University of California Transportation Center
UCTC Dissertation UCTC-DISS-2011-04

Allocation of Space and the Costs of Multimodal Transport in Cities

Eric Justin Gonzales
University of California, Berkeley
2011

Allocation of Space and the Costs of Multimodal Transport in Cities

by

Eric Justin Gonzales

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy
in

Engineering – Civil and Environmental Engineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Carlos F. Daganzo, Chair
Professor Robert B. Cervero
Professor Adib Kanafani

Spring 2011

Allocation of Space and the Costs of Multimodal Transport in Cities

by

Eric Justin Gonzales

Abstract

Allocation of Space and the Costs of Multimodal Transport in Cities

by

Eric Justin Gonzales

Doctor of Philosophy in Engineering – Civil and Environmental Engineering

University of California, Berkeley

Professor Carlos F. Daganzo, Chair

Cities worldwide face growing demand for mobility with limited transportation infrastructure. This dissertation addresses how street space should be allocated and how transport modes should be operated for different city structures. City structure is characterized by the density of travel demand and the amount of space available for transportation. Several costs are associated with transportation systems, including time, money, space, and externalities. Building on macroscopic models of traffic and transit operations in urban networks, the relationship between the costs of providing mobility with various transport modes and the structure of the city served is modeled recognizing that vehicles require space. Cities served by an individual mode (e.g., cars) and/or a collective mode (e.g., buses) are analyzed for three cases: constant demand over time (travelers can choose their mode); evening peak demand (travelers can choose their mode); morning peak demand (travelers can choose mode and departure time). In all cases, the system optimal use of space and modes which minimizes total system costs is identified along with a pricing strategy to achieve the optimum at user equilibrium. The results of this study show systematically how to allocate street space, operate transport systems, and price modes to minimize the costs of mobility for any city structure.

To my parents,
for all their love, support, and encouragement to follow my dreams.

Contents

List of Figures

List of Tables

Acknowledgments

My experience at Berkeley has exceeded my greatest expectations, and I treasure the time I have spent working, learning, and growing on this campus. I would not have been able to complete this Ph.D. without the nurturing support of many people. It has been an honor for me to work and develop friendships with all of them.

First and foremost, I want to thank my advisor, Carlos Daganzo. He has been an inspiring guide for me as I have developed my ideas, my research goals, and this dissertation. It has been a joy to work with him and learn from his example. His ability to see through to the essence of a problem and communicate his understanding with clarity and concision is a skill that I admire and aspire to develop in myself. His passion, intellect, and genuine caring make Carlos a truly remarkable advisor and friend, and I am grateful for these years we have worked together.

I also want to acknowledge the support and guidance of the other Berkeley faculty who have shaped my experience. In particular, I want to thank my committee members, Robert Cervero and Adib Kanafani, for their insights and support of my work. Through courses, seminars, and informal conversations, I am grateful for all that the Transportation Engineering faculty have done to help me develop as a teacher, a researcher, and a professional. In particular, I want to thank Michael Cassidy for his superior example as a teacher and communicator. I also want to thank Samer Madanat, Alexander Skabardonis, Joan Walker, and Mark Hansen for freely and enthusiastically sharing advice regarding research, career, and life goals.

I will always look back fondly on my experiences with my fellow students at Berkeley. I must start with my sincere gratitude to Anthony Patire and his wife Akane Matsuura for hosting me at the Transportation Engineering open house; they have been great friends since I have come to Berkeley. I am also thankful for the opportunity I had upon arriving at Berkeley to work with Nikolas Geroliminis. He has been a supportive mentor and a great friend.

The office in 416 McLaughlin Hall has been much more than a place of work. This has become a community of close friends for me. In particular I want to thank my office mates: Offer Grembek, Stella So, and Josh Pilachowski with whom I've spent many hours working, laughing, and discussing solutions to world's problems from our tower over Memorial Glade. I want to thank Josh in particular for his sense of humor and loyal friendship. I also want to thank my most recent office mates, Juan Argote, Dylan Saloner, and Yiguang Xuan, for their friendship and encouragement as a I have worked to finish this dissertation. In addition to my office mates, I am grateful for the community of colleagues and friends who have made my experience in Berkeley so enjoyable and memorable: Celeste Chavis, Vikash Gayah, Weihua Gu, Ilgin Guler, and Karthik Sivakumaran. Finally, I want to thank my friends and recent roommates, Adonis Garefalakis and Tasos Nikoleris, who have taught me about Greek culture and other ways of looking at the world.

I could not have gotten to where I am today without the love and support of my family. My parents have always encouraged me to follow my dreams and to dream big. I am thankful for all that they have done to help me make the most of the

opportunities available to me. My parents, brother, and extended family have always been there for me when I need them.

Finally, I want to thank my best friend, Eleni Christofa, for her love and companionship. She has celebrated my successes, and she has blown wind in my sail whenever I have needed a push. I cannot imagine this journey without her unwavering support.

This research was supported by the University of California Berkeley Center for Future Urban Transport, a Volvo Center of Excellence. I give additional thanks for financial support provided by a University of California Berkeley Graduate Research Fellowship, a University of California Transportation Center Dissertation Grant, and a Gordon F. Newell Memorial Fellowship. This research would not have been possible without the enthusiastic, friendly, and knowledgeable staff of ITS Berkeley's Harmer E. Davis Transportation Library.

Chapter 1

Introduction

1.1 Motivation

People all over the world need access to jobs, markets, education, healthcare, and social interaction in order to live and participate in the modern global economy. In order to get to these activities and opportunities, people use transportation to traverse space, but limitations of the transportation system translate to limitations of access. Cities exist as a solution to the access problem by bringing people and opportunities into close physical proximity in which transportation can yield greater access. Thus, transportation is of critical importance to the life and operation of cities worldwide because it is the accessibility that transportation affords that makes cities desirable and economically viable places to live and do business.

Accessibility depends both on the *city structure* and transport *mobility*. At a basic level, city structure describes how urban space is used which determines the origins and destinations of trips and the available space for the transportation systems to serve them. Mobility describes how quickly the transportation system allows people to travel across space. A city's structure can be described by the allocation of space to transportation and other uses—the amount of area provided as streets and the density of residents as well as the opportunities to which they seek access. Mobility is a consequence of the design of the transportation system and the modes which compose it. In order to operate, these modes incur many different types of costs such as time, money, and environmental externalities. As such, accessibility and its multidimensional costs are physically related to the operating characteristics of the transportation system and the structure of the city it serves.

Cities worldwide are growing rapidly as people continue to flock to urban areas seeking access to greater economic, educational, and social opportunities. This poses a challenge because the very accessibility that draws so many people together is also threatened by the ensuing congestion of the transportation system. This problem is exacerbated by the increasing rate of motorization, particularly in developing countries where a growing middle class means that more and more people can afford more comfortable, yet more costly, forms of transportation.

In order to plan effectively to maintain accessibility in the future, we need to be able to answer the following question: *What are the costs of providing accessibility for cities of different structures?* Since transport modes require physical space to serve trips, and the space available for transportation in cities is constrained, these costs must be modeled in a way that explicitly considers the spatial requirements of the transportation system. A model built on correct traffic physics which relates city structure and the costs of accessibility is need. Such a model will shed insights on how urban space should be allocated to different transportation modes and how these modes can be operated and priced to achieve efficient results.

The remaining sections of this chapter present a review of the related literature, identify the contributions of this research, and provide on overview of the structure of the following chapters of this dissertation.

1.2 Literature Review

The literature related to accessibility, the costs of transportation, and the connection between transportation and city structure is extensive. This section discusses the existing work in these areas as a foundation for the methodology and contributions of this dissertation.

First, ways of quantifying accessibility are explored, and the benefits of using an approach based on cumulative opportunities which can be accessed within a travel constraint are discussed. This is followed by a review of the various works to assess the costs of transportation systems, recognizing the different dimensions of those costs incurred. Empirical work examining how city structure is related to the performance of urban transportation systems shows that a systematic relationship exists. However, a theory of the urban physics underlying the connection between city structure and the costs of providing accessibility has not yet been developed.

A review of the work on traffic operations and the economics associated with pricing and allocating space to different transport modes is also presented. Although systems with a single transport mode have been studied at the level of individual roads and for urban networks, multimodal networks have not been studied systematically with realistic traffic physics.

1.2.1 Quantifying Accessibility

Accessibility is an important indicator of social welfare. A number of studies have shown that greater accessibility is associated with improved economic opportunity and social equity (Kain, 1968; Wachs & Kumagai, 1973; O'Regan & Quigley, 1998). The concept of accessibility is widely used in the fields of transportation and city planning and generally refers to a measure describing the ability of people to reach opportunities such as employment (Handy & Niemeier, 1997; Harris, 2001; Hanson, 2004). The two most common ways to measure accessibility are by weighting opportunities by an impedance function or by counting the cumulative number of opportunities

that can be reached within a time or cost constraint (Koenig, 1980). The meaning of accessibility, however, depends on how it is measured.

The first mathematical impedance model was proposed by Hansen (1959), expressing accessibility between analysis zones as the number of opportunities in the destination zone divided by an increasing function of travel time or distance called the impedance function. This qualitatively makes sense because the greater the distance between an origin and destination, the less accessible one is from the other. This model is the basis of the gravity model commonly used in traffic forecasting, but a major weakness is that these impedance functions require calibrated variables that lack physical meaning.

An alternative way to think about accessibility is in terms of time-space constraints as described by Hägerstrand (1970). By recognizing that a person's activities, and travel between them, can be expressed as a path in time and space, feasible activity paths are constrained by physical, social, and institutional limitations. Burns (1979) use the concept of a prism to illustrate how travel time and spatial separation determine accessibility. Although this aggregated approach has been criticized for giving equal weight to near and far opportunities (Pirie, 1979), the cumulative opportunity measure of accessibility depends only on the choice of a time constraint which has an easily understandable physical meaning. More recently, Hägerstrand's time-space framework has been applied to study accessibility with geographic information systems (Miller, 1991; Kwan *et al.*, 2003). However, these studies do not realistically account for the effects of congested traffic on travel time.

The cumulative opportunity measure of accessibility depends, of course, on the time constraint chosen. One candidate for the constraint is the travel time budget proposed by Zahavi & Talvitie (1980) which reported consistent travel time expenditures for one-way commute trips worldwide. This time budget is estimated to be about 30 minutes each way for trips where travel is at least at the speeds achieved by motorized modes (Zahavi & Ryan, 1980). Goodwin (1981) observes that if this universal travel time budget does indeed exist, travel time savings could not be measured as the benefit of transportation improvements because people will travel further as vehicle speeds increase. This supports the notion that accessibility is a benefit to be weighed against the costs of the transportation system. The existence of a travel time budget is disputed (Mokhtarian & Chen, 2004), and variation in travel times across city sizes (Gordon *et al.*, 1989) and over time (Tanner, 1981; Toole-Holt *et al.*, 2005) has been observed.

Efforts to model peoples' realized travel patterns is limited by the enormous data requirements of activity-based models (Axhausen, 1998; Doherty, 2003). Pendyala *et al.* (2002), however, shows that there is consistency in the time-space constraints for individuals even though their actual travel patterns may vary considerably within this accessible space. Therefore, even if one cannot predict the specific trips that will be made, by increasing mobility one can be assured that accessibility is increased.

