cost that users experience as a result of the operating decisions. The approach makes use of Vickrey's (1969) congestion theory to model the dynamics of the DRT system in the equilibrium condition and approximate the generalized cost for users when the operating capacity is inadequate to serve the time-dependent demand over the peak period without excess delay. The efficiency of the DRT system can be improved by optimizing one of three parameters that define the agency's operating decision: 1) the operating capacity of the system, 2) the number of passengers that have requested a pick-up and are awaiting service, and 3) the distribution of requested times for service from the DRT system. Schedule management strategy and dynamic pricing strategies are presented that can be implemented to manage demand and to reduce the total cost of the DRT system by keeping the number of waiting requests optimized over the peak period. Results of different optimization scenarios are also compared in a numerical example.

Another important decision that users make in the network is to choose routes of their trips. Rational users tend to minimize the cost of their trips by choosing the routes with the

shortest travel times; however, they are unable to predict the exact travel times in the network. Instead, they consider the average duration and the reliability of travel time in their route choice decision-making procedure according to the estimations from their previous experiences in the network. However, the relative importance of these factors not only depends on the purpose of the trip, but also varies from one person to another. The cumulative result of the individual route choice decisions of the heterogeneous users eventually leads to an equilibrium condition in which no one has incentive to switch to another route. Accordingly, route choice decision-making of the travelers can be considers as a choice problem with heterogeneous user preferences. In this research, we use the concept of the efficient frontier to model the route choice behavior of the users under travel time variability in the network when there is a heterogeneity associated with preferences of the users regarding risk. The efficient frontier of route choice (EFRC) has specific properties which are employed in a mathematical formulation of the problem. A solution method is also proposed, which employs the analytical properties of the EFRC to provide an efficient numerical solution to the traffic assignment problem. We use the proposed method to solve a numerical example for this problem.

1.2 Literature Review

There are a variety different of decisions that transportation system users make in the network to keep their costs minimized. In their decision-making procedures, users account for different factors in the network; however, the relative importance of these factors may

vary among the heterogeneous users with different trip purposes. In this research, we adapt the concept of the efficient frontier from portfolio theory (Markowitz, 1952), in finance to consider the effect of heterogeneity in preference of the users in the decision-making procedure of transportation network users.

One of these decisions that travelers need to make ahead of their travel is what time to start their trips. This problem is addressed in the congestion theory of the bottlenecks (Vickerey, 1969), while its different aspects have been widely studied in the literature. In this research, we use the primary properties of the efficient frontier to propose an analytical solution to the morning commute problem when there is a heterogeneity associated with preferences of the users. Interestingly enough, the congestion theory can be shown to have applications in modeling systems with limited capacity and time-dependent demand that can be compared to a queueing system. Demand Responsive Transit (DRT) service is one of these systems. In this respect, we make use of the congestion theory to model and optimize the operation of the DRT system.

Route choice is another important decision that users make in the network according to their own preferences. There is a large body of research on a variety of different traffic assignment models that provide more realistic approximations of the route choice behavior of the users in the transportation network. We also employ the concept of the efficient frontier to account for the heterogeneity in risk sensitivity of the users in modeling the route choice behavior of the users in network under travel time variability.

In the following section, we review the literature on the efficient frontier in modern finance theory, the morning commute problem, demand responsive transit systems, the route choice problem.

1.2.1 Efficient Frontier in Modern Finance Theory

Modern finance theory refers to a set of innovations occurred in finance starting in the 1950s for pricing the financial instruments, largely stocks and bonds. The primary components of modern finance theory are *portfolio theory*, *the capital assets pricing model*, and *options pricing theory* (Boucher, 2014).

In formulating a theory of investments that accounts for the trade-off that investors consider between maximizing expected return and minimizing risk, Markowitz (1952) introduced the concept of an *efficient frontier* of investment choices as part of *Portfolio Theory*. According to Portfolio Theory, individuals in a free market with diverse levels of sensitivity to risk may choose different assets depending on the relative levels of risk and return, while no one will invest in assets that have both higher levels of risk and lower levels of return than other assets. The cumulative result of these individual decisions is a set of dominant assets that make up the efficient frontier of the investment choice. The characteristics of efficient frontiers were later discussed and verified in Merton (1972). The concept of the efficient frontier can be adapted for the purpose of modeling bicriterion choice problems in transportation.

A second element of modern finance theory is the *Capital Asset Pricing Model* (CAPM), introduced by Sharpe (1964) to address the risk-return tradeoff for an individual asset in the free market. This theory defines a relative measure of dependency of the asset to the market to determine the appropriate rate of return of the asset. Westin (1973) also employs the CAPM to explain how this measure can be used for evaluating the industrial investment projects under risk. This model is less relevant to transportation networks, because the mechanism for pricing assets in a market is not the same as the mechanism of traffic congestion on a link. We will make use of more realistic network performance measures based on the transportation literature.

A third component of modern finance theory is *option theory*, which addresses option contracts under risk (Black and Scholes, 1973; Merton, 1973). *Real option theory* was introduced by Myers (1977) to apply concepts from option theory to evaluate the alternative courses of action in real investment projects. This application to real decision-making has a strong potential application in modeling transportation systems, because travelers may consider their options to change their decisions in response to real traffic conditions in the network. The applications of the real option theory in analyzing transportation systems is beyond the concept of this research, but it can be good extension for future research in this area.

In this research, we first adapt the concept of the efficient frontier as the equilibrium solution of the bicriterion choice problem with a linear cost function where there is heterogeneity associated with the relative importance of its components. The efficient

frontier can be proven to have specific characteristics that can be generally employed in modeling the bicriterion choice problems in transportation.

1.2.2 Morning Commute Problem

Heterogeneity of preferences among people is an important aspect of choice problems, which should not be overlooked in modeling the decision-making procedure of the users in the transportation network. In reality, heterogeneous users may consider different weights for different factors, which results in a variety of different choices by users facing the same conditions. In this respect, models that take the heterogeneity of user preferences into account can provide a more accurate approximation of the solution of a choice problem.

One important decision that users in a network make is what time to start traveling. Vickrey (1969) introduced a model of congestion dynamics based on a first-in, first-out single bottleneck problem with fixed capacity. Commuters are assumed to have identical preferences for deviating from their preferred scheduled arrival at their destination; i.e. the difference between wished and actual departure times from the bottleneck. When the demand exceeds the capacity of the system it is not physically possible for everyone to be on time. Commuters seek to minimize the combined cost of travel time, queueing delay, and schedule penalty by choosing when to start their trip. Accordingly, the result would be a user equilibrium condition in which no traveler can reduce the travel cost by changing

the start time of his/her own trip. The underlying assumption here is that commuters wish to arrive to their destination punctually, so there is a schedule penalty associated they experience in their commutes with arriving early or late. Hendrickson and Kocur (1981) elaborated the model by considering a cumulative distribution of wished bottleneck departure times for the commuters. Smith (1984) proved the existence of the equilibrium condition in which no commuter has incentive to unilaterally change his/her arrival time when the cumulative wished curve is S-shaped and schedule penalty function is smooth and convex. Daganzo (1985) also shows that such an equilibrium is unique.

In addition to the heterogeneity associated with the schedule preferences of the commuters, the penalty factor that they consider for arriving early or late to their destination also varies with the purpose of the trips among the population of the commuters. The reality that commuters do not all have identical schedule penalty preferences has been empirically shown in the literature (Ott et al, 1980; Small, 1982). In this respect, Henderson (1974, 1977, 1981) first accounted for the heterogeneity in schedule penalty preferences of the travelers in a dynamic congestion model by considering two demand groups with different penalty factors. Newell (1987) considers different schedule penalty factors for non-identical travelers with a continuous distribution of wished departure times in the morning commute problem and concludes that the commuters subject to higher schedule penalty choose their travel time closer to their preferred departure time, while more flexible commuters tend to travel at the edge of the rush. Arnott et al. (1988) derives a similar set of equilibrium results for a finite number of user groups with different schedule penalty

factors, but the same wished departure time. In another study, these authors investigated the effect of the heterogeneity of user preferences on their route choice decisions based on a reduced form of Vickrey's bottleneck model (Arnott et al., 1992). The existence and uniqueness of the equilibrium solution of the single bottleneck problem with multiclass users, who have heterogeneous schedule penalty preferences, is also proved in Lindsey (2004). Nagurney and Zhang (2012) also develops the projected dynamical system which is capable of accounting for the constraints in variational inequality problem. Recent studies have proposed solutions to the heterogeneous problem based on a linear complementary formulation for the single bottleneck problem (Ramadurani et al., 2010) and taking steps to generalize the solution for a discrete set of schedule preferences to an infinitely large number of preference groups (Qian and Zhang, 2011). Liu and Nie (2013) also proposes a semi-analytical solution to the single bottleneck problem with general heterogeneous commuters. Their proposed method employs the analytical solution of the bottleneck problem to present an equivalent assignment model that can include the closed-form cost functions, and they formulate and solve the problem in form of a variatinal inequality problem (VIP). The effect of heterogeneity on equilibrium travel patterns has been shown to affect demand patterns based on observations (Gonzales and Christofa, 2013). In this research, we adapt the concept of the efficient frontier from portfolio theory to propose a general analytical solution for the single bottleneck problem with heterogeneous commuters.

The concept of the efficient frontier was first time introduced as a part of portfolio theory in Markowitz (1952) in order to formulate a theory of investment decisions that accounts for the trade-off that investors consider between maximizing return and minimizing risk. The concept of the efficient frontier can be easily adapted for the purpose of modeling bicriterion problems in transportation where users account for a trade-off between components of the costs in the equilibrium condition. On this basis, Dial (1996) employs the concept of the efficient frontier to propose a bicriterion route choice model with heterogeneous travel demand. In this research, we employ the concept of efficient frontier to account for the heterogeneity of schedule penalty among users to analytically model equilibrium arrival of commuter to the bottleneck. Then, we use this model to propose a dynamic toll strategy that optimizes the system by minimizing the total cost for the commuters.

Pricing policies have been widely studied as an effective demand management strategy for optimizing the arrival of the commuters to the single bottleneck. Arnott et al. (1990) proposes a dynamic pricing pattern that minimizes the total cost for the users by charging a toll equal in value to the difference between the user equilibrium and system optimum costs at each point in time to entirely eliminate the delay. The transition procedure from a flat toll to optimal dynamic toll strategy has also been studied in the literature (Deakin et al., 2011; Barnes et al., 2012). Heterogeneity in value of time (VOT) of the commuters is a key factor in pricing the bottleneck that has been widely explored in the literature. Cohen (1987) first accounted for such heterogeneity in dynamic pricing the bottleneck by

considering two demand groups with different values for time and schedule deviations. This analysis has been elaborated by allowing for heterogeneity in wished departure times as well as the relative importance of the early and late departures (Arnott et al., 1994). Van den Berg and Verhoef (2011) also extends these studies by considering a continuous distribution for the value of time of the users and also a price sensitive demand function. Xiao et al. (2011) studies the effect of the single-step peak toll on the equilibrium solution of the single bottleneck problem, and characterizes the optimum solution of the problem when the value-of-time has a distribution over the population of the commuters.

The morning commute problem has been extensively studied as well for a single bottleneck with a transit mode (Tabuchi, 1993; Braid, 1996; Huang, 2000; Danielis and Marcucci, 2002; Kraus, 2003). Gonzales and Daganzo (2012) presents the user equilibrium solution of the morning commute problem with competing modes, and proposes dynamic toll and fare strategies that can optimize the system by avoiding delay in the bottleneck.

In this part of the research, we employ the concept of the efficient frontier to propose an analytical model for the equilibrium arrival of the users to the bottleneck, where the heterogeneity in preferences of the users are represented by independent probability distributions of the schedule penalty factors over the population of the commuters. The proposed analytical solution can be also used to approximate the equilibrium condition and optimal pricing strategy of the bottleneck for a general S-shaped wished departure curve and any independent distributions of the schedule penalty factors. We employ the proposed model to derive a closed form solution for the equilibrium arrival of the commuters to the

bottleneck assuming the uniform distributions of the schedule penalty preferences are given. A numerical example is also provided as an application of the proposed model for different probability distribution of user preferences. We also propose a method to retrieve independent distributions of preferences of the users from a given joint distribution. The proposed solution of the problem can be also inversely employed to approximate the distribution of preferences from observed arrival time of the users to the bottleneck. In general, solving the morning commute problem analytically provides us intuitions about travel behavior of the heterogeneous commuters, which can be crucial in predicting the change in the overall condition of the bottleneck in case of prospective alteration in the characteristics of the demand or the properties of the bottleneck.

The results of the research can also have applications in modeling and optimizing the network or any other system that can be modeled as a queueing system on an aggregated level.

1.2.3 Demand Responsive Transit Systems

Vickrey's (1969) congestion theory also have applications in modeling transportation systems with limited capacity and time-dependent demand as queueing systems. Demand responsive transit (DRT) service is a type of such systems that can be modeled and optimized by adapting the congestion theory. Thus, the analytical model proposed for the morning commute problem with heterogeneous user preferences can be also adapted for

modeling a DRT system with a state-variable capacity and heterogeneous time-dependent demand.

DRT systems are a class of transit services in which a fleet of vehicles dynamically changes routes and schedules in order to accommodate demand within a service area. A DRT system naturally has flexibility in providing service, which allows it to adapt to variations in the demand. This property of DRT makes it possible to eliminate the access distance for transit users by providing a curb-to-curb trip. The trade-off is that each user of the system must wait for a vehicle to pick them up and spend time in the system while the transit vehicles divert to pick up and drop off other passengers. The quality of service that DRT provides its users can be improved by reducing the waiting and in-service times of the users, as well as earliness and lateness that they experience in arriving to their destinations.

Improvements in the quality of service also tend to raise the operating cost of the service. From the agency's point of view, however, the high operating cost should be justified by the benefits for users; otherwise, the agency's operating cost will be reduced at the expense of a decline in the quality of the service. The resulting inconvenience to the users due to inadequacy of the service quality (e.g., delay in both pickup and service, and lack of punctuality reflected in the earliness or lateness of arrival) contributes to the total costs that users experience in order to use the service. The inherent tradeoff between the quality of the DRT service and its associated operating cost suggests that costs to the agency and users should be balanced through the design and management of the system.

As a result, the optimal balance can be introduced as a condition of the system in which the total cost of the agency and users is minimized.

There are many variables that have a direct or indirect influence on the performance of DRT services. These factors can be classified in three groups: network, operation, and demand characteristics. A large body of research uses different approaches to study the effects of these factors on the performance of the service. These approaches can be classified generally into two groups: simulations and approximate mathematical methods.

Simulation is the appropriate method when the goal is to achieve a high level of precision in results by including all the specific details of the DRT system in the model. This method has been widely used to assess the effects of different operating factors, such as zoning strategies and time windows, on the productivity of the DRT service. Dessouky et al. (2005) employs the simulation method to evaluate the effect of these factors on the total trip miles, deadhead miles, and fleet size of paratransit service in Los Angeles, California. A similar method is also adopted for Houston, Texas, to compare the impacts of centralizing and decentralizing the operations (Quadrifoglio and Shen, 2010) and using transfers for inter-zonal communication (Shen and Quadrifoglio, 2011, 2012). Other simulation studies have addressed schedule reliability (Fu, 1999) and the effect of travel time variations on optimal routing (Fu and Teply, 1999). Although the precision of the model is the main promise of the simulation method, it can be only fulfilled under the condition that detailed data is available as input.

Often, detailed data are not available, and precise vehicle routing solutions may not be particularly useful since there are stochastic variations in the specific locations and times of requested trips. For many types of systematic analysis, it is possible to formulate the problem analytically by replacing detailed data with concise summaries. Although such simplification in the problem can decrease the precision of the results, the ability to obtain an exact analytical solution of the problem can reveal properties of the system by identifying the relationships between involved variables. Having a clear understanding of the tradeoffs at work enables us to quickly obtain insights about the effects of variations in different factors on performance of the system (Daganzo et al., 2012).

There is a large body of research on analytical models of DRT systems. Daganzo (1978) presents an analytical model to approximate the capacity of a DRT system for three different pickup and drop-off strategies. This model accounts for the effect of the order that the operator serves the requests in approximating the capacity of service as a function of demand, operation, and network variables. Rahimi et al. (2014) adapts this analytical model to approximate the fleet size, VHT, and VMT of the DRT system, and calibrates this model for the ADA paratransit service in New Jersey. The study also shows that the operating cost of the service is represented well by a linear combination of these components. Many studies in recent years have used analytical approaches to model the relationships between the parameters and the performance of DRT systems. Nagurney et al. (2002) models the general equilibrium condition in a competitive supply chain network. Fu (2003) extends Daganzo's (1978) model to include additional variables, and uses the model to optimize

total time and fleet size subject to a service quality constraint. Diana et al. (2006) proposes a stochastic model to account for uncertainty in demand. Other models have been developed to estimate the length of near-optimal tours (Daganzo, 1984; Figliozzi, 2008) and to include constraints on tour duration and pickup time windows (Figliozzi, 2009). These models consider demand as exogenous, and seek to optimize the supply of DRT service that at least meets that capacity of the system. There remains a need to identify ways to manage demand that are appropriate for the dynamics of DRT operations.

In real systems, peaks in demand occur during certain times of the day when the rate of requests exceeds the operating capacity. As a result, users of the system can choose to adapt the times when they travel in order reduce the time spent waiting or riding in vehicles, but early or late arrival at their destination may be associated with costs of schedule deviation. In this research, we describe the dynamic equilibrium that is associated with oversaturated conditions in which the demand rate exceeds the operating capacity of the DRT system. The oversaturated DRT system has commonalities with the fixed-capacity bottleneck problem. Vickrey's (1969) congestion model provides a useful structure for analyzing the dynamic equilibrium and approximating the users' costs in DRT systems. As a result, the efficiency of the service can be enhanced by minimizing the total cost for the agency and users by optimizing the operating capacity of the service as well as number of waiting requests.

Conventional paratransit services, such as those in compliance with the Americans with Disabilities Act (ADA) of 1990, provide complementary transit service for people with

disability, typically schedule trips for users in the order that they are requested. By limiting the number of requests per time that can be booked, the operator limits the scheduled demand for trips to the capacity of the system. There are two problems with the current operating strategy. First, service preference is given to users who know their schedule well enough in advance that they can reserve a booking early, while later requests may be forced to incur greater deviations from their preferred schedule due to lack of availability in peak periods. Second, ADA requires that complementary paratransit for people with disabilities schedule a pick up within one hour of the initial requested pick-up time, so very large peaks in demand give operators no choice but to run more vehicles and employ more drivers.

The objective of this part of the research is to present a model and optimization approach for DRT service that is used to minimize the total cost to the agency and users combined. An analytical model for DRT systems based on Rahimi et al. (2014) and Daganzo (1978) is employed to approximate the components of the operating cost of the DRT system: fleet size, total vehicle hours traveled (VHT), and total vehicle miles traveled (VMT) in the network. Given the service area of the DRT system, these components of the agency cost can be approximated as functions of the number of waiting passengers that have requested service and the maximum rate that operators can serve passengers per time (i.e., operating capacity). The total operating cost for the agency can be estimated as a linear combination of these components (Rahimi et al., 2014).

In addition to the expenses to the agency for running the service, it is also necessary to account for the costs that users endure to use the service. To this end, Vickrey's (1969)

congestion theory is adapted to approximate the costs that the DRT users experience for the service when the operating capacity of the system is inadequate to meet the demand. In this case, the user equilibrium can be conceived as the result of competition between DRT users who are each minimizing their own travel costs, which include the waiting time to be picked up, the traveling time in the vehicle, and the cost for arriving earlier or later than preferred. In an equilibrium condition, no one has an incentive to change his/her own travel time. However, it is still possible to reduce the total costs of the system by optimizing the DRT operations and managing the temporal distribution of demand. The capacity of the service and the number of passengers awaiting pick-up are decision variables that can be maintained at optimal levels over time to minimize the total cost of the system. Since demand tends to peak during certain times of the day, an effective demand management strategy that can spread the demand uniformly over time has a key role in optimizing the operations of a DRT system.

In this part of the research, we model the dynamics of DRT system scheduling and operations and identify a management strategy to incentivize users to adapt their request times to be more uniform over time. The system optimum problem is formulated to minimize the total cost of the agency and users in three possible scenarios: optimizing the operating capacity of the service, optimizing the number of waiting requests, and optimizing both together. In these scenarios, the system optimum problem also has an analytical solution when the distribution of the wished request time is known. A dynamic pricing policy can be implemented as an effective strategy to improve the efficiency of the

DRT system by forming a uniform distribution of the demand and avoiding the underutilization of the optimal capacity during off-peak times. As a result, the total cost of the agency and users can be minimized by choosing optimal values for the system capacity and the number of waiting requests. Meanwhile an appropriate demand management strategy that can make the demand uniform is required to keep the system optimized over time. We also provide a numerical example of a DRT system and an optimal pricing strategy as well as a sensitivity analysis on the results.

1.2.4 Route Choice Problem

The route choice decision-making procedure can be considered to have two separate stages. In the first stage, travelers need to observe different routes to collect information regarding performance of different choices. In the second stage, this information is used to evaluate the available route choices. We seek to identify the equilibrium when a heterogeneous population of travelers simultaneously chooses routes based on travel time and reliability. Conventional traffic assignment models simply presume that the trip duration is the only important measure in the assessment of route performance. With perfect information, users are assumed to able to anticipate the exact trip durations for each route. As a result, the interaction between users that choose the route with the shortest travel time will lead to the classic user equilibrium (UE) in which no one can improve his/her travel time by changing his/her own route. In the user equilibrium condition, all the used routes between the origin

and destination have identical travel times while routes with longer travel times remain unused (Wardop, 1952). In spite of the simplicity of the theory, the accuracy of the equilibrium model can be improved by elaborating the underlying assumptions.

First of all, there are aspects other than the duration of the trip that travelers also take into account in their choice evaluation. The importance of such aspects to travelers not only depends on the purpose of the trip, but also varies with the taste or preference of the traveler, from one person to another (Beckmann et al., 1956; Dafermos, 1983). Accordingly, different people may choose different routes to travel from the origin to the destination. Jan et al. (2000) explores a comprehensive GPS dataset from households in Lexington, Kentucky, to demonstrate the consistency of the choices made by the same driver over time, even though these route choices vary among different drivers traveling between very similar origins and destinations. Interestingly enough, it is also concluded that these choices are often quite different from the shortest path. Furthermore, the future cannot be anticipated with certainty, because the travel time is a stochastic process in nature. Travel time variability is due to two main factors: 1) variability of the capacity and 2) variability of demand flows (Lo and Tung, 2003; Jia et al., 2011).

According to the anchoring phenomenon, introduced in Tversky and Kahneman (1974), travelers will be unable to predict the exact travel times that they will experience. Instead, they start making their choices according to their presumptions regarding the route travel times and adjust their estimations by experiencing different choices over time. However, Fujii and Kitamura (2000) shows how gaining information and experience can improve

such estimation. Laboratory experiments conducted to study the learning effect and dynamic adjustment indicate that passing time can diminish the influence of prior experiences on travelers. In other words, more recent experiences have a deeper influence on the travelers (Iida et al., 1992; Polak and Oladeinde, 2000). The day-to-day learning and adjustment process based on real-time information can be modeled in framework of the stochastic assignment procedure (Polak and Hazelton, 1998; Tian et al., 2012; Lu et al., 2014; Ding et al., 2013; Ding and Gao, 2013). In this respect, impacts of real-time information on route choice behavior of users as well as cost of the system are also studied in stochastic time-dependent networks under generalized user equilibrium condition (Gao et al., 2010; Gao, 2012; Gao and Huang, 2012).

Mahmassani and Chang (1987) uses simulation to study the day-to-day adjustments in the route choice and departure time of travelers in response to their most recent experiences. Accordingly, it suggests that such adjustments finally lead to the boundedly rational user equilibrium (BRUE), in which all the travelers are satisfied with their route choices. Although these routes may be different from the shortest path, the travelers are still not incentivized to change their routes to improve their travel times (Lou et al., 2010). Fact that the BRUE problem may not have a unique solution can be interpreted as the variability of the equilibrium flows. As a result, a stochastic traffic assignment model, which treats the travel time as variable, and recognizes taste variation among the users, can improve the consistency of the route choice with the reality of individual travel behaviors.

There is a large body of research on stochastic traffic assignment models that analyzes the route choice behavior of the users from different perspectives. The stochastic user equilibrium model (SUE), presented in Daganzo and Sheffi (1977) is one of the first models that proposes an extension to the user equilibrium by adding a random component to the travel cost functions. Such randomness can represent both errors in system performance and user perceptions. As a result, the probability of the route choice is evaluated by the probability that the route is the shortest path. De Palma et al. (1983) also proposed a stochastic extension to the dynamic user equilibrium model (DUE) for a single bottleneck. Dafermos (1980) accounts for the interaction between different links by formulating the equilibrium route choice problem as set of variational inequalities. The proposed model is also extended by considering the demand elastic in a multimodal network (Dafermos, 1982). Dafermos and Nagurney (1984b) shows that the equilibrium condition continuously depends on the travel demand as well as the route cost functions. Ben-Akiva et al. (1986) used it as the base model to expand the results for a network with multiple bottlenecks located in parallel and series patterns. Ben-Akiva et al. (2012) also proposes a dynamic traffic assignment model to account for effects of long queues and spillbacks according to the real traffic conditions in congested networks. Although some of these models include the inherent randomness associated with travel time, the ability of experienced users to estimate travel time variability is still not explicitly recognized.

In another study, Mirchandani and Soroush (1987) suggests that not only is there uncertainty with route travel times, but travelers are also unable to predict such travel times

accurately. In this framework, the study proposes a generalized traffic equilibrium model (GSTEP), which includes the randomness both in route travel times and perception of the users. To take these effects into account, the route travel time is presumed to have a probability distribution and to vary among different users. Nagurney and Zhang (1997) also proposes a route choice adjustment model formulated as a dynamic system. Nevertheless, none of these models recognizes the risk of arriving late associated with the travel time uncertainty in different routes. Watling (2006) proposes the late arrival penalized user equilibrium (LAPUE) as a general equilibrium model that also takes the schedule delay into account in the route choice problem. However, users are able to minimize the associated risk by choosing from a portfolio of routes. Tian and Gao (2013) develops a probabilistic version of the priority heuristic process to model the route choice behavior of the users under risk in travel times. Levinson and Zho (2013) suggests that users tend to reduce the combination of their journey time and lateness. On this basis, it presents a stochastic traffic assignment model in which users are assumed to choose multiple routes for the same trip on different days, to keep their general trip costs minimized. This model is also supported by GPS data collected in metropolitan Minneapolis–St. Paul, Minnesota. However, the majority of research on traffic assignment problems has been based on the assumption of similarity of taste among the network users, in spite of inherent diversity of preferences among the travelers. Including the effect of heterogeneity in sensitivity of the users to different factors can reveal the unseen motivation of different drivers for taking different routes.

Multi-objective choice models have a variety of applications in the transportation field. In this regard, Dafermos (1972) considers individuals with different cost functions in proposing a multiclass traffic assignment model. On this basis, a multiclass toll pattern is also designed to optimize the network (Dafermos, 1973). In another paper, the same author proposes a multicriteria model that accounts for the effect of heterogeneity in preference of users on their route choice in the equilibrium condition of the transportation network, and shows the existence and uniqueness of the solution (Dafermos, 1983). Dial (1996) proposes a bicriterion traffic assignment model to take both the durations and the costs of the trips into account in forecasting travelers' route choices. It adapts the concept of the efficient frontier to account for heterogeneity of users' values of time, and it generalizes the Frank-Wolfe (1956) algorithm to solve this problem. Recent studies have proposed generalized cost functions that are nonlinear combinations of travel time and tolls (Chen et al., 2010) and use probit-based bicriterion dynamic stochastic user equilibrium models to account for heterogeneity of value of time (Zhang et al., 2013). Nagurney (2000) shows that objectives that decision makers follow in the network might also be in contrast with each other, and rise in one may result in drop in others. Nagurney and Dong (2002) also considers a combination of the travel time and travel cost as the generalized cost of the routes in developing a multiclass multicriteria network equilibrium model. Wang, Jia et al. (2014) assumes that travelers choose their routes among the first several choices with the lowest general disutility (travel time and cost), so it proposes a rank-dependent bicriterion

equilibrium model for the route choice problem when there is stochasticity associated with both the criteria measurements and the subjective preferences, simultaneously.

Route choice modeling with travel time risk is another application of bicriterion traffic assignment models. In this respect, different approaches are proposed in the literature to include the effect of the risk in the travel time in route choice modeling of users. Lo and Tung (2003) assumes the degradation in capacity of the links due to incidents as the primary cause of travel time variations in the network. On this basis, it proposes probabilistic user equilibrium model to account for such uncertainty on long-term route choice behavior of the users in the network. In addition to uncertainty of capacity, the stochastic nature of the route choice decision of individual users, on an aggregate level, causes variations in travel times in the network. In this regard, Chen and Zhou (2010) develops a mean-excess traffic equilibrium model that takes both reliability and uncertainty associated with travel time into account in route choice modeling. Wu and Nie (2011) also considers the heterogeneous sensitivity of users to risk for traffic assignment by linking the first-, second-, and third order stochastic dominance to concepts of insatiability, risk-aversion, and ruin-aversion within the framework of utility maximization. In another study, Nie (2011) considers travelers that choose routes to minimize the travel time budget required to guarantee their on-time arrivals with a certain level of confidence, and proposes a multiclass percentile user equilibrium traffic assignment model. Chen et al. (2011) proposes a solution algorithm for the multiclass reliability-based user equilibrium problem. Xiao and Lo (2013) studies route choice behavior of adaptive drivers in a stochastic network, and the proposed model

optimizes the expected prospect of the choices with acceptable travel time and arrival time. Pothering and Gao (2013) accounts for the heterogeneity in risk attitudes of the users by calibrating a multiclass user equilibrium model using laboratory data. Wang, Ehrgott, and Chen (2014) proposes a general travel time reliability bi-objective user equilibrium model (TTR-BUE), which simultaneously incorporates the travel time budget model (Lo et al., 2006) and late arrival penalty model (Watling, 2006). This model has the advantage that it can identify a range of possible solutions based on the rational behavior of the users, regardless of the distribution of their preferences. Recent work has characterized equilibrium route choices when routes have variability of travel time performances but users are identical. Tan et al. (2014) assumes the link travel times are random variable functions of the link flows under assumption that the variation in travel times are the result of certain exogenous factors. So, route choice behavior of the users has been studied for different types of generalized cost functions under travel time risk. As a result, the Pareto-efficient route flow pattern is introduced as the equilibrium assignment of the users to the non-dominated routes in terms of the mean and standard deviation of travel time. On this basis, the general geometric properties of the mean-standard deviation (ES) indifference curve has been related to the definition of the generalized cost function as a combination of the mean and standard deviation of the travel time. This body of work addresses the need to account for heterogeneity in networks with travel time variability, but there remains a need for an analytical approach to determine the assignment of traffic across routes in a

network when travelers have heterogeneous preferences and the travel times on individual routes exhibit variability.

In this part of the research, we propose a traffic assignment model that includes the effect of travel time reliability on the user's route choice. In the proposed model, we consider a statistical distribution for route travel times. Research shows that travelers consider a generalized cost for their trips in the network that can be approximated as a linear combination of its mean and standard deviation, regardless of the shape of the distribution (Fosgerau and Karlström, 2010; Fosgerau and Engelson, 2011). Both mean and standard deviation of the travel times are assumed to be functions of traffic flows in the network, while there is also a relationship between the mean and standard deviation of the travel time (Fosgerau, 2010; Mahmassani et al., 2013; Noland et al., 1998). Network users are also assumed capable of estimating mean and standard deviation of travel times, while the relative importance of these variables varies among the heterogeneous travelers with different trip purposes.

In this problem, network users are assumed to seek the route that best matches their preferences while the heterogeneity of such preferences among the users is captured in form of a probability distribution of their risk sensitivity. The cumulative result of individual decisions leads to an equilibrium condition in which users have no incentive to change their choices. Such equilibrium can be represented by the concept of the efficient frontier.

The problem of investment decision-making in a financial market has many parallels with the travelers' decision-making in a transportation network. Thus, we use the concept of the efficient frontier to accounts for the fundamental heterogeneity of preferences among traveler in their route choice decision-making procedure. As a result, the concept of the efficient frontier of the route choice (EFRC) is defined based on the distribution of preferences and physically accurate link performance models. In the definition of the EFRC, each user can just choose a single route from a discrete set of choices for his/her trip, as opposed to the portfolio set of different choices in the investment problem. As a result, we define the EFRC as the convex hull that connects the discrete set of the cumulative route choices result in an increasing order of their equilibrium travel time standard deviations in the standard deviation-expected travel time plane. Thus, the concept of the EFRC can be employed to develop a bi-objective traffic assignment model under travel time variability with heterogeneous travelers' sensitivity to risk. The EFRC has specific characteristics, which makes it possible to propose a new formulation and solution algorithm for this problem. A numerical example of a sample network is also included, which compare the result of proposed traffic assignment method with that of conventional approaches.

The proposed model benefits from a basis on the analytical results directly derived from the concept of the efficient frontier. So, the proposed model can provide intuition about the primary characteristics of the solution in the equilibrium condition. In addition, the proposed model can reveal the route choice ranking of the heterogeneous users according

to their own preferences, which makes it possible to predict the consequences of prospective changes in the network on the route choice behavior of the users. The main advantage of the proposed solution algorithm also is that using the known properties of the EFRC as the equilibrium solution of the problem eliminates the need to explicitly find the cheapest path for each user with different sensitivities to travel time risk. As a result, the proposed algorithm can be very efficient for solving the route choice problem under travel time variability when the heterogeneity in the demand is extensive. Although in this part of the research, the focus is on the problem with a single origin-destination network, the solution method can be employed in the existing equilibration algorithms to solve the multi origin-destination problems (Nagurney, 1999).

1.3 Research Contribution

Heterogeneity in preferences of the users is an aspect of the choice problems in transportation that has been mostly overlooked in formulating the transportation choice problems in the literature. In this research, we adapt the concept of the efficient frontier from portfolio theory (Markowitz, 1952) in finance, and introduce the result as an appropriate tool for modeling the choice problems in transportation. In the equilibrium condition of a bicriterion choice problem, it is known that no one can reduce one of the components of his/her costs by changing his/her choice without a rise in the other component. As a result, the population of the heterogeneous users will be distributed along

the efficient frontier of dominant choices. In this respect, we employ the concept of the efficient frontier to represent the equilibrium condition of the choice problems with objectives linearly combined in a generalized cost function, when there is heterogeneity associated with the relative importance of these objectives. We show that the efficient frontier in the equilibrium condition has primary properties that remain general in different problems in transportation.

We first use the concept of the efficient frontier to derive an analytical solution for the morning commute problem with a general S-shaped wished departure curve when the schedule penalty preferences of the users have a probability distribution over the population of the commuters. On this basis, an optimal pricing strategy is proposed that can eliminate the total delay by avoiding the formation of the queue in the bottleneck. We also propose a method to retrieve independent distributions of user preferences from a given joint distribution. The proposed analytical solution of the problem can be inversely used to approximate the distribution of preferences from the empirical data of observed arrival time of the commuters to the bottleneck. The results can also have applications in modeling and optimizing the network or any other transportation system that can be modeled as a queueing system on an aggregated level. In this respect, we show that the DRT system with state-variable capacity and time-dependent demand can be modeled as a queueing system. Thus the analytical solution of the morning commute problem can be extended for analyzing the DRT system when there is a heterogeneity associated with the preference of users. Considering the heterogeneity in preferences of the users not only

allow us to draw a more accurate picture of the system in the equilibrium condition, but also enable us to propose a more effective dynamic pricing pattern for optimizing the system. Moreover, deriving analytical solutions for these problems can provide intuitions about the role of different factors and the interaction between them in the equilibrium condition of the system.

In addition, we use the concept of the efficient frontier to propose a traffic assignment model for route choice behavior of the heterogeneous users in a network under travel time variability. The analytical basis of the proposed model can provide intuitions about the primary characteristic of the equilibrium solution of the problem. Moreover, it can reveal the route choice ranking of users with different sensitivity to reliability of travel time, which can be critical for assessing consequences of any prospective changes in the network or preferences of the users.