1.2.2 Costs of Transportation

Providing accessibility via a transportation system involves costs associated with vehicles, infrastructure, and operations. Numerous studies have compared the costs of investing in different transport modes. A notable early study is Meyer *et al.* (1965) which compares the monetary costs of using different modes to serve commuter demand along a corridor in a monocentric city. The work is interesting but limited because it does not consider the effects of traffic congestion and is confined to a corridor analysis. Several studies have since compared the monetized cost of using motorized modes such as cars, buses, and commuter rail to serve commuter demand in a monocentric city (Mohring, 1972; Keeler & Small, 1975; Boyd *et al.*, 1978).

Considering only the monetary or economic costs of different modes is incomplete accounting. Without considering the full costs of transportation investments, Mishan (1967) observes that we will tend to over-endorse modes which have large negative external costs. Likewise, underrating the benefits of non-motorized modes such as walking and cycling leads to underinvestment in infrastructure for alternatives (Litman, 2003).

In recent years, there has been a greater push to quantify the external costs of transportation. The results show that there are significant monetary and non-monetary costs associated with motor-vehicle use in terms of infrastructure as well as environmental and social impacts (Murphy & Delucchi, 1998; Delucchi, 2007). The magnitude of environmental externalities may be large in some cases (Wadhwa & Wirasinghe, 2003), and the value of accounting for the full costs in designing transportation systems and deciding how to allocate road space to modes is now being recognized (Currie *et al.*, 2007). Chester & Horvath (2009) present work to quantify the full life-cycle environmental impacts of passenger transportation modes and represent an example of the type of environmental accounting that should be considered.

One dimension of cost that has received considerable attention in the literature is time. Just as individuals have a budget of money from which they can choose how much to spend on various goods and services, DeSerpa (1971) points out that activities require time as well, and everyone ultimately has a time budget of 24 hours per day to use for sleep, activities, and the travel in between. This idea has been extended to consider the amount of time an individual must work in order to afford the resources to pay for travel. As far back as the mid 19th century, Thoreau claimed in *Walden* (1854) that "the swiftest traveller is he who goes afoot" when the time that a person must work to purchase a train ticket is included in the calculation of speed.[1] This phenomenon has been called time pollution (Whitelegg, 1993) and describes the idea that the technologies designed to save time are sustained by resources which take time to acquire.

[1] In his essay *Energy and Equity*, Illich (1974) argued with rough calculations that the average American man spends a quarter of his waking hours driving, maintaining, and working to pay for his car. If this time is included in the calculation of speed, travel by car would be no faster than by bicycle.

Typically, in order to account for the costs and benefits of a project a generalized cost in monetary units is considered. This requires time expenditures to be converted to money by a value of time. Studies to estimate value of time span decades (Beesley, 1965; McKean *et al.*, 1995; Hensher, 2001), but this value can differ greatly from person to person, and using a single value can be problematic (Daganzo, 1997). Furthermore, value of time is not observable, so it can only be estimated with imperfect information (Sharp, 1967). Therefore, the selection of a value of time for computing generalized costs always carries a bias, so the value of time should be treated more like a policy variable than a physical parameter to be estimated. This supports the notion that costs in different dimensions should be accounted separately before determining politically how they should be compared and traded off against one another.

1.2.3 Connecting Transportation and City Structure

So far, we have looked at how accessibility is quantified and how transportation systems influence the cost and city space required to provide this accessibility. But the performance of the transportation system itself influences the structure of the city it serves. The connection between land use and transportation is the subject of extensive empirical research. Pushkarev & Zupan (1977) account for the costs of transportation systems, acknowledging that land use and city structure are important factors in determining mode costs. Studies of the cost of developing new rail transit systems in North America tend to claim that not enough people will use the systems to make them cost effective (Pickrell, 1985). This boils down to the inherent connection between city structure and transportation, because the density of travel demand in a city will affect the cost at which trips can be served.

To investigate the aggregate impact of land use on transportation, Kenworthy & Laube (1999, 2001) collect data characterizing cities in terms of population density and income, transportation infrastructure, and the performance of the transportation system. They identify empirical relationships which support the notion that cities are physical systems that behave with consistency. Laube *et al.* (1999) use empirical evidence to argue that urban mobility can be described systematically as a consequence of urban form and transportation infrastructure. More recently, Cameron *et al.* (2003) use dimensional analysis to look for physical relationships between the values describing city structure and transportation performance. The strength of this empirical work is that it is focused on physical dimensions that have universal meaning. However, the models are constructed to describe the data collected and are therefore missing a theory to describe the underlying physical connections in a way that demonstrates causal relationships.

Disagreements about how cities should be developed and served by transportation underscore the need for an understanding of how these factors relate to accessibility. While Newman & Kenworthy (1989) argue that denser cities operate more efficient transportation systems, others conclude that the market chooses to develop at low densities and that this is more efficiently served by private automobiles (Gordon &

Richardson, 1997; Levinson, 1998). Disputes about the most *desirable* form of city structure ultimately depend on the objective, which in the end must be determined politically and not academically. Harris (1967) notes that optimizing cities is problematic because there are multiple competing objectives and how these should be balanced against one another depends on policy goals which will differ from person to person and place to place around the world. Therefore, models of city structure and transportation should present each of the costs associated with providing accessibility separately. Then the choice of how to design transportation to serve cities can be made transparently based on the policy goals of decision makers. This will be the approach taken in this thesis.

1.2.4 Single Mode Systems

Systems of single modes are well understood at the level of a single road. Recent work has advanced our understanding of urban networks serving single modes, particularly our ability to model the dynamics of traffic congestion on networks serving cars.

Single Roads

There is an extensive body of literature on traffic operations and congestion on individual roads. Queues develop at bottlenecks when the arriving demand of vehicles exceeds the capacity, and the dynamics of this system can be modeled with kinematic wave theory (Lighthill & Whitham, 1955; Richards, 1956) and queuing theory (Newell, 1971). These methods describe the dynamic nature of car traffic on a road, and thus can be used to model the evolution of queues over time in response to exogenous demand.

Economic models have also been developed to model the user equilibrium and system optimum travel patterns accounting for the dynamics of congestion at a bottleneck. The morning commute problem for a single mode was introduced in Vickrey (1969) which considers a population of car commuters who must use a single route with a fixed capacity bottleneck to get to work at a desired time. When the demand exceeds the capacity of the bottleneck it is not possible for everyone to travel on time, so each commuter chooses when to travel in order to minimize the sum of his or her own cost of travel, delay, and penalty associated with schedule deviation. A unique user equilibrium exists even for a population with a general distribution of wished departure times (Hendrickson & Kocur, 1981; Smith, 1984; Daganzo, 1985).

The bottleneck model of the morning commute has been studied extensively for the case where all commuters are identical and wish to depart the bottleneck at a common time. For example, Arnott *et al.* (1990b) proposes an optimal time-dependent pricing scheme (or fine toll) to eliminate the equilibrium queuing delay. Related work investigates a system with elastic demand (Arnott *et al.*, 1993).

Urban Networks

Economic models have been applied to study the allocation of urban space to development and transportation. This body of work is based on Alonso (1964) which looks at the trade-off between land rents and the cost of transportation to access a central business district. Recognizing that transportation infrastructure requires space itself, Solow & Vickrey (1971) analyze traffic patterns in an idealized linear city to find an economic equilibrium allocation of space to balance land value and congestion costs. These models have been developed to identify the costs and externalities of transportation (Solow, 1972, 1973; Wheaton, 1998) and identify equilibrium and optimum urban land use patterns (Anas *et al.*, 1998; Anas & Xu, 1999; Rossi-Hansberg, 2004).

Urban economic studies on the allocation of space are important in recognizing that transportation competes with other land uses for urban space. However, economic literature overlooks the fact that while transportation infrastructure takes up space, so do the vehicles themselves. By assuming that the speed of traffic at a location on the network is a function only of the flows at that location, the spillover effects responsible for traffic congestion in cities are ignored. Lago (2003) shows that using these models with and without spillover effects can lead to very different conclusions.

Recent advances have been made in modeling urban networks that serve cars, accounting for realistic traffic physics. It has been shown both theoretically (Daganzo & Geroliminis, 2008) and empirically (Geroliminis & Daganzo, 2008) that there is a consistent relationship between the average vehicle flow on a network and the average vehicle density. This relationship is called the Macroscopic Fundamental Diagram (MFD). The MFD has a strong advantage over more disaggregate approaches to modeling traffic networks, because this relationship depends only on the characteristics of the network and is insensitive to the details of the origin-destination tables which are difficult if not impossible to attain. Although the current theory only considers one mode at the city scale, the MFD connects the physical road space to performance of the transportation system based on realistic physics of urban traffic and congestion. This serves as an important spring-board to build a connection between the allocation of space to transport modes in cities.

If the average length of trips on a network is not changing over time, then the MFD defines the rate at which trips can exit the network as a function of the vehicle accumulation in the network (Daganzo, 2007). The rate that cars exit the network is analogous to the discharge flow at a bottleneck, and since the network capacity depends on the number of vehicles in the system, a network can often be macroscopically modeled as a single bottleneck with state-dependent capacity. The network capacity is a function of the number of vehicles in the network and decreases as queues grow on the streets. The congestion resulting from this reduced capacity has been called hypercongestion (Small & Chu, 2003).

Although the morning commute problem has been studied for very simple networks of parallel routes between an origin and destination (Arnott *et al.*, 1990a), an important extension of the bottleneck model is the morning commute problem on urban networks where origins and destinations are distributed across space. Geroliminis

& Levinson (2009) employs a macroscopic method to analyze the user equilibrium and examine pricing strategies for the morning commute problem in a city with a single mode (cars).

The analysis of transportation systems at the network level has not been exclusively focused on cars. There is also a body of literature looking at how public transit networks should be structured and operated. A macroscopic approach to model transit system structure was adopted to consider how the design of a transit system affects the costs for users and agencies (Holroyd, 1967; Newell, 1979; Daganzo, 2010). Although strictly speaking Wirasinghe *et al.* (1977) considers the design of a system with trains and buses, these are part of a single transit system.

1.2.5 Multiple Modes

The literature on multimodal transportation systems is not as developed as for a single mode. It has long been recognized that some modes use road space more efficiently than others. Navarro *et al.* (1985), for example, compares the space per person required by non-motorized modes with automobiles and buses in an urban environment. There are more detailed studies that evaluate traffic operations and the economics of single roads serving multiple modes. However, the existing work at the network level is limited to urban economic studies, which use static traffic models that do not correctly account for the spatial requirements of transport modes.

Single Road

Studies since Sparks & May (1971) have evaluated the effectiveness of priority lanes for high occupancy vehicles (HOVs) to move people rather vehicles. More recent work has brought this analysis into the urban environment to compare alternatives for dedicating urban road space to HOVs or buses. There are varying degrees to which space can be shared or separated by mode based on the range of vehicle types allowed to use a lane (Black *et al.*, 1992). Currie *et al.* (2004) promote the idea of full cost accounting for deciding whether or not to dedicate road space to transit service. These detailed comparisons are site-specific and based on disaggregated microsimulation, so they cannot be scaled-up to look at the performance of the transportation system across entire neighborhoods or cities. An exception to this is Eichler & Daganzo (2006) which looks at how a lane can be dedicated intermittently to allocate space to buses only when dedicated space is needed to serve them. This work is based on traffic theory that can be applied generally, and it effectively shows how a non-integer number of lanes can be dedicated to a transport mode.