1.4 Dissertation Organization

This dissertation is organized as follows. Chapter 2 adapts the concept of the efficient frontier from portfolio theory, and demonstrates its specific properties for a linear cost function. Chapter 3 uses these specific properties to propose an analytical model for the equilibrium arrival of the commuters to the bottleneck when there is a heterogeneity associated with their preferences regarding schedule deviations and penalty factors. On this basis, we propose a dynamic optimal pricing strategy to minimize the total cost of the

traveler by avoiding formation of the queue in the bottleneck. Chapter 4 shows that a demand responsive transit (DRT) system with a state-variable capacity and time-dependent demand can be modeled as a queueing system. Hence, the analytical solution of the morning commute problem can be also adapted to optimize the operation of the DRT system by implementing the proposed pricing strategy. Chapter 5 employs the concept of the efficient frontier to model the route choice behavior of the users with heterogeneous sensitivity to reliability of travel time in a network under travel time variability. Finally, Chapter 6 includes a summary of contributions, conclusions, and proposal of the future extension.

CHAPTER 2

EFFICIENT FRONTIER OF BICRITERION CHOICE PROBLEM WITH HETEROGENEOUS USER PREFERENCES

Decision-making is a multi-criterion optimization process by nature in which individuals tend to minimize their costs or maximize their benefits according to their own preferences. However, the heterogeneity of preferences among users makes the problem complicated. In this section, we adapt the concept of the efficient frontier to represent the equilibrium solution of bicriterion choice problems with linear cost functions in transportation. In the equilibrium condition, no one can reduce both components of the generalized cost by changing his/her choice as drop in one component of the cost corresponds to rise in another one and vice versa. Thus, the population of the users with heterogeneous preferences regarding the relative importance of components of the generalized cost will be distributed along the efficient frontier of dominant choices in the equilibrium condition. Here, we show that this efficient frontier has three specific characteristics that remain general for all of the problems. In the next chapters, we use these characteristics of the efficient frontier to model different bicriterion choice problems with heterogeneous user preferences in the transportation network.

In a decision-making procedure in which different choices come with different costs, users tend to minimize their own disutilities by making choices with lowest possible cost. However, in the equilibrium condition that there is a trade-off between components of the

generalized cost, a drop in one of them will result in a rise in another. In this trade-off, the relative importance between these costs plays an important role in optimizing the balance between these costs. On this basis, the generalized cost function of the choice problem can be defined as the linear combination of these costs, weighted by their relative importance, which varies among the individuals as follows:

$$C_{p,i} = C_i^1 + \alpha_p C_i^2 \tag{2.1}$$

where, C_i^1 and C_i^2 represent the components of the cost associated with the choice i, which can represent different costs in different bicriterion problems. The coefficient α_p denotes the relative importance of these costs for users in the preference group p. To include the heterogeneity of the preferences among the users, we may presume a discrete (green) or continuous (blue) probability distribution over the demand, as illustrated in the Fig 1. As a result, $C_{p,i}$ is the generalized cost of choice i for the users in group p. With this definition of the generalized cost function, users with higher values for α_p give more importance to the second component of the cost, while their sensitivity to the second component diminishes with decrease in α_p.

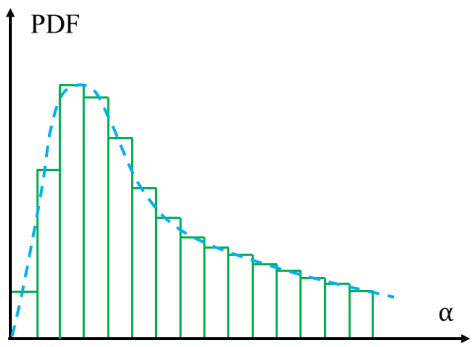


Figure 2.1. Hypothetical probability distribution of the user preferences

In this case, heterogeneous users with different values for α_p find different choices minimizing their generalized costs according to their own preferences. As a result, the concept of the efficient frontier can be employed to represent the set of dominant choices for non-identical users with heterogeneous values for α_p.

2.1 The Efficient Frontier of Choices

The concept of the efficient frontier was introduced in Portfolio Theory by Markowitz (1952) to represent the cumulative result of individual bicriterion decisions of the heterogeneous investors in a free market. The market offers the investor a variety of assets with different levels of return; nonetheless, there is always a risk associated with the investment in the market. So, heterogeneous investors who tend to optimize the balance between risk and return of their investments may find different assets to maximize their benefits according to their own preferences. In this respect, more conservative investors

will prefer assets with a lower level of risk in spite of their lower returns. In contrast, risk-takers will invest in assets with higher levels of return, although their investments might be subject to higher levels of risk as well. However, no one will invest in assets with higher level of risk and lower level of return dominated by an alternative asset with a lower level of risk and higher level of return. As illustrated in Figure 2.2a, the heterogeneity of preferences among the individuals causes them to invest in different assets along the efficient frontier (green points), while no one is inclined to invest in the assets dominated by the efficient frontier (pink points).

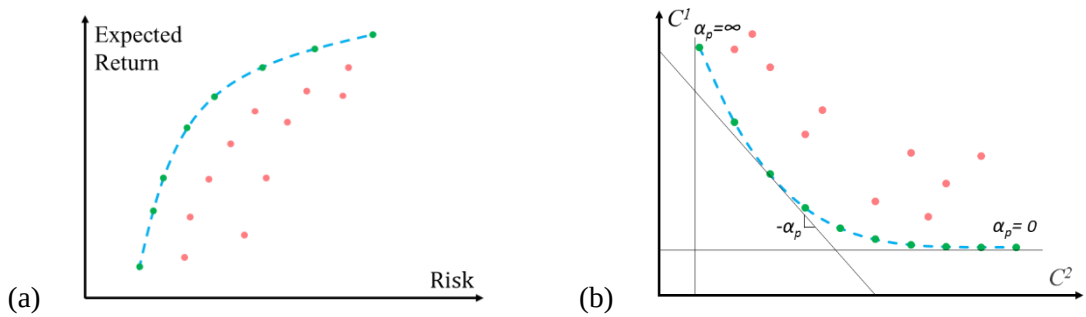


Figure 2.2. The efficient frontier of (a) the dominant assets in a free market (b) the choices with heterogeneous preferences

The analogy between the bicriterion problems in finance and transportation disciplines makes it possible to adapt the theory by substituting the concepts of the risk and return by the components of the users' generalized cost for modeling the bicriterion decision making problems in transportation.

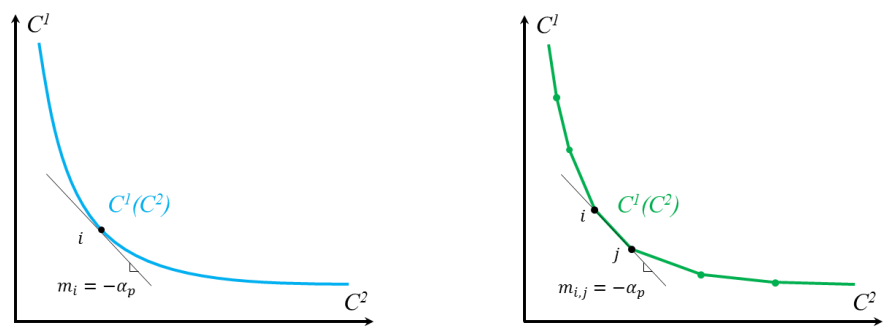
Definition (Efficient frontier of choices). The efficient frontier of choices is the convex hull of the equilibrium choice set in the C^2- C^1 plane that have the minimum generalized cost according to (2.1) for the users with heterogeneous preferences regarding the relative importance of the components of the cost (α).

Recognizing that the rational users tend to minimize their costs by seeking for the choice with the lowest cost, they may make variety of different - but dominant - choices due to heterogeneity in their presences. In the other words, each user is looking for a specific optimal balance between these costs that can minimize his/her generalized cost according to his/her preferences regarding relative importance of these criteria. However, as illustrated in Figure 2.2b, there might be choices (pink points) that are always dominated by their alternatives (green points) with lower associated costs in all the components, which no one with any preferences would not choose them. Accordingly, the cumulative result of individual decisions with heterogeneous preferences regarding the relative importance of the criteria can be well presented by the efficient frontier of the choices.

When the generalized cost of the heterogeneous users can be approximated as a linear combination of components of the costs as presented in (2.1), the efficient frontier of the choices can be proven to have a set of primary characteristics that remains general for the solution of the different bicriterion problems. Once these characteristics are identified, they can reveal the key properties of the solutions of the bicriterion problems.

2.1.1 Differentiable Efficient Frontier of Choices

In cases that the choice set of the users has a continuous nature, e.g., arrival time to the bottleneck, and user preferences have a continuous distribution, then the efficient frontier, $C^1(C^2)$, is a continuous and differentiable function as illustrated in Figure 2.3a, with following characteristics:



(a) Differentiable efficient frontier (b) Piecewise linear efficient frontier

Figure 2.3. Efficient frontier of continuous set of choices

Proposition 2.1 (monotonicity). The efficient frontier, $C^1(C^2)$, is a non-increasing function of C^2.

Proof. The proof is made by contradiction. Let $C^1(\cdot)$ represent the continuous differentiable efficient frontier function. According to the definition of decreasing monotonicity, the efficient frontier will not be a decreasing monotone function if and only if there exist at least two points on the efficient frontier satisfying the following conditions (see Figure 2.4a):

$$C_1^2 < C_2^2 \tag{2.2}$$

$$C^1(C_1^2) < C^1(C_2^2) \tag{2.3}$$

Multiplying both sides of the inequality (2.2) by any nonnegative α_p and adding the result to the inequality (2.3), we get the following inequality:

$$C^1(C_1^2) + \alpha_p\, C_1^2 < C^1(C_2^2) + \alpha_p\, C_2^2\ , \quad \forall \alpha_p \geq 0 \tag{2.4}$$

According to the definition of (2.1), left and right sides of the inequality (2.4) represent generalized costs of the points 1 and 2, respectively.

$$C_{p,1} < C_{p,2}, \quad \forall \alpha_p \geq 0 \tag{2.5}$$

Inequality (2.5) indicates that point 2 is always dominated by point 1, for any nonnegative value of α_p. This is in direct contradiction with the definition of the efficient frontier as set of dominant choices, so point 2 cannot be on the efficient frontier. Therefore, if (2.2) holds, then $C^1(C_1^2) \geq C^1(C_2^2)$ on the efficient frontier; thus, decreasing monotonicity of the efficient frontier is guaranteed. ■

Proposition 2.2 (convexity). The efficient frontier, $C^1(C^2)$, is a convex function of C^2.

Proof. The proof is made by contradiction as well. By definition, a convex function must satisfy the following:

$$\theta C^1(C_1^2) + (1-\theta)C^1(C_2^2) \geq t(\theta C_1^2 + (1-\theta)C_2^2), \quad \forall C_1^2, C_2^2, \ \forall\theta \in [0,1]. \tag{2.6}$$

Therefore, the efficient frontier will not be convex if and only if $\exists C_1^2, C_2^2$, and $\exists\theta \in [0,1]$ such that

$$\theta C^1(C_1^2) + (1-\theta)C^1(C_2^2) < t(\theta C_1^2 + (1-\theta)C_2^2), \tag{2.7}$$

In other words, if there exist at least at one point on the efficient frontier that lies above a linear combination of these two points, the convexity condition is violated (see Figure 2.4b).

To derive the generalized costs of points 1 and 2, we can use the following algebraic equality:

$$\alpha_p\theta C_1^2 + \alpha_p(1-\theta)C_2^2 = \alpha_p(\theta C_1^2 + (1-\theta)C_2^2)\ , \quad \forall C_1^2, C_2^2, \theta, \alpha_p \tag{2.8}$$

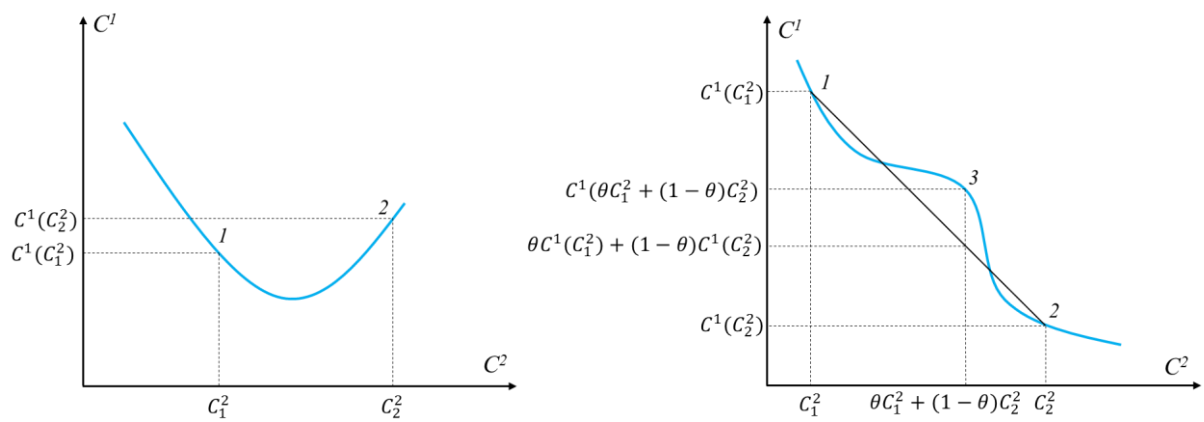
We hold the inequality (2.7) by adding the left and right sides of (2.8) to the left and right sides of (2.7), respectively. According to the definition of the generalized cost in (2.1), the left side of this inequality is a weighted sum of the generalized cost of points 1 and point 2. The right side is also the generalized cost of a third point associated with $C_3^2 = C_1^2 + (1-\theta)C_2^2$, which we denote simply as $C_{p,3}$. Following from the conditions of (2.7) and (2.8) that $\forall\alpha_p$, $\exists C_1^2$, C_2^2 , and $\exists\theta \in [0,1]$, we have:

$$\theta C_{p,1} + (1-\theta)C_{p,2} < C_{p,3} \tag{2.9}$$

According to the definition of the efficient frontier, a rational user with preference α_p who chooses between points 1 and 2 in the absence of better alternatives will choose the route with the lower generalized cost of $C_{p,\min} = \min\{C_{p,1}, C_{p,2}\}$. Using the identity, $C_{p,\min} = \theta C_{p,\min} + (1-\theta)C_{p,\min}$, we can see that

$$C_{p,min} = \theta C_{p,min} + (1-\theta)C_{p,min} \leq C_{p,1} + (1-\theta)C_{p,2} < C_{p,3} \tag{2.10}$$

because $C_{p,\min} \leq C_{p,1}$ and $C_{p,\min} \leq C_{p,2}$. So, existence of a θ for which condition (2.7) holds means that there exist at least a point on the efficient frontier with a generalized cost lower than point 3, for any value of α_p. This contradicts the inherent dominancy of the efficient frontier, thus the efficient frontier must be convex everywhere. ■



(a) (b)

Figure 2.4. (a) monotonicity of the efficient frontier; (b) convexity of the efficient frontier

Proposition 2.3 (geometric property). The slope of the differentiable efficient frontier at each point is $m_i = -\alpha_p$ of the users choosing that point (see Figure 2.3a)

Proof. According to the definition of the efficient frontier as a set of dominant choices, users pick the choices that minimize their generalized cost according to their own preferences. Since the efficient frontier is defined continuous and differentiable, the first derivative of the generalized cost of each choice *i* on the efficient frontier for demand group p with respect to C_i^2 should be equal to zero. By substituting the definition of the generalized cost from equation (2.1), we can rewrite the first order condition as follows:

$$\frac{\partial C_{p,i}}{\partial C_i^2} = \frac{\partial C_i^1}{\partial C_i^2} + \alpha_p = 0 \tag{2.11}$$

As a result, the slope of the efficient frontier at choice i, m_i, picked by group p, can be determined as below:

$$m_i = \frac{\partial C_i^1}{\partial C_i^2} = -\alpha_p \tag{2.12}$$

■

Corollary 2.1 (assignment order). Labeling the choices in an increasing order of their C^2, the heterogeneous users pick these choices in a decreasing order of their α_p. According to the monotonicity property (Proposition 2.1), C^1 will be a decreasing function of C^2. In the meantime, the convexity property (Proposition 2.2) ensures that the negative slope is

increasing (i.e., becoming less steep). Then, the geometric property (Proposition 2.3) implies that the assignment starts with the greatest α, where the negative slope is steepest, and progresses sequentially to the lowest α, where the slope is flattest (See Figure 2.3a).

Now that the general properties of the efficient frontier are revealed, we extend the results for a discrete distribution of α.

2.1.2 Piecewise Linear Efficient Frontier of Choices

In the case that the distribution of the user preferences is discrete, according to the Proposition 2.3, the efficient frontier,$C^1(C^2)$, turns out to be a piecewise linear function, as illustrated in Figure 2.3b, with very similar characteristics as the differentiable function. For one thing, the efficient frontier is still monotonically deceasing exactly as demonstrated in Proposition 2.1. For another, the piecewise linear efficient frontier is also a convex function as showed in Proposition 2.2. Finally, the pricewise efficient linear efficient frontier has the same geometric property presented in Proposition 2.3. However, since α has a discrete distribution, the slope of the segment *(i,j)* of efficient frontier ($m_{i,j}$), which includes a range of choices from *i* to *j*, will be equal to $-\alpha_p$ of the group of users picking these choices, as depicted in the Figure 2.3b.

$$m_{i,j} = -\alpha_p \tag{2.13}$$

Accordingly, the assignment order that explained in Corollary 2.1 also holds for a discrete distribution of α, and the heterogeneous users pick these choices in a decreasing order of their α_p.

Now that these specific characteristics of the efficient frontier are known, they can be employed to analytically model the single bottleneck problem with distributions in the schedule penalty factors of the commuters as a bicriterion choice problem with heterogeneous user preferences.

CHAPTER 3

MORNING COMMUTE PROBLEM WITH HETEROGENEOUS USER PREFERENCES

Choosing the starting time of a trip is one of the important decisions that users need to make ahead of their trips according to their own preferences. Rational users of a bottleneck tend to minimize the generalized cost of their trips by adjusting their arrival times to the bottleneck. Thus, the cumulative result of the individual decisions of the users leads to the user equilibrium condition in which no one can reduce his/her cost anymore by changing the starting time of his/her trip. This problem is first introduced in Vickrey's (1969) congestion theory based on a first-in, first-out single bottleneck model with fixed capacity, and elaborated in the literature by considering heterogeneity in schedule preferences of the users (Hendrickson and Kocur ,1981; Smith, 1984; Daganzo, 1985). In this section, we employ the concept of the efficient frontier to propose an extension to the user equilibrium and optimum pricing models of the bottleneck by accounting for the heterogeneity in schedule penalty preferences of the users in modeling the travel behavior of the commuters. In addition, the proposed analytical solution for the morning commute problem can have applications in modeling and optimizing the aggregated networks that can be modeled as queueing systems (Daganzo, 2007; Geroliminis and Daganzo, 2008; Gonzales and

Daganzo, 2012). An applicable method is also proposed to retrieve independent distributions of preferences of the users from a given joint distribution. The proposed analytical solution of the problem can also be inversely used to approximate the distribution of user preferences from the empirical data from observations.

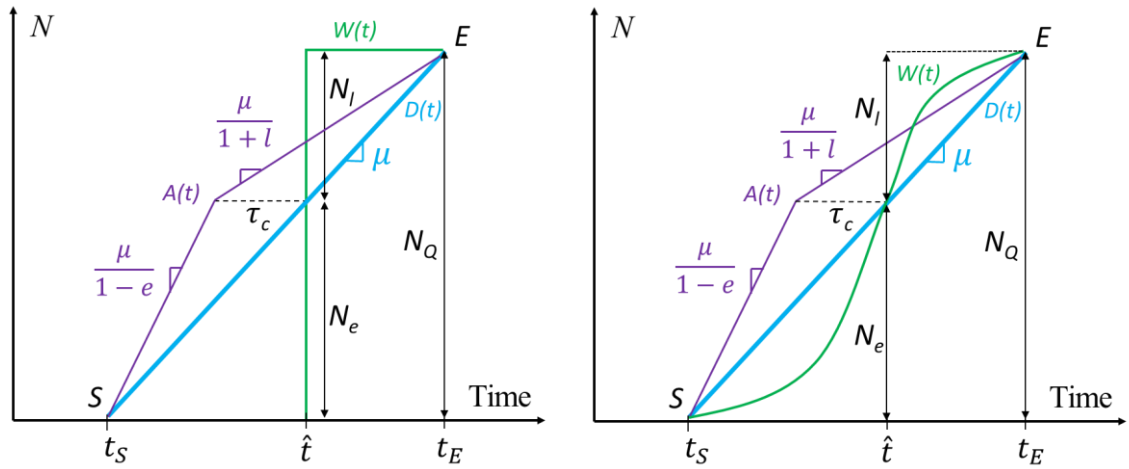
3.1 Trip Scheduling Problem with Heterogeneous Traveler Preferences

The trip scheduling problem for a single bottleneck was introduced in Vickrey (1969), which explains how users adjust the timing of their travel to minimize their own costs when time-dependent demand exceeds a bottleneck's fixed capacity, μ. As a result of queuing, a commuter who arrives at the bottleneck will experience a combination of queueing delay and schedule penalty, depending on whether the departure time from the bottleneck is before or after their preferred schedules, in addition to the free flow travel time of the bottleneck. Thus, the total cost for the commuter N, with earliness penalty factor e_p and lateness penalty factor l_p, can be expressed as a linear combination of the free flow travel time (τ_F), delay (τ_D), and earliness (τ_E) or lateness (τ_L) that this commute experiences:

$$C_p(N) = \tau_F + \tau_D(N) + e_p\ \tau_E(N) + l_p\ \tau_L(N) \tag{3.1}$$

Here, the schedule penalty factors e_p and l_p recognize the weights of the schedule deviation relative to queueing delay depending on the preferences of the heterogeneous commuters commuter denoted by p.

The pattern of arrivals and departures from the bottleneck may be represented by cumulative counts of the number of passengers to arrive by time t, $A(t)$, and the number to depart by time t, $D(t)$. So the delay can be represented as the horizontal distance between arrival and departure curves. As illustrated in the Figure 3.1a and 5b, such delay increases from zero at start of the peak period, t_s, to its maximum value, τ_c, at time $\hat{t}$, then it decreases back to zero at the end of end of the peak, t_E. The schedule deviation is the difference between the time that a commuter wishes to have departed the bottleneck as described by a cumulative wished curve, $W(t)$, and the actual departure. Thus, the earliness and lateness are a consequence the cumulative arrivals and resulting queues. If commuters choose when to start their trips to arrive at the bottleneck at a time t, the user equilibrium is defined by the $A(t)$ that allows no commuter to reduce his/her own generalized cost by changing his/her own arrival time. Daganzo (1985) proves that a unique user equilibrium exists when $W(t)$ is S-shaped, with slope that exceeds bottleneck capacity between two time points and is less outside. Assuming that demand is homogenous with identical earliness (e) and lateness (l) penalty factors, Fig 4a, b illustrate queueing diagrams of the well-known conventional user equilibrium respectively for a stepwise and an S-shaped wished and departure curve.



(a) Stepwise wished curve (b) S-shaped wished curve

Figure 3.1. Queueing diagram of bottleneck user equilibrium

Two conditions describe the user equilibrium when commuters have homogeneous schedule penalties. First, the arrival curve is piecewise linear with specific slopes in the user equilibrium condition.

$$\frac{dA(t)}{dt} = \begin{cases} \mu/(1-e), & \text{for commuters who depart early} \\ \mu/(1+l), & \text{for commuters who depart late} \end{cases} \tag{3.2}$$

Second, the proportion of the commuters who experience earliness, N_e, to those who are late, N_l, equals the proportion of lateness penalty to earliness penalty.

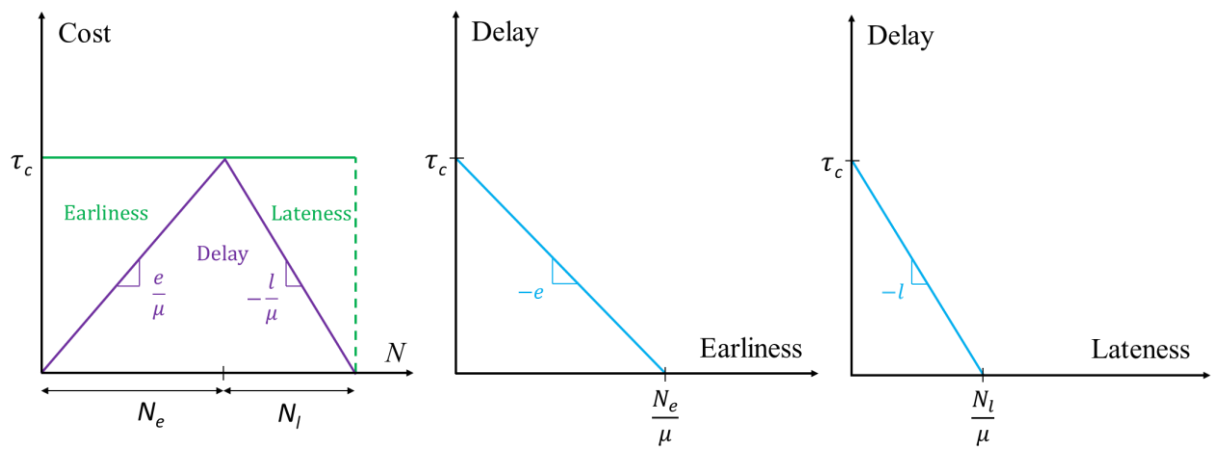
$$\frac{N_e}{N_l} = \frac{l}{e} \tag{3.3}$$

In the equilibrium condition, no one can reduce his/her generalized cost by changing his/her arrival time to the bottleneck. The underlying assumption of the congestion theory is that all commuters have identical schedule penalties for departing early and late from the bottleneck. However, empirical evidence shows that in reality there is a heterogeneity associated with the schedule penalty factors that commuters account for them in their trips (Small, 1982; Gonzales and Christofa, 2013). To include such heterogeneity of the users' preferences, we employ the concept of the efficient frontier of choices, to propose an analytical solution to the morning commute problem with heterogeneous users. First, in section 3.1.1, we explain the methodology under the simplifying assumption that commuters have a stepwise wished curve as depicted in Figure 3.1a. Then, in section 3.1.2, we show that a same approach can be followed in case that the departure curve has a general smooth S-shape as illustrated in Figure 3.1b. On this basis, we propose a general formulation of the model when there is heterogeneity associated with both schedule and schedule penalty preferences of the bottleneck commuters.

3.1.1 Stepwise Wished Curve

In a simplified case of the morning commute problem of Figure 3.1a for users with an identical wished departure time ($\hat{t}$) and the same schedule penalty factors as in Arnott et al.

(1990), the generalized cost of the trips remains equal among the homogenous commuters in the equilibrium condition, as depicted in Figure 3.2a. It is worth pointing out that the delay that commuters experience decreases linearly with their earliness and lateness with slopes $-e$ and $-l$, respectively, as depicted in the Figure 3.2b and 6c. In addition, it can be shown that earliness and lateness of the users vary linearly with their departure order from bottleneck with slope *1/μ* (See Figure 3.1a). The result is depicted in Figure 3.2a with delay increasing linearly with slope *e/μ* for early commuters to its maximum value τ_c and then decreasing with slope *-l/μ* to zero for late commuters. Meanwhile, the generalized cost of travel remains constant for all the commuters.



(a) Variation of commuters' generalized cost

(b) Delay-earliness relationship

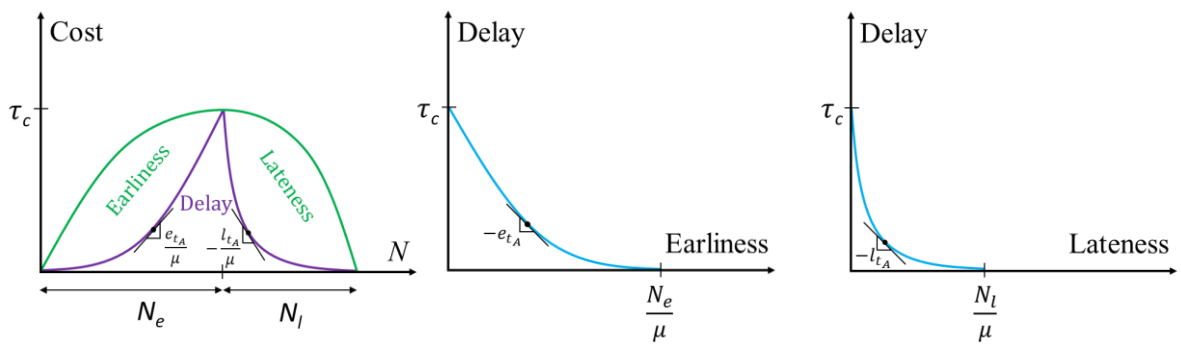
(c) Delay-lateness relationship

Figure 3.2. Variation of the components of the generalized cost for homogenous commuters in user equilibrium

To account for the heterogeneity of the user preferences, we may consider a probability distribution for the schedule penalties of the commuters like the one plotted in Figure 2.1. In this case, the morning scheduling problem can be viewed as a choice problem with heterogeneous user preferences. The heterogeneous commuters seek to minimize the generalized cost of their trips according to their own preferences by adjusting their arrival time to the bottleneck. In this problem, the commute cost of equation (3.1) can be compared to the generalized cost function (1), where the associated delay and schedule deviation are the components of the generalized cost that users experience to commute through the bottleneck. In this respect, each commuter chooses the arrival time to the bottleneck that minimizes the combined delay and schedule deviation that he or she experiences according to his/her own schedule penalty factors.

The cumulative result of the individual bicriterion decisions in the equilibrium condition can be well represented by the efficient frontier of choices. Accordingly, the population of the heterogeneous commuters will be distributed along the efficient frontier of arrival times in the equilibrium condition. As a result, the concept of the efficient frontier can be employed to demonstrate the relationships of delay with earliness and lateness of the equilibrium arrival of the users as illustrated in Figure 3.3b and 7c. In these figures, each point on the efficient frontier corresponds to a specific arrival time, which is represented by its associated delay and schedule deviation. So, the slope of the efficient frontier, $\tau_D(\cdot)$ as function of τ_E or τ_L, at each point equals to the negative value of the schedule penalty of the user who chooses the corresponding arrival time (t_A) of this point,

$-e_{t_A}$ or $-l_{t_A}$ for earliness and lateness periods, respectively. In the other words, commuters with lower schedule penalty factors tend to arrive to the bottleneck closer to the beginning (t_S) and the end (t_E) of the peak period, where they will experience a lower delay although they may significantly deviate from their schedule. In contrast, commuters with higher schedule penalty factors prefer to arrive to the bottleneck in the middle of the peak period to arrive to their destinations with a lower deviation from their schedules in spite of higher delays that they experience in the bottleneck. Ignoring the free flow travel time of the users, the variations in generalized cost of equation (3.1) for heterogeneous commuters can be illustrated as in Figure 3.3a.



(a) Variation of commuters' generalized cost

(b) Delay-earliness relationship

(c) Delay-lateness relationship

Figure 3.3. Variation of the components of the generalized cost for heterogeneous commuters in user equilibrium

As shown in the Chapter 2, the efficient frontier of choices has specific properties that are general for all problems. In the next section, we use these properties to propose an analytical solution to the morning commute problem with an S-shaped wished curve and heterogeneous user schedule penalty factors.

3.1.2 S-Shaped Wished Curve

To account for the heterogeneity in the schedule preferences of the users, we may assume a smooth S-shaped wished curve for the commuters. Meanwhile, we account for the heterogeneity of the schedule penalty factors of the users by considering probability distributions for earliness and lateness penalty factors. Heterogeneous users choose arriving times to the bottleneck that make the combination of the delay and schedule deviation of their commutes minimized according to their own preferences. The cumulative result of the individual decisions will be an equilibrium condition in which no commuter will have an incentive to change his/her arrival time to the bottleneck to be earlier or later. Accordingly, the concept of the efficient frontier can be employed to represent the equilibrium condition of the bottleneck with heterogeneous commuters. Thus, we use specific properties of the efficient frontier, $\tau_D(\cdot)$ as function of τ_E or τ_L, to demonstrate the relationship of arrival times of the commuters to the bottleneck with their schedule deviation penalty factors. On this basis, we can derive the arrival distribution of the heterogeneous commuters over time in the equilibrium condition.

In Section 3.1.2.1, we first use the main properties of the efficient frontier, presented in Chapter 2, to specify the properties of the equilibrium arrival of the heterogeneous commuters to the bottleneck. On this basis, in Section 3.1.2.2, we derive an analytical solution for the equilibrium arrival of the commuter to the bottleneck given the PDFs of the schedule deviation penalty factors. In section 3.1.2.3, we show that the proposed analytical solution derived based on the properties of the efficient frontier is necessary and sufficient for the equilibrium conditions of the bottleneck derived according to Daganzo (1985).

3.1.2.1 Efficient Frontier of Arrival Choices

The main objective of this part of the research is to derive the arrival distribution of the heterogeneous users over time who find different arrival times minimizing the generalized cost of their commutes according to their own schedule deviation penalty preferences. In this respect, commuters can be tagged with their arrival times to the bottleneck as well as their earliness or lateness penalty factors. Once we figure out the relationship between arrival times of the commuters with their schedule deviation penalty factors, we can derive the equilibrium solution for the arrival of the commuters to the bottleneck. For this to happen, let variable t_A denotes the arrival time of the commuters to the bottleneck. So $N_{t_A} = A(t_A)$ is the cumulative arrivals of the commuters by time t_A while the earliness (or lateness) penalty factor of the commuter who arrived to the bottleneck at time t_A is

represented by e_{t_A} (or l_{t_A}). On this basis, we can derive the relationship between these variables in the equilibrium condition using the concept of the efficient frontier.

In this section, we use the main properties of the efficient frontier, $\tau_D(\cdot)$, according to Chapter 2, to demonstrate the properties of the equilibrium arrivals of the heterogeneous commuters to the bottleneck. According to the Propositions 2.1 and 2.2, the efficient frontier is a monotonically decreasing convex function. However, we need to specify the geometric property of the efficient frontier in Proposition 2.3 for the morning commute problem to identify the differential relationship between the delay and schedule deviation that users experience in the equilibrium condition, as demonstrated in Proposition 3.1. Knowing the relationship between schedule deviation for the users and their departure times from the bottleneck in the equilibrium condition, we derive the differential relationship between delay that users experience and their departure times as an equivalent for the geometric property of Proposition 3.1, as presented in Corollary 3.1. Moreover, we use the result of the Corollary 3.1 to adapt the assignment order rule of the Corollary 2.1 specifically for the morning commute problem in Corollary 3.2. In Section 3.1.2.2, we use the results of the Corollaries 3.1 and 3.2 to derive the analytical solution for the morning commute problem given the PDFs of the schedule deviation penalty factors.

Proposition 3.1 (geometric property). Each point on the efficient frontier, $\tau_D(\tau_E)$ (or $\tau_D(\tau_L)$), represents the combinations of delay and earliness (or lateness) with minimum

costs, which directly corresponds to choosing an arrival time (t_A) to the bottleneck. On this basis, the slope of the efficient frontier at each point equals to $-e_{t_A}$ (or $-l_{t_A}$) for the early (or late) commuter N_{t_A} who chooses arrival time t_A.

$$m_{t_A} = \begin{cases} \frac{\partial \tau_{D,N_{t_A}}}{\partial \tau_{E,N_{t_A}}} = -e_{t_A}, & \text{for commuters who depart early} \\ \frac{\partial \tau_{D,N_{t_A}}}{\partial \tau_{L,N_{t_A}}} = -l_{t_A}, & \text{for commuters who depart late} \end{cases} \tag{3.4}$$

Proof. Rational users tend minimize their generalized cost according to their own preferences. In the equilibrium condition, heterogeneous users arrive to the bottleneck in an order in time that each of them experiences the minimum possible generalized cost according to his/her preference. So, the first derivative of the generalized cost for commuters at each arrival point in time, t_A, and earliness penalty factor e_{t_A}, with respect to *N* should be equal to zero.

$$\frac{\partial C_{t_A}(N_{t_A})}{\partial N} = \frac{\partial C_{t_A}(N_{t_A})}{\partial \tau_{E,N_{t_A}}} \cdot \frac{\partial \tau_{E,N_{t_A}}}{\partial N} = 0 \tag{3.5}$$

According to the definition of the schedule deviation, the earliness of commuter *N* can be approximated as the difference between the wished and actual departure times:

$$\tau_{E,N} = W^{-1}(N) - D^{-1}(N) = t_{w,N} - \frac{N}{\mu}, \text{ for the commuter who departs early} \tag{3.6}$$

In this relation, $W^{-1}(.)$ and $D^{-1}(.)$ are the inversed cumulative wished curves, which denote the time corresponding to commuter N, respectively. Here, $W^{-1}(N)$ has a fixed value for the commuter N, $t_{w,N}$. Moreover, since bottleneck has a fixed capacity, the departure time of the commuter N can be derived as $D^{-1}(N) = N/\mu$. As a result the first derivative of the earliness of the commuter N with respect to N can be derived as bellow:

$$\frac{\partial \tau_{E,N_{t_A}}}{\partial N} = -\frac{1}{\mu} \tag{3.7}$$

By substituting the definition of the generalized cost from equation (3.1), we can rewrite the first order condition (3.5) as follow:

$$-\frac{1}{\mu}\left(\frac{\partial \tau_{D,N_{t_A}}}{\partial \tau_{E,N_{t_A}}} + e_{t_A}\right) = 0 \tag{3.8}$$

As a result, the slope of the efficient frontier at each arrival point in time, $m_{t_A} = \partial \tau_{D,N_{t_A}} / \partial \tau_{E,N_{t_A}}$, is negatively proportional to the earliness penalty factor of the commuters who choose that arrival time in the equilibrium condition, e_{t_A}. A very similar conclusion also can be drawn for the commuters who depart the bottleneck later than they wished as generalized in condition (3.4).■

Corollary 3.1 (equivalent geometric property). As it is illustrated in Figure 3.4, in the equilibrium condition, the first derivative of the delay that a commuter experiences with

respect to his departure time, $D^{-1}(N) = t_D$, equals to e_{t_A} (or $-l_{t_A}$) of that commuter in the earliness (or lateness) period.

$$\frac{\partial \tau_{D,N_{t_A}}}{\partial t_D} = \begin{cases} +e_{t_A}, & \text{for commuters who depart early} \\ -l_{t_A}, & \text{for commuters who depart late} \end{cases} \tag{3.9}$$

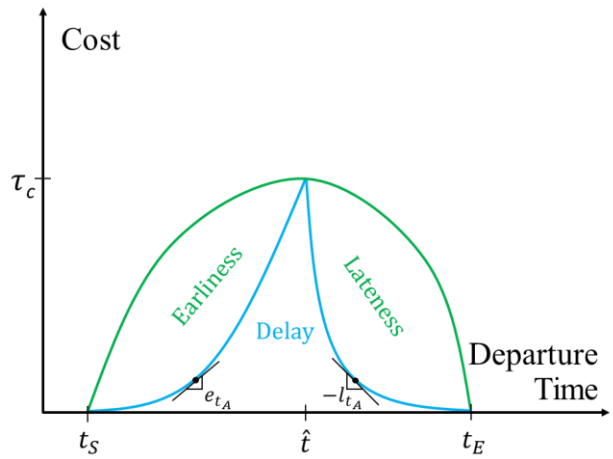


Figure 3.4. Variation of components of the generalized cost over time for heterogeneous commuters in user equilibrium

Proof. According to the differentiation chain rule we have:

$$\frac{\partial \tau_{D,N_{t_A}}}{\partial t_D} = \frac{\partial \tau_{D,N_{t_A}}}{\partial \tau_{E,N_{t_A}}} \cdot \frac{\partial \tau_{E,N_{t_A}}}{\partial t_D} \tag{3.10}$$

For one thing, we have the value of first term ($\partial \tau_{D,N_{t_A}} / \partial \tau_{E,N_{t_A}}$) from equation (3.4) of Proposition 3.1. For another we need to determine the value of the second term

($\partial\tau_{E,N_{t_A}}/\partial t_D$). By substituting $W^{-1}(N)$ with $t_{w,N}$ and also $\mathrm{D}^{-1}(N)$ with t_D in the definition of earliness in equation (3.6), the first derivative of the earliness with respect to t_D can be derived as below:

$$\frac{\partial\tau_{E,N_{t_A}}}{\partial t_D} = -1 \tag{3.11}$$

Thus, the condition (3.9) for the early commuters can be derived by substituting the equations (3.4) and (3.11) in the equation (3.10). A similar result also can be derived following the same approach for late commuters. ■

Corollary 3.2 (assignment order). As it can be inferred from Figure 3.4, early commuters arrive to the bottleneck in an increasing order of their earliness penalty factor from time t_S to $\hat{t}$. In contrast, late commuters arrive to the bottleneck in a decreasing order of their lateness penalty factor from time $\hat{t}$ to t_E. In other words, commuters with lower schedule penalty factors are more sensitive to delay and tend to minimize their generalized costs by adjusting their arrival times closer to the beginning (t_S) and end (t_E) of the peak period. On the contrary, commuters with higher schedule penalty factors have less sensitivity to delay and prefer the arrival times closer to the rush of the peak period ($\hat{t}$).

According to Propositions 2.1 and 2.2, the efficient frontier of arrival choices is a monotonically decreasing convex function. Thus, it can be concluded from comparing Proposition 3.1 with Corollary 3.1 that the delay first monotonically increases to its

maximum value, τ_c , from time t_S to $\hat{t}$, and then monotonically decreases back to zero from time $\hat{t}$ to t_E. However, this function remains convex in the earliness part as well as the lateness part, which means that the slope of delay curve is monotonically increasing and decreasing in earliness and lateness parts, respectively. According to the Corollary 3.1, this slope at each arrival time, t_A, equals to the $\mathrm{e}_{t_\mathrm{A}}$ or $-\mathrm{l}_{t_A}$ of the early or the late commuter(s) who chooses that arrival time, respectively. As a result, it can be concluded that commuters arrive to the bottleneck in an increasing and a decreasing order of their schedule penalty factors in the earliness and lateness periods, respectively.

As it is mentioned Daganzo (1985), the earliness penalty factor of the users can vary between 0 and 1, and the lateness penalty factor is nonnegative. Accordingly, as illustrated in Figure 3.4, the slope of delay curve monotonically increases from zero at time t_S to unity at time $\hat{t}$. At this point, the slope of the delay drops to negative infinity, then it again monotonically increases back to zero to time t_E.

3.1.2.2 Equilibrium Arrival of the Heterogeneous Commuters

Rational users minimize their own generalized costs by adjusting their arrival times to the bottleneck in order to balance delay and schedule deviation that they experience during their commutes through the bottleneck. The cumulative result of these individual decisions is the user equilibrium condition in which the population of the commuters will be distributed along the efficient frontier of the arrival choices. Here, we use the concept of

the efficient frontier to derive an analytical solution for the morning commute problem given the PDFs of the schedule deviation penalty factors of the users. To this end, we first we use the introduced properties of the efficient frontier, presented in Section 3.2.1, to determine the relationship of the arrival times of the commuters to the bottleneck with their schedule deviation penalty factors. On this basis, we derive the cumulative arrival of the heterogeneous commuter to the bottleneck as well as their generalized cost of commute according to their own preferences in the equilibrium condition.

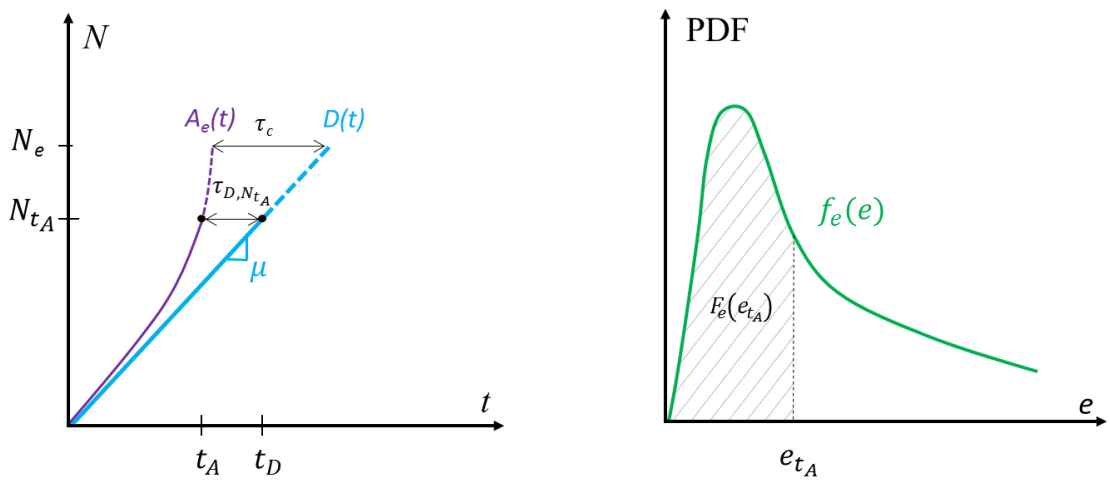
In this respect, we combine the equivalent geometric property of the Corollary 3.1 with the assignment order rule of the Corollary 3.2 to determine the equilibrium arrival times of the heterogeneous commuters to the bottleneck given the PDFs of their schedule deviation penalty factors. Corollary 3.1 relates the variations of delay at each point in (departure) time to the schedule deviation penalty factor of the user that experiences that delay in the equilibrium condition.

According to condition (3.9), the first derivative of the delay for each early user with respect to his/her departure time equals to his/her earliness penalty factor, e_{t_A}. By employing the differentiation chain rule we can rewrite the condition (3.9) for an early commuter, N_{t_A}, as below:

$$\frac{\partial \tau_{D,N_{t_A}}}{\partial t_D} = \frac{\frac{\partial \tau_{D,N_{t_A}}}{\partial t_A}}{\frac{\partial t_D}{\partial t_A}} = e_{t_A} \tag{3.12}$$

For one thing, we need to derive $\partial\tau_{D,N_{t_A}}/\partial t_D$ to substitute it back into equation (3.12).In this respect, the delay of this early commuter can be approximated as the horizontal distance between the cumulative arrival and departure curves, as illustrated in the Figure 3.5a:

$$\tau_{D,N_{t_A}} = D^{-1}(N_{t_A}) - A_e^{-1}(N_{t_A}) = \frac{N_{t_A}}{\mu} - t_A \tag{3.13}$$



(a) Variation of delay in the queueing diagram

(b) Probability distribution of earliness penalty factor

Figure 3.5. Relation between the equilibrium arrivals of the early commuters with the distribution of the earliness penalty factor

In the equation (3.13), we need to approximate the value of the cumulative arrival of heterogeneous commuters at time t_A, $N_{t_A} = A(t_A)$. According to the Corollary 3.2, early commuters arrive to the bottleneck in an increasing order of their earliness penalty factor

in the equilibrium condition. So, arrival of a commuter with earliness penalty factor e_{t_A} to the bottleneck at time t_A can be interpreted as arrival of the all commuters with earliness penalty factors less than e_{t_A} ahead of this user. Given the probability density function (PDF) of the earliness penalty factor of the early commuters, $f_e(e)$, the cumulative proportion of the early commuters who arrive before time t_A , N_{t_A}/N_e , is approximated as the cumulative probability density function (CDF) of the earliness penalty factor at e_{t_A}, $F_e(e_{t_A})$, which represents the area under the $f_e(e)$ up to e_{t_A}, as illustrated in Figure 3.5b. As a result, the cumulative number of the arrivals at time t_A can be approximated as cumulative probability distribution of the earliness penalty factor at e_{t_A} times the total number of early commuters:

$$N_{t_A} = N_e F_e\left(e_{t_A}\right) = N_e \int_0^{e_{t_A}} f_e(e)de \tag{3.14}$$

By substituting N_{t_A} from equation (3.14) in equation (3.13), delay of the commuter can be approximated as below:

$$\tau_{D,N_{t_A}} = \frac{N_e \int_0^{e_{t_A}} f_e(e)de}{\mu} - t_A \tag{3.15}$$

Accordingly, the first derivative of the delay to the arrival time can be derived as follows:

$$\frac{\partial \tau_{D,N_{t_A}}}{\partial t_A} = \frac{N_e}{\mu} \dot{e}_{t_A} f_e\left(e_{t_A}\right) - 1 \tag{3.16}$$

where, $\dot{e}_{t_A} = de_{t_A}/dt_A$.

For another, we need to derive $\partial t_D/\partial t_A$ to substitute it back into equation (3.12). According the definition of delay from equation (3.13), we can relate t_D and t_A as follows (See Figure 3.5a):

$$t_D = t_A + \tau_{D,N_{t_A}} \tag{3.17}$$

By taking the first derivative of the equation (3.17) with respect to t_A we will have:

$$\frac{\partial t_D}{\partial t_A} = 1 + \frac{\partial \tau_{D,N_{t_A}}}{\partial t_A} \tag{3.18}$$

By substituting $\partial t_D/\partial t_A$ from equation (3.18) in equation (3.12), it can be rewritten as below:

$$\frac{\frac{\partial \tau_{D,N_{t_A}}}{\partial t_A}}{1+\frac{\partial \tau_{D,N_{t_A}}}{\partial t_A}} = e_{t_A} \tag{3.19}$$

Now, we plug $\partial \tau_{D,N_{t_A}}/\partial t_D$ from equation (3.16) into equation (3.19):

$$\frac{\frac{N_e}{\mu}\dot{e}_{t_A} f_e\left(e_{t_A}\right)-1}{\frac{N_e}{\mu}\dot{e}_{t_A} f_e\left(e_{t_A}\right)} = e_{t_A} \tag{3.20}$$

By substituting $\dot{e}_{t_A}$ with its equivalent term de_{t_A}/dt_A and a little algebraic manipulation equation (3.20) can be rewritten as below:

$$dt_A = \frac{N_e}{\mu}(1 - e_{t_A})f_e\left(e_{t_A}\right)de_{t_A} \tag{3.21}$$

Taking the integral of both sides of the equation (3.21), the analytical solution of the morning commute problem with heterogeneous user preferences is given by the following:

$$t_A = \frac{N_e}{\mu}\int_0^{e_{t_A}}(1 - e)f_e(e)de \tag{3.22}$$

Accordingly, the equilibrium arrival of the heterogeneous users to the bottleneck can be determined based on their own penalty preferences for the early departure from the bottleneck.

Numerical methods can be employed here to estimate the arrival of the heterogeneous commuters using this relation specifically when it is not analytically simple to take the integral of the probability density function of the earliness factors analytically. However, the cumulative arrival of the heterogeneous commuters also can be analytically derived using relation (3.22), when the integral has a definite solution. In this case, we simply need to solve equation (3.22) for e_{t_A} and substitute the result in equation (3.14). As a result, the

cumulative arrival of heterogeneous early commuters in the equilibrium condition can be approximated as below:

$$A_e(t_A) = N_e F_e\left(\hat{e}_{t_A}\right) \tag{3.23}$$

where, $\hat{e}_{t_A}$ denotes the solution of equation (3.22) as a function of t_A.

On this basis, we also can derive the distribution of the earliness factor of the equilibrium arrivals by inverting the function (3.23):

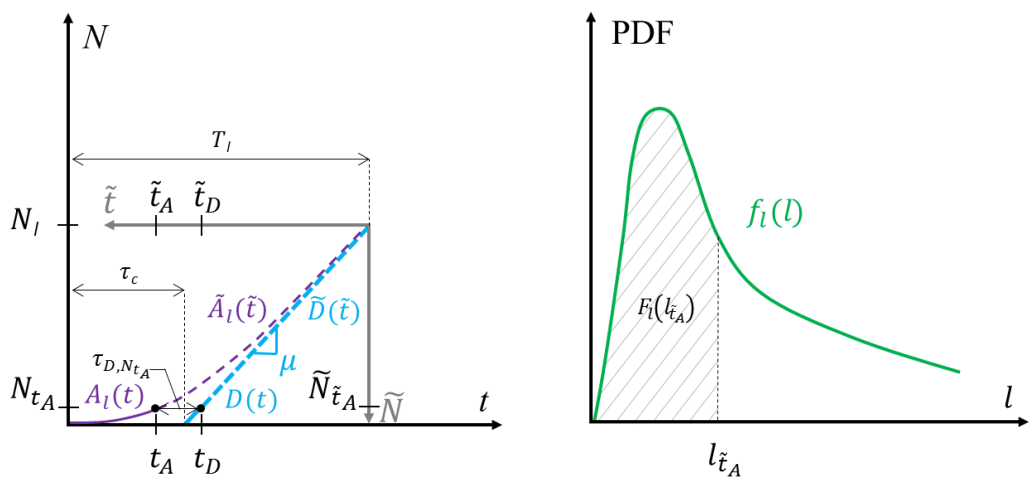
$$e_{t_A=} F_e^{-1}\left(\frac{N_{t_A}}{N_e}\right) \tag{3.24}$$

By substituting the equations (3.6), (3.13), and (3.24) in the generalized cost function (3.1), we can rewrite the generalized cost function of early commuters as below:

$$\mathrm{C}_{t_A}\left(\mathrm{N}_{t_A}\right) = \tau_F + \left(\frac{N_{t_A}}{\mu} - A_e^{-1}\left(N_{t_A}\right)\right) + F_e^{-1}\left(\frac{N_{t_A}}{N_e}\right) \cdot \left(t_{w,N_{t_A}} - \frac{N_{t_A}}{\mu}\right),\ N_{t_A} \in [0, N_e] \tag{3.25}$$

Similar results also can be derived for the equilibrium arrival of the heterogeneous commuters who depart the bottleneck later than they wished. In this respect, having an auxiliary coordination system for the queueing diagram with origin at (t_E, N_Q) and both

axes in the opposite directions with the original system as depicted in Figure 3.6a, can significantly simplify the solution of the problem. Thus, we define the new axes as $\tilde{t} = T_l - t$ and $\widetilde{N} = N_l - N$, where the length of the lateness period is denoted by $T_l = A^{-1}(N_l)$. In this system, $\tilde{A}_e(\tilde{t})$ and $\widetilde{D}(\tilde{t})$ also denote cumulative arrival and departure curves of late commuters, respectively. Notice as well that the slopes of the curves remain exactly the same in the auxiliary coordinate system due to inversion of the both axes in this system. So, here, we first employ the auxiliary coordination system to derive the equilibrium arrival of the late commuters, and then rewrite the solution in the original coordination system.



(a) Variation of delay in the queueing diagram

(b) Probability distribution of lateness penalty factor

Figure 3.6. Relation between the equilibrium arrivals of the late commuters with the distribution of the earliness penalty factor

The equilibrium arrival of the late commuters in the auxiliary coordinate system is very similar to the early commuter problem. So, with same line of reasoning for the relation (3.22), the arrival time of the late commuters in the auxiliary coordinate can be formulated as below:

$$\tilde{t}_A = \frac{N_l}{\mu} \int_0^{l_{\tilde{t}_A}} (1 + l) f_l(l) dl \tag{3.26}$$

where, $l_{\tilde{t}_A}$ denotes the lateness penalty factor of the commuter who arrives to the bottleneck at time $\tilde{t}_A$ in the auxiliary coordination system, corresponding to t_A in the original system. $f_l(l)$ and $F_l(l)$ also represents PDF and CDF of the lateness penalty factor and of the late commuters, respectively, as illustrated in the Figure 3.6b.

Accordingly, the cumulative arrival of heterogeneous late commuters in the equilibrium condition can be approximated as below:

$$\tilde{A}_l(\tilde{t}_A) = N_l F_l\left(\hat{l}_{\tilde{t}_A}\right) \tag{3.27}$$

where, $\hat{l}_{\tilde{t}_A}$ denotes the solution of equation (3.26) as a function of $\tilde{t}_A$.

As a result, we can rewrite the equation (3.27) in the original coordination system as below:

$$A_l(t_A) = N_l - \tilde{A}_l(T_l - t_A) = N_l\left(1 - F_l\left(\hat{l}_{t_A}\right)\right) \tag{3.28}$$

Here, $\hat{l}_{t_A}$ represents the solution of the equation (3.26) by plugging $\tilde{t}_A = T_l - t_A$.

Moreover, we can derive the distribution of the lateness factor of the equilibrium arrivals following similar approach as for the earliness factor as below:

$$l_{t_A} = F_l^{-1}\left(\frac{\tilde{N}_{t_A}}{N_l}\right) \tag{3.29}$$

The generalized cost of the late commuters can be approximated similar to generalized cost of early commuters, presented in equation (3.25), as follows:

$$\mathrm{C}_{\tilde{t}_A}\left(\widetilde{\mathrm{N}}_{\tilde{t}_A}\right) = \tau_F + \left(\tilde{A}_l^{-1}\left(\widetilde{\mathrm{N}}_{\tilde{t}_A}\right) - \frac{\tilde{N}_{t_A}}{\mu}\right) + F_l^{-1}\left(\frac{\widetilde{\mathrm{N}}_{\tilde{t}_A}}{N_l}\right) \cdot \left(t_{w,\widetilde{\mathrm{N}}_{\tilde{t}_A}} - \frac{\widetilde{\mathrm{N}}_{\tilde{t}_A}}{\mu}\right),\ \widetilde{\mathrm{N}}_{\tilde{t}_A} \in [0, N_l] \tag{3.30}$$

3.1.2.3 Equilibrium Condition of the Efficient Frontier

Previously, we derived an analytical solution for the morning commute problem given the PDFs of the schedule deviation penalty factors of the heterogeneous users. In this section, we show that the proposed analytical solution of the problem is necessary and sufficient for satisfying the equilibrium conditions of the bottleneck derived according to Daganzo (1985).

In the equilibrium condition, no commuter should have an incentive to change his/her arrival time to the bottleneck to be earlier or later. To generalize the first equilibrium condition, we consider the effect of changing arrival time for a single commuter with arrival time t_A, which is earlier than his/her wished time $t_{w,1}$, as illustrated in Figure 3.7a. So, we denote the earliness penalty factor of this commuter with e_{t_A}. The effect of a small shift in the arrival time by Δt earlier will result in departing the bottleneck $\dot{A}_e(t_A)\Delta t/\mu$ earlier in time, where $\dot{A}_e(t_A)$ denotes the slope of the arrival curve at time $t_A \in (t_s, \hat{t})$. In equilibrium, earliness must increase in the same amount as the reduction in delay for the commuter. Accordingly, the resulting change in delay and earliness of the commuter can be calculated as follows:

$$\Delta\tau_D = \tau_{D,2} - \tau_{D,1} = +\Delta t - \frac{\dot{A}_e(t_A)\Delta t}{\mu} \tag{3.31}$$

$$\Delta\tau_E = \tau_{E,2} - \tau_{E,1} = +e_{t_A}\frac{\dot{A}_e(t_A)\Delta t}{\mu} \tag{3.32}$$

In the equilibrium condition, users should not be able to reduce their generalized cost by changing their arrival times. Thus, a marginal shift in arrival time by Δt should not affect the generalized cost for the user. So, the following condition must hold:

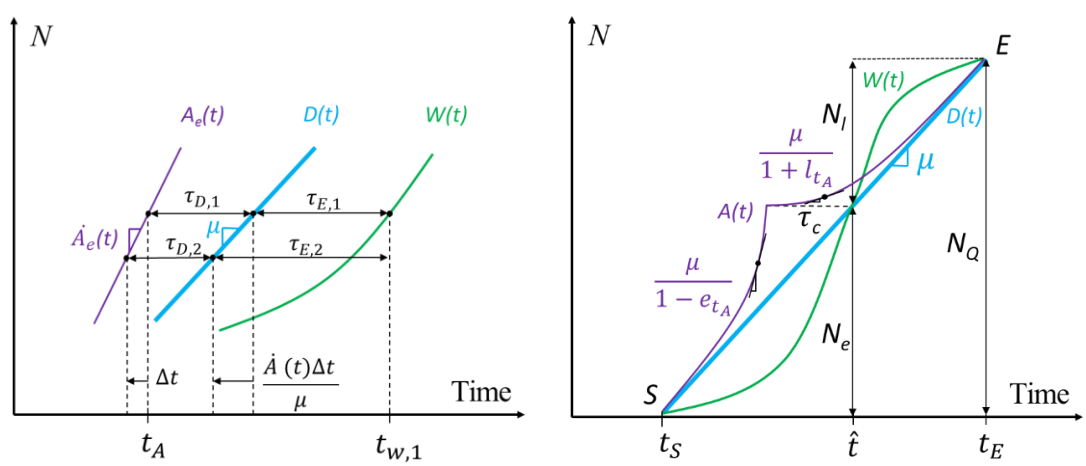
$$\Delta\tau_D + \Delta\tau_E = 0 \tag{3.33}$$

By substituting (3.31) and (3.32) into (3.33), $\dot{A}_e(t_A)$ can be solved in terms of e_{t_A} as illustrated in the Figure 3.7b:

$$\dot{A}_e(t_A) = \frac{\mu}{1-e_{t_A}}, \forall t_A \in (t_s, \hat{t}), e_{t_A} \in [0,1) \tag{3.34}$$

The same approach can also be followed to derive the slope of the second part of the equilibrium arrival curve $\dot{A}_l(t_A)$ in terms of the lateness penalty factor of the commuter who arrives to the bottleneck at time t_A, l_{t_A}, and departs the bottleneck later than wished, as illustrated in Figure 3.7b:

$$\dot{A}_l(t_A) = \frac{\mu}{1+l_{t_A}} \qquad , \forall t_A \in (\hat{t}, t_E), l_{t_A} \in [0, \infty) \tag{3.35}$$



(a) Variation of delay and earliness for a user resulting from a shift of Δt in the arrival time.

(b) Queueing diagram of bottleneck user equilibrium

Figure 3.7. Bottleneck user equilibrium of heterogeneous commuters with an S-shaped wished curve

The schedule penalty factors are inherently defined as nonnegative, while the earliness penalty factor also needs to be less than one to remain within its domain in equation (3.34). On the basis of assignment order rule in Corollary 3.2, the slope of the arrival curve generally starts increasing monotonically from μ when $e_{t_A} = 0$ at point S, and goes to infinity as $e_{t_A} \to 1$ when delay is at the maximum possible value, τ_c. Then, this slope drops to near zero at the beginning of the lateness period where l_{t_A} is greatest. The slope of arrival curve monotonically increases towards μ at point E as $l_{t_A} \to 0$.

In the Proposition 3.2, we show that the first equilibrium condition of (3.34) and (3.35) hold if and only if the arrival time choices of the commuters are along an efficient frontier with the specific properties discussed in Section 3.1.2.2.

Proposition 3.2 (equilibrium condition). The first equilibrium condition of equations (3.34) and (3.35) holds if and only if the heterogeneous commuters are distributed along the efficient frontier of arrival choices according to the property (3.9).

Proof.

Part 1: Equations (3.34) and (3.35) hold if property (3.9) governs.

We first prove this argument for the early commuters and then extend the proof for the late commuters.

According to the relation (3.17), we can take the first derivative of delay with respect to the departure time using the differentiation chain rule, where $t_A = A^{-1}(N)$:

$$\frac{\partial \tau_{D,N_{t_A}}}{\partial t_D} = 1 - \frac{\partial A_e^{-1}(N_{t_A})}{\partial N} \cdot \frac{\partial N}{\partial t_D} \tag{3.36}$$

For one thing, with a little algebraic manipulation in equation (3.34) we have:

$$\frac{\partial A_e^{-1}(N_{t_A})}{\partial N} = \frac{1}{\dot{A}_e(t_A)} = \frac{1-e_{t_A}}{\mu} \tag{3.37}$$

For another, the slope of the cumulative departure curve equals to the fixed capacity of the bottleneck:

$$\frac{\partial N}{\partial t_D} = \frac{\partial D(t_D)}{\partial t_D} = \mu \tag{3.38}$$

By substituting the equations (3.37) and (3.38) in (3.36), we can derive the property (3.9) for the early commuters. With same line of reasoning we can derive the property (3.9) for the late commuters as well.

Part 2: Property (3.9) holds if equations (3.34) and (3.35) govern.

Equation (3.23) is directly concluded from condition (3.9).So, following from equation (3.23), we can take the first derivative of the arrival curve with respect to time using the differentiation chain rule:

$$\dot{A}_e(t_A) = N_e \frac{\partial \mathrm{F_e}(e_{t_A})}{\partial e_{t_A}} \cdot \frac{\partial e_{t_A}}{\partial t_A} = N_e \mathrm{f_e}(e_{t_A}) \dot{e}_{t_A} \tag{3.39}$$

By solving the equation (3.21) for $\mathrm{f_e}(e_{t_A})$ and substituting the result in the equation (3.39), we can conclude the equation (3.34) for the early commuters. Likewise, equation (3.35) can be derived following a similar approach for the late commuters.■

The second equilibrium condition (3.3) also can be generalized for heterogeneous commuters. In this respect, the peak period demand of the bottleneck, N_Q, should split between the earliness and lateness periods such that the maximum delay of the both periods is the same, τ_c. The delay of the commuters N_e and N_l, is horizontal the difference between the arrival and departure curves in the earliness and lateness periods (See Figure 3.5a and 6a), respectively, which equals:

$$\tau_c = \frac{N_e}{\mu} - A_e^{-1}(N_e) = \tilde{A}_l^{-1}(N_l) - \frac{N_l}{\mu} \tag{3.40}$$

With a little manipulation, we can conclude from (3.40) that in the equilibrium condition the combined duration of the earliness and lateness periods equals the time it takes for the queue to disappear from the bottleneck:

$$A_e^{-1}(N_e) + \tilde{A}_l^{-1}(N_l) = \frac{N_q}{\mu} \tag{3.41}$$

As a result, the second condition (3.41) sheds light on the relation between N_e and N_l in the equilibrium condition, where the schedule penalty factors have probability distributions over the population of the users.

The generalized first and second equilibrium conditions proposed in this research can determine the equilibrium arrival of the heterogeneous users to the single bottleneck. Given the PDFs of the schedule penalty factors, $\mathrm{f}_e(e)$ and $\mathrm{f}_l(l)$, the arrival time of the early and late commuters can be numerically estimated in terms of N_e and N_l using the relations (3.22) and (3.26), respectively. Knowing the distribution of the arrivals, it is possible to estimate the values of $A_e^{-1}(N_e)$ and $\tilde{A}_l^{-1}(N_l)$ in terms of of N_e and N_l, respectively. Thus the relation between N_e and N_l in the equilibrium condition can be determined by plugging the results into the equation (3.41). Although this problem can be solved numerically for any given distribution of schedule penalty factors, it has a closed form solution when that the integrals of relations (3.22) and (3.26) have definite solutions. In the following section, we use these results to propose a dynamic pricing pattern that can optimize the system by avoiding the delay. In section 3.3, a closed form solution of the morning commute problem is derived when the probability distributions of schedule penalty factors are assumed to be uniform over the population of the users.

3.2 System Optimum and Pricing

Rational users tend to minimize the generalized cost of their own commutes by adjusting their arrival time to the bottleneck. The cumulative result of the individual decisions leads to the user equilibrium condition in which no one can improve his/her cost by shifting his/her own arrival time individually. In the user equilibrium condition, each commuter experiences a combination of delay and schedule penalty, in addition to the free flow travel time, for travelling through the bottleneck. Nonetheless, the total generalized cost for the commuters can be minimized by avoiding unnecessary delay in the bottleneck. As a result, such a system can be optimized by dynamically charging the users as much as the delay they expect in the user equilibrium condition in order to eliminate the delay entirely (Arnott et al., 1990; Gonzales and Daganzo, 2012).

In the general case that schedule penalty factors of the users have a probability distribution over the population of the commuters, equation (3.9) can demonstrate the variation of delay over time in equilibrium as illustrated in Figure 3.4. By plugging the values of $N = \mu t$ and $\widetilde{N} = N_Q - \mu t$ in equations (3.24) and (3.29), and substituting the results in equation (3.9), we can derive the dynamic pricing pattern that can optimize the system by avoiding delay, as illustrated in Figure 3.8:

$$\dot{\$}(t) = \begin{cases} F_e^{-1}\left(\frac{\mu t}{N_e}\right), & t \in (t_s, \hat{t}) \\ -F_l^{-1}\left(\frac{N_Q - \mu t}{N_l}\right), & t \in (\hat{t}, t_E) \end{cases} \tag{3.42}$$

where, $\dot{\$}(t)$ denotes the variation rate of the optimal price unit of time over the peak period.

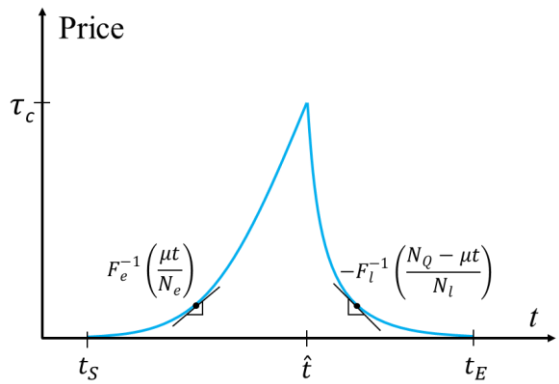


Figure 3.8. The system optimal dynamic pricing of the bottleneck with heterogeneous demand

As explained in the Corollary 3.1, such slope monotonically increases from zero at time t_S to unity at time $\hat{t}$, then it drops to negative infinity but monotonically increases back to zero at time t_E.

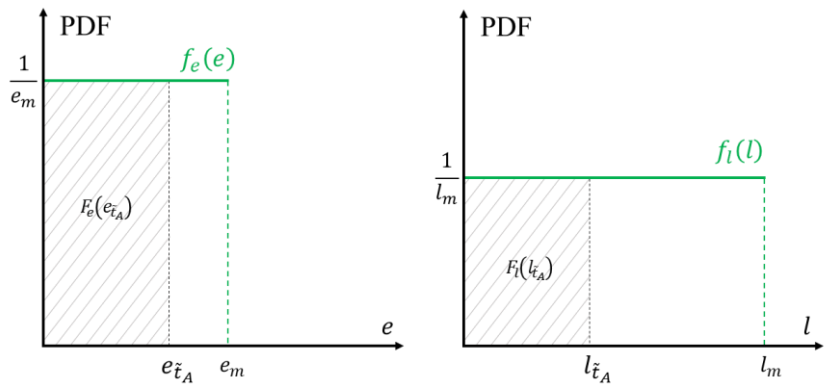
As explained in Geroliminis and Levinson (2009) and Gonzales and Daganzo (2012), system optimum prices for single bottlenecks also apply to pricing networks. A bottleneck is typically assumed to have a constant or nearly constant capacity that is independent of the length of the queue that forms. The dynamics of congestion in networks differ from queueing at bottlenecks, because the capacity of a network to serve vehicle trips is a function of the accumulation vehicles (or the vehicle density) in the network. When queues in the network are long enough to spill back and block upstream intersections, the result is a reduction in network-wide flows as represented by the characteristic downward sloping

branch of the macroscopic fundamental diagram (Daganzo, 2007; Daganzo and Geroliminis, 2008). In this respect, Zheng et al. (2012) combines the macroscopic congestion theory with an agent-based simulator to propose a dynamic cordon pricing scheme that accounts for the complexity in the travel behavior of the users in an urban network. Zheng et al. (2015) also presents a more efficient and equitable area-based pricing scheme for a bi-model network based on macroscopic fundamental diagram by considering the heterogeneity in income level and value of time of the travelers. Moreover, the macroscopic model of the traffic congestion has applications in dynamic pricing the limited parking in a multimodal urban networks (Zheng and Geroliminis ,2014). System optimum prices for bottlenecks eliminate queues by incentivizing users to travel when there is bottleneck capacity available to accommodate their trip. The same holds for networks. By incentivizing users to spread their travel over time, optimal bottleneck prices can be applied to networks in order to prevent vehicle accumulations from reaching the congested conditions associated with queue spillbacks. In this way, flows on the network can be maintained at efficient uncongested traffic states so that the network behaves like a single bottleneck with fixed capacity. In this respect, Zheng et al. (2012) com

3.3 Closed Form Solution for Uniformly Distributed Schedule Penalty Factors

In this section, we use the proposed equilibrium model to derive the closed form solution for a special case of the morning commute problem at a single bottleneck with

heterogeneous travel demand. In this case, both schedule penalty factors of the commuters are assumed to have uniform PDF functions, distributed from zero to their maximum values e_m and l_m, while e_m also needs to be less than one, as illustrated in Figure 3.9a,b. However, it is worth pointing out that the proposed analytical approach remains general for the distributions of schedule penalty factors for which the integrals of the relations (3.22) and (3.26) are well defined functions.