Economic models for single roads serving multiple modes, such as Mohring (1979), are still being extended (Arnott & Yan, 2000; Ahn, 2009). Like the static models described for a single mode, these do not consider the physical evolution of congestion over time and therefore collapse the dynamics out of the problem.

The morning commute problem, which is dynamic, has been studied for systems with cars and transit to identify equilibrium patterns and optimal pricing schemes

(Tabuchi, 1993; Braid, 1996; Huang, 2000; Danielis & Marcucci, 2002). This work has been limited in two main ways. First, commuters have been assumed to share an identical desired bottleneck departure time, and second, only unrealistically simple families of transit mode cost functions have been considered. Existing models, for example, do not recognize that transit operations reduce the remaining capacity for cars, and the frequency of real transit service is adapted to the number of transit riders. Furthermore, unlike the case of the single bottleneck with distributed demand, the literature does not provide a system optimum solution with two modes, and whether it can be achieved with pricing.

Urban Network

Detailed, disaggregate models of multi-modal networks (Ferrari, 1999; Li *et al.*, 2007) suffer from the same drawbacks as disaggregate models of single mode networks. The economics literature has long recognized that multiple modes can be used on city streets (Sharp, 1967), and this work is still being extended (Mogridge, 1997; Kitamura *et al.*, 1999; Ferrari, 2005). A weakness of the economic literature for multiple modes in networks is that the models are static, and therefore cannot recognize the inherently dynamic nature of traffic conditions on urban networks.

1.3 Research Contribution

The literature review shows that extensive work has been done to understand the costs associated with the transportation systems we use to provide accessibility in cities. Traffic operations and the economics of single mode systems are well understood at the level of a single road, and recent advances have been made in modeling urban networks with a single mode. Tools now exist to look at cities with single modes in a macroscopic way, recognizing the spatial requirements of vehicles and the dynamic nature of transportation on urban networks. However, the literature on multimodal systems is either limited to looking at individual roads or does not account for the dynamic nature of traffic congestion.

Cities are complex and chaotic systems. Much of the work on transportation systems in cities is site-specific, but there is value in understanding the basic underlying relationships that apply to all cities and all transport modes in general. The road space in real cities can be used by multiple modes. No work puts together all the pieces to describe the systematic physical relationships that connect city structure, transportation, and the resulting costs of providing access. This must be done with recognition that vehicles themselves require space. What is needed is a theory of urban physics which relates the costs of multimodal transportation systems to the structure of cities.

The contribution of this dissertation is in advancing the understanding of urban networks with multiple modes. This is done using macroscopic models with correct physics which allow us to recognize that vehicles require space. This macroscopic

approach can be used to to develop functions describing the various costs and spatial requirements of any mode operating in isolation.

The research method is to start by analyzing the physics of the simplest case before complicating the model by relaxing assumptions. Even an idealized model provides insights for how the costs of transport systems depend on the characteristics of a city. As a building block, it is shown how modes should operate together in an idealized city with constant demand that does not vary with time. Then, more realistic cities in which the demand is peaked in time are considered in order to incorporate the dynamics of transportation operations in an urban network.

The results of this research are normative models for how road space should be allocated to different modes, and how these modes should be priced in order to minimize the total system cost of providing accessibility in cities.

1.4 Dissertation Organization

This dissertation is organized as follows. Chapter 2 presents the methodology for relating the costs and spatial requirements of individual modes to city structure. Chapter 3 shows how total system costs are minimized in idealized cities with constant (time-independent) demand, focusing on two modes: cars and transit. Then, the analysis is extended to the more realistic case of cities with peaked (time-dependent) demand. Chapter 4 presents the evening commute in which the start time of trips is determined exogenously, and commuters can choose their mode. Chapter 5 presents the morning commute in which both the mode and travel time are chosen by commuters. First, the morning commute problem is presented for a single bottleneck that can serve two modes with demand that is distributed over time. Then, it is shown how the results for the single bottleneck apply to networks. Pricing strategies to obtain system optimum allocation of space and use of modes are discussed in each chapter. Finally, Chapter 6 includes a summary of contributions, conclusions, and some directions for future work.

Chapter 2

City and Transport Cost Models

Cities are complex and intricate systems which are impossible to model in perfect detail. The approach taken in this dissertation is to look at cities and their transportation systems at a macroscopic level. Rather than trying to capture the detailed patterns of land use and the fine geometry of the transportation network, we will look instead at aggregated, neighborhood-level, characteristics. The two primary characteristics of city structure which are the focus of study in this dissertation are the demand density and the available road space for transportation.

A city viewed through this macroscopic lens resembles a flattened plane in which the population, demand patterns, and transportation infrastructure appear the same everywhere. This aggregate approach provides a clean look at the character of different cities and transportation systems. From this perspective, we may treat a city in an idealized way as if it were made up of identical individuals (e.g., same value of time, trip length) distributed uniformly across space and served by a road network that is a dense uniform grid. As a starting point, we consider a city that has constant demand over time. This assumption is relaxed in the subsequent chapters.

This chapter presents the methodology for using a macroscopic approach to model transport costs. Section 2.1 describes the core elements of city structure that determine transport costs. Section 2.2 presents the macroscopic model used to relate road space to the performance of the transportation system. Then, Section 2.3 lays out the basic structure for modeling the costs associated with any transport mode. Individual and public transit mode cost functions are described in Section 2.4. Finally, Section 2.5 assembles the assumptions and models described in the preceding sections to present the methodology for combining costs which is used in the remaining chapters of this dissertation.

2.1 City Structure

City structure describes how space in cities is used. Urban space is allocated to many different types of uses. For example land can be developed with structures where people are housed or employed, or it can be kept open as parks and other green

space. The patterns of land use are important determinants of the costs associated with transportation systems. Two characteristics of city structure are particularly important in determining the costs of transport systems: demand density and available road space. The city structure also affects the character of the demand, such as the length of trips made.

2.1.1 Demand Density and Road Space

Urban development consists of housing for residents (origins) and the activities and opportunities they access (destinations). The demand for travel per unit area of city λ (trips/m^2·sec) is based on the overall population density D (ppl/m^2) and the average rate at which each person makes trips δ (trips/ppl·sec):

$$\lambda = D\delta. \tag{2.1}$$

At the macroscopic level, we are not concerned with the details of each origin destination pair. Instead, we will assume that each of the trips making up λ have similar characteristics and are distributed uniformly across space so that the city is translationally symmetric.

Land which is developed with buildings or left open as green space varies for different types of cities. The area in a city that remains available for roads (if needed) per area of city is described by R (white area in Figure 2.1). This road space for moving vehicles may take up a significant amount of the surface area in cities. On Manhattan's Upper West Side, over one third of the surface space is devoted to streets. The road space per person, however, is much less than in suburban Pleasant Hill, California. The demand for travel must be served within the available road space or with modes like subways which do not operate at the surface.

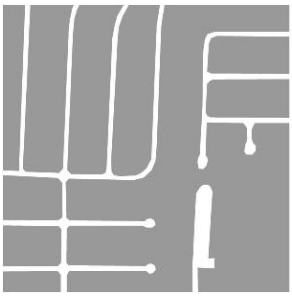
A significantly greater amount of space is often dedicated to parking in cities. With the exception of curb parking which could be eliminated to add an extra lane for moving traffic, most parking is provided off-street in garages or parking lots. This space is another characteristic of city structure, but it will not be a focus of analysis in this dissertation.

2.1.2 Accessibility and Trip Length

The purpose of the transportation system is to provide accessibility for the residents of a city. The accessible number of opportunities associated with a trip length d is affected by the city structure. Each trip is associated with a reachable area a which is proportional to d^2. The reachable area is important, because opportunities are scattered across the area of a city, so the accessibility associated with d is the total number of opportunities located within this area. The shape and size of the reachable area depends on the network geometry, and for a flat city with a dense grid, the area is a diamond (see Figure 2.2) where $a = 2d^2$.

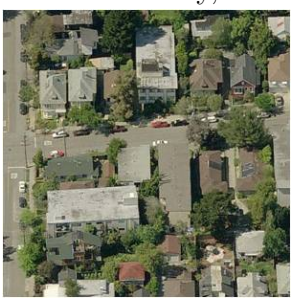
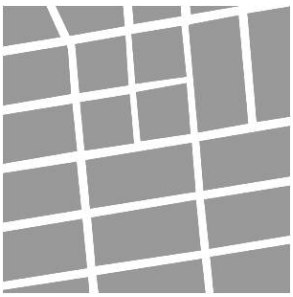
The average reachable area $\bar{a}$ depends on the distribution of the lengths of individual trips. If the trip lengths follow a single-parameter distribution, then $E(d^2)$ is

Pleasant Hill, California

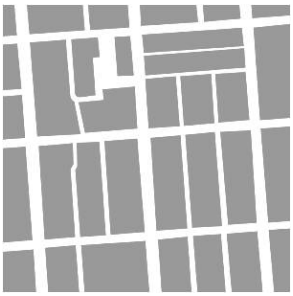
$D = 2{,}300 \ \frac{\text{ppl}}{\text{km}^2}$
$R = 0.11 \ \frac{\text{m}^2}{\text{m}^2}$

North Berkeley, California

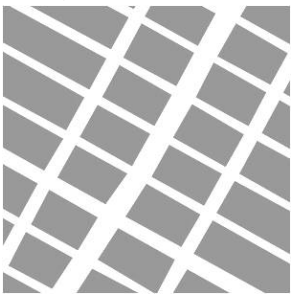
$D = 5{,}900 \ \frac{\text{ppl}}{\text{km}^2}$
$R = 0.20 \ \frac{\text{m}^2}{\text{m}^2}$

Mission District, San Francisco, California

$D = 16{,}400 \ \frac{\text{ppl}}{\text{km}^2}$
$R = 0.26 \ \frac{\text{m}^2}{\text{m}^2}$

Upper West Side, New York, New York

$D = 64{,}000 \ \frac{\text{ppl}}{\text{km}^2}$
$R = 0.35 \ \frac{\text{m}^2}{\text{m}^2}$

Figure 2.1. Examples of population density, D, and road space per unit area, R, in four neighborhoods (Source: Microsoft Virtual Earth, 2009)

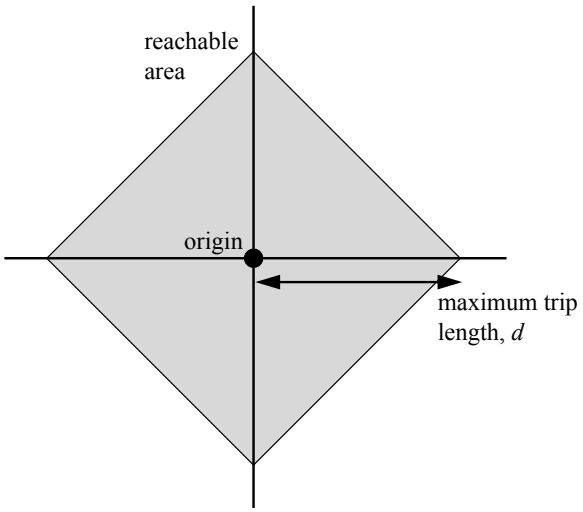


Figure 2.2. Diamond-shaped area reached by making a trip up to length d on a dense, orthogonal grid network

proportional to $E(d)^2$, where the average trip length is $E(d) = \bar{d}$. They are related by a factor of 2 for a negative exponential distribution or $4/3$ for a uniform distribution.