(a) Earliness penalty factor (b) Lateness penalty factor

Figure 3.9. Uniform probability distributions of the schedule penalty factors

In this case that distributions of the schedule penalty factors are defined by uniform PDF functions, the result of the first equilibrium condition in relations (3.22) and (3.26) can be further simplified to approximate the arrival time of early and late commuters, respectively:

$$t_A = \frac{N_e}{\mu e_m}\left(e_{t_A} - \frac{e_{t_A}^2}{2}\right), 0 \leq e_{t_A} \leq e_m \tag{3.43}$$

$$\tilde{t}_A = \frac{N_l}{\mu l_m}\left(l_{t_A} + \frac{l_{t_A}^2}{2}\right),\ 0 \le l_{t_A} \le l_m \tag{3.44}$$

On this basis, the cumulative arrival of the early and late commuters in the equilibrium condition can be derived according to the relations (3.23), (3.27), and (3.28) as follows:

$$A_e(t_A) = \frac{N_e}{e_m}\left(1 - \sqrt{1 - \frac{2\mu t_A e_m}{N_e}}\right) \tag{3.45}$$

$$\tilde{A}_l(\tilde{t}_A) = \frac{N_l}{l_m}\left(-1 + \sqrt{1 + \frac{2\mu \tilde{t}_A l_m}{N_l}}\right) \tag{3.46}$$

$$A_l(t_A) = N_l - \frac{N_l}{l_m}\left(-1 + \sqrt{1 + \frac{2\mu (T_l - t_A) l_m}{N_l}}\right) \tag{3.47}$$

where, the length of the lateness period, T_l, can be derived in this problem as below:

$$T_l = A^{-1}(N_l) = \frac{N_l}{\mu}\left(1 + \frac{l_m}{2}\right) \tag{3.48}$$

The generalized cost for early and late commuter also can be approximated by substituting the inverse cumulative arrival functions derived from relations (3.45) and (3.46) and inverse CDF of the uniform schedule factor according to the Figure 3.9a,b in relations (3.25) and (3.30):

$$\mathrm{C}_{t_A}\left(\mathrm{N}_{t_A}\right) = \tau_F + \left(\frac{e_m N_{t_A}^2}{2N_e\mu}\right) + \frac{e_m N_{t_A}}{N_e} \cdot \left(t_{w,N_{t_A}} - \frac{N_{t_A}}{\mu}\right),\ N_{t_A} \in [0, N_e] \tag{3.49}$$

$$\mathrm{C}_{\tilde{t}_A}\left(\widetilde{\mathrm{N}}_{\tilde{t}_A}\right) = \tau_F + \left(\frac{l_m \widetilde{N}_{t_A}^2}{2N_l\mu}\right) + \frac{l_m \widetilde{N}_{t_A}}{N_l} \cdot \left(t_{w,\widetilde{\mathrm{N}}_{\tilde{t}_A}} - \frac{\widetilde{\mathrm{N}}_{\tilde{t}_A}}{\mu}\right),\ \widetilde{\mathrm{N}}_{\tilde{t}_A} \in [0, N_l] \tag{3.50}$$

In addition, the relation between N_e and N_l can be revealed through the second equilibrium condition by substituting the inverse cumulative arrival functions derived from (3.45) and (3.46) in equation (3.41):

$$\frac{N_e}{N_l} = \frac{l_m}{e_m} \tag{3.51}$$

Following from above, the maximum delay that commuters experience in the equilibrium condition can be approximated according to the equation (3.40):

$$\tau_c = \frac{e_m N_e}{2\mu} = \frac{l_m N_l}{2\mu} \tag{3.52}$$

As explained in Section 3.2, such delay can be entirely avoided in the system optimum by charging the commuters a dynamic price equal to the cost of the delay they expect in the user equilibrium condition, according to the equation (3.42):

$$\dot{\$}(t) = \begin{cases} \frac{e_m \mu t}{N_e}, & \mathrm{t} \in (t_s, \hat{t}) \\ -\frac{l_m(N_Q - \mu t)}{N_l}, & \mathrm{t} \in (\hat{t}, t_E) \end{cases} \tag{3.53}$$

As it can be inferred from equation (3.53), the time-variant price monotonically rises from zero, at time t_S, to its maximum value τ_c, at time $\hat{t}$, meanwhile its rate also increases from zero to e_m. At this point, it rapidly decreases with slope of negative infinity back to zero, where its slope goes to zero as well, at time t_E.

3.4 Numerical Example

To provide a numerical example, we employ the proposed model to solve the morning commute problem for a single bottleneck with a fixed capacity of 100 vehicles per minute. This bottleneck is assumed to have a peak period demand of 10,000 vehicles over 100 minutes with a smooth s-shaped wished curve as illustrated in Figure 3.10. To observe the effect of heterogeneity in the preferences of the commuters, the equilibrium arrival of the users is plotted for three different distribution scenarios of the schedule penalty factors in Figure 3.10.

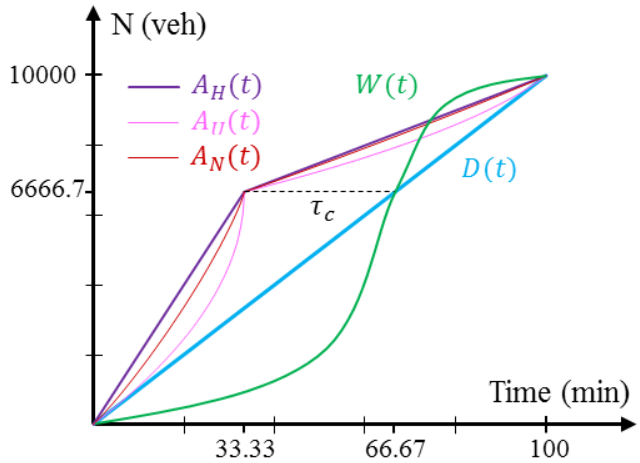


Figure 3.10. Queueing diagram of bottleneck user equilibrium in three different heterogeneity scenarios

In the first scenario, commuters are assumed to have homogeneous schedule penalty factors $e = 0.5$ and $l = 1$. In this case, the equilibrium arrival of the homogenous users (purple curve), $A_H(t)$, turns to a piecewise curve with the slopes of 200 veh/min and 50 veh/min in the earliness and lateness periods, respectively. So, the total delay that users experience during their commutes can be approximated as the area between the arrival and departure curves, which in the first scenario equals to 2777.5 hours. In the second scenario, we assume the distributions of the schedule penalty factors uniform over time as illustrated in Figure 3.9a,b where $e_m = 1$ and $l_m = 2$. In this scenario, the equilibrium arrival of the commuter with uniform distribution of schedule penalty factors (pink curve), $A_U(t)$, can be analytically approximates according to (3.45) and (3.47). Finally, in the third scenario, the schedule penalty factors of the users are assumed to have truncated normal distributions with average values $\bar{e} = 0.5$ and $\bar{l} = 1$, standard deviations $\sigma_e = \sigma_l = 0.12$, and the same

upper and lower bounds as the aforementioned uniform distributions in the second scenario. In this case, the equilibrium arrival of the users with both normally distributed schedule penalty factors (red curve) , $A_N(t)$, can be estimated using the relations (3.22) and (3.26). As illustrated in Figure 3.10, the arrival curve of the users with normal distributions of the schedule penalty factors falls between the arrival curves of the users with homogenous and uniform distributions of schedule penalty factors all the points over time. Accordingly, the total delay of the commuters reduces by 33% and 9% in the second and third scenarios, respectively; however, the maximum delay remains exactly the same, $\tau_c = 33.3$ min, in all scenarios.

3.5 Probability Distributions of the Schedule Penalty Factors

In this research, we proposed an analytical solution to the morning commute problem when there is a heterogeneity associated with the schedule penalty factors of the users. The underlying assumption here is that the probability distributions of the earliness and lateness penalty factors are known independently for early and late commuters, respectively. Nevertheless, it may be more realistic to assume a joint probability distribution function for the schedule penalty factors that defines the proportion of (e, l) pairs among the population of the commuters. In this section, we propose a method to retrieve the independent probability distributions of earliness and lateness penalty factors from a given joint distribution.

Rational users tend to minimize their own cost by adjusting their arrival time to the bottleneck. Assuming that values of the both schedule penalty factors vary over the population of the heterogeneous commuters, they choose their arrival time to the bottleneck by comparing the minimum possible costs of the earliness and lateness periods according to their own schedule preferences. In this respect, we define the equivalent schedule penalty factor curve as a set of penalty factor pairs and each point represents an equivalent earliness and lateness penalty factors with equal generalized cost in the earliness and lateness periods, respectively. This relation between the equivalent schedule penalty factors can be determined by equalizing the generalized cost functions (3.25) and (3.30), and expressing $N = N_e F_e(e)$ and $\widetilde{N} = N_l F_l(l)$ in the result in terms of e and l. Since we are writing this equation in terms of schedule penalty factors, it is now possible to denote the generalized cost for commuters with $C(\cdot)$ for the purpose of simplification:

$$C(N_e F_e(e)) = \mathrm{C}(N_l F_l(l)\ ) \tag{3.54}$$

This equation reveals the relationship between the equivalent schedule penalty factors in the equilibrium condition. For example, in case that schedule penalty factors have uniform probability distributions as assumed in Section 3.3, equation (3.54) can be simplified to a linear relation between the equivalent schedule penalties factors as below:

$$l_{eq}(e) = \frac{l_m}{e_m} e \tag{3.55}$$

where, $l_{eq}(e)$ represents the equivalent lateness penalty factor of a earliness penalty factor.

As a result, for a given value of e_o, commuters with $l > l_{eq}(e_o)$ will depart earlier than they wished to minimize their generalized cost. In contrast, commuters with $l < l_{eq}(e_o)$ prefer late departures in order to keep their generalized costs minimized in the equilibrium condition. Assuming the joint PDF of the schedule penalty factors given, $p_{e,l}(e, l)$, we can derive $f_e(e)$ and $f_l(l)$ by projecting the area under $p_{e,l}(e, l)$ for early and late commuters, as illustrated in the Figure 3.11.

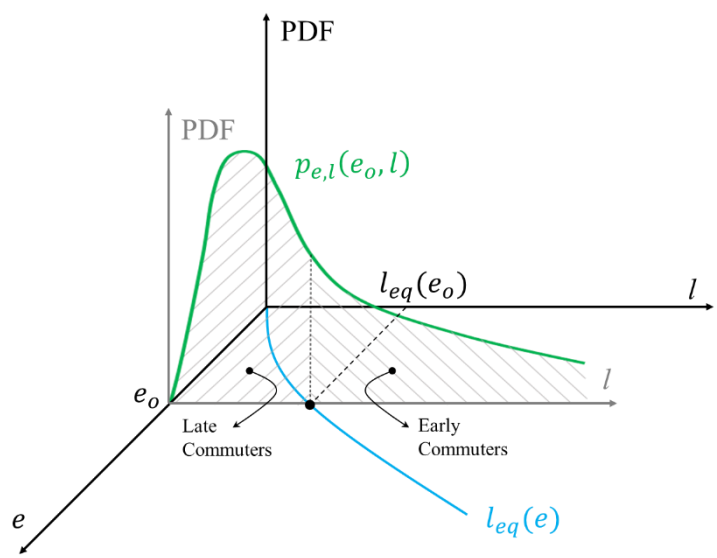


Figure 3.11. Joint probability distribution of the schedule penalty factors

On this basis, the value of the $f_e(e_o)$ is proportional to the area under $p_{e,l}(e_o, l)$ for $l > l_{eq}(e_o)$, which represents the early commuters. Likewise, $f_l(l_o)$ is also proportional to the

area under $p_{e,l}(e, l_o)$ for $e > l_{eq}^{-1}(l_o)$, which represents the late commuters. Notice as well that we are splitting a joint distribution function $p_{e,l}(e, l)$ into two parts. Thus it is necessary to divide each of these areas by the total volume under $p_{e,l}(e, l)$ for both early and late commuters, to convert these numbers back to probabilities to derive the independent PDFs, $f_e(e)$ and $f_l(l)$:

$$f_e(e) = \frac{\int_{l_{eq}(e)}^{l_m} P_{e,l}(e,l)dl}{\int_0^{e_m} \int_{l_{eq}(e)}^{l_m} P_{e,l}(e,l)dlde} \tag{3.56}$$

$$f_l(l) = \frac{\int_{l_{eq}^{-1}(l)}^{e_m} P_{e,l}(e,l)de}{\int_0^{l_m} \int_{l_{eq}^{-1}(l)}^{e_m} P_{e,l}(e,l)dedl} \tag{3.57}$$

As a result, we can use these relations to retrieve the probability distributions of earliness and lateness factors independently for the commuters who depart the bottleneck early and late, respectively. Having the PDFs $f_e(e)$ and $f_l(l)$, it becomes possible to employ the proposed model to analytically solve the morning commute problem with heterogonous user preferences.

The proposed analytical solution of the morning commute problem can be also inversely used to approximate the distribution of schedule penalty factors of the heterogeneous users from empirical data of users' cumulative arrivals to the bottleneck. Having the cumulative arrival curve of the heterogeneous users to the bottleneck in the

equilibrium condition, $A(t)$, we can employ the proposed equilibrium model to approximate the actual PDFs of the independent schedule penalty factors, $f_e(e)$ and $f_l(l)$, using the empirical data from the bottleneck. Observations of traffic approaching the San Francisco-Oakland Bay Bridge show that queues regularly develop for traffic heading westbound across the bridge into San Francisco each morning. An empirical analysis of these queueing patterns has been presented in Gonzales and Christofa (2013). Analysis of traffic flows across a series of loop detectors shows that the queue and corresponding queueing delay increases in a convex way at the beginning of each rush and decreases in a convex way at the end of each rush in a pattern similar to the shape in Fig. 7. Such data can be used to estimate the distribution of earliness and lateness penalty factors.

In this respect, we use equation (3.14) to relate the CDF of the earliness penalty factor of the heterogeneous users to their cumulative arrivals to the bottleneck. Relation (3.34) also determines the earliness penalty factor of a commuter as function of his/her arrival time to the bottleneck. As a result, the CDF, and subsequently PDF, of the earliness penalty factor of the users can be approximated using the empirical data from equilibrium arrival of the heterogeneous users to the bottleneck by combing the relations (3.14) and (3.34). Distribution of the lateness penalty factors of the users also can be estimated using the empirical data in a very similar way.

As the first step, we replace $N_{t_A} = A_e(t_A)$ in the equation (3.14) to derive the CDF of the earliness penalty factors of the heterogeneous commuters in term of the cumulative arrival function of the users to the bottleneck as follows:

$$F_e\left(e_{t_A}\right) = \frac{A_e(t_A)}{N_e} \tag{3.58}$$

Then, we relate the arrival time of the heterogeneous users to their schedule preferences by solving equation (3.34) for t_A as below:

$$t_A = \dot{A}_e^{-1}\left(\frac{\mu}{1-e_{t_A}}\right) \tag{3.59}$$

where, $\dot{A}_e^{-1}(\cdot)$ denotes the inverse arrival function for the early arrivals. In case that the cumulative arrival of the heterogeneous commuters can be estimated using the empirical data as a differentiable function $A_e(\cdot)$, CDF of earliness penalty preferences of the early commuters can be approximated by plugging t_A from equation (3.59) into relation (3.58) as below:

$$F_e\left(e_{t_A}\right) = \frac{A_e\left(\dot{A}_e^{-1}\left(\frac{\mu}{1-e_{t_A}}\right)\right)}{N_e} \tag{3.60}$$

Alternatively, when the cumulative arrival of commuters is a discrete dataset that cannot be estimated with a differentiable function, the arrival time t_A corresponding to e_{t_A} can be numerically calculated at each point using the empirical arrival time data of the

heterogeneous commuters to the bottleneck according to the equation (3.59). Subsequently, $F_e\left(e_{t_A}\right)$ can be estimated for previously determined t_A based on the relation (3.58).

Similar steps can be also followed to derive a same set of results for the distribution of the lateness penalty preferences of the users. In this respect, equation (3.60) can be adapted to approximate the CDF of the lateness penalty factor of the late commuters as follows:

$$F_l\left(l_{t_A}\right) = 1 - \frac{A_l\left(\dot{A}_l^{-1}\left(\frac{\mu}{1+l_{t_A}}\right)\right)}{N_l} \tag{3.61}$$

where, $\dot{A}_l^{-1}(\cdot)$ denotes the inverse arrival function for the late arrivals.

As a result, the PDF of the schedule penalty factors can be derived by taking the first derivation of their CDFs. Approximating PDF of the heterogeneous users quantifies the heterogeneity in the preferences of users, which can be useful for modeling systems with similar characteristics of demand. Thus, the approximated PDFs can be used in analyzing the effects of alteration in characteristics of demand and system on the equilibrium condition. Moreover, designing a dynamic pricing strategy that can optimize the system by minimizing the total costs strictly depends on accurate approximation of PDF of the schedule penalty preferences of the commuter. Hence, it is of great importance to have an accurate approximation of the heterogeneity in preferences of the users based on the empirical data from the network.

3.6 Summary

This paper adapts the concept of the efficient frontier from Portfolio Theory to propose an analytical solution for the equilibrium condition of the linear bicriterion choice problem when there is a heterogeneity associated with relative importance of such criteria among the users. More specifically, we study the morning commute problem of a single bottleneck with a fixed capacity and time-dependent demand when there is heterogeneity associated with the schedule penalty factors of the users. When the bottleneck capacity is insufficient to fulfill the time-dependent demand, users experience a combination of delay and schedule deviation in their commutes, while the relative importance of these factors varies among the heterogeneous users. The rational tendency of users to minimize their own generalized cost by adjusting their arrival times to the bottleneck eventually leads to the user equilibrium condition in which no one can improve his/her generalized cost by individually shifting his/her arrival time. This bicriterion choice problem has strong similarities with the investment problem in the free market, which makes it possible to adapt the concept of the efficient frontier from finance to propose an analytical solution to the morning commute problem with heterogeneous users' preferences.

In this research, we first showed that the efficient frontier of the choices has specific properties that remain general for different problems. Given the independent PDFs of the schedule penalty factors, we employed such properties to analytically derive the equilibrium arrival of the heterogeneous commuters with a general s-shaped wished

departure curve to the bottleneck. On this basis, a dynamic pricing pattern is also presented for the bottleneck with heterogeneous demand that can optimize the system by avoiding delay. The proposed dynamic pricing strategy is also adapted to improve the performance of the network. Deriving an analytical solution for the morning commute problem with heterogeneous user preferences also provides intuition regarding the effect of different factors on the travel behavior of the user in the equilibrium condition. The intuition obtained from the analytical solution of the problem can be crucial for predicting the change in the equilibrium condition in case of alternations in characteristics of the demand or system. To provide an example of the proposed model, we derived a closed form solution of equilibrium arrivals of commuters as well as an optimal dynamic pricing for the bottleneck assuming that schedule penalty factors have uniform probability distributions. However, the proposed model can be employed to solve the morning commute problem for any given probability distributions of the schedule penalty factors and a general s-shaped wished departure curve. Finally, we provide an explanation on how independent PDFs of the schedule penalty factors can be retrieved from their general joint PDF. It is also shown that the analytical results can be inversely employed to approximate the distributions of the schedule penalty factors from the empirical data on arrival time of the heterogeneous users to the bottleneck. Such information can be vital in analyzing the equilibrium condition and designing a dynamic pricing strategy that can effectively optimize the system.

The result of this research can be extended for other transportation systems with limited capacity and time-dependent demand that can be modeled as queueing systems. For

example, Vickrey's theory has been used to propose an analytical model for a demand responsive transit (DRT) service that provides a door-to-door service for users with time-dependent demand (Amirgholy and Gonzales, 2015b). As a result, the analytical solution proposed in this paper for the morning commute problem can be also employed to include the heterogeneity in user preferences of a DRT system. Accordingly, the dynamic pricing strategy presented in Section 4 can be also implemented to optimize the DRT system by minimizing the total cost of the operator and DRT service users. Furthermore, in this research we adapted the concept the efficient frontier to include heterogeneity in schedule penalty factors of the users in the analytical model we propose for the morning commute problem with a general distribution of the wished departure times. Nonetheless, heterogeneity of the users' value of time is another key factor that can be included in model to improve the effectiveness of the dynamic pricing policy. In this respect, the reality that non-identical users consider different values for the time should be taken into account to propose an optimal pricing strategy that can entirely eliminate the delay and minimize the total generalized cost for the commuters.

As the next step, we propose an extension to this model to also account for the heterogeneity in value of time of the commuters. In this way, we can determine the optimal pricing of the bottleneck by considering heterogeneities in relative importance of the deviation from the schedule as well as the value of the time to the users in the analytical model. In general, the concept of the efficient frontier has been identified an appropriate

tool for modeling the equilibrium condition of multi-criterion choice problems with heterogeneous preferences regarding the relative importance of the conflicting criteria.

CHAPTER 4

DEMAND RESPONSIVE TRANSIT SYSTEMS WITH TIME-DEPENDENT DEMAND

Demand responsive transit (DRT) systems are a class of transit services in which a fleet of vehicles dynamically changes routes and schedules in order to accommodate demand within a service area. A DRT system naturally has flexibility in providing service, which allows it to adapt to variations in the demand. This property of DRT makes it possible to eliminate the access distance for transit users by providing a curb-to-curb trip. Nonetheless, the quality of the service depends on the operating capacity of the DRT system as well as spatiotemporal characteristics of the demand. The operating capacity of the DRT system can be defined as the maximum number of requests that can be served in unit of time. In this part of the research, we show that the DRT system with limited operating capacity and time-dependent demand can be modeled as a queueing system. On this basis, we propose a model and optimization approach for DRT service that is used to minimize the total cost to the agency and users combined.

An analytical model for DRT systems based on Rahimi et al. (2014) and Daganzo (1978) is employed to approximate the components of the operating cost of the DRT system: fleet size, total vehicle hours traveled (VHT), and total vehicle miles traveled (VMT) in the network. Given the service area of the DRT system, these components of the agency cost can be approximated as functions of the number of waiting passengers that

have requested service and the maximum rate that operators can serve passengers per time (i.e., operating capacity). The total operating cost for the agency can be estimated as a linear combination of these components (Rahimi et al., 2014).

In addition to the expenses to the agency for running the service, it is also necessary to account for the costs that users endure to use the service. To this end, Vickrey's (1969) congestion theory is adapted to approximate the costs that the DRT users experience for the service when the operating capacity of the system is inadequate to meet the demand. In this case, the user equilibrium can be conceived as the result of competition between DRT users who are each minimizing their own travel costs, which include the waiting time to be picked up, the traveling time in the vehicle, and the cost for arriving earlier or later than preferred. In an equilibrium condition, no one has an incentive to change his/her own travel time. However, it is still possible to reduce the total costs of the system by optimizing the DRT operations and managing the temporal distribution of demand. The capacity of the service and the number of passengers awaiting pick-up are decision variables that can be maintained at optimal levels over time to minimize the total cost of the system. Since demand tends to peak during certain times of the day, an effective demand management strategy that can spread the demand uniformly over time has a key role in optimizing the operations of a DRT system.

The objective of this Chapter is to model the dynamics of DRT system scheduling and operations and to identify a management strategy to incentivize users to adapt their request times to be more uniform over time. A dynamic pricing policy can be implemented as an

effective strategy to improve the efficiency of the DRT system by forming a uniform distribution of the demand and avoiding the underutilization of the optimal capacity during off-peak times. As a result, the total cost of the agency and users can be minimized by choosing optimal values for the system capacity and the number of waiting requests. Meanwhile an appropriate demand management strategy that can make the demand uniform is required to keep the system optimized over time.

The analytical solution of the morning commute problem presented in Chapter 3 can be used with the proposed model for the operation of DRT system in this chapter to account for the heterogeneity in schedule penalty preferences of the users. Accordingly, the dynamic pricing strategy proposed for the DRT system will be adapted as well.

4.1 Modeling Tools

Agencies possess limited equipment, crews, and facilities with which to operate DRT services, and the costs of acquiring additional resources can be very expensive so agencies have an incentive to use their resources as efficiently as possible. On the other hand, when the demand rate exceeds the operating capacity of the system, users must to tolerate higher delay, in-service travel time, earliness, and lateness due to lack of adequate capacity in the system. To improve the quality of the service, an agency needs to increase its operating capacity, which may raise its operating cost as well. Therefore, it is of great importance for

the decision makers to have an accurate approximation of the potential operating cost and users' cost in order to optimize the balance between them.

4.1.1 Agency Operating Cost Model

The expenses that the operator incurs to run a DRT service can be categorized into three parts: costs attributed to fleet size, VHT, and VMT. Rahimi et al. (2014) adopts a continuum analytical model based on Daganzo (1978) to approximate these components for three prevalent loading-unloading operating strategies:

- *Strategy 1*: The operator aims to minimize the total distance traveled by finding the next closest pickup or drop-off point from each stop.
- *Strategy 2*: The operator alternates between pickup and drop-off phases in order to minimize the variance of the riding time. To achieve this, each vehicle starts by collecting n_v requests in the pickup phase and then delivers them to their destinations in the drop-off phase.
- *Strategy 3*: Each pickup is followed by the closest drop-off point and vice versa. In this operating strategy the vehicle first picks up n_v requests and then continues to alternate between the closest pickup and the closest drop-off.

On this basis, the total operating cost of the different strategies can be estimated as a summation of aforementioned components. In this part of the research we employ this analytical model to estimate the total agency operating cost as a function of operating

capacity of the system, μ, as well as number of pickup requests awaiting service, n_w. The maximum rate that requests can be served depends on the fleet size and other characteristics of the system, so we call this rate the capacity for the given operating conditions.

The required fleet acquisition and its associated expenses are the key capital component of agency cost for running the service. Given the fixed and variable costs per vehicle, the fleet cost can be estimated as a linear function of the fleet size. In this regard, the following analytical model is used to approximate the required fleet size for each of the operating strategies:

$$M_i(\mu, n_w) = \mu\left(b + \frac{1}{2v} k_i r\sqrt{A}\right) \tag{4.1}$$

where M_i is the fleet size (number of vehicles) that the agency needs to operate to provide the operating capacity μ (number of requests served per unit of time) in the service area of size A (units of area) in each operating strategy i. Here, b represents the boarding-alighting time (unit of time), which represents the total time that it takes the operator to load and unload a passenger at the start and in the end of a single requested passenger trip, respectively. The average moving speed in the network of vehicles traveling between the loading/unloading stops is represented by v (distance per unit of time), which depends on a variety of factors like characteristics of the network, properties of the vehicles, and even the type of service they are providing to the users. All other network and operation

characteristics that remain fixed in the short run, like the geometry of the network and routing of the vehicles, are reflected in a unitless network travel parameter, r. The value of this parameter can be estimated through calibration of the model using available data from the same (or similar) service operation. To keep the structure of the model general for all three operating strategies, we define the parameter k_i in terms of n_w and n_v for each of the different strategies as below:

$$k_i = \begin{cases} \frac{1}{\sqrt{2n_w}} & i = 1 \\ \frac{1}{\sqrt{n_w}} + \frac{\sqrt{2+4n_v}-1.45}{n_v} & i = 2 \\ \frac{1}{\sqrt{n_w}} + \frac{1}{\sqrt{n_v}} & i = 3 \end{cases} \tag{4.2}$$

Given the values of its parameters, this model demonstrates the relationship between the operating capacity of the system, the number of awaiting requests, and the fleet size as the main decision variables of the system. Having two variables out of three, this model can be employed to approximate the third. For simplicity, we consider here that the required fleet size for running the service can be approximated as a function of the operating capacity that it needs to provide and the number of awaiting requests. In Section 4.2.3, we use the same model to approximate the operating capacity of the system based on available fleet size and number of requests waiting for the service.

The total hours of driving time are associated with costs like driver wages and benefits. Thus, the analytical model can be used to approximate the VHT operated to serve a total demand N_Q within a time period using each of the different strategies as follows:

$$VHT_i(n_w) = M_i \frac{N_Q}{\mu} = N_Q \left(b + \frac{1}{2v} k_i r \sqrt{A}\right) \tag{4.3}$$

Note that VHT is independent of the operating capacity of the system and depends only on the number of waiting requests.

Expenses like fuel cost, maintenance, and vehicle depreciation costs directly depend on the total distance that vehicles travel in the network. The VMT of the service can be approximated for the different loading-unloading strategies as below:

$$VMT_i(n_w) = \frac{1}{2} N_Q k_i r \sqrt{A} \tag{4.4}$$

This expression shows that VMT also depends only on the number of waiting requests.

To sum up, the total operating cost of the agency in strategy i, AC_i, can be represented as a linear function of these components minus the total fare that all users pay for the service. The total fare for all customers in scenario i is represented by P_i. This expression can be also interpreted as the minimum budget that the agency needs to run such a service. It is worth pointing out that the parameters of the model, α_j and β_j, can be estimated

through calibration. In principle, these parameters are interpreted as the fixed and variable unit costs of providing each component, indexed by j.

$$AC_i(\mu, n_w) = [\alpha_1 + \beta_1 M_i(\mu, n_w)] + [\alpha_2 + \beta_2 VHT_i(n_w)] + [\alpha_3 + \beta_3 VMT_i(n_w)] - P_i \tag{4.5}$$

Note that the agency operating cost turns out to be an increasing function of the operating capacity (μ) and a decreasing function of the number of waiting requests (n_w) when the total demand (N_Q) is fixed. Consequently, cuts in operating capacity are one way to reduce costs, but changing the number of waiting requests also has an effect on the agency's costs.

4.1.2 User Cost Model

In an undersaturated system, the demand rate always remains below the capacity of the system, so customers can be served at their desired times without delay. However, this may not always be the case in reality due to the high cost of providing adequate capacity. In fact, the incentives for agencies to reduce operating costs as much as possible will tend to affect the quality of the service making oversaturated conditions an expected outcome. In this case, users will experience delay in pickup, longer in-service times, and earliness or

lateness in delivery as unpleasant consequences of excessive requests for the operating capacity of the system.

This problem is comparable to the morning commute problem in traffic flow theory in many ways. In both problems, the system has a limited capacity to serve users, where the delay, earliness, and lateness are undesirable consequences of a demand rate that is higher than the capacity of the system. Users may be expected to account for the delay and in-service time as well as the earliness or lateness that they experience relative to their preferred travel schedule when making travel decisions. Accordingly, rational users will attempt to reduce their costs by adjusting the times they request to travel. The cumulative result of individual decisions will eventually lead to an equilibrium condition in which no one has incentive to change his/her request time. Due to these similarities, Vickrey's (1969) congestion theory is adopted as a basis for modeling the equilibrium in the oversaturated condition.

Users of a DRT service have different desired service times, which can be represented as a distribution of demand over time. However, this demand rate may exceed the capacity of the system at some points making delays inevitable. Assuming that experienced users have perfect information regarding the generalized cost associated with the service at each point in time, they will adapt their request times by advancing or postponing their requests to keep their costs minimized. This generalized cost has different components that users experience when served by the DRT system. In the oversaturated condition, it is impossible to serve all users at their requested time due to inadequacy of the operating

capacity of the system, so users will experience some delay for pickups at their origins. After the passengers are picked up, in-service time is experienced until the DRT system delivers the passengers to their destinations.

While in-service time for the users depends on the operating strategy that the agency follows for processing the requests, it is worth noting that the variations of in-service time in each strategy are small enough to be neglected. We use the average time that users spend onboard the vehicles (i.e., in-service) to calculate the total time in-service. Therefore, the in-service time of the users can be treated as a fixed part of the cost that they pay for service and does not impact their scheduling decisions. Finally, requests may be delivered to their destinations earlier or later than their desired times, and have to accept the penalty for such deviation from the preferred schedule. Accordingly, the generalized cost for all users in each strategy, UC_i, can be formulated as below:

$$UC_i(\mu, n_w) = VOT\left(T_D(\mu) + T_{S,i}(\mu, n_w) + eT_E(\mu) + lT_L(\mu)\right) + P_i \tag{4.6}$$

where VOT denotes the value of time for users, $T_D(\mu)$ is the delay, $T_{S,i}$ is the in-service time for operating strategy i, T_E is the earliness, T_L is the lateness, and e and l represent the relative cost of earliness and lateness times in equivalent units of travel time. The interaction between users will eventually result in an equilibrium condition in which no one will be able to reduce his/her own costs by unilaterally changing his/her requested

time. Note that a flat fare will have no effect on the request times that users choose to minimize their costs.

Figure 4.1 illustrates the cumulative counts of passengers using the DRT system in the equilibrium condition. The S-shaped curve, $W(t)$, depicts the cumulative distribution of wished request times. The slope of this curve at each point in time represents the wished demand rate of the users. Here, we assume that the wished demand rate continuously increases such that it exceeds the operating capacity of the service at time t_1. It continues to increase to its maximum value, which exceeds μ, then starts decreasing such that it drops below the operating capacity after time t_2. The performance of the system, with operating capacity of μ, is also illustrated by the cumulative service curve, $S(t)$, which describes the times when passengers are picked up by DRT vehicles.

For operations, the agency is assumed to collect users' requests and add them to the pool of n_w requests waiting to be picked up. Nonetheless, the available capacity may be insufficient to serve the users at the rate their requests for pickup are made, and delays will ensue. On the flip side, users are assumed to be able to predict the delay and delivery times so that they can consider the inevitable earliness or lateness when choosing a request time. The request curve, $R(t)$, represents the cumulative distribution of the individual user requests when each minimizes their own travel cost. The user equilibrium condition exists when no one has an incentive to alter his/her own requested time. In this condition, the number of waiting requests at each point in time can be approximated as the vertical

distance between the request curve and the service curve. As indicated in Figure 4.1, the number of requests awaiting service does not remain constant in the equilibrium condition; in contrast, it varies with the rise in demand over time.

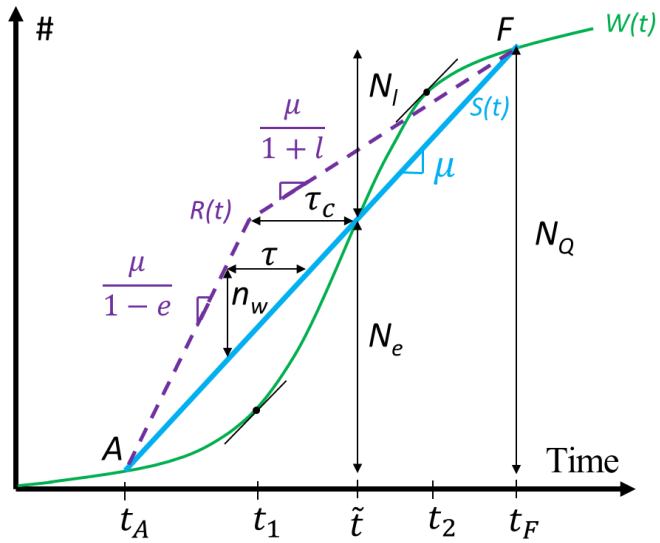


Figure 4.1. User equilibrium queueing diagram for the S-shaped wished curve

When the wish-curve has a smooth S-shape, Daganzo (1985) proves that a unique user equilibrium is associated with two conditions. First, the proportion of the users who experience earliness, N_e, to users who are late, N_l, equals the proportion of lateness penalty to that of earliness.

$$\frac{N_e}{N_l} = \frac{l}{e} \tag{4.7}$$

Second, the request curve is piecewise linear with specific slopes in user equilibrium condition. The first N_e users make their requests early, and next N_l requests are made late.

$$\frac{dR(t)}{dt} = \begin{cases} \frac{\mu}{1+e}, & \text{for users who are served early} \\ \frac{\mu}{1-l}, & \text{for users who are served late.} \end{cases} \tag{4.8}$$

In the equilibrium condition, components of the user cost can be approximated according to the queueing diagram. The delay that each user experiences for the service is the difference between his/her requested and actual pickup times, τ (i.e., the horizontal distance between the request and service curves in Figure 4.1. Delay in the service starts at time t_A and increases linearly to its maximum value, τ_c, at time $\tilde{t}$. Then it linearly decreases back to zero at time t_F. Accordingly, the total delay that all the users experience waiting for the service can be approximated as the area between $R(t)$ and $S(t)$ in Figure 4.1.

$$T_D(\mu) = \frac{N_Q^2 el}{2\mu(e+l)} \tag{4.9}$$

Earliness and lateness are defined as the time gap between actual and wished delivery of users to their destinations. So, if a request delivers a user to his/her destination before or after the wished time, an early or late schedule penalty will be included as part of the

user cost. The total earliness for all users is approximated by the area between $S(t)$ and $W(t)$ from t_A to $\tilde{t}$, and the total lateness is approximated by the area between $S(t)$ and $W(t)$ from $\tilde{t}$ to t_F as shown in Figure 4.1. In the simplified case that the wished curve has an inverse Z-shape with slope ω, as depicted in Figure 4.2, the total earliness and lateness that users experience can be approximated as below:

$$T_E(\mu) = \frac{1}{2} N_Q^2 \left(\frac{l}{e+l}\right)^2 \left(\frac{1}{\mu} - \frac{1}{\omega}\right) \tag{4.10}$$

$$T_L(\mu) = \frac{1}{2} N_Q^2 \left(\frac{e}{e+l}\right)^2 \left(\frac{1}{\mu} - \frac{1}{\omega}\right) \tag{4.11}$$

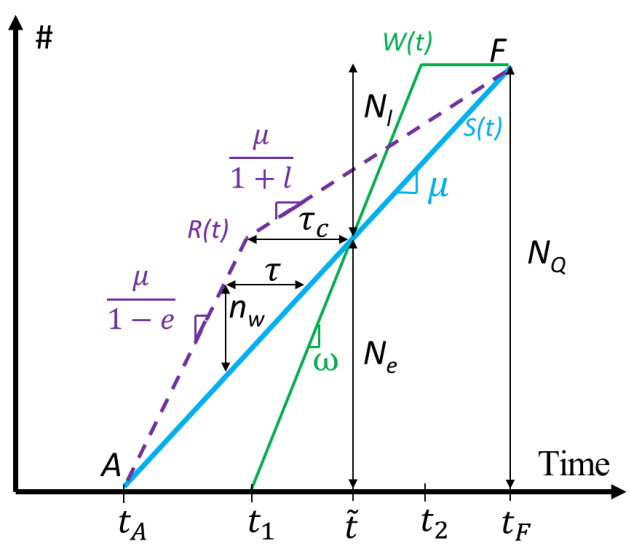


Figure 4.2. User equilibrium queueing diagram for the Z-shaped wished curve

Daganzo's (1978) analytical model can be generally employed to approximate the total time that users spend in service for each of the three operating strategies as below:

$$T_{S,i}(\mu, n_w) = \begin{cases} N_Q \frac{n_w M}{\mu} & i = 1 \\ N_Q \frac{n_v M}{2\mu} & i = 2 \\ N_Q \frac{M(n_v - 0.5)}{\mu} & i = 3 \end{cases} \tag{4.12}$$

Note that the components of the total cost for users are decreasing functions of the operating capacity of the DRT service, μ. In other words, any increase in the operating capacity improves the quality of the service for the users by reducing their costs. The distribution of the requests over time is another key factor that affects the total cost for the users, because increases in N_Q or ω lead to increased user cost. The next section addresses the problem of minimizing the total cost for the agency and users in order to obtain an optimal balance between the expense of operation and quality of the DRT service.

4.2 System Optimum

The inherent tradeoff between the cost and quality of the DRT service necessitates holding a balance between the agency and user costs. The optimal balance between the operation cost and quality of the service can be determined by minimizing the total cost for the agency and users. In this respect, the operating capacity of the service can be considered

as the decision variable of the system because it implies both the operating cost and the user cost. Moreover, the efficiency of the system can be further improved by having a uniform distribution of requests with an optimal number of requests awaiting service. To this end, we first optimize the operating capacity of the system to minimize the total costs for the users and the agency in the equilibrium condition. Next, we show how a uniform distribution of demand with the optimal number of waiting requests can further reduce the total cost. Lastly, we consider a special case in which the agency has no plan to change its fleet size (a realistic constraint) and can improve the efficiency of the service only by incentivizing an optimal demand pattern. Then, in Section 4.3, we present demand management strategies that can be implemented to achieve the optimal distribution of demand over time.

4.2.1 Scenario I: Optimal Operating Capacity

In the equilibrium condition illustrated in Figure 4.1, increasing the operating capacity of the service can reduce the total delay, earliness, and lateness of the users. Equation (4.6) shows that the user cost is a decreasing function of μ, and equation (4.5) shows that the agency cost is an increasing function of μ. As a result, the efficiency of the system can be enhanced by optimizing the capacity of the service in order to keep the sum of the agency and user cost minimized:

$$\mu_i^* = \arg\min_{\mu} TC_i(\mu, \bar{n}_w) = \arg\min_{\mu}\big(AC_i(\mu, \bar{n}_w) + UC_i(\mu, \bar{n}_w)\big) \tag{4.13}$$

where TC_i denotes the total cost of the system in different strategies, which can be approximated by substituting the values of agency cost from (4.5) and user cost from (4.6). The fare that users pay and that the agency collects is cancelled out of (4.13) because this is a transfer from users to the agency and does not affect the total combined cost to both. Here, $\bar{n}_w$ is the average number of requests awaiting service over time. In the equilibrium condition, the number of waiting requests varies over time. Since both the agency and user costs are functions of n_w, the exact number of waiting requests is replaced with its average value in equation (4.13) to simplify the problem. It will later be shown that for the optimal control n_w should be held constant. The average number of requests awaiting service can be approximated using Little's (1961) formula to multiply the average waiting time by the rate that requests are picked up:

$$\bar{n}_w = \frac{\tau_c}{2}\mu = \frac{N_Q el}{2(e+l)} \tag{4.14}$$

This problem has a closed-form solution when the wished curve has a specific shape like an inverse Z-shape as illustrated in Figure 4.2. In this case, the components of the agency cost, (4.1), (4.3), and (4.4), and components of the users cost, (4.9), (4.10), and (4.11), can be substituted into the objective function of problem (4.13). As a result, the optimization problem can be expressed as below:

$$\mu_i^* = \arg\min_{\mu} \left(C_{i,1}\mu + \frac{C_{i,2}}{\mu} + C_{i,3} \right) \tag{4.15}$$

where $C_{i,1}$ and $C_{i,2}$ represent the combined coefficients of μ and $1/\mu$, respectively, in the objective function for strategy i after combining terms. $C_{i,3}$ denotes the terms in the objective function that are independent of μ. The objective function is a convex function of the operating capacity of the service, μ. According to the first order condition, the optimal operating capacity in each scenario, μ_i^*, can be approximated as follows:

$$\mu_i^* = \sqrt{\frac{C_{i,2}}{C_{i,1}}} \tag{4.16}$$

By substituting the optimal capacity back into the objective function of the problem, the minimum total cost of the system in each strategy, TC_i^{min}, can be approximated as below.

$$TC_i^{min} = TC_i(\mu_i^*, \bar{n}_w) = 2\sqrt{C_{i,1}C_{i,2}} + C_{i,3} \tag{4.17}$$

From a practical point of view, the optimal operating capacity can be used to approximate the optimal fleet size for running the system efficiently in each strategy.

$$M_i^* = M_i(\mu_i^*, \bar{n}_w) \tag{4.18}$$

Optimizing the operating capacity of the system can improve the efficiency of the system by minimizing its total cost. However, the efficiency of the system can be further improved by optimizing the distribution of the requests as well as number of requests awaiting service.

4.2.2 Scenario II: Optimal Operating Capacity and Number of Waiting Requests

In the oversaturated condition, the operating capacity of the service is insufficient to cover the demand in the peak period, so users may adjust their request times to reduce the cost of their trip. The cumulative result of individual decisions leads to an equilibrium condition in which no one has an incentive to change his/her request time. According to condition (4.8), in the equilibrium condition the rate of requested trips will have a stepwise distribution in the peak period. As illustrated in Figures 1 and 2, the slope of the first part of the equilibrium request curve exceeds the capacity of the system ($\mu/(1-e) > \mu$), while in the second part of this slope drops below the operating capacity of the system ($\mu/(1+l) < \mu$). Although the average demand rate through the congested period is equal to the average service rate, the uneven distribution of request times causes avoidable delay. The same number of travelers could be served without excess delay if $R(t)$ were linear with constant slope μ. We now consider the total cost for the agency and users by assuming that it is possible to make the distribution of demand uniform over time. In the next section,

demand management strategies that achieve this constant distribution of requested times through the rush will be presented.

Consider the same case of a general S-shaped cumulative curve of wished pick-up times, $W(t)$, with maximum demand rate exceeding μ, as shown in Figure 4.3. If the distribution of actual requests is forced to follow a constant rate λ, the cumulative request curve, $R(t)$, is linear. If $R(t)$ has slope $\lambda = \mu$, as shown by the dashed line in Figure 4.3, then the system will operate at capacity throughout the rush while maintaining a steady number of waiting requests. The resulting performance of the DRT system is depicted by $S(t)$ (the solid line in Figure 4.3), which allows the number of requests to accumulate up to n_w, and then starts to serve trips at the capacity μ.

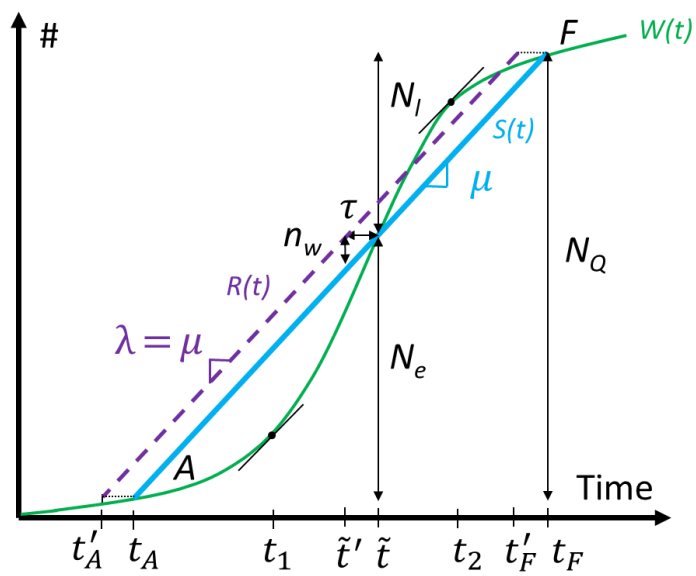


Figure 4.3. Queueing diagram for a uniform distribution of demand

In this system, all of the users experience the same delay, τ, in the peak period. So, the total delay of the users in this condition is the area between the request and service curves, and can be approximated as follows:

$$T_D(\mu, n_w) = N_Q \frac{n_w}{\mu} \tag{4.19}$$

The total delay is a decreasing function of μ and an increasing function of n_w. The total earliness of the system is still represented by the area between the service curve and the wished curve from t_A to $\tilde{t}$, and the total lateness by the area between the curves from $\tilde{t}$ to t_F in Figure 4.3. Note as well that total earliness and lateness are decreasing functions of the operating capacity of the system, just as they are for the user equilibrium. For a system with fixed capacity, the total earliness and lateness in system optimum is the same as in the user equilibrium (Daganzo, 1985). The earliness and lateness for the DRT system are independent of the number of the waiting requests, and the total in-service time is still approximated by (4.12).

A uniform distribution of demand with a request rate equal to the capacity of the system makes it possible to limit the number of users awaiting service. The goal is not to drive n_w to 0, however, because low values of n_w are associated with high agency costs, based on (4.5), and high in-service times for Strategy 1, based on (4.12). This is phenomenon is due

to the fact that having a high number of waiting requests, n_w, allows flexibility for each vehicle to travel a shorter distance and spend less time to make its next pick-up. Passengers on-board experience less in-service time because there is less distance traveled out of the way to pick up other passengers in the DRT vehicle. In selecting the optimal value of n_w, an inherent tradeoff exists with total delay balanced against the operating cost for the agency and the in-service time for the users.

A tradeoff also exists between the costs for the agency and users. This should be taken into account to determine the optimal operating capacity of the service. The fleet expenses and operating cost of the agency is an increasing function of operating capacity of the service, while the total delay, earliness, and lateness of the system decrease with a rise in the operating capacity of the service. Accordingly, the total costs for the agency and the users in each strategy can be minimized by optimizing the operating capacity of the service and number of waiting requests as below:

$$\left(\mu_i^*, n_w^{i*}\right) = \arg\min_{\mu, n_w} TC_i(\mu, n_w) = \arg\min_{\mu, n_w}\left(AC_i(\mu, n_w) + UC_i(\mu, n_w)\right) \tag{4.20}$$

where the costs for the agency and users can be approximated according to the equations (4.5) and (4.6), respectively. As before, the components of the agency cost can be approximated using equations (4.1), (4.3), and (4.4). In this scenario, the total delay for the users can be approximated by substituting (4.19) into (4.6). When the wished curve has a

specific inverse Z-shape with slope ω, the total cost of the system in the problem (4.20) can be formulated as a function of μ and n_w. The total earliness and lateness of the system also can be approximated according to equations (4.10) and (4.11), respectively.

To solve this problem, we start by treating n_w as a parameter in the objective function to derive the optimal μ as a function of n_w.

$$\mu_i^* = \arg\min_{\mu} \; TC_i(\mu, n_w) = \left(D_{i,1}(n_w)\mu + \frac{D_{i,2}(n_w)}{\mu} + D_{i,3}(n_w)\right) \tag{4.21}$$

where $D_{i,1}(n_w)$ and $D_{i,2}(n_w)$ denote the coefficients of μ and $1/\mu$ in the total cost of the system as functions of n_w using strategy i. $D_{i,3}(n_w)$ represents the part of this cost that is a function of n_w but independent of μ.

Since the total cost of the system is a convex function of the operating capacity of the service, the optimal value of μ can be approximated according to the first order condition as below:

$$\mu_i^* = \sqrt{\frac{D_{i,2}(n_w)}{D_{i,1}(n_w)}} \tag{4.22}$$

By substituting the optimal operating capacity of the service from (4.22) into (4.20), the total cost of the system can be formulated solely as a function n_w. Accordingly, the optimization problem can be reformulated as follows:

$$n_w^{i^*} = \arg\min_{n_w} TC_i(\mu_i^*, n_w) =$$

$$\arg\min_{n_w} 2\sqrt{D_{i,4}n_w + D_{i,5}\sqrt{n_w} + \frac{D_{i,6}}{\sqrt{n_w}} + D_{i,7}} + D_{i,8}n_w + D_{i,9}\sqrt{n_w} + \frac{D_{i,10}}{\sqrt{n_w}} + D_{i,11} \quad (4.23)$$

where terms $D_{i,4}$ through $D_{i,11}$ denote the parameters of the objective function of the problem for each strategy i.

The total cost of the system is a convex function of the number of waiting requests. As a result, the objective function can be minimized by finding the optimal number of requests awaiting service. Since the objective function of this problem is both continuous and differentiable over n_w, the global optimum solution of this problem can be calculated quickly and precisely with the help of simple numerical methods like Newton's method. Since the number of the requests awaiting service remains fixed over time, the average time that users spend in the vehicles, as calculated by (4.12), remains constant over time as well. It is worth pointing out that this conclusion is consistent with our initial assumption that the in-service time is fixed over time.

The minimum total cost of the system in each strategy can be estimated by plugging in μ_i^* and $n_w^{i^*}$ back into (4.20):

$$TC_i^{min} = TC_i\left(\mu_i^*, n_w^{i^*}\right) \quad (4.24)$$

The optimal fleet size of the operator also be approximated by plugging μ_i^* and $n_w^{i^*}$ into (4.1):

$$M_i^* = M_i\left(\mu_i^*, n_w^{i^*}\right) \tag{4.25}$$

4.2.3 Scenario III: Optimal Number of Waiting Requests with a Fixed Fleet Size

Many agencies face constrained budgets and facilities that can make it difficult to change the fleet size for the DRT service. Thus, it is useful to consider the efficiency improvement that can be achieved if the distribution of demand and operation of the service are optimized while holding the fleet size fixed. In this case, the fleet size (M) and its associated cost remains fixed, and the operating capacity of the system in each strategy, μ_i, becomes a function of the number of waiting requests. This function can be approximated by solving equation (4.1) for μ.

$$\mu_i(n_w) = M\left(b + \frac{1}{2v}k_i r\sqrt{A}\right)^{-1} \tag{4.26}$$

The total delay that users experience in this scenario can be approximated by substituting (4.26) into (4.19):

$$T_D(n_w) = N_Q\frac{n_w}{M}\left(b + \frac{1}{2v}k_i r\sqrt{A}\right) \tag{4.27}$$

Like the first and second scenarios, the total earliness and lateness can be approximated as the area between $R(t)$ and $S(t)$. As before, the total in-service time for the users can also be approximated using (4.12), and the total cost for the users can be approximated by substituting these components into (4.6). The operating cost for the agency expressed by equation (4.5), with the components of this cost are approximated by equations (4.1), (4.3), and (4.4).

As a result, the number of requests awaiting service can be optimized to improve the efficiency of the service by minimizing the total costs of the agency and the users as below:

$$n_w^{i^*} = \arg\min_{n_w} TC_i(n_w) = \arg\min_{n_w}\big(AC_i(n_w) + UC_i(n_w)\big) \tag{4.28}$$

Under the simplifying assumption that the wished curve has an inverse Z-shape with slope ω, the total earliness and lateness of the system can be approximated by substituting (4.26) into (4.10) and (4.11), respectively.

$$T_E(\mu) = \frac{1}{2}N_Q^2\left(\frac{l}{e+l}\right)^2\left(\frac{1}{M}\left(b + \frac{1}{2v}k_i r\sqrt{A}\right) - \frac{1}{\omega}\right) \tag{4.29}$$

$$T_L(\mu) = \frac{1}{2}N_Q^2\left(\frac{e}{e+l}\right)^2\left(\frac{1}{M}\left(b + \frac{1}{2v}k_i r\sqrt{A}\right) - \frac{1}{\omega}\right) \tag{4.30}$$

In this case, the problem expressed in (4.28) can be simplified by substituting the components of the agency and users' costs into its objective function as follows:

$$n_w^{i^*} = \arg\min_{n_w} TC_i(n_w) = \left(E_{1,i} n_w + E_{2,i}\sqrt{n_w} + \frac{E_{3,i}}{\sqrt{n_w}} + E_{4,i}\right) \tag{4.31}$$

where $E_{1,i}$, $E_{2,i}$, and $E_{3,i}$ represent the coefficients of n_w, $\sqrt{n_w}$, and $1/\sqrt{n_w}$ in the objective function, respectively. $E_{4,i}$ denotes the part of the objective function that is independent of n_w. The total cost of the system is a convex function of the number of waiting requests, so the first order condition can be used to identify the value of $n_w^{i^*}$ that minimizes the total cost in (4.31).

$$\frac{dTC_i(n_w)}{dn_w} = 2E_{1,i} + \frac{E_{2,i}}{2\sqrt{n_w^{i^*}}} - \frac{E_{3,i}}{2\sqrt{n_w^{i^{*3}}}} = 0 \tag{4.32}$$

By multiplying this equation by $\sqrt{n_w^{i^{*3}}}$, it can be expressed as a cubic function of $\sqrt{n_w}$ that can be solved analytically using Carano's Method. The function can also be solved quickly and precisely with numerical methods.

Like the second scenario, the number of waiting requests and the average in-service time for the users remains constant over the peak period, which is consistent with our initial assumption regarding the average in-service times. As a result, the optimal operating

capacity of the service can be approximated by substituting $n_w^{i^*}$ into (4.26). The minimum total cost of the system can be estimated by substituting the optimum number of waiting requests into the objective function in (4.28).

$$TC_i^{min} = TC_i\left(n_w^{i^*}\right) \tag{4.33}$$

Up to this point, it has been assumed that it is possible to have a uniform distribution of demand instead of an equilibrium stepwise distribution. However, this cannot happen without implementing an effective demand management strategy. In the following section, two different strategies are presented that can be implemented to make the demand uniform over time.

4.3 Demand Management Strategies

As explained in the previous section, the underlying assumption in optimizing the number of requests awaiting service in the peak period is that demand can be distributed uniformly across the period. However, the distribution of the demand in the equilibrium condition will naturally tend to be stepwise in the peak period, which is the cumulative result of the rational behavior of the individual users minimizing their own costs. Accordingly, demand management strategies that can make the demand uniform over time can fulfill the key role in enhancing the efficiency of the DRT system.

4.3.1 Schedule Management Strategy

A simple strategy to make the schedule of requests uniform over time is to simply stop accepting requests once the capacity has been reached. This strategy has been widely used in DRT systems such as ADA paratransit services to form a uniform demand. In this strategy, the operator accepts requests in a first-in, first-served order, and puts them in a reservation list. To keep the distribution of requests uniform over time, the number of requests per time added to the reservation list should not exceed the operating capacity of the service. If a user's requested time is not available they will have to book travel at an earlier or later time. The problem is that this strategy incurs schedule penalties on users who are not able to book their trips well in advance. Furthermore, regulations require that rescheduling not exceed one hour difference from the initial request, so the operator can still be stuck having to invest in additional resources to serve the peak demand. Alternatively, it is possible to implement a dynamic pricing strategy to make the distribution of demand uniform over the peak period, while users maintain freedom in choosing their own requested times.

4.3.2 Dynamic Pricing Strategy

Dynamic pricing is another strategy that can be implemented to achieve a uniform distribution of requests by incentivizing the users to adjust their request times. In general,

pricing is known to be an effective strategy for managing the demand for optimized systems although technical issues have often limited its implementation. However, this may not be the case for DRT systems where the technical infrastructure already exists, but agencies are choosing to charge flat fares. Alternatively, a dynamic fare pattern can be charged to make the distribution of the requests uniform over time.

In the equilibrium condition, demand has a stepwise distribution over the peak period. As illustrated in the Figures 2 and 3, the equilibrium demand rate of the first part of the peak period is higher than the operating capacity of the service. On the contrary, this rate drops below the capacity of the system in the second part, while the average weighted demand rate of the whole peak period equals the operating capacity of the system. Accordingly, a dynamic pricing strategy that decreases the demand rate of the first part of the peak period and increases that of the second part to fill the gap between these demand rates and make the distribution of the requests uniform over the peak period. Dynamic pricing strategies have been widely studied as an effective tool for optimizing traffic flow at bottlenecks in transportation networks. Here, we adapt the theory to draw similar conclusions for pricing DRT systems during the peak period.

4.3.2.1 Optimum Pricing Strategy

In Vickrey's (1969) congestion theory, the delay that users experience during their travels is an unnecessary consequence of the congestion at a bottleneck that can be avoided. A

dynamic price that charges users the equivalent monetary cost of the delay that users tolerate in the equilibrium condition (see Figure 4.1) can eliminate the delay by making the distribution of the demand uniform over the peak period. The optimum price for using the bottleneck should first arise from the minimum off-peak price, p_o, at time t_A with rate e up to the maximum value, T_c, at time $\tilde{t}$. Then the price should fall with slope $-l$ to the same minimum value in the end of peak period, at time t_B. Gonzales and Daganzo (2012) show that such a pricing strategy is optimal even when the capacity of the bottleneck is not fixed.

In transportation networks, tolls can be charged at bottleneck locations, when users are served, so the theory of the dynamic pricing of bottlenecks has been developed under the assumption that users are charged at bottlenecks. In contrast, users in DRT systems are typically charged a fare when they are picked up. This difference makes it necessary to adapt the theory of dynamic pricing for the DRT system. As a result, it can be shown that a dynamic pricing pattern, as described above can make the distribution of demand of the DRT system uniform over the peak period.

The dynamic prices for the DRT service must be set so that no user has any incentive to change their requested time to be earlier or later. It is useful to consider a single request and the effect of changing the request time on the delay and earliness. An example of a user who makes a request at time t, which is earlier than their wished time t_w, is illustrated in Figure 4.4. This user pays a price $p(t)$ based on when they request service, then they spend some time waiting to be picked up and transported to their destination. The effect of

a small shift in requested time by Δt earlier will result in being served $\dot{R}_e(t)\Delta t/\mu$ earlier in time, where $\dot{R}_e(t)$ denotes the slope of the request curve for all $t \in (t_A, \tilde{t})$. The resulting changes in the individual's experienced delay (τ_D) and earliness (τ_E) can be calculated as follows:

$$\Delta\tau_D = \tau_{D,2} - \tau_{D,1} = +\Delta t - \frac{\dot{R}_e(t)\Delta t}{\mu} \tag{4.34}$$

$$\Delta\tau_E = \tau_{E,2} - \tau_{E,1} = +e\frac{\dot{R}_e(t)\Delta t}{\mu} \tag{4.35}$$

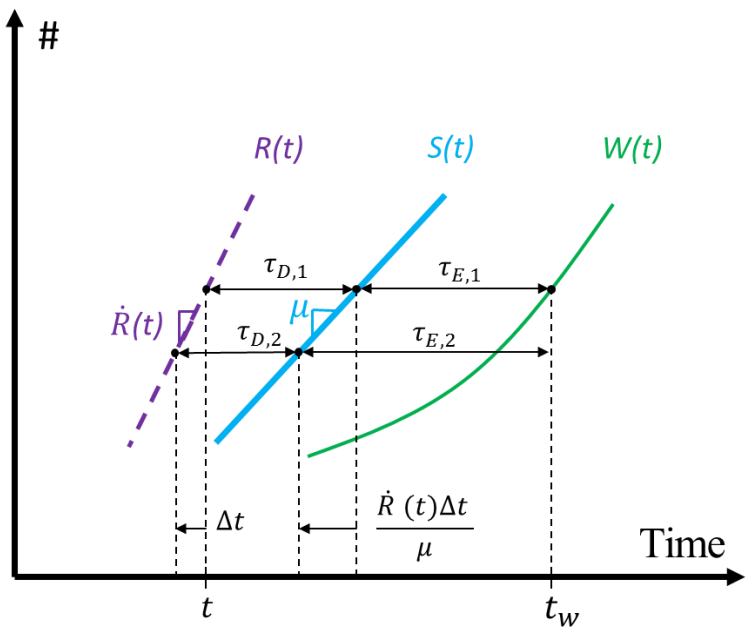


Figure 4.4. Variation of delay and earliness for a user resulting from a shift of Δt in the request time

There will be also a change in the price of the service (p) associated with this shift in request time. Assuming that price of the service varies smoothly over time, this change can be expressed as:

$$\Delta p = p_2 - p_1 = -\dot{p}_e(t)\Delta t \tag{4.36}$$

where $\dot{p}_e(t)$ is the rate of price increase per unit time at time t for the N_e requests that are earlier than wished.

In the equilibrium condition, a marginal shift in request time by should not affect the generalized cost for the user, so the following condition must hold:

$$\Delta UC = \Delta\tau_D + \Delta\tau_E + \Delta p = 0 \tag{4.37}$$

otherwise the user can achieve a lower cost by changing his/her request. By substituting (4.34), (4.35), and (4.36) into (4.37), $\dot{R}_e(t)$ can be solved in terms of $\dot{p}_e(t)$:

$$\dot{R}_e(t) = \frac{(1-\dot{p}_e(t))\mu}{1-e} \tag{4.38}$$

A similar line of reasoning can be used to show that that the slope of the second part of the request curve, $\dot{R}_l(t)$ is:

$$\dot{R}_l(t) = \frac{(1-\dot{p}_l(t))\mu}{1+l} \tag{4.39}$$

where $\dot{p}_l(t)$ denotes the rate of price increase per unit time at time t for the N_l requests that are later than wished. Note that (4.38) and (4.39) are not the same as the equilibrium arrival

curve in bottleneck model, because prices are charged at the requested time instead of the service time.

Following from (4.38) and (4.39), the dynamic price that increases from the base off-peak price, p_o, with rate e from time t'_A to $\tilde{t}'$, and then decreases with slope $-l$ back to p_o at time t'_F can make the distribution of the requests uniform and equal to the operating capacity of the service.

$$\dot{p}_e(t) = e \qquad \text{for} \qquad t \in (t'_A, \tilde{t}') \tag{4.40}$$

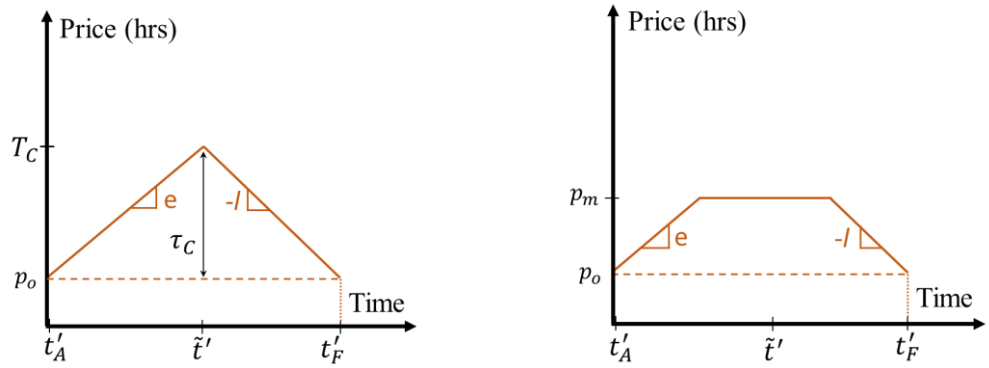
$$\dot{p}_l(t) = -l \qquad \text{for} \qquad t \in (\tilde{t}', t'_F) \tag{4.41}$$

These prices are illustrated in Figure 4.5(a). Although $\dot{R}_e(t)$ and $\dot{R}_l(t)$ differ from the conventional bottleneck model, the effective dynamic prices follow the same pattern just shifted from the service times to the request times. In general, the total price that users pay to the agency to use the service in each scenario can also be approximated as a function of the area under its pricing graph (Φ_i):

$$P_i = VOT \cdot \Phi_i \, \mu_i^* \tag{4.42}$$

Implementing an effective dynamic pricing strategy motivates users to distribute their requests uniformly over the peak period, which helps the agency reduce the cost of operating the system and reduces the waiting time for users of the system. To keep the

optimum number of requests awaiting service at $n_w^{i^*}$, the agency needs to operate by waiting for the optimum number of requests to accumulate and then serve the requests at its optimum capacity.



(a) Optimum Pricing Strategy (b) Constrained Pricing Strategy

Figure 4.5. Comparison of dynamic pricing strategies over the peak period

4.3.2.2 Constrained Pricing Strategy

In practice, an operator may have not have complete flexibility to charge any prices they want. For example, regulations may restrict the price by an upper bound. ADA paratransit services are intended to provide comparable transit service for people with disabilities, so charging fares that exceed fares on conventional public transit would undermine the role of the service. There is usually no lower limit on the price that can be charged, so prices may still be varied within the bounds to improve the efficiency of the system. An external constraint can be included in the optimization problem to keep the optimal price within the

allowable range, and the result would be the most effective dynamic pricing strategy that is feasible.

As depicted in Figure 4.5(b), the constrained pricing strategy can be derived by restricting the most effective pricing pattern to the upper limit. Accordingly, the dynamic price rises from time t = 0 to t'_A with constant rate e up to the maximum allowable limit, p_m. Then, it stays fixed until it can decrease with rate $-l$, such that the price returns to zero at time t'_F. Adding a new constraint to an optimization problem can never lead to an improvement in the objective function. In view of that, although implementing such a pricing strategy can help enhance the efficiency of the system, it will not achieve a uniform distribution of the requests. As a result having a non-uniform distribution of the requests causes at least some users to experience more delay in comparison to a system with a uniform distribution of the demand.

4.4 Numerical Example

To provide a numerical illustration of this problem, in this section we employ the proposed analytical model to optimize different operating strategies for a DRT service in different optimization scenarios. This DRT system is assumed to provide a curb-to-curb service for its users in an area of $A = 500$ mi^2 with a network travel parameter of $r = 1$, and a fixed peak demand of $N_Q = 150$ requests. It is assumed that the fixed demand is uniformly distributed over 3 hours of the peak period with a fixed demand rate of $\omega = 50$ requests

per hour. The value of time of the users is assumed to be \$20 per hour with an earliness penalty factor of $e = 0.5$, and a lateness penalty factor of $l = 1.5$. Here, the total boarding and alighting time is $b = 15$ minutes per request, and the average moving speed of the vehicles in the network is assumed be to $v = 40$ mph, while each vehicle has $n_v = 4$ requests onboard on average in each point in time.

In this case, the agency and users' cost of the DRT system for different strategies can be approximated following from the equations (4.5) and (4.6), as below:

$$AC_i(\mu, n_w) = 1000 + 500M_i(\mu, n_w) + 1.5VHT_i(n_w) + 0.5VMT_i(n_w) - P_i \quad (4.43)$$

$$UC_i(\mu, n_w) = 20\left(T_D(\mu) + T_{S,i}(\mu, n_w) + 0.5T_E(\mu) + 1.5T_L(\mu)\right) + P_i \quad (4.44)$$

On this basis, we first optimize this DRT system under different scenarios, then we propose an optimal pricing strategy to keep the system optimized over peak period, according to Section 5.2.1. Next, we perform a sensitivity analysis on the results of different operating strategies under different scenarios to show how these results may change with the variations in the level of demand, value of time for the users, and the fleet size (in Scenario III).

4.4.1 System Optimum and Dynamic Pricing Strategy

As explained in the Section 4.3.2, the total cost for the agency and users of a DRT service in different operating strategies can be minimized by optimizing the capacity of system as well as number of awaiting requests. In Scenario I, we optimize the operating capacity of the system by solving problem (4.15) for this numerical example. The optimal values of the operating capacity and waiting requests in Scenario II can also be approximated by solving problems (4.21) and (4.23). Finally, in Scenario III, we assume that the agency only has $M = 5$ vehicles available for running the DRT service, so we optimize the number of waiting requests by solving problem (4.31). Given the optimal vales of these decision variables, it becomes possible to approximate the optimal fleet size, the agency and users' costs, and the minimum total costs in each of the different strategies. Table 4.1 summarizes the results for the different DRT operating strategies under the three different optimization scenarios.

In Scenario I, with optimal operating capacity but variable number of awaiting requests, the second operating strategy with the largest fleet size, highest agency cost, and the lowest users' cost (excluding the price) has the lowest total cost among the operating strategies under the first optimization scenario. Since Scenario I optimizes the operating capacity while allowing users to make requests in an unrestricted user equilibrium, the number of awaiting request varies over the peak period. As explained in Section 4.2.1, we approximate

the number of waiting requests in the first scenario with its average value using equation (4.14).

To further simplify the problem, we can set a flat price to zero, which indeed has no effect on the optimized solution or the total cost of the DRT system. In fact, the price, whether it is flat or dynamic, is a portion of the cost that will be canceled out in summing up the agency and users' costs in the total cost function. A pricing strategy will be effective in changing the system performance only if it can make the distribution of the demand more uniform over the peak period by incentivizing the users to adjust their request times. Otherwise, a flat price is an endogenous component of the costs that is transferred from the users to the agency.

It is also worth pointing out that the users' cost and the total cost of Strategy 1 in Scenario I is significantly higher than those of all other strategies in these three scenarios. The reason for this surprisingly high users' cost in this case is the inefficiency of the operations. In Strategy 1, vehicles always choose the nearest point as their next stop, regardless whether it is an origin or destination. Without an effective demand management strategy, this causes a rise in number of requests waiting for the service during the peak period. Such an accumulation of requests simply increases the chance of pickups relative to drop-offs, and consequently the in-service time is elongated. It can be generally concluded that Strategy 1 is not very efficient in Scenario I, although its agency cost might be relatively low.

In Scenario II, both the operating capacity and the number of waiting requests are optimized, and the total cost of each operating strategy is reduced to its lowest possible value. Interestingly, Strategy 1, which has the highest total cost in Scenario I, has the lowest total costs among all of the scenarios and strategies in this scenario. This strategy provides the most optimal service quality with the lowest delay, in-service time, and the cost for the users (excluding the price) using the smallest fleet size and the lowest number of waiting requests among all of the other cases, which explicitly demonstrates the concept system optimization.

Table 4.1. System optimization results of the numerical example

		Optimal Design Variables			Optimal Costs ($)											
					Agency Cost				Users' Cost				(Excluding Price)			
SO	i	μ_i^* *(Pax/hr)*	n_w^{i*} *(Pax)*	M_i^* *(Veh)*	α	*Fleet*	*VHT*	*VMT*	*Delay*	*In-Service*	*Earliness*	*Lateness*	AC_i^*	UC_i^*	TC_i^{min} *($)*	P_i ($)
Scenario I	1	34.28	28.13†	9.85	10,000	4,923	646	112	2,462	24,238	581	194	15,681	27,474	43,155	0††
	2	26.04	28.13†	12.96	10,000	6,481	1,120	744	3,241	2,987	1,165	388	18,345	7,781	26,126	0††
	3	27.62	28.13†	12.22	10,000	6,110	996	577	3,055	4,646	1,026	342	17,683	9,068	26,751	0††
Scenario II	1	20.19	1.19	8.71	10,000	4,355	970	544	177	1,538	1,868	623	15,869	4,205	20,074	4,178
	2	19.64	7.19	10.79	10,000	5,395	1,236	898	1,098	3,296	1,957	652	17,529	7,003	24,532	4,296
	3	21.42	8.87	10.36	10,000	5,180	1,088	701	1,242	5,078	1,688	563	16,969	8,571	25,540	3,938
Scenario III	1	12.15	1.50	5.00†††	10,000	2,500	926	484	370	1,851	3,941	1,314	13,910	7,476	21,386	6,942
	2	8.86	5.52	5.00†††	10,000	2,500	1,269	942	1,869	3,384	5,873	1,958	14,711	13,085	27,796	9,519
	3	10.04	6.64	5.00†††	10,000	2,500	1,121	745	1,985	5,231	5,040	1,680	14,366	13,936	28,302	8,407

† Average value approximated using relation (14)
†† No dynamic pricing
††† Fixed fleet size

In Scenario III, the objective is to reduce the total cost of the DRT system by optimizing the number of waiting requests for a given fleet size. In this example, we assume that the agency possesses 5 vehicles to serve the demand. In this case, Strategy 1 is the most efficient, with the lowest agency and users' costs (excluding the price). Notice as well that all of the components of the agency and users' cost for Strategy 1 are at their minimum possible values, which emphasizes the dominance of Strategy 1 for this scenario.

As explained in Section 4.3, implementation of Scenarios II and III inevitably requires an effective demand management strategy that can make the distribution of requests uniform over the peak period. To this end we propose a dynamic pricing strategy that can keep the number of waiting requests in Scenarios II and III optimized over time. Such optimal pricing strategies can be designed by charging each of the users a price equal to the cost of the delay they experience in the user equilibrium condition, according to criteria (4.40) and (4.41).