In order to translate the reachable area into an expression for accessibility, the distribution of opportunities over space can be expressed as a density. Cities can be characterized by densities of jobs, schools, hospitals, or any other opportunity of interest. Using population as a proxy for opportunities, the cumulative number of opportunities that can be reached within a trip length d is the product of $\bar{a}$ and D. The average accessibility per trip, A (reachable opportunities), is:

$$A = D\kappa\bar{d}^2 \tag{2.2}$$

where κ is a dimensionless constant including the constant that relates a to d^2 based on network geometry and a factor which relates $E(d^2)$ to $E(d)^2$ based on the distribution of trip lengths. The value of κ may also include a constant describing the number of opportunities per population (e.g., jobs per population). This physical definition shows how the cumulative opportunity measure of accessibility is related to the city structure and length of trips. From this point on, we will consider trip length as a parameter of the demand, because minimizing the cost of serving trips of length d is the same as minimizing the costs of providing accessibility A.

2.2 Traffic on Urban Networks

Now that we have described the road space available for surface transportation, R, as a property of city structure, let us look at how R relates to the performance of the street network. Just as a macroscopic approach can be used to look at city structure, traffic on urban streets is modeled macroscopically. This method allows us

to recognize that all vehicles that use the streets require space. First, the performance of a street network serving only cars is described. Then, a way to model networks serving multiple modes is presented.

2.2.1 Macroscopic Fundamental Diagram

Traffic on individual city streets is chaotic and unpredictable, and to model these streets microscopically requires huge data collection. Recent work suggests that there is a consistent relationship between the average network vehicle density and average network flow called a Macroscopic Fundamental Diagram (MFD) (Geroliminis & Daganzo, 2008). This relationship is a property of the network and does not depend on the demand pattern.

Figure 2.3 shows macroscopic traffic data from measurements in Yokohama and simulations of San Francisco and Nairobi. Symbols indicate different cities (squares are Yokohama, diamonds are San Francisco, and circles are Nairobi), and different shades represent different days. In the simulations of San Francisco and Nairobi, dramatically different origin-destination tables were used for the different simulation scenarios shown, yet the macroscopic relationships remain robust. Note that the macroscopic data for each of these three cities is plotted on axes which are normalized to control for network size. The MFD shows the average flow and density per lane on the network. Clearly, different network structures across different cities can create very different traffic outcomes.

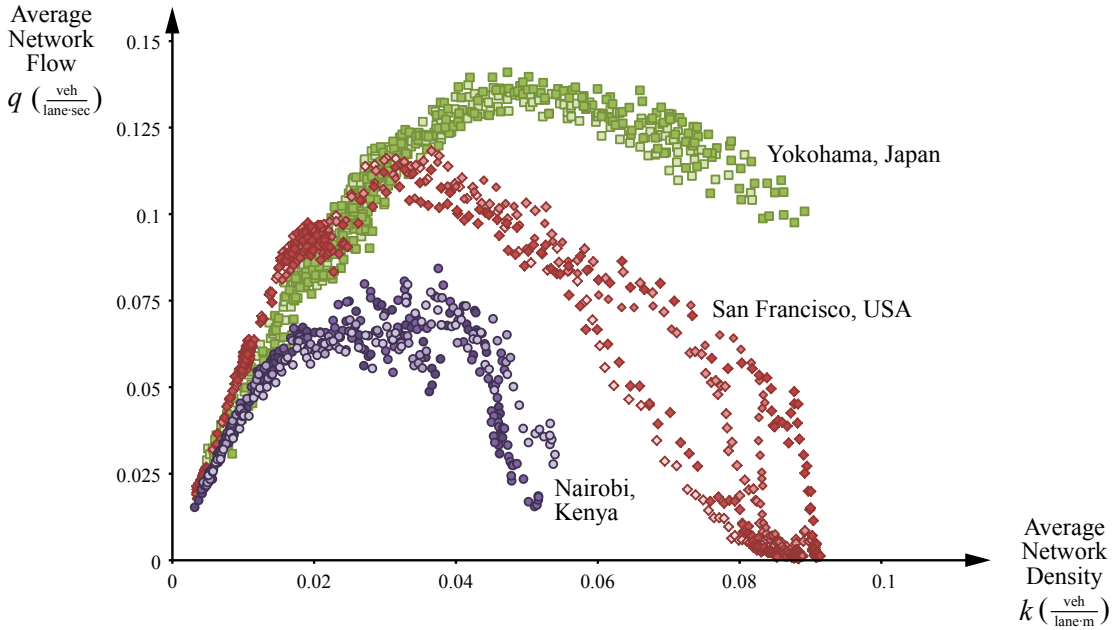


Figure 2.3. Experimental Macroscopic Fundamental Diagrams (MFDs) from measurement in Yokohama (Geroliminis & Daganzo, 2008) and simulations in San Francisco (Geroliminis & Daganzo, 2007) and Nairobi (Gonzales *et al.*, 2011)

If the street network is redundant, homogeneous, uniformly loaded, and minimally affected by turning vehicles, Daganzo & Geroliminis (2008) shows how the MFD can be predicted analytically using variational theory.[1] This analytical MFD is a concave function $Q(k)$ that provides an upper bound for the average network flow, q (veh/sec·lane), associated with any average network density, k (veh/lane·m). A generic concave MFD is illustrated in Figure 2.4.

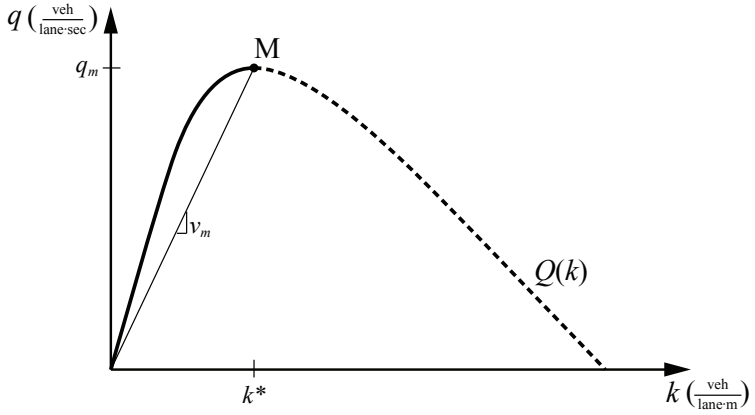


Figure 2.4. A generic concave Macroscopic Fundamental Diagram (MFD)

The maximum network flow q_m (at point M in Figure 2.4) consistently occurs at the same critical vehicle density k^* regardless of the demand pattern. This is verified by observation of Figure 2.3, especially for Yokohama and San Francisco which have consistent peaks. If the goal is to maximize mobility, then this is done by maximizing the network flow. Any flow below q_m is associated with two densities; one less than k^*, the other greater. Traffic states to the right of M (dashed line in Figure 2.4) are congested because the same flows could be served to the left of M with fewer cars on the road. Therefore, a well-managed network should never allow the density to exceed k^*.

The average network speed for a traffic state, including stops for traffic signals and queuing, is represented on the MFD as the slope from the origin to the traffic state. For example, the average speed at which the vehicle can traverse the network when the flow is maximized, v_m, is given by:

$$v_m = \frac{q_m}{k^*}. \tag{2.3}$$

Since congested traffic states serve the same traffic flow with greater vehicle density, they are also associated with slower speeds.

[1]The shape of the analytical MFD depends on the average block length, signal phasing and offsets, maximum discharge rate per lane, free flow speed of vehicles, and the jam density for a single lane.

2.2.2 Vehicle Footprints

Every vehicle requires road space to move efficiently through the city. This *footprint* of required space can be thought of as the area of road that must be rented for the duration of a trip in order for a vehicle to move on the network. Thus, the footprint has components of physical area and time.

For cars, the area component is described by the lane width, w, and the average spacing of vehicles, which is the inverse of the vehicle density. According to the MFD, this area is w/k^* at maximum network flow (point M in Figure 2.4). The time component is the duration of a trip, t_m, because the area required by a car is occupied for this time. The footprint, r, is given by:

$$r = \frac{wt_m}{k^* c} \tag{2.4}$$

where c is the number of trips per car, or the car's passenger occupancy. Note that the area required for a car to move is much greater than its physical dimensions.[2]

The travel time is determined by dividing the length of a trip, d, by v_m. If the length of a trip is taken into account, the critical vehicle density and speed can be removed from (2.4) by substituting (2.3), and the footprint is:

$$r = \frac{wd}{q_m c}. \tag{2.5}$$

Clearly, the footprint is minimized when the flow on the network is maximized. Every other feasible network flow is associated with either a low-density uncongested traffic state or a high-density congested state. Both require the same road space per vehicle because the smaller area associated with greater density is paired with longer trip durations due to slower speeds. Congestion is a suboptimal use of space, and it is doubly wasteful by imparting delays.

If we assume that the block length and signal phasing for a network are fixed, then the same MFD should apply if the size of the network is increased uniformly (e.g., by widening all roads). Based on the individual footprints expressed by (2.5), the minimum road space, R (as defined in Section 2.1.1), to serve demand λ with cars is:

$$R = \lambda r = \frac{\lambda wd}{q_m c}. \tag{2.6}$$

This expression could also be manipulated to identify the maximum λ that can be served with a given R.

The MFD approach applies to individual modes like cars, but public transit systems also require space. Since a transit system can be centrally controlled by an operating agency, its footprint depends on how the agency chooses to provide service. This relationship is described in greater detail in Section 2.4.3.

[2]For example, if San Francisco has lane widths of 3 m, then a car requires about 100 m^2 of road to move ($k^* \approx 0.03$ veh/lane·m from Figure 2.3). This is substantially greater than the 10 m^2 size that the car itself occupies.

2.2.3 Network Exit Function and Multimodal Networks

The MFD reveals the footprint of street space required by cars when traffic is in a steady state, but it can also be used to model the dynamics of traffic conditions in a network if demand is changing over time. When the average trip length is fixed, the MFD defines a consistent function relating the number of cars in the network to the discharge flow of cars exiting the network (Daganzo, 2007; Geroliminis & Daganzo, 2008). We call this second relationship the Network Exit Function (NEF). This relationship describes the state-dependent discharge rate from a network as a function of the number of vehicles in the network.