With the absence of a dynamic pricing strategy in Scenario I, there will be a time-variant delay associated with service as illustrated for different operating strategies in Figure 4.6(a). As indicated in this figure, the length of the peak period varies among the different strategies based on their optimal operating capacities. In view of that, Strategy 1 has the highest operating capacity and the shortest peak period, while the Strategy 2 has the lowest operating capacity and the longest peak period in Scenario 1. However, in all strategies, this delay first rises with slope $e = 0.5$ at start of the peak period at time t'_{A_i}, up

to its maximum value, τ_c, at time $\tilde{t}'$. Then it falls with a negative slope of $l = -1.5$ until the end of the peak period at time t'_{F_i}.

(a) Scenario I (b) Scenario II (c) Scenario III

Figure 4.6. Time-variant delay (a) and optimal pricing (b and c) for different operating strategies

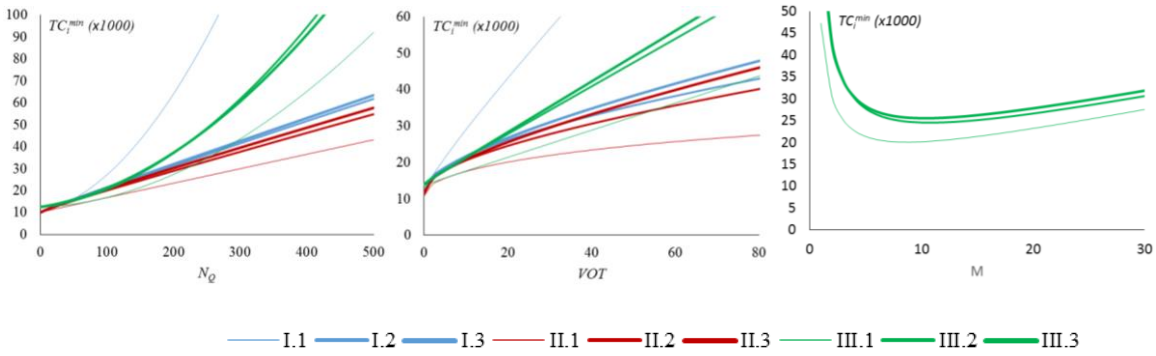
Like Scenario I, the lengths of the peak periods in Scenarios II and III directly depend on the operating capacity of the strategies. On this basis, Strategies 1 and 3 of Scenario II turn out to have the shortest and longest peak periods in Figure 4.6(b), although the lengths of these peak periods are relatively close across the different strategies. In Scenario III, as illustrated in Figure 4.6(c), Strategy 1 still has the longest peak period due to its lowest operation capacity, while Strategy 2 has the shortest peak period. In all of these pricing strategies, the optimal price in the first part of the peak period linearly increases from zero,

at time t'_{A_i}, with slope $e = 0.5$ up its maximum value, τ_c, at time $\tilde{t}'$. Then it linearly decreases back to zero with a negative slope of $l = -1.5$, at time t'_{F_i}.

4.4.2 Sensitivity Analysis

In the previous section, we minimized the total cost of the DRT system for the numerical example by optimizing the decision variables of different operating strategies under different optimization scenarios. However, it is useful to see how the minimum total cost of the service changes with variations in the demand and the value of time of its users. We summarize here the results of the sensitivity analysis on the numerical example optimized in the previous section.

Thanks to the analytical solution of the problem, variations in the total cost of the system can be accurately approximated for wide ranges of input parameters, as plotted in Figure 4.7. Figure 4.7(a) depicts the total cost with variations in the total number of the waiting requests during the peak period. It can be inferred from this Figure 4.that the total costs of the different strategies in the different scenarios are relatively close when demand is low. However, as the demand starts growing, the total costs of the some of the cases, in particular I.1, rises relatively faster than the others. This rise is relatively slow for some other cases, in particular II.1.



(a) Demand (b) Value of time (c) Fleet size

Figure 4.7. Variations of the total cost with (a) demand (b) value of time, and (c) fleet size

A similar pattern appears for the variation of the total cost with users' value of time. As shown in Figure 4.7(b), the total costs of all cases are relatively close for the lower values of *VOT*. However, some of them rise faster, such as I.1, while some others rise slower, such as II.1. Finally, Figure 4.7(c) plots the total costs of different strategies with variations in the fleet size in the Scenario III, in which the agency is assumed to possess a fixed number of vehicles. In this case, the total cost of Strategy 1 (III.1) always remains lower than that of the other strategies.

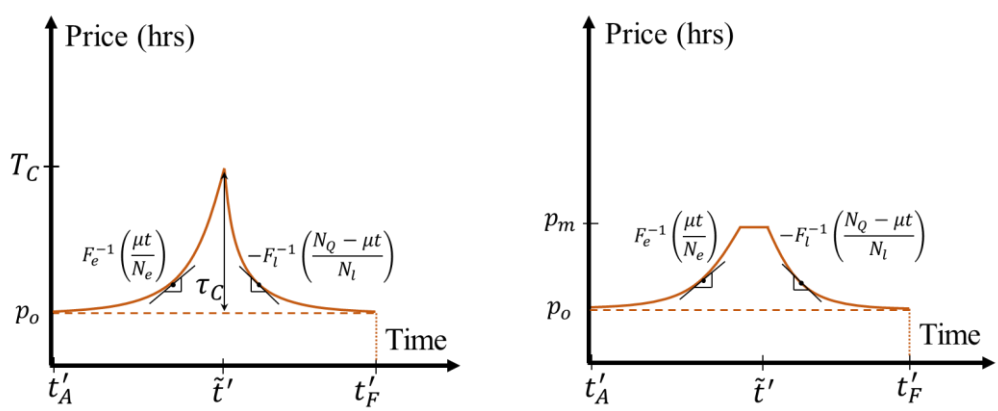
Figure 4.7(c) also reveals the importance of the fleet size optimization by indicating that, even for a given number of waiting requests, the total cost of all operating strategies can be minimized by choosing the optimal fleet size.

4.5 DRT System with a Heterogamous Demand

In this chapter, we proposed a model and an optimization method for the operation of the DRT systems. In this respect, the generalized cost of the users and the operation cost of the agency are approximated using analytical models. On this basis, the operation of the DRT system is optimized by minimizing the total cost of the users and agency using the proposed schedule management strategies. To approximate the agency cost, we employed an analytical model to formulate the operation cost of the DRT system. To approximate the users' cost, we first showed that a DRT system with a state-dependent capacity and a time-dependent demand can be modeled as queueing system. In the sense of that, we adapted the Vickrey's (1969) congestion theory to approximate the components of the cost that users experience in the system. For the purpose of simplicity, we assumed that the relative importance of schedule deviation remains constant among the homogenous users. However, the reality is that the preferences of the users varies among the heterogeneous users. To relax such simplifying assumption, we may account for the heterogeneity in preferences of the users by adapting the concept of the efficient frontier for modeling such queueing system.

On this basis, the approximation of the components of the users' cost can be generalized by plugging the analytical solution of the morning commute problem proposed in Section 3.1.2.2 into the user cost model presented in Section 4.1.2 to account for the heterogeneity in schedule penalty preferences of the users. Subsequently, the solution of the optimization

problems of Scenarios I, II, and III will be updated as well, although the general formulations of optimization problems (equations (4.13), (4.20), and (4.28)) remain untouched. Having the analytical solution of the morning commute problem with heterogeneous user preferences, it becomes possible to generalize the variation of the waiting time that users experience over time in the equilibrium condition, according to Figure 3.4. In this respect, the effeteness of dynamic pricing strategies of Figure 4.5a, b will be improved by considering the heterogeneity in schedule penalty preferences of the user according to the Figure 3.8. In this respect, the proposed optimal and constrained pricing strategies for the DRT system can be generalized as illustrated in Figure 4.8a, b.



(a) Optimum Pricing Strategy (b) Constrained Pricing Strategy

Figure 4.8. Dynamic pricing strategies with heterogeneous user preferences.

4.6 Summary

The inherent trade-off between the operating cost and the quality of service of a DRT system necessitates optimizing the operations to balance them. In this part of the research, an analytical model based on Daganzo (1978) is employed to approximate fleet size, VHT, and VMT of the DRT system. Accordingly, the operating cost for the agency is estimated as a linear combination of these components. The users are also subjected to costs of using the service. When the operating capacity of the system is inadequate to cover the demand, users of the system incur costs of delay, earliness, and lateness.

In response to the costs associated with excessive demand, rational users adjust their request times to keep their own generalized cost minimized. The cumulative result of the individual decisions of each user leads to an equilibrium condition in which no one has an incentive to change his/her own requested time. In this respect, we adapt Vickrey's (1969) congestion theory to model the DRT system, and approximate the delay, earliness, and lateness of the users in the equilibrium condition. In addition, the total time that users spend in service can be approximated using the analytical model from Daganzo (1978). As a result, the efficiency of the DRT system can be optimized by minimizing the total cost for the agency and users, where the operating capacity of the system or the number of waiting requests or both can be considered as the decision variable(s) of the problem. This part of the research presents optimizations for three scenarios: allowing only the operating capacity to change, allowing both to change, or holding the fleet size fixed. In each

scenario, the general problem with an S-shaped wished curve is formulated mathematically. The analytical solution is presented for the simplified case with an inverse Z-shaped wished curve.

In order to achieve optimal efficient operations of the DRT system, the demand should be spread as uniformly as possible over the rush period. Two demand management strategies are presented to spread the requests uniformly over the peak period in order to maintain an optimal number of waiting requests. In the schedule management strategy, the operator keeps the demand within the operating capacity of the service by only allowing a limited number of requests to be scheduled per time. A dynamic pricing strategy is proposed that incentivizes users to change their requested travel times without other restrictions from the operator. The most effective pricing strategy can make the distribution of the requests uniform over the peak period when the price is unbounded. A constrained dynamic pricing strategy is also proposed, which distributes the demand as uniformly as possible by changing the prices only within an allowable range. DRT systems that serve peaked demand can be optimized to balance the costs for the agency and users.

Although these models have been developed based on the operation of demand responsive transit systems, the principles can apply to other demand responsive systems (such as movement of goods) in which the cost of operations must be balanced against the quality of service for customers. The key to addressing dynamics of demand is to use management strategies that are also dynamic. Prices that vary by the time of day can play a key role in achieving efficient system performance.

CHAPTER 5

ROUTE CHOICE PROBLEM UNDER TRAVEL TIME VARIABILITY WITH HETEROGENEOUS USER PREFERENCES

Route choice is another important decision that users make for their trips in a network. Rational users tend to reduce the costs of their trips by choosing the routes with the minimum travel cost. Thus, the cumulative result of the individual decisions is the user equilibrium condition in which no one can reduce his/her cost by switching to another route. The duration of the trip is one of the important components of the travel cost, which is also correlated with number of other components of the cost like fuel consumption. Conventional traffic assignment models simplify the route choice problem by making the assumption that travel time is the only influential factor in route choice behavior of the users, which can be precisely predicted by travelers in the network. However, research shows that travelers can estimate the average travel time as well as its variations for different routes based on their previous experiences in the network. In this part of the research, we study the route choice behavior of the users in the network under travel time variability, while there is a heterogeneity associated with risk sensitivity of the users. The concept of the efficient frontier is used to represent the equilibrium solution of the route choice problem. On the basis of the specific properties of the efficient frontier, we propose a mathematical formulation of the route choice problem under travel time variability. A

solution algorithm is also designed that uses the primary characteristics of the equilibrium condition to assign the heterogeneous demand to the network. The efficiency of the proposed solution method is also comparted with a classic smoothing assignment method in a numerical example. The proposed model can also have broader applications in modeling decision making procedure of the travelers in the transportation network.

5.1 Generalized Cost Function

Route choice is a multi-criterion decision-making process by nature in which individuals tend to minimize their costs or maximize their benefits according to their own preferences. However, it is the heterogeneity of preferences among users that makes the equilibrium problem complicated. The generalized cost associated with the trip encompasses different components of the cost that users experience to get to the destination. In addition, a risk always comes along with this cost, which should be taken into account as well. Beckmann et al. (1956) describes how the uncertainty in the travel cost gives rise to the risk that should be taken into account in travel cost estimation:

"…the cost of transportation on a road includes not only the operating cost of a vehicle over the length of the road, but also such things as the travel time and the risks incurred. … Under Risk Cost we shall include the losses from accidents in terms of life, health, and

property, as well as the irritation from the threat of such accidents, which is particularly manifest under conditions of road congestion." (Beckmann et al., 1956).

However, the uncertainty in cost estimation is not restricted to the monetary costs. As explained in the previous section, the variation in the trip duration is another source of risk involved in the decision making procedure. Travel time is a primary component of the travel cost, and it is also correlated with fuel consumption and other costs that are incurred as part of their travel. Nevertheless, it is not possible for users to predict the exact route travel times that they will experience, because they can only estimate average trip durations according to their previous experiences. Hence, the variation in the trip durations is another unpleasant factor that travelers consider in route choice (Abdel-Aty et al., 1995; Noland, 1999). The variation in travel time poses a risk of schedule delay to the users who desire to arrive to their destinations punctually and can freely choose when to start their trips and their route. Based on Vickery (1969), these users account for the earliness or lateness penalty cost that they may experience due to arrival at their destinations ahead of or behind schedule. From a statistical perspective, the variation of route travel time can be measured by its standard deviation, regardless of the shape of the distribution (Fosgerau and Engelson, 2011). However, the cost of deviation from the schedule is not always symmetric, since it may change with the purpose of the trip and vary among the people.

The value of travel time reliability represents the level of sensitivity of the user to risk. In this regard, more conservative users place higher importance on the travel time variation;

in contrast, risk takers show less consideration of it. The variation in the trip duration determines the reliability of the route choice, so it should be included in the generalized cost function. As a result, the linear combination of the expected travel time and its standard deviation is employed as the generalized cost function to measure the disutility of a chosen route, based on Small's (1982) scheduling model, regardless of the form of the travel time distribution (Fosgerau and Karlström, 2010; Shahabi et al., 2013). The disutility experienced by a member of risk group p traveling on route i is given by:

$$DU_i^p = t_i + \gamma_p \, s_i \tag{5.1}$$

where, t_i represents the expected travel time on route i, and its standard deviation is denoted by s_i. Here, we keep the simplifying assumption that the travel time of the links are independent of each other, so t_i and s_i^2 can be calculated as the summation of link expected travel times and associated variances, respectively (Lo et al., 2006; Watling 2006; Tan et al, 2014). Moreover, in this cost function, γ_p is the risk sensitivity parameter which reflects the relative importance of the travel time variability for the users in group p who share the same risk sensitivity. In other words, the risk sensitivity can be interpreted as the equivalent delay that the user is willing to tolerate to reduce one unit of time in the variation of his/her trip duration.

According to Fosgerau and Karlström (2010), the value of the risk sensitivity parameter can be determined based on the marginal utilities of earliness and lateness for any given distribution of the trip duration. However, this value does not apply equally to all travelers due to the heterogeneity of user preferences. To include such heterogeneity, γ can have a discrete or continuous probability distribution over the travel demand. Figure 5.1 illustrates a hypothetical probability distribution of the risk sensitivity of users.

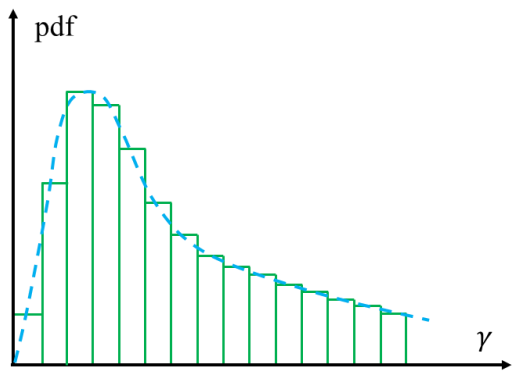


Figure 5.1. Probability distribution of the risk sensitivity parameter

Statistically speaking, the expectation and standard deviation of the route travel time can be defined as a function of these variables in the constituting links. Of special interest here is that the expected travel time and its standard deviation are not independent of each other; in contrast, there appears to be a relationship between them, especially for travel time in networks (Herman and Lam, 1974; Richardson and Taylor, 1978; Fosgerau, 2010; Mahmassani et al., 2013). Both of these variables can be defined as functions of the

traffic flows (Noland et al., 1998). On this basis, in this part of the research we consider the standard deviation of travel time as a function of the traffic flow. Therefore, we can use a change of variables to express travel time as a function of the standard deviation based on the relationship implied by the traffic flow. Here, we consider a general relationship between these variables in which the expected travel time is an increasing function of the standard deviation of the travel time, and standard deviation of travel time in an increasing function of the flows in the network. By using the linear cost function (5.1) from the literature, the EFRC can be shown to be decreasing convex function with a specific geometric property as explained in the following section.

5.2 Representing Equilibrium with the Efficient Frontier of Route Choice

In this section, we first demonstrate the route choice problem under travel time variability with heterogeneous user preferences to risk using a set of complementary conditions. Then, we explain how the concept of the efficient frontier can be adapted from portfolio theory (Markowitz, 1952) in finance to model the equilibrium of the route choice problem. For that purpose, we first consider the simplified case of the continuous differentiable efficient frontier to demonstrate the specific characteristics of the EFRC. Then, we extend the results to the realistic case of the piecewise efficient frontier by relaxing the simplifying assumptions. The identified properties of the EFRC are used in upcoming sections for modeling and solving the route choice problem with heterogeneous user preferences.

In the route choice decision making process, the rational user tends to minimize his/her general travel cost by making choices to reduce both the average trip duration and reliability of the routes according to his/her previous experiences. Due to variation of the taste among people, different users may find that different routes best match with their preferences. The aggregate results would be an equilibrium condition in which all users have specific route choices in accordance with their sensitivity to the risk, which leave them no incentive to switch to another route. In this section, we first present the equilibrium conditions for a set of heterogeneous travelers choosing among diverse routes. Then the equilibrium conditions are used to introduce the concept of the efficient frontier as the representation of equilibrium route choices.

In the equilibrium condition, users who share the same value for the reliability experience the same general travel costs, but this cost may vary over groups with different preferences. The traffic assignment problem for a single origin-destination network can be formulated as a set of Karush-Kuhn-Tucker complementary conditions, the solution of which represents the equilibrium route flows.

$$x_i^p\left(DU_i^p - \lambda_p\right) = 0 \qquad \forall i, p \tag{5.2}$$

$$DU_i^p - \lambda_p \geq 0 \qquad \forall i, p \tag{5.3}$$

$$\sum_i x_i^p = d_p \qquad \forall p \tag{5.4}$$

$$x_i^p \geq 0 \qquad \forall i, p \tag{5.5}$$

In this formulation, x_i^p denotes the flow of the sensitivity group p in route i, while d_p is the number of travelers in group p. λ_p is the dual variable for the flow conservation constraint for the group p. In the equilibrium condition, λ_p represents the minimum disutility associated with traveling in the network for group p, which is identical for the people in the same demand group.

Heterogeneity of the user preferences regarding travel time reliability leads to asymmetric route choice of the users. So, in a similar condition, different travelers with different sensitivities to risk may choose different routes, which might be different from the shortest path as defined by the expected travel time. In this regard, more conservative users (higher γ) prefer the routes with lower standard deviation in travel time and tolerate higher expected travel time. In contrast, risk-takers (lower γ) choose the routes with lower expected travel time and accept a higher level of variation in their trip durations. However, no one take routes with higher expected travel time and higher level of risk as long as there exists at least one alternative route that dominates these routes by offering better expected travel time, reliability, or both.

Definition (EFRC). The efficient frontier of route choices (EFRC) is the convex hull of the equilibrium route choice set in the $s - t$ plane that have the minimum generalized cost according to (5.1) for the users with heterogeneous preferences regarding the relative importance travel time variability.

On the *s*-*t* plane (See Figure 5.2), each fixed point represents the expected travel time and its standard deviation for a route, which in this illustration are independent of the route flows. A line with slope $-\gamma_p$ represents a set of points in the *s*-*t* space with constant disutility for a person in group p. Disutility increases moving away from the origin as travel time and variability increase. At equilibrium, the travelers in group p will choose the point(s) where the disutility is lowest, and this disutility is λ_p. No user in group p will choose a route with greater disutility than λ_p, as constrainted by (5.2) and (5.3). The upper envelope of the constant disutility lines across all p defines a convex curve called the efficient frontier of route choices (EFRC), which is shown by the dashed curve. Only the portfolio of dominant routes, which are along this frontier, will be used in the equilibrium condition, and users will be distributed among these routes based on their preferences for risk. Figure 5.2 illustrates the concept the EFRC for the single origin-destination problem.

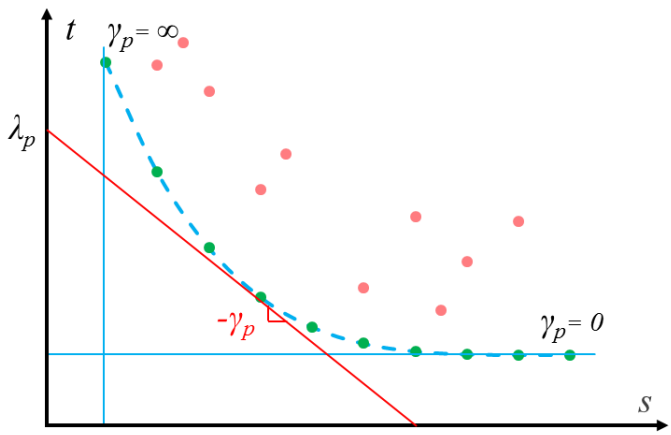


Figure 5. 2. Route choice efficient frontier concept shown by dashed line

In this framework, the conventional user equilibrium can be interpreted as a special case of the EFRC when all users are absolutely risk seekers and seek the route with the shortest expected travel time, regardless of the associated risk. In this case that travelers have no sensitivity to the risk ($\gamma_p = 0$), the EFRC turns out to be a horizontal line in the *s*-*t* plane. As the risk sensitivity of the users goes to infinity ($\gamma_p = \infty$), the solution of the problem tends to be a vertical line on the *s*-*t* plane. In realistic cases, there is heterogeneity among users in their sensitivity to risk, which results in the general shape of the EFRC. The EFRC is the solution of this bi-objective route choice problem, which not only provides intuitions about the route choice behavior of the users in the equilibrium condition, but also ranks the route choice of the users according to their sensitivity to risk. This information is crucial for updating the traffic assignment of the users in case a change occurs in structure or performance of the network, with no need to solve the whole problem again. The EFRC

has specific properties that can shed light on the equilibrium route choice of the users. To identify these properties, we can simplify the problem to a differentiable efficient frontier by assuming infinitely many routes in the *s*-*t* plane and a continuous distribution for the risk sensitivity of the users. These simplifying assumption can help demonstrating the properties of the efficient frontier. Then, we relax the simplifying assumptions to extend the results to the realistic case of the piecewise linear efficient frontier.

5.2.1 Differentiable Efficient Frontier of the Route Choice

For the first step, we consider a couple of simplifying assumptions to keep the EFRC continuous and differentiable in order to demonstrate the specific properties of the EFRC. For one thing, we assume that there is an infinite number of routes connecting the single origin-destination pair in the network with known *s*-*t* relationships that cover the entire *s*-*t* plane. For another, we assume that the risk sensitivity parameter γ has a continuous probability distribution over the population of the heterogeneous users. In this case, the EFRC turns to a differentiable curve as illustrated in Figure 5.3. In this figure, t of the routes are depicted as increasing leaner function of their s just for the purpose of simplifying the illustration. However, this is not a general assumption in this problem, and the results remain general for any relationship for t as an increasing function of s for the routes of the network.

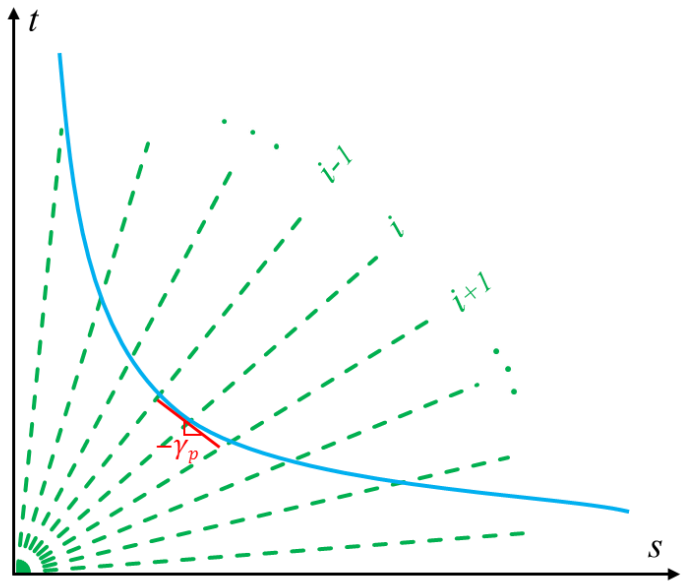


Figure 5. 3. Differentiable EFRC for infinite number of routes and continuous distribution of γ

In Figure 5.3, straight dashed lines represent the performance of the routes and, in general, characteristics of the network. The equilibrium travel time expectation and standard deviation of the routes can be determined as the intersection of the route performance curves with the EFRC which is depicted by continuous line. The shape and the location of the EFRC in the *s*-*t* plane depends on the characteristics of the routes in the network, the total demand, and the distribution of the risk sensitivity parameter. Nonetheless, the EFRC always has certain properties that are general for all problems. Furthermore, it can be shown that there is a relation between the shape of the EFRC and the distribution of γ, which can shed light on the solution of the equilibrium route choice problem.

Proposition 5.1 (monotonicity). The EFRC, $t(s)$, is a non-increasing function of s (See Proposition 2.1 for proof).

Proposition 5.2 (convexity). The EFRC, $t(s)$, is a convex function of s (See Proposition 2.2 for proof).

Proposition 5.3 (geometric constraint). When there are an infinite number of routes and a continuous distribution of γ, the slope of the EFRC at its intersection with route i is $m_i = \partial t(s_i)/\,\partial \mathrm{s_i} = -\gamma_p$ for the users choosing this route as illustrated in Figure 5.3

Proof. In the equilibrium condition all the users choose the route with least disutility. If there are an infinite number of routes so that the disutility is continuous and differentiable across all s_i, the first derivative of the disutility for demand group p with respect to x_i^p should be equal to zero.

$$\frac{\partial DU_i^p}{\partial x_i^p} = \frac{\partial DU_i^p}{\partial s_i} \cdot \frac{\partial s_i}{\partial x_i^p} = 0 \tag{5.6}$$

By substituting the definition of the disutility (5.1) into the equation (5.6), we may rewrite this equation as below:

$$\frac{\partial s_i}{\partial x_i^p}\left(\frac{\partial t(s_i)}{\partial s_i} + \gamma_p\right) = 0 \tag{5.7}$$

Knowing that s_i strictly increases with x_i^p, the slope of the EFRC at its intersection with route i, m_i, chosen by group p, can be determined as follows:

$$m_i = \frac{\partial t(s_i)}{\partial s_i} = -\gamma_p \tag{5.8}$$

■

Corollary 5.1 (assignment order). If routes are labeled in an increasing order of their equilibrium standard deviation of travel times, the demand fills the routes in a decreasing order of the risk sensitivity. The decreasing monotonicity property (Proposition 5.1) preserves the order of s_i and t_i, so the expected travel time will be in decreasing order because the slope is always non-positive. The convexity property (Proposition 5.2) ensures that the negative slope is increasing (i.e., becoming less steep). Then, the geometric constraint (Proposition 5.3) implies that the assignment starts with the greatest γ, where the negative slope is steepest, and progresses sequentially to the lowest γ, where the slope is flattest.

Now that the general properties of the EFRC is demonstrated using the differentiable efficient frontier, we may relax the simplifying assumptions to extend the results to the realistic case of the piecewise linear efficient frontier.

5.2.2 Piecewise Linear Efficient Frontier of the Route Choice

In urban transportation networks, there are a finite number of routes, n, connecting the origin-destination pairs for the travel demand with either a discrete or continuous distribution of γ values. In this case, EFRC, as the convex hull of the equilibrium route choice sets of the heterogeneous users in the $s - t$ plane, becomes a piecewise linear curve with similar properties to the differentiable case, as illustrated in the Figure 5.4, the performance of the routes is plotted by the straight dashed lines, which are simplified for the illustration, and the results hold for general relationships between t_i and s_i. The piecewise linear EFRC is depicted by the continuous curve. Proposition 5.1 and Proposition 5.2 still hold as stated in Section 5.2, while Proposition 5.3 must be adapted for discrete and continuous distribution of γ in the following sections.

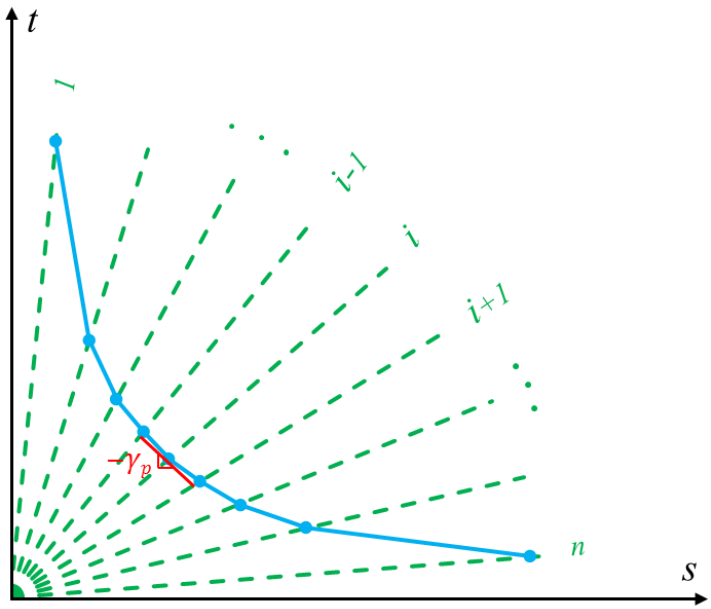


Figure 5.4. Piecewise linear EFRC for finite number of routes and discrete distribution of γ

Proposition 5.4 (geometric constraint). For n routes that are labeled sequentially in order of increasing order of standard deviation of equilibrium travel time s_i, route i has the lowest disutility for the user with risk sensitivity γ_p if and only if $-\gamma_p$ falls between the slopes of the EFRC at its intersection with route i. The first and the last routes are just bounded from one side.

$$p \in P_i \longleftrightarrow \begin{cases} m_{i,i+1} \geq -\gamma_p & , \ i = 1 \\ m_{i-1,i} \leq -\gamma_p \leq m_{i,i+1} & , \ 1 < i < n \\ m_{i-1,i} \leq -\gamma_p & , \ i = n \end{cases} \tag{5.9}$$

where, $m_{i,j}$ denotes the slope the piece of the equilibrium EFRC between routes i and j, and P_i denotes the set of sensitivity groups that choose route i.

Proof. Since the piecewise EFRC is not differentiable, we make use the fundamental concept of the minimum disutility in route choice. It follows from the complementary conditions (5.2) and (5.3) that a traveler with risk sensitivity γ_p chooses route i, if and only if, route i offers a lower generalized cost than any other route j.

$$p \in P_i \longleftrightarrow DU_i^p \leq DU_j^p \quad , \quad 1 \leq j \leq n \tag{5.10}$$

The disutility of route i and route j for users with the sensitivity γ_p can be substituted with the definition from (5.1), to obtain the condition below:

$$p \in P_i \longleftrightarrow t(s_i) + \gamma_p\, s_i \leq t(s_j) + \gamma_p\, s_j \quad , \quad 1 < j < n \tag{5.11}$$

By the solving the inequality for γ_p, we get two cases, depending on whether j is greater than or less than i:

$$-\gamma_p \leq \frac{t(s_j) - t(s_i)}{s_j - s_i} \quad , \quad j > i \tag{5.12}$$

$$-\gamma_p \geq \frac{t(s_i) - t(s_j)}{s_i - s_j} \quad , \quad j < i \tag{5.13}$$

The right side of (5.12) and (5.13) represent the slope connecting points i and j. For the route $i-1$ and route $i+1$, expression (5.11) can be rewritten using (5.12) and (5.13) as follows:

$$p \in P_i \longleftrightarrow \frac{t(s_i)-t(s_{i-1})}{s_i-s_{i-1}} \leq -\gamma_p \leq \frac{t(s_{i+1})-t(s_i)}{s_{i+1}-s_i} \quad , \quad 1 < i < n \tag{5.14}$$

According to the definition of the slope of the piecewise EFRC, the following result can be concluded.

$$p \in P_i \longleftrightarrow m_{i-1,i} \leq -\gamma_p \leq m_{i,i+1} \quad , \quad 1 < i < n \tag{5.15}$$

For the first and the last route, condition (5.13) can be modified as below:

$$p \in P_i \longleftrightarrow \begin{cases} m_{i,i+1} \geq -\gamma_p \ , \ i = 1 \\ m_{i-1,i} \leq -\gamma_p \ , \ i = n \end{cases} \tag{5.16}$$

because the slopes at the edge are bounded from only one side. ■

Corollary 5.2 (general geometric constraint). Considering the convexity property of the EFRC, condition (5.25) can be generalized as below:

$$p \in P_i \longleftrightarrow m_{1,2} \leq m_{2,3} \leq \cdots \leq m_{i-1,i} \leq -\gamma_p \leq m_{i,i+1} \leq \cdots \leq m_{n-2,n-1} \leq m_{n-1,n}$$

(5.17)

Corollary 5.3 (assignment order). Similar to Corollary 5.1 for the differentiable EFRC, by keeping the routes in order of increasing standard deviation at the equilibrium condition, and combining Propositions 5.1, 5.2, and 5.4, demand groups should be assigned to the routes in a decreasing order of their sensitivity to the risk at equilibrium.

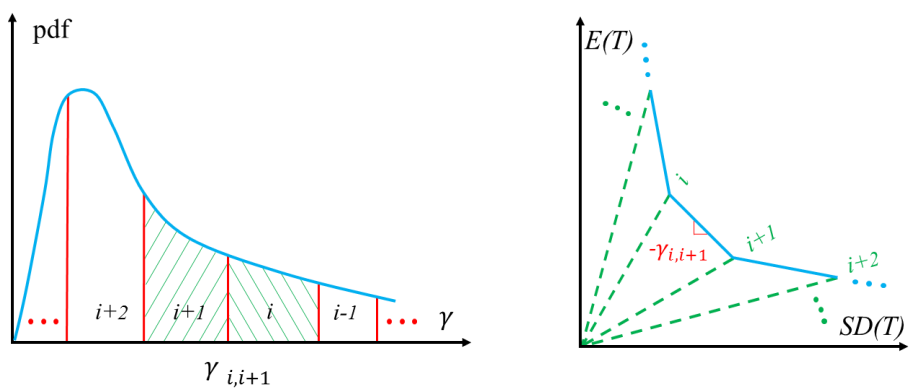
Corollary 5.4a (dual geometric constraint discrete case). Constraint (5.9) holds for all the groups choosing 2 consecutive routes, if the slope of the piece of the EFRC between these routes will be higher than $-\gamma_p$ of all groups choosing the first route and also lower than all $-\gamma_p$ of the of the groups choosing the second route, and vice versa. Accordingly, condition (5.9) holds for all routes and groups if and only if the condition (5.18) holds for all the routes.

$$sup_{p\epsilon P_i}\{-\gamma_p\} \leq m_{i,i+1} \leq inf_{p\epsilon P_{i+1}}\{-\gamma_p\} \tag{5.18}$$

In the case that the travel demand of the users is considered to have a continuous distribution of γ values, the piecewise linear EFRC will hold the same characteristics presented in Proposition 5.4 as well as Corollaries 2 and 3. Corollary 5.4a, however, can be simplified owing to the continuous nature of the γ distribution.