A generic concave MFD (as shown in Figure 2.4) describes the average network flow $q = Q(k)$ for all possible vehicle densities. Daganzo (2007) shows that the MFD can be used to derive the NEF which expresses the flow of cars exiting the network, f (veh/sec), as a function of the total number of vehicles circulating in the network, n (veh):

$$f = F(n) = \frac{l}{d} Q\left(\frac{n}{l}\right) \tag{2.7}$$

where l (lane·m) is the total length of the network. Note that the exit function, $F(n)$, is simply a rescaling of the MFD, $Q(k)$, to account for the size of the network and length of trips (see the bold curve in Figure 2.5). We will study this system assuming that the instantaneous exit flow depends only on the number of vehicles in the network at that time.[3]

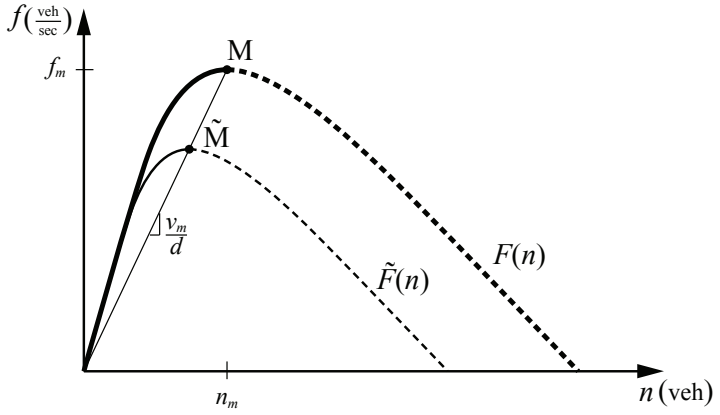


Figure 2.5. A generic concave Network Exit Function (NEF).

The maximum feasible exit flow is associated with point M in Figure 2.5. The duration of trips of length d, t_m, is the reciprocal of the slope from the origin to the traffic state on the NEF, v_m/d. This is analogous to the slope on the MFD which represents traffic speed.

[3]This assumption holds when traffic is in a steady state. Transitions between steady states are not instantaneous but have durations comparable to a trip time (Daganzo, 2007). The effect of these transitions is small if the traffic conditions change slowly.

In reality, transit services often share the same street space as other vehicles, so deploying buses will reduce the remaining capacity available for cars. If road space is dedicated to transit uniformly across the network, the effect should be the same as reducing the network size for cars to $\alpha < 1$ times the original network length. For fully dedicated transit lanes, α will be directly related to the length of lanes from which cars are banned, reducing the network length for cars from l to αl. For buses and trams operating in mixed traffic lanes, α must account for the losses due to conflicts between the different types of vehicles. The result is that the capacity of each individual street to serve cars is reduced to an average of α times its original.

Since the change in network size is uniform and none of the other determinants of network capacity have been altered, the MFD as described by $Q(k)$ should remain unchanged. However, the NEF for cars when transit is operated, $\tilde{F}(n)$, is scaled by α. So from (2.7),

$$\tilde{F}(n) = \frac{\alpha l}{d} Q\left(\frac{n}{\alpha l}\right) \tag{2.8}$$

which is shown in Figure 2.5. Note that the point M associated with the maximum exit flow moves along the ray with slope v_m/d towards the origin (to point $\tilde{M}$) so the travel time per trip associated with maximum exit flow does not change.

The NEF describes how the size of the street network relates to the rate that car trips can be served. In a multimodal network where modes are deployed efficiently and independently, the footprint associated with one mode is not affected by the operations of another. In this way, multimodal systems can be modeled by accounting for each mode in isolation as is described in Section 2.4. Their costs can then be combined as described in Section 2.5.

2.3 General Cost Model

There are many types of costs associated with transportation systems, and these depend on the structure of the city, the characteristics of the demand, and the properties of the transport modes used. Section 2.3.1 presents a general methodology for building cost functions based on the physical relationships between mode and trip characteristics. Section 2.3.2 discusses how these costs can be combined into generalized cost functions or held as constraints, depending on the policy objectives of a city.

2.3.1 Physical Components and Costs

Any trip from an origin to a destination can be broken down into a series of segments which depend on the mode used (see Figure 2.6). For example, a trip from home to work using a bus system requires an access segment to get from home to the bus stop, some time waiting for the vehicle to arrive, some time riding in the vehicle, and then an egress segment after alighting the bus to reach the destination. If the trip involves a transfer, there may be additional access, waiting, and in-vehicle segments.

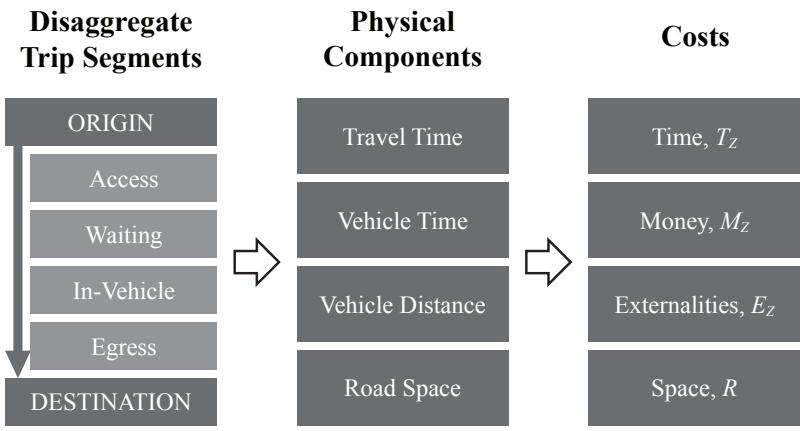


Figure 2.6. Structure of the cost model

Each of these segments is associated with unavoidable physical components which depend on the mode used. A trip by any mode requires the user to spend travel time, and a vehicle must be owned and stored for some time and operated over distance. The vehicle also occupies a footprint of road space on the network. Vehicles while not in-use require parking infrastructure to store them, and whether or not the vehicle is in use, there are resources and costs associated with manufacturing and owning the vehicle that can be amortized over time.

Each of these physical components are associated with costs which can be broadly categorized as time, money, externalities, and space. Each of these costs have different physical units, but they can all be related analytically to how modes are deployed to serve demand. Since combining these costs into generalized cost functions embeds policy decisions in the analysis, we start by building functions for each type of cost individually.

Time

Every mode of transportation requires time to traverse distance. The time required to make a trip is more than just the time moving in-vehicle. The door-to-door trip time using any mode can be broken into component parts: access time, waiting time (in the case of scheduled modes), and time moving toward the destination. The time cost, T_Z, is the sum of all the components of travel time.

Money

Transportation modes also require financial resources to operate. The money cost per trip includes cost of vehicles per time including vehicle purchase and insurance, c_t, cost of mode operation per distance including fuel and maintenance, c_d, and cost of required paved infrastructure, c_i. These unit costs are mode-specific and apply to

the various physical components to make up the monetary cost:

$$M_Z = c_t V_t + c_d V_d + c_i (R + R_p) \tag{2.9}$$

where V_t is the vehicle hours of ownership, V_d is the vehicle distance traveled, R is the road space required for moving trips, and R_p is the space required for parked vehicles. While the physical components are basic properties of the mode, the cost coefficients may vary across different parts of the world as the availability of resources and the costs of doing business vary.

Some of the monetary costs are paid directly by the user proportionally to how much they travel (e.g., gasoline to travel by car) while other costs are incurred by public institutions which receive funding from taxes paid by everyone (e.g., construction and maintenance of streets and sidewalks). It is important not to double count, so monetary costs such as bridge tolls or transit fares will not be counted as mode costs in our analysis because these are politically determined transfers between users and institutional providers which do not affect the total system cost of transportation. The costs incurred by users, however, will likely affect the modes people use to travel, so pricing can be used as a mechanism to achieve optimal use of the transportation system.

Externalities

Externalities are the broadest category of costs associated with transportation and include many environmental impacts but can be extended to consider such aspects as safety and human health outcomes. Travel time itself is not directly responsible for external costs, but vehicle operations and paved infrastructure are connected to the consumption of energy and emissions of pollutants, for example. The energy consumption is roughly proportional to the emission of greenhouse gases, particulate matter and other pollutants.

A simplistic model of how these relate to the transportation system is similar to the expression for monetary costs:

$$E_Z = e_t V_t + e_d V_d + e_i (R + R_p) \tag{2.10}$$

where e_t, e_d, and e_i are mode-specific coefficients describing how the external costs of interest relate to the vehicle use and infrastructure requirements. Although this research is not focused on conducting life-cycle analysis of transportation, policy decisions can be made taking into account environmental effects by counting these as yet another dimension of cost.

Space

As described in Section 2.2, moving vehicles need space on streets to operate. The required space for moving vehicles, R, is the aggregation of footprints associated with all trips in the city.

2.3.2 Generalized Cost Functions and Constraints

Each of the costs associated with the transportation system (right column of Figure 2.6) may be combined in generalized cost functions or held as constraints, depending on the policy objectives of a city. There are many ways to do this. In this dissertation, we will focus on the problem of minimizing the generalized system cost Z which includes time, money, and externalities, subject to constrained road space R. This type of problem may be of interest to a city which is already built, and where streets cannot be easily widened.

The generalized cost requires converting all costs into the same units. If Z is expressed in units of time (per area and time of analysis), then:

$$Z = T_Z + \frac{1}{\beta} M_Z + \frac{\gamma}{\beta} E_Z \tag{2.11}$$

where T_Z is the time cost, M_Z is the monetary cost, E_Z is the externality of interest (for concreteness we will focus on greenhouse gas emissions), and β and γ are politically determined parameters to relate different types of cost. Using generalized cost in units of time will be useful in later discussions of pricing. It is common in economics literature to relate time and money by a value of time β which may be expected to increase with wealth. Recently there have been efforts to put a price on carbon, γ, and although it is difficult to do this on a scientific basis, this is already being done as a policy.[4] Then, the costs are structured in a way that can be minimized subject to a maximum road space requirement, R.

The methods used to study this particular optimization problem can be applied to any other formulation of generalized costs and constraints. An alternative, for example, may be to minimize the costs subject to a maximum quantity of greenhouse gas emissions. Such a formulation will have a similar structure mathematically, and could be used to identify ways to meet greenhouse gas emission targets.

2.4 Comparing Modes in Isolation

Before modeling jointly deployed transport modes in cities, cost models for modes operating in isolation are constructed using the methodology proposed in Section 2.3. As discussed at the beginning of this chapter, the macroscopic approach adopted for modeling cities and transportation systems is consistent with the idea of an idealized city which is completely uniform and symmetric.

As a starting point, we look at the costs of transportation in cities which are time-independent (demand is constant over time). The cost models for each mode are built on the assumptions that demand and the transportation network are uniformly distributed over space, modes operate independently (vehicle footprints are independent

[4]The Chicago Climate Exchange trades carbon futures based on the cost of carbon off-set programs, and carbon taxes are beginning to be implemented. However, there is an enormous range of estimates for the marginal social cost of greenhouse gas emissions ranging from 5–155 \$/tCO_2-eq (IPCC, 2007).

of other mode operations), and efficiently (traffic flows are served without congestion and transit is optimized to minimize system costs). Under these assumptions, the minimum possible costs associated with each transport mode are identified. By formulating costs in this way, we are looking at the least possible costs for each mode to serve uniform demand.

Modes are classified as either individual (private) or collective (public transit), and the fundamental physical components are modeled according to the way vehicles perform. The associated costs are presented here on a per trip basis but could also be determined on a total system-wide basis if costs are multiplied by the total number of trips. Although the approach is generic and any mode can be modeled, the focus will be on comparing two specific cases: cars and a bus transit system.