Corollary 5.4b (dual geometric constraint continuous case). Constraint (5.9) holds for the γ-range of the travelers choosing 2 consecutive routes, if the slope of the piece of the EFRC between these routes will be higher than $-\gamma_p$ of all users choosing the first route and also lower than all $-\gamma_p$ of the of the users choosing the second route, and vice versa. As it is illustrated in the Figure 5.5a for a hypothetical continuous distribution of the γ values, the upper and lower bounds of $m_{i,\mathrm{i}+1}$ in condition (5.18) will be equal to the same value, we name $-\gamma_{i,i+1}$. Accordingly, condition (5.9) holds for all routes and users if and only if condition (5.19) holds for all the routes as illustrated in Figure 5.5b.

$$m_{i,i+1} = sup_{p\epsilon P_i}\{-\gamma_p\} = inf_{p\epsilon P_{i+1}}\{-\gamma_p\} = -\gamma_{i,i+1} \tag{5.19}$$



(a) Route choice of a traffic demand with continuous distribution of γ

(b) Piecewise linear EFRC for the continuous distribution of γ

Figure 5.5. The EFRC of a continuous distribution of γ

In Figure 5.5a, the continuous curve illustrates a hypothetical probability distribution of γ values, and the vertical straight lines split this distribution between the routes in a decreasing order of the risk sensitivity, according to the Corollary 5.3. As a result, the share of each route from the total demand can be represented by the area bounded between these lines, under the distribution curve. In the equilibrium condition, such assignment of demand results in a monotonically decreasing and convex EFRC, with the segment slope between each two consecutive routes *i* and *i+1* equal to $-\gamma_{i,i+1}$, as indicated in the Figure 5.5b.

Corollary 5.5. As a direct result of Corollary 5.4a, the demand of group p splits between two or more routes if and only if the slope of the EFRC between these routes is $-\gamma_p$. It can be equivalently interpreted as the same generalized cost for these users.

$$m_{i,i+1} = \cdots = m_{j-1,j} = m_{i,j} = -\gamma_p \quad \longleftrightarrow$$

$$\cap_{k=i}^{j} P_k = \{p\} \;\; AND \; p \notin (\cup_{k<i} P_k) \cup \left(\cup_{k>j} P_k\right) \tag{5.20}$$

Corollary 5.6. Following from Corollary 5.5, each subset of routes can have at most one group in common which splits between those routes such that condition (5.20) holds, because the split group must have a $\gamma_p = -m_{i,i+1}$ for each consecutive pair of routes.

The piecewise EFRC represents the equilibrium solution of the proposed traffic assignment problem. Thus, its primary properties can be employed in a mathematical formulation of the problem.

5.3 Mathematical Formulation

The traffic assignment model proposed in this part of the research considers the heterogeneity of risk sensitivity among users while there is an inherent variation associated with route travel times. The tendency of users to minimize their own general travel costs eventually end in an equilibrium condition, which can be represented by the EFRC. The EFRC has specific properties that can be employed to formulate the traffic assignment problem, and the equilibrium solution of the problem can be derived as a feasible solution that satisfies these properties. These properties can be included as a set of constraints in the mathematical formulation. It is also worth noting that all of the properties of the EFRC are directly derived from the Karush-Kuhn-Tucker complementary conditions (5.2) and (5.3). Accordingly, the existence and uniqueness of the solution of the proposed problem can be verified.

5.3.1 General Mathematical Problem

Keeping the assumption that demand groups are labeled in decreasing order of their sensitivity to risk, the results of the Corollaries 3, 4a, and 5 can be adapted to formulate the traffic assignment problem with a discrete distribution of demand as below:

$$x_i^p x_j^{p+1}\left(m_{i,j} + \gamma_p\right) \geq 0 \quad , \quad \forall i, j, p \tag{5.21}$$

$$x_i^p x_j^{p+1}\left(m_{i,j} + \gamma_{p+1}\right) \leq 0 \quad , \quad \forall i, j, p \tag{5.22}$$

$$x_i^p x_j^p\left(m_{i,j} + \gamma_p\right) = 0 \quad , \quad \forall i, j, p \tag{5.23}$$

$$x_i^p x_j^{p+1}\left(s_i - s_j\right) \leq 0 \quad , \quad \forall i, j, p \tag{5.24}$$

$$\sum_i x_i^p = d_p \quad , \quad \forall p \tag{5.25}$$

$$\sum_{p \epsilon P_i} x_i^p = x_i \quad , \quad \forall i \tag{5.26}$$

$$m_{i,j} = \frac{t(s_i) - t(s_j)}{s_i - s_j} \quad , \quad \forall i, j \tag{5.27}$$

$$x_i^p \geq 0 \quad , \quad \forall i, p \tag{5.28}$$

In this mathematical problem, constraints (5.21) and (5.22) restrict the slopes of each segment of the EFRC to its upper and lower bounds in accordance with Corollary 5.4a. These constraints allow $x_i^p, x_j^{p+1} > 0$ if and only if the slope of the slope of segment (i, j) of the EFRC falls between negative values of risk sensitivity of groups p and $+1$, $-\gamma_p \leq m_{i,j} \leq -\gamma_{p+1}$. Equation (5.23) also determines the slope of the EFRC when the demand of one group splits between two or more routes, on the basis of Corollary 5.5.

Corollary 5.3 is also included by the constraint (5.24) which ensures that demand groups fill the routes in a decreasing order. The set of constraints (5.21) through (5.24) guarantees the primary properties of the EFRC in the equilibrium condition (see Propositions 5.1, 5.2, and 5.4). Moreover, equation (5.25) also assures the conservation of flow in the network. In addition to these constraints, which reflect the properties of the solution, there are a couple of definitions included in the model. Equation (5.26) defines the route flow as the summation of flows of different groups in that route. The geometric slope of each segment of the EFRC is also defined by equation (5.27). Lastly, a non-negativity constraint is included by inequality (5.28).

For a continuous distribution of γ values, a similar traffic assignment problem can be formulated by adapting Corollaries 3 and 4b to split the continuous distribution of the demand between the routes as follows:

$$\delta_{i,j}\left(m_{i,j} + \gamma_i^U\right) = 0 \quad , \quad \forall i,j \tag{5.29}$$

$$\delta_{i,j}\left(\gamma_i^U - \gamma_j^L\right) = 0 \quad , \quad \forall i,j \tag{5.30}$$

$$\sum_{i,j} \delta_{i,j} = n - 1 \tag{5.31}$$

$$\left(s_i - s_j\right)\left(\gamma_i^U - \gamma_j^U\right) \leq 0 \quad , \quad \forall i,j \tag{5.32}$$

$$\sum_i \left(F_\gamma(\gamma_i^U) - F_\gamma(\gamma_i^L)\right) = 1 \quad , \quad \forall i \tag{5.33}$$

$$\left(\gamma_i^U - \gamma_j^U\right)\left(\gamma_i^L - \gamma_j^L\right) \geq 0 \quad , \quad \forall i,j \tag{5.34}$$

$$\left(F_\gamma(\gamma_i^U) - F_\gamma(\gamma_i^L)\right) D = x_i \quad , \quad \forall i \tag{5.35}$$

$$\delta_{i,j} \in \{0,1\} \quad , \quad \forall i,j \tag{5.36}$$

$$\gamma_i^U \geq \gamma_j^L \geq 0 \quad , \quad \forall i \tag{5.37}$$

In this formulation of the problem, constraints (5.29), (5.30), and (5.31) determine the slopes of the EFRC segments according to the Corollary 5.4b, where γ_i^L and γ_i^U are the decision variables of the problem and respectively represent the lower and upper bounds of the share of route *i* from the continuous distribution of γ. In this set of constraints, equation (5.29) makes sure that if routes *i* and *j* are next to each other in the s-t plane, the slope of the EFRC segment between them, $m_{i,j}$, is exactly equal to the upper and lower bounds of the shares for routes i and j, which are equal in the equilibrium condition as illustrated in the Figure 5.5a. Constraints (5.30) and (5.31) also define the value of the dummy variable $\delta_{i,j}$ by setting its value equal to 1 if routes *i* and *j* are right next to each other such that $\gamma_i^U = \gamma_j^L = \gamma_{i,j}$; otherwise, $\delta_{i,j}$ remains zero (See condition (5.19)). Constraint (5.32) also assures that the travel demand of the heterogeneous users fill the routes in a decreasing order of γ to reflect the Corollary 5.3. As a result, the primary characteristics of the EFRC in the equilibrium condition (See Propositions (5.1), (5.2), and (5.4)) are covered by constraints (5.29) through (5.32). In addition, constraint (5.33) guarantees the flow conservation of the demand by restricting the summation of route shares to 1, while constraint (5.34) makes sure that their shares will not overlap (See Fig 6 where, $F_\gamma(\cdot)$ denotes the cumulative distribution function of γ. Equation (5.35) also defines

the route flows, in which D is the tot al travel demand of the users. Finally, $\delta_{i,j}$ is defined as dummy variable in the constraint (5.36), while the non-negativity constraint (5.37) sets the upper bound of the share of each route higher than its lower bound. In this problem, definition of the slopes of the EFRC segments is the same as the equation (5.27).

5.3.2 Two-Stage Formulation

The mathematical formulation of the problem presented in Section 5.3.1 can be simplified in the case that the order of the routes is known. Accordingly, the problem can be formulated as a two-stage model for which the upper stage problem is to put the routes in an increasing order of their standard deviation of the travel times, in a ccordance with the Corollary 5.3. Given the order of the routes for a discrete distribution of γ, the lower stage problem splits the distribution of the demand between the routes such that the equilibrium solution meets the primary properties of the EFRC as below.

$$x_i^p\left(m_{i-1,i}+\gamma_p\right) \leq 0 \quad , \quad \forall i,p \tag{5.38}$$

$$x_i^p\left(m_{i,i+1}+\gamma_p\right) \geq 0 \quad , \quad \forall i,p \tag{5.39}$$

$$m_{i-1,i} \leq m_{i,i+1} \quad , \quad \forall i \tag{5.40}$$

$$\sum_{p\epsilon P_i} x_i^p = x_i \quad , \quad \forall i \tag{5.41}$$

$$\sum_i x_i^p = d_p \quad , \quad \forall p \tag{5.42}$$

$$x_i^p \geq 0 \quad , \quad \forall i,p \tag{5.43}$$

Constraints (5.38) and (5.39) again determine the boundaries of the slope of each segment of the EFRC on the basis of Corollary 5.4a. It is worth pointing out that there is no need to add a separate constraint to include Corollary 5.5 in this formulation of the problem, because it is redundant to the first two constraints. Constraint (5.24) also can be replaced by constraint (5.40) which directly represents the Proposition 5.2. The conservation of demand constraint is guaranteed by the equation (5.41). Constraint (5.42) also defines the route flows. Finally, inequality (5.43) represents the non-negativity constraint. In this problem, the slope of the segments of the EFRC, $m_{i,i+1}$, is defined according to equation (5.26), where $j=i+1$.

The lower stage problem can be significantly simplified by considering a continuous distribution for γ values over the travel demand of the users as follows:

$$m_{i,i+1} + \gamma_{i,i+1} = 0 \quad , \quad \forall i \tag{5.44}$$

$$\gamma_{i-1,i} \geq \gamma_{i,i+1} \geq 0 \quad , \quad \forall i \tag{5.45}$$

$$\left(F_\gamma(\gamma_{i-1,i}) - F_\gamma(\gamma_{i,i+1})\right) D = x_i \, , \quad \forall i \tag{5.46}$$

Here, constraint (5.44) reflects Corollary 5.4b by determining the slopes of the EFRC segments according to condition (5.19) as illustrated in Fig 6. Here, routes are presumed to be labeled in an increasing order of their travel time standard deviations, so the non-negativity constraint (5.45) assures that the continuous distribution of demand fills these

routes in a decreasing order of γ according to the Corollary 5.3. As a result, we may substitute γ_{i+1}^{U} and γ_{i}^{L} with their equivalent value $\gamma_{i-1,i}$ as the decision variable of the problem, according to the condition (5.19), where $\gamma_{i+1}^{U} = \gamma_{i}^{L} = \gamma_{i,i+1}$. In this formulation of the problem, constraints (5.44) and (5.45) fulfill all the primary properties of the piecewise linear EFRC demonstrated in Propositions (5.1), (5.2), and (5.4). Additionally, equation (5.46) assigns the routes their shares of the total demand (See Fig 6).

This formulation of the problem is of great importance for proposing a solution algorithm for this problem, since there exists a variety of different algorithms that can efficiently relabel the routes in an increasing order of the standard deviation of their travel times. In practice, we just need to solve this simplified version of the problem. On this basis, the next section proposes an efficient solution algorithm for this bi-objective route choice problem.

5.4 Solution Algorithm

The convex combination method that can be used for solving the multi-objective problems typically involves finding the best "movement directions" and optimizing the "movement steps" in each iteration, for each demand group separately (Dial, 1996). Consequently, the efficiency of the method clearly declines as the number of the sensitivity groups increases up to the point the distribution of γ can be considered continuous over the population of the travelers. However, the specific properties of the EFRC can be employed to design a

more efficient solution algorithm to solve this bi-objective traffic assignment problem when γ is continuously distributed.

We now propose a bi-stage algorithm which can solve the route choice problem by building the EFRC for both continuous and discrete distribution of demand. The upper stage algorithm labels the routes in increasing order of the standard deviations of their travel times, while the lower stage algorithm iterates to build the EFRC using its known characteristics, given the order of the routes. Whenever building the EFRC requires an alteration in the order of the routes, the algorithm returns back to the upper stage algorithm for label rearrangement, then the lower stage algorithm starts constructing the EFRC for the updated order of the routes. In this way, the proposed algorithm simply iterates over the paths to modify the assignment of the heterogeneous demand according to the insights from the properties of the EFRC until all of the equilibrium conditions are met. Consequently, more extensive heterogeneity of the demand may require a higher number of iterations over the paths before equilibrium is achieved. However, the heterogeneity of the demand does not complicate the assignment modification procedure of the algorithm as its steps remain the same for any number of demand sensitivity groups. This property of the proposed algorithm can be viewed as an advantage over other solution methods, like convex combination, that require finding the best directions and optimal movement step sizes for all the demand sensitivity groups one by one. Notice as well that there exists only one feasible order of the routes in which the unique EFRC can be rebuilt by the solution algorithm, so the equilibrium is identified when conditions (5.38) through (5.43) or (5.44)

through (5.46) are met for discrete or continuous distribution of γ, respectively. It is also worth pointing out here that whether the γ values follow a discrete or continuous distribution does not change the general procedure of the solution algorithm; however, the steps of the lower stage algorithm should be adapted accordingly. The main steps of this solution algorithm for both discrete and continuous distributions of γ are summarized in the flowchart of Figure 5.6.

This method starts with a feasible initial solution for the problem. Then it alternates between the upper stage and lower stage algorithms until the convergence criterion is met.

5.4.1 Initialization

The initial solution to the route choice can be any feasible distribution of the demand between the routes. However, the properties of the EFRC shed light on the general form of the solution which can be used to define an appropriate initial condition to reduce the number of iterations needed to converge. In this case, the algorithm starts with a monotonically decreasing and convex EFRC in which the groups are assigned to the routes in a decreasing order of their risk sensitivity. In this respect, the steps of the proposed initialization method are listed as below:

Step 0. For a discrete distribution of γ, label the demand groups in decreasing order by their sensitivity to risk. For a continuous distribution of γ demand is in decreasing order by

definition. In addition, label the routes in an increasing order of standard deviations of their travel times when there is no flow in the routes.

Step 1. Assign an initial feasible assignment of demand to each route. Although any feasible assignment of traffic flows may be used, a good starting point is to identify the equilibrium for a homogeneous population in which all users' risk sensitivity is equal to the population average (γ_{avg}). By making the population homogeneous, the traffic assignment problem becomes a conventional user equilibrium problem, which can be solved using any available method. As a result, the initial EFRC can be depicted as a straight line with a negative slope in the *s*-*t* plane which meets both monotonicity and convexity constraints (see the Propositions 5.1 and 5.2).

Step 2. Split the demand between the routes in a decreasing order, as explained in Corollary 5.3, such that the aggregate flows become equal to the initial flows derived in Step 1.

The result can be used as an initial condition for the upper stage problem.

5.4.2 The Upper Stage Algorithm

The upper stage algorithm is designed to heuristically label the routes according to the re-ordering of the routes by increasing standard deviation of travel time after updates in route flows. The algorithm steps are summarized as follows:

Step 0. Label the routes in an increasing order of their updated standard deviations. To keep the new labels in order, it is necessary that the total flow on each route remains constant. So, the demand should be reassigned to the routes in accordance with the Corollary 5.3 such that the new assignment results in same total flows on each route but with the new order of the routes.

Step 1. Update the expected travel time and the standard deviation of the routes.

As a result of the upper stage algorithm, the monotonicity condition is met (Proposition 5.1), while the demand groups are assigned in a decreasing order (Corollary 5.3). Since routes are labeled based on a new order, their travel time expectations and standard deviations should be updated accordingly. The lower stage algorithm also uses this travel time information to check and analytically modify the assignment to satisfy the geometric constraint (Corollary 5.4a or 4b).

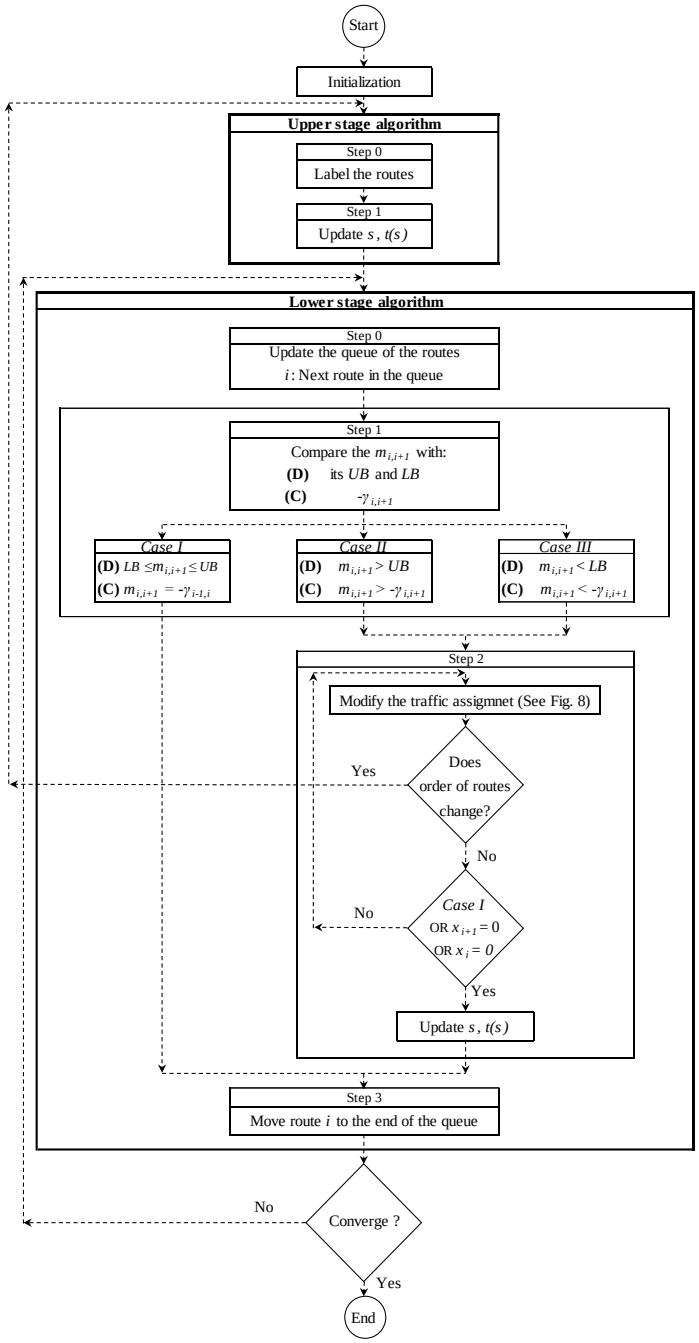


Figure 5.6. Flowchart representation of the heuristic algorithm for discrete (D) and continuous (C) distributions of γ

5.4.3 The Lower Stage Algorithm

The lower stage problem first checks that the current traffic assignment satisfies the geometric condition, route by route. If not, it iterates to analytically modify the assignment by systematically shifting the splits of the demand between the routes according to the properties of the EFRC. If such a change causes a swap in the route orders because standard deviation of two routes swap order, the algorithm diverts the iteration to the upper stage algorithm. The fundamental procedure of the lower stage algorithm is similar for both discrete and continuous distributions of the γ values; however, some minor adaptations on the details of the algorithm should be made accordingly. The steps of the lower stage algorithm for both discrete (D) and continuous (C) distributions are listed below:

Step 0. Put the routes in a queue in order of their current labels, then select the first route in the queue to evaluate the slope of the corresponding segment of the EFRC.

Step 1. Compare the slope of the selected segment of the EFRC with:

(D) the lower and upper bounds defined by condition (5.18).

(C) the split point that routes i and $i+1$ share on a continuous γ distribution, $\gamma_{i,i+1}$, defined by condition (5.19).

Three cases are possible:

Case I.

(D) The selected slope falls between its lower and upper bounds,

(C) The selected slope equals to $\gamma_{i,i+1}$,

which means the current assignment in this route needs no further modification in this iteration. Go to Step 3.

It is also worth pointing out that *Case I* also includes the condition route $i+1$ drops exactly on top of point i in the $s-t$ plane, which indeed cause the segment $(i, i+1)$ get eliminated from EFRC , and obviously $m_{i,i+1}$ becomes indeterminate.

Case II.

(D) The slope is higher than the upper bound.

(C) The slope is higher than $\gamma_{i,i+1}$.

In this case, the assignment should be modified to reduce the slope up to the point that it comes to the *Case I*. Go to Step 2.

Case III.

(D) The slope is less than the lower bound.

(C) The slope is less than $\gamma_{i,i+1}$.

So the slope should be increased by modifying the assignment such that the *Case I occurs*. Go to Step 2.

Step 2. Start with modifying the assignment of the demand in *Cases II* and *III* as below:

(D) In *Case II* (*Case III*), shift the entire demand of the group with highest (lowest) sensitivity to risk in route *i+1* (route *i*) from this route to route *i* (route *i+1*) to decrease (increase) the slope of the segment. Next, check if the slope of the segment exceeds its new upper bound (drops below its new lower bound) and goes to the *Case III* (*Case II*). In this condition, the demand of the last shifted group, *q*, should split between routes *i* and *i+1* such that $m_{i,i+1} = \gamma_q$, and then check the order of the routes. Otherwise, just check the order of the routes. The mechanism of shifting demand insightfully with the help of the identified properties of the EFRC and with no need to incremental adjustments can significantly improve the efficiency of the method in comparison to convex combination methods that need to make the adjustments incrementally. Details of procedure of modifying the assignment of the discrete demand are summarized in the flowchart of the Figure 5.7a.

(C) In *Case II* (*Case III*), decrease (increase) $\gamma_{i,i+1}$ up to the point that one of the following conditions occurs: *(5.1)* $m_{i,i+1} = -\gamma_{i,i+1}$; *(5.2)* $\gamma_{i+1,i+2} = \gamma_{i,i+1}$ ($\gamma_{i,i+1} = \gamma_{i-1,i}$), which means that there is no flow left in the route *i+1* (*i*) to transfer to route *i* (*i+1*). Given the

cumulative distribution of γ, $\gamma_{i,i+1}$ can be approximated for each couple of consecutive routes in C*ase III* by solving equation (5.19) for this sub-problem using available analytical or numerical methods. Figure 5.7b graphically demonstrates the procedure of modifying the assignment for a continuous distribution of demand.

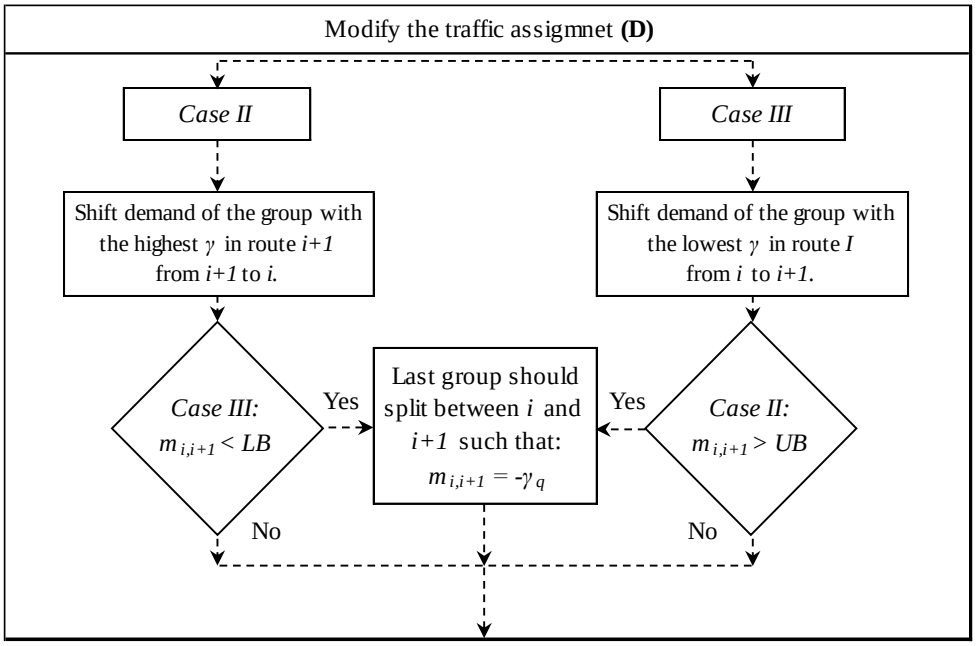
The adjustment procedure for both discrete and continuous distributions of γ will be terminated whenever the order of routes changes, in which case the algorithm returns to the upper stage algorithm. However, if the adjustment procedure is accomplished with no need to change in the order of the routes, check that one of the following conditions is met:

i) Condition of the *Case I* (See Step 1 of the lower stage algorithm).

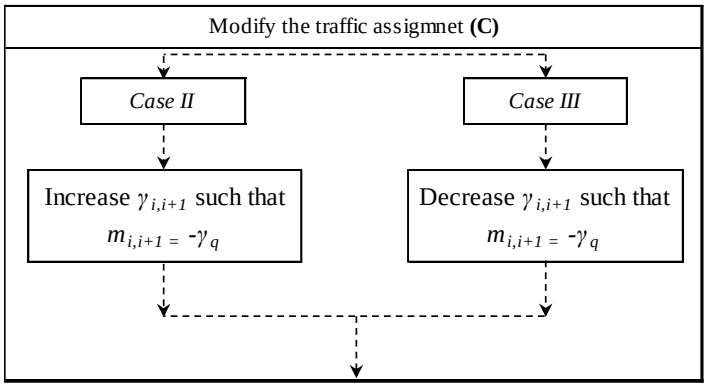
ii) $x_{i+1} = 0$ ($x_i = 0$) in the *Case II* (*Case III*).

If at least one of these conditions is met, the algorithm continues to the next step by updating the expected travel times and standard deviations. Otherwise, it repeats the Step 2 again.

Step 3. Move the selected route to the end of the queue.



(a)



(b)

Figure 5.7. Flowchart representation of procedure of modifying the traffic assignment for a (a) discrete (b) continuous distributions of γ

5.4.4 Convergence

In the equilibrium condition, no one will be able to reduce the generalized cost of the trip by switching his/her route. In case of homogeneity in preferences of the users, all the used routes should have the same generalized cost, while the more expensive routes remain unused. Nonetheless, when there is a heterogeneity associated with preferences of the users, different routes will have different cost for different users even in the equilibrium condition, which makes it impossible to use this criterion for measuring the convergence of the algorithm. Instead, we may consider the summation of the relative differences in the generalized cost for users in different groups between two consecutive iterations as the convergence measure of the algorithm. This procedure can be summarized in the following steps:

Step 0. Calculate the total generalized cost for the distribution of demand.

Step 1. If the relative change in the total generalized cost of the last iteration in comparison to the previous one meets the convergence criterion, then the equilibrium solution is considered to be achieved. Otherwise, the next iteration will start from the lower stage algorithm.

5.5 Numerical Example

To provide a numerical example of the proposed solution method, we employ this algorithm to solve a sample problem and compare its results with that of the convex combination method. In this problem, the single origin-destination network is assumed to be connected with 4 routes that have different travel time characteristics. The travel demand is also classified into 4 sensitivity groups. Thus, both the high stage and low stage algorithms are employed to estimate the equilibrium solution of the traffic assignment problem for this single origin-destination network.

The communicating routes are assumed to provide a variety of different choices for the users. For simplicity of presentation, we assume a linear relationship between the expected travel time and its standard deviation for all the routes. We also assume that the standard deviation of travel time increases linearly with the average flow ($\bar{x}_i$) in the route. Although these linear relationships help simplify the presentation of this example, the proposed algorithm is applicable for any increasing function. The following equations represent these relationships for route i in this sample problem:

$$t(s_i) = a_i + b_i . s_i \qquad , \ 1 < i < 4 \tag{5.47}$$

$$s_i(\bar{x}_i) = c_i . \bar{x}_i \qquad , \ 1 < i < 4 \tag{5.48}$$

where the values of the parameters of these relations for this numerical example are provided in Table 5.1. Travel demand is also assumed to have a discrete distribution of the risk sensitivity parameter (γ) as presented in Table 5.2.

Table 5.1. Parameters of the route travel time functions

Routes, i	Parameters		
	a_i	b_i	c_i
1	*1*	*2*	*1*
2	*2*	*1*	*1*
3	*4*	*3*	*1*
4	*3*	*4*	*1*

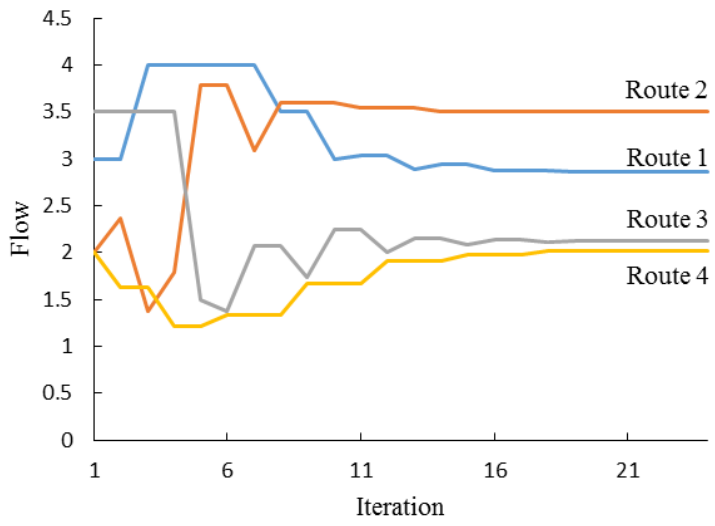
Table 5.2. Distribution of demand

Group, p	Demand, x_p	γ_p
1	*3*	*7*
2	*2*	*5*
3	*2*	*3*
4	*3.5*	*1*

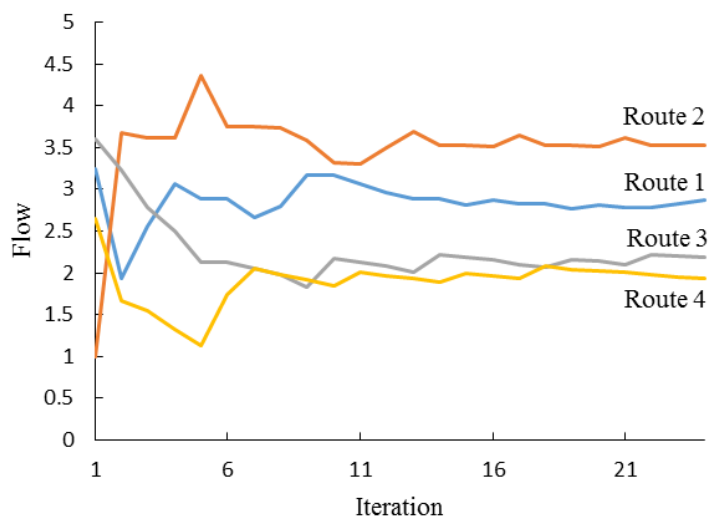
The proposed solution algorithm can solve the sample problem very fast. To initialize the problem, it would be more efficient to use the solution of the equivalent user equilibrium problem as the initial condition of the algorithm as explained in the Section 5.4.1. Nonetheless, in this example, we skip this optional step to show the efficiency of the

method in solving such problems. So, here we start with the initial condition of demand groups 1, 2, 3, and 4 each using the routes 1, 2, 4, and 3, respectively. The initial solution can be wisely chosen based on the general properties of the equilibrium solution instead of a random solution. Then, the proposed algorithm modifies the assignment iteratively using the insights directly derived from general properties of the EFRC, and with no need to find the cheapest paths for each sensitivity group. As a result, the proposed method can approach to the convergence measure of 0.003% and estimate the equilibrium flows with precision of 0.01 in just 24 iterations, while the comparable solution with the same level of precision requires at least 2636 iterations of an iterative assignment method with variable smoothing factor (See Fisk, 1980; Patriksson, 1994). Figure 5.8a, b compares variations of the route flows in the first 24 iterations of these algorithms to indicate the efficiency of the proposed method in solving this traffic assignment problem.

The route choice equilibrium for the numerical example is illustrated by the EFRC in Figure 5.9. The increasing straight dashed lines represent the performance of the routes in the sample network. The equilibrium expected travel times and the standard deviations are also provided in Figure 5.9. The monotone decreasing and convex EFRC is depicted by the continuous line segments which have slopes as labeled in the legend. The equilibrium route flows and the associated generalized cost for different groups are presented in Tables 5.3 and 5.4, respectively.



(a)



(b)

Figure 5.8. Variations of route flows in the first 24 iterations of (a) the proposed solution algorithm (b) iterative method with variable smoothing factor

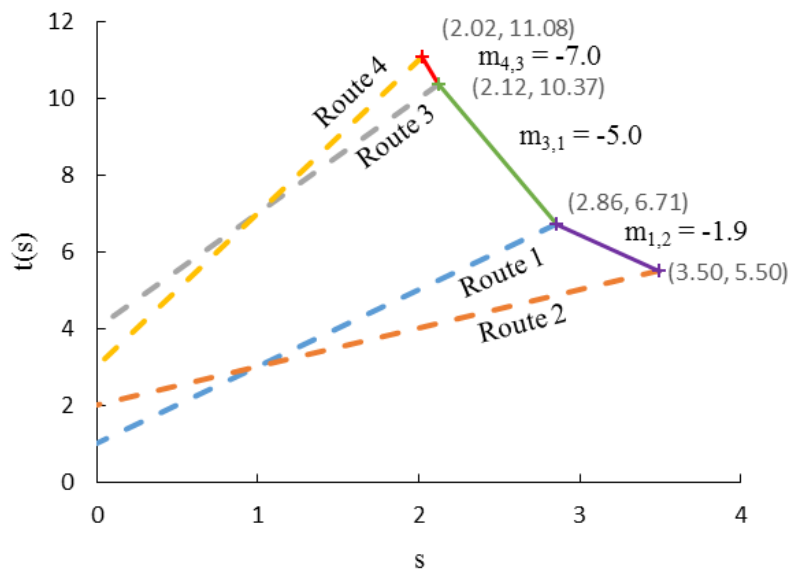


Figure 5.9 . The EFRC of the sample problem

Table 5.3. Equilibrium route flows, x_i^p, of different groups in the sample network

Group, p	Route, i				Total, x_p
	1	2	3	4	
1			0.98	2.02	3.00
2	0.86		1.14		2.00
3	2.00				2.00
4		3.50			3.50
Total, x_i	2.86	3.50	2.12	2.02	10.50

Table 5.4. Equilibrium route disutilities for different groups, DU_i^p, in the sample network

Group, p	Route, i				Minimum,
	1	2	3	4	λ_p
1	26.7	30.0	25.2	25.2	25.2
2	21.0	23.0	21.0	21.2	21.0
3	15.3	16.0	16.7	17.1	15.3
4	9.6	9.0	12.5	13.1	9.0

In this example, route 4 is the least risky route in the equilibrium condition. Routes 4 and 3 both offers the same lowest travel cost to the users in group 1, who are the most conservative demand group with risk sensitivity of $\gamma_1 = 7$. Consequently, the demand of group 1 splits between routes 4 and 3, and the slope of this segment of the EFRC is $m_{4,3} = -7$. Indeed, this slope coincides with the result of Corollary 5.5. Similarly, routes 3 and 1 have the lowest generalized cost for the users in the group 2 with $\gamma_2 = 5$, so the demand of this group is split between routes 3 and 1 resulting in slope $m_{3,1} = -5$ for this segment of the EFRC. However, more risk seeking users of groups 3 and 4 choose routes 1 and 2 with lower expected travel times, in spite of higher associated variability. Thus, the slope of EFRC between routes 1 and 2, meets the lower bound of -3 and the upper bound of -1, in accordance with constraint (5.18). It is worth pointing out that in the equilibrium condition, the demand groups fill the routes in a decreasing order of their risk sensitivity as explained in Corollary 5.3, while no subset of routes shares more than one group, as claimed in Corollary 5.6.