2.4.1 Individual Modes: Cost Components

Individual modes are those for which people travel at the time of their own choosing and using their own vehicle. This includes modes such as walk, bicycle, or car. All modes require the passenger's time for travel as well as road space to provide mobility. Except for walking, a vehicle is also associated with travel and must be used for the duration of the trip. First, we identify the physical components related to the operating characteristics of these modes. Then, these components are associated with costs which make up the generalized cost function and road space requirement.

Travel Time

Each trip using an individual mode begins with an access segment from the origin to where the vehicle is parked followed by the travel time spent in-vehicle and finally ends with an egress segment from the parked vehicle to the destination. The access is typically by walking, but the formulation below is general to any access mode. The total travel time required per trip is the sum of access and in-vehicle time:

$$\bar{T} = t_a + \frac{d}{v_m} \tag{2.12}$$

where t_a is the time required to access the vehicle at the beginning and egress at the end of a trip.

Vehicles

With the exception of walking (in which case the vehicle is the human body itself) a vehicle must be purchased or rented. If each trip is associated with ψ vehicle hours (when in use and parked), then

$$\bar{V}_t = \psi \tag{2.13}$$

which has units of vehicle time per trip. This term will be used to prorate fixed costs associated with vehicle production and ownership across trips.

Additionally, vehicles are used to traverse distance, and the length of each trip d represents the vehicle distance traveled.

$$\bar{V}_d = d \tag{2.14}$$

This term relates to variable costs that are associated with the vehicle distance per trip.

Road Space

As described in Section 2.2.2, vehicles such as cars need an area several times the size of the vehicle in order to move at network capacity. Space is also required for vehicles to park when they are not moving, and for the access portion of the trip which is typically walking. The road space required during the trip is the sum of the footprint for the in-vehicle travel, given by (2.5), and the road space required for access:

$$\bar{R} = \frac{wd}{q_m c} + \frac{w_a v_a}{q_a c_a} t_a, \tag{2.15}$$

where the subscript a denotes values associated with the access mode: lane width, w_a, average speed, v_a, lane capacity, q_a, occupancy, c_a, and access time, t_a. The road space for access is expressed in terms of the time a mode is used.

The minimum space for parked vehicles can be determined from the number of vehicles per person and the amount of time these vehicles are not being used. A vehicle is in use for t_m each trip, but it must be parked during the remaining time even when it is not used. Prorating the space for parking to each trip, the minimum space for parking required per car trip if no space is wasted is:

$$\bar{R}_p = \left(\psi - \frac{d}{v_m}\right) r_p \tag{2.16}$$

where r_p is the area of a parking spot and the required space for the vehicle to maneuver in and out of it. This is a very optimistic case, because a flat city with time-independent demand has no days and nights. In a time-dependent city where people sleep at night, much more space for parking would be required to park all of the vehicles simultaneously.

2.4.2 Individual Modes: Generalized Cost

The generalized cost of a trip with an individual mode can be constructed as described by (2.11) and the components described in the Section 2.4.1. Note that for individual modes, the travel time, vehicle operation, and minimum space required depends only on the trip length d and are independent of demand λ. Therefore, the total generalized cost for all trips is proportional to demand.

For concreteness, we will focus on a system of cars. Then the total generalized cost for cars serving car demand λ_C is:

$$Z_C(\lambda_C) = \alpha_C \lambda_C \tag{2.17}$$

where α_C is a coefficient that includes all of the various parameters and policy variables described above. The explicit function which describes α_C in terms of travel time, vehicle operation, and footprint is described in detail in Appendix B.

The road space that these car trips require has a similar form. If we are concerned only with the spatial constraint for moving vehicles, then the road space required per car trip $r_C = \bar{R}$, as described by (2.15), is:

$$R_C(\lambda_C) = r_C \lambda_C \tag{2.18}$$

In this dissertation we will suppose that parking space is provided off-street so that parking does not compete for space with moving traffic. However, the space for parking is considered in terms of the cost of infrastructure.

2.4.3 Public Transit Modes: Cost Components

While individual modes allow people to travel at the time of their own choice, public transit modes are collective and require trips to be consolidated in time and space so that vehicles carry multiple passenger trips. There are many ways to structure transit services, and any of these structures can be systematically analyzed using the same methodology employed above.

Suppose that in the uniform city, we are given a demand for transit, λ_T, which will be served with a grid bus network. The fewest necessary design parameters for this system are the route/stop spacing, s, which determines spatial coverage, and the service headway, H, which determines temporal coverage. The system structure is illustrated in Figure 2.4.3, and it provides uniform service everywhere in the city similar to the grid structure proposed in Holroyd (1967).

Travel Time

The travel time for a trip by a transit mode involves three components: access time, waiting time, and in-vehicle time. Assuming that the demand is uniformly distributed across space and riders access the nearest station using the dense grid of streets, the average access time at the beginning and end of the trip combined is:

$$t_a = \frac{s}{v_a}. \tag{2.19}$$

If travelers must wait half a headway on average for each vehicle they board and allow an extra headway to be sure they arrive to their appointment at the destination on time, the total waiting time is:

$$t_w = 2H \tag{2.20}$$

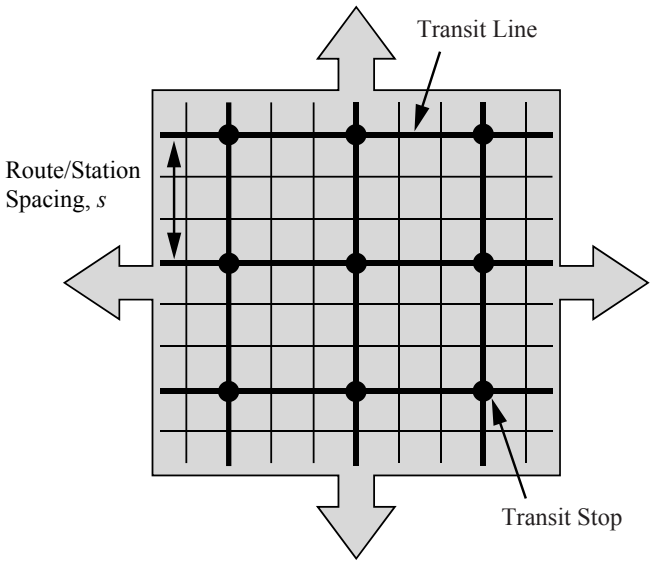


Figure 2.7. Structure of a simple 2-dimensional transit system

for a trip with one transfer. If the spacing $s \ll d$, and the possible destinations are uniformly distributed on the frontier of the reachable area, then nearly every trip will involve a transfer.

The in-vehicle travel time involves modeling the average speed v of the transit vehicles taking into account the time lost for making stops. The time in-vehicle, t_m is proportional to the pace of a transit vehicle $1/v$. This includes the time it takes to traverse distance at cruising speed v_m, the fixed loss time per stop, y, and the time required per boarding and alighting passenger, x, counted per unit distance:

$$t_m = \frac{d}{v} = d\left(\frac{1}{v_m} + \frac{y}{s} + \frac{x\lambda_T sH}{2}\right). \tag{2.21}$$

The number of boarding and alighting passengers per stop is a consequence of both s and H.[5] For transit services where boarding and alighting is efficient and the loss time per stop is relatively independent of the number of riders (e.g., systems with pre-paid fares such as bus rapid transit or metro), x becomes very small, and the last term becomes insignificant.

The total travel time is the sum of these three components, and substituting

[5]The number of passenger trips per time served by each stop is λs^2, and $\lambda s^2 H$ are served in each headway. These passengers are equally likely to travel in any of 4 directions, and each trip involves 2 vehicles because of the transfer. So, the number of passengers boarding and alighting a vehicle at each stop is $\frac{2}{4}\lambda s^2 H$.

(2.19), (2.20), and (2.21) into the expression, we get that

$$\begin{aligned} \bar{T} &= t_a + t_w + t_m && (2.22) \\ &= \frac{s}{v_a} + 2H + d\left(\frac{1}{v_m} + \frac{y}{s} + \frac{x\lambda_T sH}{2}\right). && (2.23) \end{aligned}$$

So, the travel time on a transit mode is a function of the design parameters, s and H, and the demand, λ_T.

Vehicles

The total vehicle hours of transit operations per area-time is simply the expression for the number of transit vehicles per area, which is $4/sHv$.[6] Dividing this by the transit demand and using the average transit speed v implied by (2.21), the vehicle hours per transit trip is:

$$\bar{V}_t = \frac{4}{sH\lambda_T}\left(\frac{1}{v_m} + \frac{y}{s} + \frac{x\lambda_T sH}{2}\right). \tag{2.24}$$

The vehicle distance traveled is simply the product of $\bar{V}_t$ and the average speed of a transit vehicle, v:

$$\bar{V}_d = \frac{4}{sH\lambda_T}. \tag{2.25}$$

Note that unlike individual modes, the vehicle operations associated with an individual public transit trip depend on the total transit demand. This is the consequence of sharing vehicles among multiple trips. Furthermore, since the mode operations are determined by s and H, the vehicle operations are independent of the length of individual trips.

Road Space

The footprint of a transit trip is most easily represented by considering the entire footprint of the transit system and dividing it by the transit demand. Each of the $4/sHv$ vehicles requires a lane of width w and a clear headway h in order to operate without disruption from other traffic, so the transit vehicle requires a length of hv clear lane. The footprint per transit passenger trip is:

$$\bar{R} = \frac{4wh}{sH\lambda_T} + \frac{w_a s}{q_a c_a} \tag{2.26}$$

where the second term is the space required by the access mode following (2.5). In this case the length of the access trip is $s/2$ at the beginning and end of the trip, and this total distance determines the access footprint. For systems where an entire lane

[6] There are $1/s$ routes per direction in 4 directions (north, south, east, and west). On each route, the density of transit vehicles per length of route is $1/Hv$.

is dedicated to transit, the headway for clear transit operations equals the service headway ($h = H$).

Transit vehicles also require space to be parked in their maintenance yard. The required parking space is proportional to the number of transit vehicles:

$$\bar{R}_p = \frac{4}{sH\lambda_T}\left(\frac{1}{v_m} + \frac{y}{s} + \frac{x\lambda_T sH}{2}\right) r_p \tag{2.27}$$

where r_p is the area required per vehicle. The parking requirements for transit are very small compared to the requirements for individual modes, but they are included here for completeness.

2.4.4 Public Transit Modes: Generalized Cost

Just as shown for individual modes, the generalized cost of public transit can be expressed by combining each of the components described in Section 2.4.3. Each of these components depends on the demand for transit, λ_T, and the design of the system as determined by s and H, so the generalized cost, as defined by (2.11), is a function, $Z(\lambda_T, s, H)$.

An efficiently run transit system should be operated to minimize the total system costs to serve the demand. Note that Z is a convex function of s and H so the system can be optimized by setting the first derivative equal to zero. It is straightforward to determine the optimal headway endogenously by this method, and this is a reasonable assumption to make because a transit agency can adjust the service headway relatively easily. Although there is not a simple analytical solution for the optimal route/stop spacing, the optimal s is insensitive to demand. So, the physical structure of the transit network is treated as fixed, and this is also reasonable, because it is costly and difficult for a transit agency to change the alignment of routes across the network after they are built. Further details on the optimization of the transit mode are presented in Appendix B.

The general form of the transit cost function for a grid network is given by:

$$Z_T(\lambda_T) = \alpha_0 + \alpha_1\lambda_T + \sqrt{\alpha_2\lambda_T + \alpha_3\lambda_T^2} \tag{2.28}$$

where each α incorporates the various physical components when the headway has been optimized. The coefficients capture the user cost of travel time and agency costs of capital, operations, infrastructure investments as they relate to the demand for transit service. These coefficients capture the cost components as follows:

- α_0 is the fixed cost of infrastructure investment which must be paid for any $\lambda_T > 0$ independent of vehicle operations (e.g., tunneling for metro). This is only significant when the total system infrastructure footprint is independent of the service headway.

- α_1 incorporates the user costs per trip which are independent of the service headway, primarily the access time and in-vehicle riding time of customers.

- α_2 captures the cost of transit operations required to provide service at the optimal headway.
- α_3 includes the additional cost of transit operations required to maintain the optimal headway as a result of boarding and alighting loss time per passenger.

Some of the α values become zero depending on the specific operating conditions. For example, a Bus Rapid Transit (BRT) system that shares streets with traffic and has a loss time per stop that is independent of the number of boarding and alighting passengers has only coefficients of α_1 and α_2 (see Appendix B).

The road space required for the transit system has a similar structure:

$$R_T = r_1\lambda_T + \sqrt{r_2\lambda_T + r_3\lambda_T^2} \tag{2.29}$$

where each r is a coefficient of road space and depends on the various physical components. Similarly, r_3 depends on x and therefore is not significant for modes like BRT. Detailed expressions for each of the cost and road space components are presented in Appendix B.

Other transit structures such as radial or hybrid networks can also be modeled in a similar way (Daganzo, 2010). The important feature of cost functions for public transit systems are that they typically exhibit economies of scale in that greater demand reduces the cost per trip. This is expected, because greater demand decreases the optimal headway so that all users enjoy more frequent service, and the transit infrastructure and vehicles are shared among more people.

2.4.5 Comparison of Modes

As presented in the previous sections, the generalized costs and spatial requirements of modes can be modeled based on their physical operating characteristics in a network. These models show how each mode operates in isolation.

For different types of trips, different modes may be more appropriate than others. A comparison of individual modes is shown in Figure 2.8 for different trip lengths and values of time. For short trips, the travel times for slow modes do not amount to much, so their lower cost makes them competitive. As trips get longer, faster modes like cars become more competitive even when they are more costly. Transit systems are a little difficult to compare in this way, because the costs per trip depend on demand.

The α values in (2.17) and (2.28) depend on characteristics of the trips as described in Appendix B and capture the user cost of travel time and agency costs of capital, operations, and infrastructure investments as they relate to the demand for transit service. We are interested in cases where transit has the potential to be cost-competitive with cars ($\alpha_C > \alpha_1 + \sqrt{\alpha_3}$).[7] Table 2.1 shows typical values of α

[7]This occurs whenever the generalized cost of access and in-vehicle riding time for a transit trip as well as the small contribution of an individual's boarding and alighting time to vehicle operating costs is less than the complete generalized cost of a car trip including vehicle depreciation and other out-of-pocket costs. This condition generally holds unless β or d are very large.

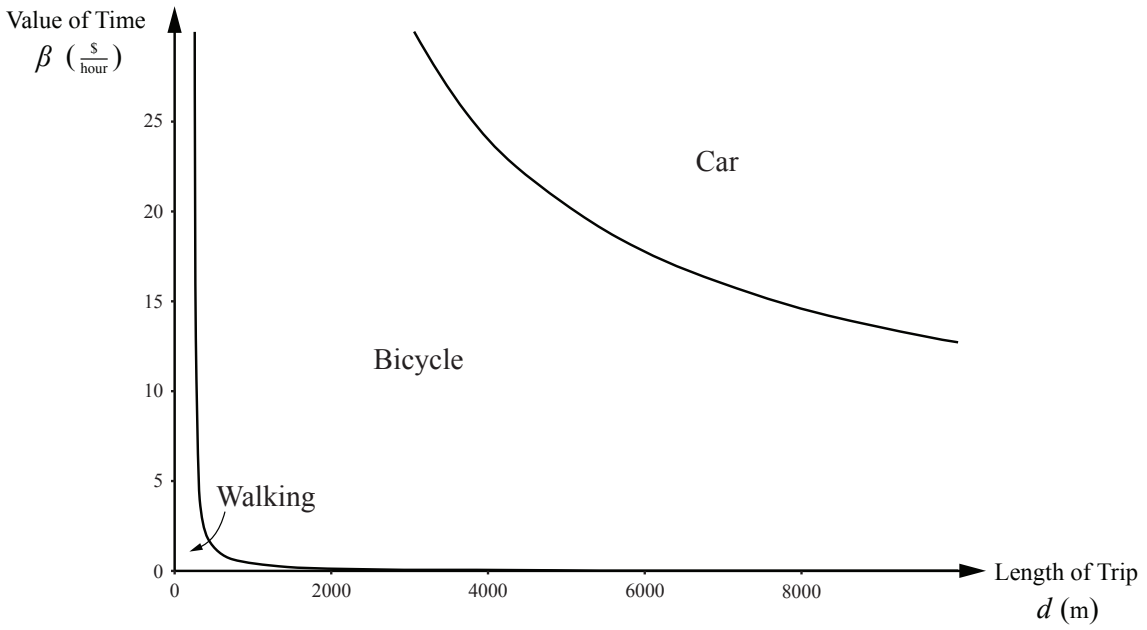


Figure 2.8. Comparison of individual modes over trip length and value of time showing which mode provides the minimum generalized cost per trip

for a city where people value time at β and make trips of length d. The values are expressed in units of money rather than time to avoid confusion in the units between an hour of cost and an hour of the day. Since there is so much uncertainty in the appropriate values for γ, environmental externalities have been omitted from these values. However, the results presented in the following chapters are general, so any of the values of α can be modified to reflect different relative valuations of costs.

Table 2.1. Cost Coefficients ($\beta = 20$ \$/hour, $d = 3000$ m)

Mode	$\beta\alpha_0$ [\$/m^2·hour]	$\beta\alpha_C$, $\beta\alpha_1$ [\$/trip]	$\beta^2\alpha_2$ [\$^{2}/m^2·hour·trip]	$\beta^2\alpha_3$ [\$^{2}/trip^2]
Walk		16.7		
Bicycle		7.01		
Car		8.75		
Bus		6.27	0.00726	3.65
BRT		6.09	0.0118	
Metro	0.00465	6.27	0.00820	

Bicycles appear very cost-effective for a wide range of trip lengths and values. For trips with $d = 3000$ meters and $\beta = 20$ \$/hr, Figure 2.8 shows that bicycle is the most cost effective of the individual modes. This is also verified in Table 2.1. Bicycles appear so competitive because they are fast and very inexpensive compared

to motorized modes. However, not everyone is able to use a bicycle, and not every trip can be served by a bicycle. Furthermore, it is difficult to account for perceived costs related to inclement weather and fatigue which are associated with bicycle trips. Therefore, the comparison of modes in the remaining chapters focuses on cars and transit, because in most developed cities these are the primary modes of transportation.

Typical values for r are summarized in Table 2.2 where metro requires no surface road space since lines are built in tunnels. Input parameters used to attain the values in Tables 2.1 and 2.2 are shown in Table B.2.2. The the combined road space for moving vehicles and for parking is the sum of the road space functions with the coefficients from Table 2.2. As explained in Section 2.4.2, road space for parking is not considered as a spatial constraint in the analysis presented in this thesis so this would be consistent with using only the coefficients in the "Moving" column. However, the costs of parking infrastructure are included in the cost coefficients in Table 2.1.

Table 2.2. Road Space Coefficients ($\beta = 20$ \$/hour, $d = 3000$ m)

Mode	r_C, r_1 [m^2·hour/trip]		r_2 [m^2·hour/trip]		r_3 [m^4·hour^2/trip^2]	
	Moving	Parking	Moving	Parking	Moving	Parking
Walk	1.67	0				
Bicycle	6.87	5.75				
Car	23.5	212.4				
Bus	0.266	0.420	0.0335	0.0044	16.9	2.2
BRT	0.212	0	0.0326	0.0059		

2.5 Summary of Methodology

This chapter has presented the definitions and methodology for modeling the costs of transportation in cities. This is done by taking a macroscopic view of city structure, and using macroscopic models of urban transportation networks. Trip length is used as a proxy for accessibility, so trips are characterized by their length to account for costs. In order to analyze the relationship between transportation costs and city structure in a systematic way, we will study an idealize city that behaves according to the following assumptions:

1. Travel demand is distributed uniformly across the city, and all trips share the same characteristics (i.e., trip length and value of time). (Section 2.1)

2. The road network is a symmetric grid which is uniform across the city. (Section 2.2.1)

3. Each mode operates independently of the others (without conflicts) so that the costs and road space requirements depend only on its own demand. (Section 2.2.3)

4. Each mode is deployed efficiently so that trips are served at the least cost possible. (Section 2.4)

Under these assumptions, modes can be deployed jointly, and the total costs and road space requirements are simply the sum of those for each mode used.

In the following chapters, the focus will be on analyzing a city served by two modes: cars and buses. The total system cost function is then given by:

$$Z(\lambda_C, \lambda_T) = Z_C(\lambda_C) + Z_T(\lambda_T) \tag{2.30}$$

and the total road space required for moving vehicles is:

$$R(\lambda_C, \lambda_T) = R_C(\lambda_C) + R_T(\lambda_T). \tag{2.31}$$

These analytical models, based on the physics of urban traffic and parameters of city structure establish bounds for what is physically achievable.

First, the simple case of a city with constant demand is analyzed using the same assumptions and cost functions presented here (Chapter 3). Many insights about the effect that constrained road space has on transportation costs can be gained by comparing modes in this way. Once this simple system is understood, more realistic cities with peaked demand are considered. The evening peak case is presented for which demand is peaked in time, but travelers can change only which mode they use (Chapter 4). Then, this case is extended to the morning peak in which travelers can change both their mode and when they travel (Chapter 5).

Chapter 3

City with Constant Demand

This chapter presents a first level of analysis for how road space should be allocated and how modes should be jointly used in cities. We start by considering a city with constant, uniform demand over time as described in Chapter 2. This analysis focuses on minimizing the total generalized cost of travel by individual modes (e.g., car) and collective public transit modes (e.g., bus, metro). The relevant question is how total demand, λ, should be split between two modes (car, λ_C, and transit, λ_T), and how modes should be jointly deployed to minimize the total social cost of the transport system, $Z(\lambda_C, \lambda_T)$. Using (2.30) and (2.31), this problem can be expressed as the following mathematical program:

$$\begin{aligned} \min \quad & Z = Z_C(\lambda_C) + Z_T(\lambda_T) \\ \text{s.t.} \quad & R_C(\lambda_C) + R_T(\lambda_T) \leq R \\ & \lambda_C + \lambda_T = \lambda. \end{aligned}$$

For this analysis, we consider the space in a city available for transportation, R, as given. Only the streets needed to serve λ_C and λ_T must be built and paid for, but there is only space to build them up to R. All of the demand must be served by cars or transit. We will look at how the optimized transportation system depends on the city characteristics, λ and R.