5.6 Extensions

5.6.1 Network with Multiple Origin-Destination Pairs

The solution method proposed in this research is designed to solve the route choice problem for a single origin-destination network. Nonetheless, it can be adapted to solve the multiple origin-destination network as well. Nagurney (2013) presents two general equilibration algorithms which employ single origin-destination algorithms to solve multiple origin-destination problems. The equilibration algorithms are developed based on the relaxation method by reformulating the multiple origin-destination network into the series of single origin-destination pairs. Accordingly, these equilibration algorithms can be modified to use the proposed solution method to solve its single origin-destination sub-problems. In this framework, the equilibration algorithms iterate over the origin-destination pairs one by one to achieve the equilibrium. The first equilibration algorithm proceeds to the next origin-destination pair by accomplishing a single iteration of the proposed method, and so on, until the equilibration algorithm converges. In each iteration of this equilibration algorithm, it considers one of the single origin-destination pairs as a sub-problem and executes just a single iteration of the proposed solution method for this sub-problem. Although the equilibrium may not be achieved yet for this origin-destination pair, it will pick the next pairs over and over again until the convergence criteria are met for the origin-destination pairs. In this case, the EFRCs of the origin-destination pairs will be convex and

decreasingly monotone, while the geometric constraints are also met. In contrast, the second equilibration algorithm continues the iterations on the same origin-destination pair until the sub-problem equilibrium is achieved before proceeding to the next origin-destination pair. After the convergence criterion of the sub-problem is met, it will pick the next origin-destination pair to reassign its demand, and this procedure repeats for all the origin-destination pairs, until the equilibrium is achieved for all of them. In spite of the capability of both the equilibration algorithms in solving the multiple origin-destination problem, the first algorithm may solve the multiple origin-destination problem faster using the proposed method, since it recognizes the dependency of the solution of the origin-destination pairs to each other.

5.6.2 Mode Choice Problem

In addition to route choice, users may also be able to travel by another mode of transport. Travelers may consider the same aspects of the travel costs in comparing modes. In this respect, using a transit system is associated with an expected travel time and some variation of waiting time or in-vehicle travel time. Users with heterogeneous preferences regarding travel time risk assess the available modes and routes together in their decision making process. As a result, the share of demand that select alternative modes can be determined by including the mode choice in the proposed traffic assignment model. Figure 5.10 provides an illustration of a transit mode that operates independently of the traffic network

and is represented by a fixed point in the in the *s*-*t* plane. Alternatively, the proposed model and solution method remain applicable for any given relationship representing the transit mode.

In the Figure 5.10, the transit mode is assumed to have a fixed expected travel time while there is risk associated with the waiting time. Therefore, the transit mode will remain unused if it can be dominated by alternative routes. Otherwise, the efficient frontier of the mode and route choices will include the transit mode, which determines its share from the travel demand in the equilibrium condition.

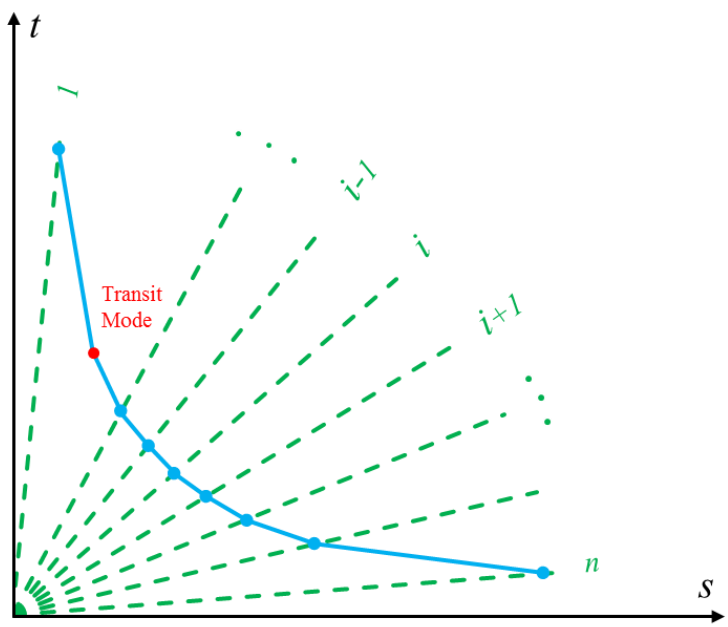


Figure 5.10. The efficient frontier of the mode and route choices

5.6.3 General Efficient Frontier of Choices

The concept of the efficient frontier can be extended to include multiple factors in the disutility function in modeling the general decision making process for heterogeneous users. Considering the relationships between the properties of each known choice, the general efficient frontier, as the equilibrium solution of the choice problem, can be defined as the set of dominant choices in a multidimensional coordinate system. The general efficient frontier is expected to have the equivalent properties which can be used in modeling and solving the multi-objective optimization problems. However, multidimensionality of the problem naturally complicates the problem as well.

5.7 Summary

In a transportation network, users seek the route with the lowest generalized cost for their travel. In this respect, the route disutility can be presumed to be a linear combination of the expected travel time and its standard deviation, which can be estimated by users based on their previous experiences. The importance of the travel time variability for users can be described by how much they value travel time reliability relative to expected travel time. To include the effect of heterogeneity of user preferences, the risk sensitivity can be considered to have a probability distribution over the traveling population.

In this part of the research, our objective is to employ the concept of the efficient frontier as a tool for solving a bi-objective route choice problem with heterogeneous user

preferences. The analogy between the decision making procedures in a free market and a transportation network makes it possible to borrow this concept from portfolio theory in finance and adapt it for modeling the bi-objective traffic assignment problem in transportation. In this framework, the equilibrium solution of the traffic assignment problem is presented by the concept of the efficient frontier of route choice (EFRC).

The EFRC also demonstrates the route choice rank of the users with different sensitivities to risk, which can be critical for predicting the effect of any prospective changes in network on route choice behavior of the travelers. The specific properties of the EFRC can provide intuitions regarding the characteristics of the equilibrium solution of the traffic assignment problem. For one thing, it is shown that the EFRC is always monotonically decreasing and convex. For another, there is a relationship between the distribution of the risk sensitivity and shape of the EFRC. As a result, these properties are used to propose a mathematical formulation of the route choice problem. On this basis, a two-stage solution algorithm is also designed to solve the traffic assignment problem under travel time variability. The proposed solution algorithm modifies the traffic assignment of the heterogeneous users using the insights from the general characteristics of the EFRC with no need to incremental modification of the traffic assignment. Thus, with a large number of sensitivity groups or even a continuous distribution of sensitivity to travel time reliability among users does not complicate the assignment modification procedure of the algorithm. As a result, the proposed algorithm can efficiently solve the route choice problem under travel time variability with heterogeneous demand.

The result of a numerical example shows that the proposed solution algorithm can solve the problem more efficiently than conventional methods. Although the numerical example employs the solution method to solve a single origin-destination problem, it can be adapted to solve a multiple origin-destination network as well. Furthermore, the efficient frontier can also be used to model the joint modal split-route choice problem. The general concept of the efficient frontier is an appropriate tool to model the general decision making procedure when there is heterogeneity associated with the relative importance of the parameters in the generalized cost function, such as the case when travelers face travel time variability in addition to expected travel time when making route choice decisions.

CHAPTER 6

CONCLUSIONS AND FUTURE EXTENSIONS

6.1 Conclusions

There are variety of different decisions that a traveler needs to make for each of his/her trips in the network, e.g. choosing starting time, destination, mode, and route of the trips. In this respect, rational users seek choices that minimize the generalized cost of their trips. Research shows that transportation network users consider different factors in their decision-making procedure, while the relative importance of these factors may vary among the heterogeneous travelers with different trip purposes. Under the realistic assumption that the components of the cost associated with choices are increasing functions of the demand for those choices, the cumulative result of the individual decisions of the rational users eventually leads to the user equilibrium condition in which no one can reduce his/her cost by changing his/her decision. In this research, we adapted the concept of the efficient frontier from portfolio theory (Markowitz, 1952) in finance to represent the equilibrium solution of the bicriterion choice problems. The efficient frontier is shown to always be a non-increasing convex hull with a specific geometric property, determined by the probability distribution of the preferences. Then, we employ the identified properties of the efficient frontier to model the equilibrium condition of different choice problems in transportation with heterogeneous user preferences.

One of the important decisions that travelers need to make in the network is what time to start their trips to keep their generalized costs minimized. This problem is first introduced in Vickrey's congestion theory, which addresses the commute problem of a single bottleneck with a time-dependent demand and fixed capacity. Insufficient capacity of the bottleneck to meet the time-dependent results in the formation of a queue, which causes users to experience a combination of delay and schedule deviation in their commutes. Rational users tend to minimize their own combination of cost by adjusting their arrival times to the bottleneck. The cumulative result of the individual decisions of the commuters would be the user equilibrium condition in the bottleneck in which no one can reduce his/her cost by switching his/her arrival time to the bottleneck. In this part of the research, we use the concept of the efficient frontier to propose an extension to the user equilibrium model by accounting for the heterogeneity in the schedule penalty preferences of the commuters. For that purpose, we make a use of properties of the efficient frontier to propose an analytical model for the equilibrium arrival of the heterogeneous commuters to the bottleneck, given the PDFs of their schedule penalty factors. On this basis, we propose a dynamic pricing pattern that can optimize the system by avoiding the formation of the queue in the bottleneck. In addition, we provide a demonstration on extracting the independent probability distributions of the schedule penalty factors from a given joint distribution. We also demonstrate how the proposed model can be inversely used to approximate the schedule penalty preferences of the heterogeneous commuters from empirical data derived by observing the arrival time of the users to the bottleneck.

Although the proposed analytical solution is derived for a single bottleneck, the results still can be extended for analyzing other transportation systems with limited capacity and time-dependent demand. For one thing, we briefly explained that the proposed analytical solution of the morning commute problem can be combined with the macroscopic network model (Daganzo, 2007; Gonzales and Daganzo, 2012; Geroliminis and Daganzo, 2008) in order to account for the heterogeneity of preferences of the users in modeling and optimizing the network on an aggregate level. For another, we showed that a demand responsive transit (DRT) service can be modeled as a queueing system, and then used the proposed analytical solution of the morning commute problem to account for the heterogeneity in preferences of the users in optimizing the operation of the DRT system.

DRT systems are a class of transit services in which a fleet of vehicles dynamically changes routes and schedules in order to accommodate demand within a service area. A DRT system naturally has flexibility in providing service, which allows it to adapt to variations in the demand. This property of DRT makes it possible to eliminate the access distance for transit users by providing a curb-to-curb trip. The inherent trade-off between the operating cost and the quality of service of a DRT system necessitates optimizing the operations to balance them. In this research, an analytical model based on Daganzo (1978) is employed to approximate fleet size, VHT, and VMT of the DRT system. Accordingly, the operating cost for the agency is estimated as a linear combination of these components. The users are also subjected to costs of using the service. When the operating capacity of the system is inadequate to cover the demand, users of the system incur costs of delay,

earliness, and lateness. In this respect, we adapt Vickrey's (1969) congestion theory to model the DRT system, and approximate the delay, earliness, and lateness of the users in the equilibrium condition. In addition, the total time that users spend in service can be approximated using the analytical model from Daganzo (1978). As a result, the efficiency of the DRT system can be optimized by minimizing the total cost for the agency and users, where the operating capacity of the system or the number of waiting requests or both can be considered as the decision variable(s) of the problem. This part of the research presents optimizations for three scenarios: allowing only the operating capacity to change, allowing both to change, or holding the fleet size fixed. In each scenario, the general problem with an S-shaped wished curve is formulated mathematically. The analytical solution is presented for the simplified case with an inverse Z-shaped wished curve. Two demand management strategies are also presented to spread the requests uniformly over the peak period in order to maintain an optimal number of waiting requests: (i) schedule management strategy, and (ii) dynamic pricing strategy. The proposed analytical solution of the morning commute problem is also employed to generalize the optimization method and dynamic pricing strategy of the DRT system by accounting for the heterogeneity in the schedule penalty preferences of the users.

Route choice is another important decision that users need to make in the network in order to minimize the generalized cost of their trips. Conventional traffic assignment models mostly simplify the generalized cost function to the travel time of the routes in the network. However, the underlying assumption of homogenous travelers who can predict

the exact travel times in the network has been also discussed in the literature. Research shows that travelers are just able to estimate the average travel times and its associated variations according to their previous experiences in the network. On this basis, a linear combination of the expected and standard deviation of travel time is used to approximate the travel cost of the users in a network under travel time variability. To account for the heterogeneity in preferences of the travelers, we considered a probability distribution for sensitivity of the user to risk in travel times. On this basis, we employed the concept of the efficient frontier to model the route choice behavior of the heterogeneous users in a network under travel time variability. The efficient frontier has specific characteristics that can be used in formulating the traffic assignment problem as set of complementary constraints. Under assumption that routes are labeled in an increasing order of the standard deviation of their travel times in the equilibrium condition, the formulation of the problem can be significantly simplified. Accordingly, we proposed a two-stage mathematical model for this problem in which the upper stage model reorders the routes according to their travel time standard deviation, while the lower stage model assigns the heterogeneous demand to the routes using insights attained through studying the properties of the efficient frontier. On the basis of the two-stage formulation of the problem, we also designed a solution algorithm to assign the traffic to the network iteratively. The proposed model can be also generalized to be used as a joint mode-route choice model.

6.2 Future Extensions

The objective of this research is to account for the heterogeneity in preferences of the users in modeling multi-criterion choice problems in transportation. For that purpose, we adapted the concept of the efficient frontier from portfolio theory (Markowitz, 1952) in finance to represent the cumulative results of the individual decisions in the user equilibrium condition. The efficient frontier is shown to have specific properties that can be employed to model different choice problem in transportation. In this respect, morning commute problem is one of the multi-criterion choice problems in which heterogeneous commuters tend to minimize the cost of their trips according to their own preferences by adjusting their arrival times to the bottleneck. In this research, we accounted for the heterogeneity in the schedule preferences and also schedule penalty preferences of the commuters in formulation of an analytical model for the morning commute problem. On this basis, we proposed a dynamic pricing strategy that can optimize the system by avoiding formation of the queue in the bottleneck. In formulating the optimal pricing strategy for the bottleneck, we simplified the problem by overlooking the heterogeneity in value of time (VOT) of the commuters. However, VOT can be considered as another aspect of the heterogeneity in preferences of the users that should be taken into account to improve the effectiveness of dynamic pricing strategy in optimizing the system. In this respect, the proposed analytical model for the morning commute problem can be extended to account for the heterogeneity in the VOT of the commuters. To include such heterogeneity in VOT

of users, VOT can be considered to have a probability distribution over the population if the commuters as illustrated in Figure 2.1. Although such heterogeneity in VOT of the users has no influence on the arrival of the commuters to the bottleneck in the user equilibrium condition, it still can affect the effectiveness of the proposed dynamic pricing strategy formulated under assumption of an identical VOT for the commuters. On this basis, considering the heterogeneity in VOT of the commuters can enhance the effectiveness of the dynamic pricing strategy in optimizing the system. In this respect, the concept of the efficient frontier can be used to generalize the optimal pricing strategy of the bottleneck by accounting for the heterogeneity in VOT of commuters in formulation of the problem. Subsequently, the dynamic pricing strategy of the DRT system can be updated as well.

REFERENCES

Abdel-Aty, M.A., Kitamura, R., Jovanis, P.P., 1995. Investigating effect of travel time variability on route choice using repeated-measurement stated preference data. Transportation Research Record, 1493, 39–45.

Abdel-Aty, M., Abdalla, M. F., 2004. Modeling drivers' diversion from normal routes under ATIS using generalized estimating equations and binomial probit link function. *Transportation*, 31(3), 327-348.

Amirgholy M., Gonzales E. J., 2015a. Efficient frontier of route choice under travel time variability. *Economics of Transportation* (under revision).

Amirgholy M., Gonzales E. J., 2015b. Demand responsive transit systems with time-dependent demand: user equilibrium, system optimum, and management strategy. *Transportation Research Part B*, DOI: 10.1016/j.trb.2015.11.006 (in press).

Amirgholy M., Gonzales E. J., 2015c. Analytical equilibrium of bicriterion choices with heterogeneous user preferences: application to the morning commute problem. *Transportation Research Part B* (under revision).

Amirgholy, M., Rezaeestakhruie, H., Poorzahedy, H., 2015. Multi-objective cordon price design to control long run adverse traffic effects in large urban areas. *NETNOMICS: Economic Research and Electronic Networking*, 1-52.

Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record 1197, 56–67.

Arnott, R., De Palma, A., Lindsey, R., 1990. Economics of a bottleneck. Journal of Urban Economics 27 (5.1), 111–130.

Arnott, R., De Palma, A., Lindsey, R., 1992. Route choice with heterogeneous drivers and group-specific congestion costs. Regional Science and Urban Economics. 22(5.1), 71–102.

Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28 (2), 139–161.

Barnes, I.C., Deakin, E., Frick, K.T., Skabardonis, A., 2012. Impact of peak and off peak tolls on traffic in the San Francisco–Oakland BayBridge corridor. 91st annual meeting of the transportation research board, 22–26 January, Number 12-4412, Washington, DC.

Beckmann, M., McGuire, C.B., Winsten, C.B., 1956. Studies in the economics of transportation, Yale University Press, New Haven, Connecticut.

Bekhor, S., Ben-Akiva, M. E., Ramming, M. S., 2006. Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research*, 144(5.1), 235-247.

Ben-Akiva, M., Bergman, M. J., Daly, A. J., Ramaswamy, R., 1984. Modeling inter-urban route choice behaviour. In *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, VNU Press, Utrecht (pp. 299-330).

Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science* (pp. 5-33). Springer US.

Ben-Akiva, M., De Palma, A., Kanaroglou, P., 1986. Dynamic model of peak period traffic congestion with elastic arrival rates. Transportation Science 20, 164–181.

Ben-Akiva, M. E., Gao, S., Wei, Z., Wen, Y., 2012. A dynamic traffic assignment model for highly congested urban networks. *Transportation research part C: emerging technologies*, *24*, 62-82.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. *Journal of Political Economy*, 18(3):637–654.

Boucher, T. O., 2014. "Engineering Economics," in Development and Economic Sciences, *Encyclopedia of Life Support Systems, UNESCO, Eolss Publishers, Oxford, UK.*

Braid, R., 1996. Peak-load pricing of a transportation route with an unpriced substitute. Journal of Urban Economics 40 (2), 179–197.

Chen, A., Oh, J., Park, D., Recher, W., 2010. Solving the Bicriteria Traffic Equilibrium Problem with Variable Demand and Nonlinear Path Costs. Applied Mathematics and Computation, 217(3.4):3020–3031.

Chen, B., Lam, W.H.K., Sumalee, A., Shao, H., 2011. An efficient solution algorithm for solving multi-class reliability-based traffic assignment problem. Mathematical and Computer Modelling, 54(5–6):1428–1439.

Chen, A., Zhou, Z., 2010. The α-reliable mean-excess traffic equilibrium model with stochastic travel times. Transportation Research Part B, 44(4.4):493–513.

Cohen, Y., 1987. Commuter welfare under peak-period congestion tolls: Who gains and who loses? International Journal of Transport Economics, 14, pp.238-66.

Dafermos, S.C., 1972. The traffic assignment problem for multiclass-user transportation networks. *Transportation science*, *6*(1), pp.73-87.

Dafermos, S.C., 1973. Toll patterns for multiclass-user transportation networks. *Transportation science*, *7*(3), pp.211-223.

Dafermos, S., 1980. Traffic equilibrium and variational inequalities. *Transportation science*, *14*(1), pp.42-54.

Dafermos, S., 1982. The general multimodal network equilibrium problem with elastic demand. *Networks*, *12*(1), pp.57-72.

Dafermos, S., 1983. A multicriteria route-mode choice traffic equilibrium model. *Bulletin of the Greek Mathematical Society*, Vol. 24.

Dafermos, S. and Nagurney, A., 1984a. On some traffic equilibrium theory paradoxes. *Transportation Research Part B: Methodological*, *18*(2), pp.101-110.

Dafermos, S., Nagurney, A., 1984b. Sensitivity analysis for the asymmetric network equilibrium problem. *Mathematical programming*, *28*(2), 174-184.

Daganzo, C.F., 1985. The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. Transportation Science 19 (5.1), 29–37.

Daganzo, C. F., 1984. The distance traveled to visit *N* points with a maximum of *C* stops per vehicle: An analytic model and an application. *Transportation Science* 18(4.4):331–350.

Daganzo, C. F., 1978. An Approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Research* 12(4.5):325–333.

Daganzo, C. F., V. V. Gayah, and E. J. Gonzales, 2012. The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO Journal on Transportation and Logistics* 1(1–2):47–65.

Daganzo, C.F., Sheffi, Y., 1977. On stochastic models of traffic assignment. Transportation Science, 11, 253–274.

Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transportation Research Part B 41 (5.1), 49–62.

Danielis, R., Marcucci, E., 2002. Bottleneck road congestion pricing with a competing railroad service. Transportation Research Part E 38 (4.5), 379–388.

De Palma, A., Ben-Akiva, M., Lefevre, C., Litinas, N., 1983. Stochastic equilibrium model of peak period traffic congestion. Transportation Science 17 (4.4), 430–453.

De Palma, A., Picard, N., 2006. Route Choice Behavior with Risk Averse Users. Spatial Evolution and Modelling, 139–178.

Deakin, E.A., Frick, K.T., Cervero, R., Skabardonis, A., Barnes, I., Kingsley, K., Rubin, J., Murakami, J., Amaro, J., Jensen, E., 2011. Bay bridge toll evaluation: final report. Technical Report, Global Metropolitan Studies, University of California, Berkeley.

Dessouky, M., F. Ordóñez, and L. Quadrifoglio, 2005. Productivity and Cost-Effectiveness of Demand Responsive Transit Systems. *California PATH Research Report*, UCB-ITS-PRR-29.

Dial, R., 1996. Bicriterion Traffic Assignment: Basic Theory and Elementary Algorithms. Transportation Science, 30(2), 93–111

Diana, M., M. M. Dessouky, and N. Xia, 2006. A model for the fleet sizing of demand responsive transportation services with time windows. *Transportation Research Part B* 40(8):651–666.

Ding, J., Gao, S., 2013. An optimal adaptive routing algorithm for large-scale stochastic time-dependent networks. In *Transportation Research Board 92nd Annual Meeting* (No. 13-4273).

Ding, J., Gao, S., Jenelius, E., Rahmani, M., Huang, H., Ma, Long, Ma, Pereira, F., Ben-Akiva, M.,2013. Routing policy choice set generation in stochastic time-dependent networks: Case studies for Stockholm and Singapore.

Figliozzi, M. A., 2009. Planning approximations to the average length of vehicle routing problems with time window constraints. *Transportation Research Part B* 43(4.4):438–447.

Figliozzi, M. A., 2008. Planning approximations to the average length of vehicle routing problems with varying customer demands and routing constraints. *Transportation Research Record: Journal of the Transportation Research Board* 2089:1–8.

Fisk, C., 1980. Some developments in equilibrium traffic assignment. Transportation Research Part B, 14, 243–255.

Fosgerau, M., 2010. On the relation between mean and variance of delay in dynamic queues with random capacity and demand. Journal of Economic Dynamics & Control, 34,598–603.

Fosgerau, M, Engelsson, L., 2011. The value of travel time variance. Transportation Research Part B, 45, 1–8.

Fosgerau, M., Karlström, A., 2010. The value of reliability. Transportation Research Part B, 44, 38–49.

Frank, M., Wolfe, P., 1956. An Algorithm for quadratic programming. Naval Research Logistics Quarterly, 3, 95–110.

Fu, L., 2003. Analytical model for paratransit capacity and quality-of-service analysis. *Transportation Research Record* 1841: 81–89.

Fu, L., 1999. Improving paratransit scheduling by accounting for dynamic and stochastic variations in travel time. *Transportation Research Record* 1666:74–81.

Fu, L., and S. Teply, 1999. On-Line and off-line routing and scheduling of dial-a-ride paratransit vehicles. *Computer-Aided Civil and Infrastructure Engineering* 14(4.5):309–319.

Fujii, S., Kitamura, R., 2000. Anticipated travel time, information acquisition and actual experience: the case of hanshin expressway route closure. Paper presented at the 79th Annual Meeting of the Transportation Research Board, Washington, DC.

Gao, S., Frejinger, E., Ben-Akiva, M., 2010. Adaptive route choices in risky traffic networks: A prospect theory approach. *Transportation research part C: emerging technologies*, *18*(5), 727-740.

Gao, S., 2012. Modeling strategic route choice and real-time information impacts in stochastic and time-dependent networks. *Intelligent Transportation Systems, IEEE Transactions on*, *13*(3), 1298-1311.

Gao, S., Huang, H., 2012. Real-time traveler information for optimal adaptive routing in stochastic time-dependent networks. *Transportation Research Part C: Emerging Technologies*, *21*(1), 196-213.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B 42 (9), 759–770.

Geroliminis, N., Levinson, D., 2009. Transportation and traffic theory. Springer Science, Chapter: Cordon pricing consistent with the physics of overcrowding, pp. 219–240.

Golledge, R. G., Garling, T., 2002. Spatial behavior in transportation modeling and planning. In K. Goulias (Ed.), *Transportation systems planning: Methods and applications* (pp. 3-1–3-21). New York: CRC Press.

Gonzales, E.J., Christofa, E., 2013. Bottleneck congestion with a constant and peak toll: San Francisco–Oakland Bay Bridge. EURO Journal on Transportation and Logistics, 3(3–4), 267–288.

Gonzales, E.J., Daganzo, C.F. 2012. Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing. Transportation Research Part B, 46(10), 1519–1534.

Hato, E., Taniguchi, M., Sugie, Y., Kuwahara, M., Morita, H., 1999. Incorporating an information acquisition process into a route choice model with multiple information sources. *Transportation Research Part C: Emerging Technologies*, 7(2), 109-129.

Hendrickson, C., Kocur, G., 1981. Schedule delay and departure time decisions in a deterministic model. Transportation Science 15 (5.1), 62–77.

Henderson, J.V., 1974. Road congestion: A reconsideration of pricing theory. Journal of Urban Economics, 1, pp. 346-55.

Henderson, J.V., 1977. Economic theory and the cities. New York, Academic Press, chapter 8.

Henderson, J.V., 1981. The economics of staggered work hours. Journal of Urban Economics, 9, pp.349-64.

Hensher, D. A., 1994. Stated preference analysis of travel choices: the state of practice. *Transportation*, 21(2), 107-133.

Herman, R., Lam, T., 1974. Trip characteristics of journeys to and from work. In Buckley, D. J. (Ed.). Transportation and Traffic Theory, Proceedings of the Sixth International Symposium on Transportation and Traffic Theory. (Elsevier: New York).

Huang, H., 2000. Fares and tolls in a competitive system with transit and highway: the case with two groups of commuters. Transportation Research Part E, 36 (4.4), 267–284.

Iida, Y., Akiyama, T., Uchida, T., 1992. Experimental analysis of dynamic route choice behavior. Transportation Research B, 26, 17–32.

Jia, A., Zhou, X., Li, M, Rouphail, N.M., Williams, B.M., 2011. Incorporating stochastic road capacity into day-to-day traffic simulation and traveler learning framework. Transportation Research Record 2254 (5.1), 112–121.

Kraus, M., 2003. A new look at the two-mode problem. Journal of Urban Economics 54 (3), 511–530.

Levinson, D., Zhu, S., 2013. A portfolio theory of route choice. Transportation Research Part C, 35, 232–243.

Little, J. D. C. 1961. A proof for the queueing formula: $L = \lambda W$. *Operations Research* 9(3):383–387.

Lindsey, R., 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. Transportation Science 38(3), 293–314.

Liu, Y., Nie, Y., 2015. A semi-analytical approach for solving the bottleneck model with general user Heterogeneity. Transportation Research Part B, 71, 56-70.

Lou, Y., Yin, Y., Lawphongpanich, S., 2010. Robust congestion pricing under boundedly rational user equilibrium. Transportation Research Part B 44 (5.1), 15–28.

Lo, H.K., Luo, X.W., Siu, B.W.Y., 2006. Degradable transport network: Travel time budget of travellers with heterogeneous risk aversion. Transportation Research Part B, 40(9):792–806.

Lo, H. K., Tung, Y. K., 2003. Network with degradable links: capacity analysis and design. *Transportation Research Part B: Methodological*, 37(4.4), 345-363.

Lu, X., Gao, S., Ben-Elia, E., 2011. Information impacts on route choice and learning behavior in a congested network: experimental approach. *Transportation Research Record: Journal of the Transportation Research Board*, (2243), 89-98.

Lu, X., Gao, S., Ben-Elia, E., Pothering, R. (2014). Travelers' day-to-day route choice behavior with real-time information in a congested risky network. *Mathematical Population Studies*, *21*(4), 205-219

Mahmassani, H., Chang, G., 1987. On boundedly rational user equilibrium in transportation systems. Transportation Science 21 (2), 89–99.

Mahmassani, H.S., Hou, T., Saberi, M., 2013. Connecting network-wide travel time reliability and the network fundamental diagram of traffic flow. 92nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Markowitz, H., 1952. Portfolio selection. The Journal of Finance 7 (5.1), 77–91.

Merton, R. C., 1972. Analytic derivation of the efficient portfolio frontier. The Journal of Financial and Quantitative Analysis, 7(4.4):1851–1872.

Mirchandani, P., Soroush, H., 1987. Generalized traffic equilibrium with probabilistic travel times and perceptions. Transportation Science 21 (3), 133–152.

Meyers, S.C., 1977. Determinants of corporate borrowing. *Journal of Financial Economics*, 5:147–175

Nagurney, A., 2000. Congested urban transportation networks and emission paradoxes. *Transportation Research Part D: Transport and Environment*, *5*(2), 145-151.

Nagurney, A., 1999. *Network Economics: A Variational Inequality Approach*. Vol. 10. Springer Science & Business Media.

Nagurney, A., Zhang, D., 1997. Projected dynamical systems in the formulation, stability analysis, and computation of fixed-demand traffic network equilibria. *Transportation Science*, *31*(2), 147-158.

Nagurney, A., Dong, J., 2002. A multiclass, multicriteria traffic network equilibrium model with elastic demand. *Transportation Research Part B: Methodological*, *36*(5), 445-469.

Nagurney, A., Dong, J., Zhang, D., 2002. A supply chain network equilibrium model. *Transportation Research Part E: Logistics and Transportation Review*,*38*(5), 281-303.

Nagurney, A., Zhang, D., 2012. *Projected dynamical systems and variational inequalities with applications* (Vol. 2). Springer Science & Business Media.

Newell, G.F., 1987.The morning commute for nonidentical travelers. Transportation Science 21(2), 74–88.

Nie, Y., 2011. Multi-class percentile user equilibrium with flow-dependent stochasticity. Transportation Research Part B, 45(10):1641–1659.

Noland, R.B., 1999. Information in a two-route network with recurrent and non-recurrent congestion. In R. Emmerink and P. Nijkamp (Ed.). Behavioural and Network Impacts of Driver Information Systems. Aldershot: Ashgate Publishing.

Noland, R.B., Small, K.A., Koskenoja, P., CHU, X., 1998. Simulating travel reliability. Regional Science and Urban Economics, 28, 535–564.

Ott, M., Slavin, H., Ward, D., 1980. Behavioral impacts of flexible working hours. Transportation Research Record, 767, pp. 1-6.

Pothering, R., Gao, S., 2013. Calibration of user equilibrium model with heterogeneous risk attitudes. In *Transportation Research Board 92nd Annual Meeting* (No. 13-3641).

Patriksson, P., 1994. The traffic assignment problem: Models and methods. Utrecht, The Netherlands. ISBN: 9067641812.

Polak, J.W., Hazelton, M.L., 1998. The influence of alternative traveler learning mechanisms on the dynamics of transport systems. Proceedings 26th European Transport Forum, PTRC, London.

Polak, J.W., Oladeinde, F., 2000. An empirical model of travellers' day-to-day learning in the presence of uncertain travel times. In M.G.H. Bell and C. Cassir (Ed.). Reliability in Transport Networks. Hertfordshire: Research Studies Press.

Qian, Z., Zhang, H.M., 2011. The morning commute problem with heterogeneous travelers: the case of continuously distributed parameters. Transportmetrica 9(2), 1–26

Quadrifoglio, L., and Ch. Shen, 2010. Performance analysis of the "zoning" strategies for ADA paratransit services. *Research Report*, SWUTC/10/169114-1.

Rahimi, M., M. Amirgholy, E. J. Gonzales, 2014. Continuum approximation modeling of ADA paratransit operations in New Jersey. Paper Number 14-4864. *Transportation Research Board 93rd Annual Meeting*, 12–16 January, Washington, D.C.

Ramadurai, G., Ukkusuri. S.V., Zhao. J., Pang, J-S, 2010. Linear complementarity formulation for single bottleneck model with heterogeneous commuters. Transportation Research Part B 44(2), 193–214

Razo, M., Gao, S., 2013. A rank-dependent expected utility model for strategic route choice with stated preference data. *Transportation Research Part C: Emerging Technologies*, 27, 117-130.

Richardson, A.J., Taylor, M.A.P., 1978. Travel Time Variability on Commuter Journeys. High-Speed Ground Transportation, 6, 77–99.

Shahabi, M., Unnikrishnan, A., Boyles, S.D., 2013. An outer approximation algorithm for the robust shortest path problem. Transportation Research Part E, 58, 52–66.

Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3)425–442.

Shen, Ch., and L. Quadrifoglio, 2012. Evaluation of zoning design with transfers for paratransit services. *Transportation Research Record: Journal of the Transportation Research Board* 2277:82–89.

Shen, Ch., and L. Quadrifoglio, 2011. The coordinated decentralized paratransit system: formulation and comparison with alternative strategies. Paper Number 11-2825. *Transportation Research Board 90th Annual Meeting*, 23–27 January, Washington, D.C.

Small, K., 1982. The scheduling of consumer activities: work Trips. American Economic Review 72 (3), 467–479.

Smith, M.J., 1984. The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. Transportation Science 18(4.4), 385–394

Sheffi, Y., 1985. Urban transportation networks: equilibrium analysis with mathematical programming methods. PRENTICE-HALL, INC., Englewood Cliffs, New Jersey.

Tabuchi, T., 1993. Bottleneck congestion and modal split. Journal of Urban Economics 34 (3), 414–431.

Tan, Z., Yang, H., Guo, R., 2014. Pareto efficiency of reliability-based traffic equilibria and risk-taking behavior of travelers. Transportation Research Part B, 66:16–31.

Thomas, T. C., Thompson, G. I., 1971. Value of time saved by trip purpose with discussion and closure. *Highway Research Record*, (369).

Tian, H., Gao, S., Fisher, D. L., Post, B., 2012. A mixed-logit latent-class model of strategic route choice behavior with real-time information. In *Transportation Research Board 91st Annual Meeting* (No. 12-2867).

Tian, H., Gao, S., 2013. A process model for route choice in risky traffic networks. *Procedia-Social and Behavioral Sciences*, *80*, 764-778.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science, New Series, Vol. 185, No. 4157, 1124-1131.

Uchida, T., Iida, Y. A. S. U. N. O. R. I., Nakahara, M., 1994. Panel survey on drivers' route choice behavior under travel time information. In *Vehicle Navigation and Information Systems Conference*, 1994. Proceedings, pp. 383-388. IEEE.

van den Berg, V. A. C., Verhoef, E. T, 2011. Congestion Tolling in the Bottleneck Model with Heterogeneous Values of Time. Transportation Research Part B 45 (5.1): 60–70.

Vickrey, W., 1969. Congestion Theory and Transport Investment. American Economic Review 56, 251–260.

Wang, G., Jia, N., Ma, S., Qi, H., 2014. A rank-dependent bicriterion equilibrium model for stochastic transportation environment. European Journal of Operational Research, 235(3)511–529.

Wang, J.Y.T., Ehrgott, M., Chen, A., 2014. A bi-objective user equilibrium model of travel time reliability in a road network. Transportation Research Part B, 66:4–15.

Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers, 1(2), 325–378.

Watling, D., 2006. User equilibrium traffic network assignment with stochastic travel times and late arrival penalty. European Journal of Operational Research, 175(3), 1539–1556.

Weston, J.F., 1973. Investment decisions using the capital asset pricing model. *Financial Management*, 2(5.1):25–33.

Wu, X., Nie, Y., 2011. Modeling heterogeneous risk-taking behavior in route choice: a stochastic dominance approach. Transportation Research Part A, 45:896–915.

Xiao, F., Qian, Z., Zhang, H.M., 2011. The morning commute problem with coarse toll and nonidentical commuters. Network and Spatial Econmics 11(2), 343–369.

Xiao, L., Lo, H.K., 2013. Adaptive vehicle routing for risk-averse travelers. Transportation Research Part C, 36:460–479.

Zhang, K., Mahmassani, H.S., Lu, C., 2013. Dynamic pricing, heterogeneous users and perception error: Probit-based bicriterion dynamic stochastic user equilibrium assignment. Transportation Research Part C, 27:189–204.