We first consider a city where there is unlimited space available for potential road infrastructure ($R = \infty$) in Section 3.1. Then in Section 3.2, realistic cases where road space is constrained by existing buildings or other protected land ($R < \infty$) are considered. Section 3.3 shows how the user equilibrium differs from the system optimum, and Section 3.4 shows how prices should be set to achieve system optimum. Section 3.5 discusses how parameter inputs which describe trip and mode characteristics affect the results, and Section 3.6 summarizes the contributions of this chapter.

3.1 System Optimum: Unlimited Road Space

Suppose that mode costs are given by generalized cost functions of the form (2.17) and (2.28). Suppose also that streets are managed so that the spatial requirements of one mode are not affected by the trips made with another as described in Chapter 2. In other words, the cost of serving a trip when modes share the right of way is the same as if the modes were operated in isolation so that the least cost of each mode is a function only of its own users. Thus, the total cost of transportation per area-time when the demand is (λ_T, λ_C) is given by (2.30), $Z(\lambda_T, \lambda_C) = Z_T(\lambda_T) + Z_C(\lambda_C)$, and the optimal deployment of each mode is determined by the modal split that minimizes this cost.

Since $Z_T(\lambda_T)$ is concave and $Z_C(\lambda_C)$ is linear, Z is concave in (λ_T, λ_C). Therefore, the contours of constant cost plotted on the λ_C versus λ_T plane are always convex to the origin. From (2.17) and (2.30), the iso-cost contours $Z(\lambda_T, \lambda_C) = K$, are the family of functions

$$\lambda_C(\lambda_T) = \frac{K}{\alpha_C} - \frac{Z_T(\lambda_T)}{\alpha_C}. \tag{3.1}$$

Note that K increases moving away from the origin, displayed by thin gray lines in Figure 3.1. The total demand for travel, $\lambda = \lambda_C + \lambda_T$, are represented by lines with slope -1 (heavy black lines in Figure 3.1) with demand increasing moving away from the origin.

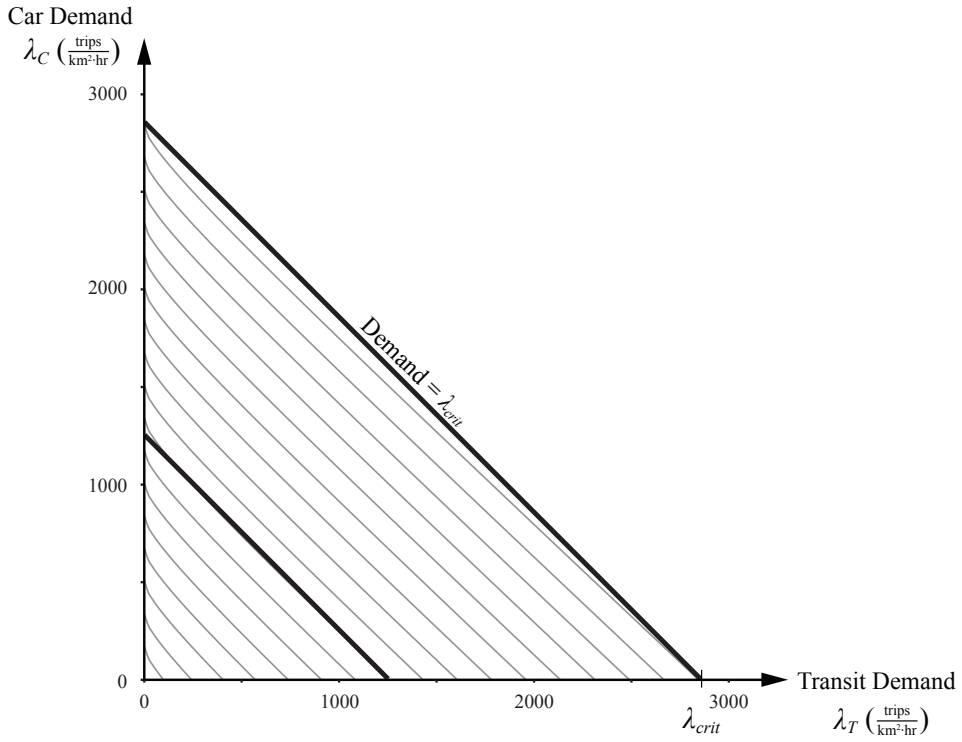


Figure 3.1. Cost contours and demand for a city with cars and transit

The mode split which minimizes $Z(\lambda_T, \lambda_C)$ for a total demand λ corresponds to the point on the demand line which lies on the lowest cost contour (closest to the origin). Since the contours are convex, this will always happen at an extreme point on either the vertical axis ($\lambda_C = \lambda$) or the horizontal axis ($\lambda_T = \lambda$). All other points along the total demand line represent transport service with a mixture of modes and cannot be optimal. Therefore, since the optimal modal split will be with all cars or all transit, it is sufficient to compare the cost of the two modes. This comparison can be done systematically for all values of λ by considering the difference of the cost of serving all trips by transit and by car, $Z_T(\lambda) - Z_C(\lambda)$, as in Proposition 1.

Proposition 1. *If $Z_T(\lambda) - Z_C(\lambda)$ is unimodal, concave, and it is non-negative and increasing at $\lambda = 0$, then $\exists \lambda_{crit} > 0$ such that*

$$\begin{aligned} Z_T(\lambda) - Z_C(\lambda) &> 0 \quad \textit{for } \lambda < \lambda_{crit}, \\ Z_T(\lambda) - Z_C(\lambda) &= 0 \quad \textit{for } \lambda = \lambda_{crit}, \\ Z_T(\lambda) - Z_C(\lambda) &< 0 \quad \textit{for } \lambda > \lambda_{crit}. \end{aligned}$$

Proof. Since $Z_T(\lambda) - Z_C(\lambda)$ is a unimodal concave function, then there must be a demand $\lambda_{crit} > 0$ where $Z_T(\lambda) = Z_C(\lambda)$; see Figure 3.2. This implies that for values of $\lambda < \lambda_{crit}$, the cost difference is positive and car serves the demand at lower generalized cost than transit. Likewise, for $\lambda > \lambda_{crit}$, the cost difference is negative so transit serves the demand at the lowest cost. □

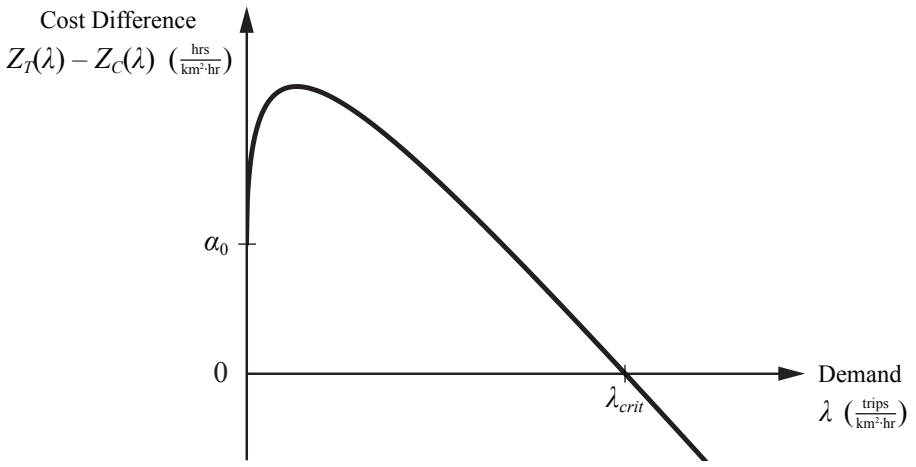


Figure 3.2. Cost difference between transit and car for different λ

Since $Z_C(\lambda)$ is linear-increasing and $Z_T(\lambda)$ is concave-increasing, and at $\lambda = 0$ these functions are non-negative, then $Z_T(\lambda) - Z_C(\lambda)$ is concave and unimodal whenever $\alpha_C > \alpha_1 + \sqrt{\alpha_3}$. So, Proposition 1 applies. This property is also shown in Figure 3.1 where the demand line associated with λ_{crit} intersects the same cost contour at both axes. Lower demands intersect the lowest cost contour on the vertical axis (all car), while greater demands intersect the lowest cost contour on the horizontal axis (all transit).

3.2 System Optimum: Limited Road Space

Now suppose that the city has a finite amount of space available for roads, R. This can be approximately accomplished with technologies such as intermittent priority (Eichler & Daganzo, 2006) so that capacity is not wasted by interactions between modes. The space required by the modes must not exceed the space available:

$$R_C(\lambda_C) + R_T(\lambda_T) \leq R \tag{3.2}$$

The road space requirements for each mode are given by (2.18) and (2.29).

The solution for a city with cars and bus rapid transit ($\alpha_0 = 0$ and $\alpha_3 = 0$) is presented in Section 3.2.1. The general solution is represented graphically by a single figure. The effect of changing the transit technology is discussed in the following sections for cases when $\alpha_3 > 0$ (Section 3.2.2) and $\alpha_0 > 0$ (Section 3.2.3).

3.2.1 Car and Bus Rapid Transit (BRT) System

For the case of car and BRT, illustrated in Figure 3.3, both modes require road space, and the constraint takes the form:

$$\lambda_C \leq \frac{R}{r_C} - \frac{R_T(\lambda_T)}{r_C}. \tag{3.3}$$

For any given R, the boundary of the space constraint is a curve $\mathcal{R}$ which, like the cost contours, is convex-decreasing on the (λ_T, λ_C) plane. These curves also move away from the origin as R is increased, but they are much flatter because the transit system's footprint is much smaller than that of car.

Solution with Road Space Constraint

Consider now what happens when constraint (3.2) is added to the optimization problem. Since $Z(\lambda_T, \lambda_C)$ is concave and the feasible mode splits for λ lies on a line segment in the $(\lambda_T,\ \lambda_C)$ space, the optimal mode split must still be at an extreme point at the edge of the feasible region. Figure 3.3 shows a road space constraint boundary $\mathcal{R}_1$ associated with $R = R_1$. For realistic transit services, $dR_T(\lambda_T)/d\lambda_T < r_C$, so the slope of $\mathcal{R}_1$ is always flatter than the demand line. Note that $\mathcal{R}_1$ can cross each possible demand line no more than once, and therefore the feasible mode splits on each demand line form a contiguous segment.

We will use a superscript to denote the demand line passing through a point, so $\lambda^{(\mathrm{A})}$ and $\lambda^{(\mathrm{B})}$ are defined as the total demands associated with points A and B, respectively. Then, for example, the feasible demand segment for $\lambda^{(\mathrm{B})}$ is $\overline{\mathrm{BC}}$ in Figure 3.3. For $\lambda \leq \lambda^{(\mathrm{A})}$, the feasible demand segment extends from the vertical axis (all car) to the horizontal axis (all transit). When $\lambda > \lambda^{(\mathrm{A})}$, the feasible demand segment extends only to $\mathcal{R}_1$ where the road space is used at capacity by a mix of modes. In order to determine the least cost operating strategy, we only need to compare the cost

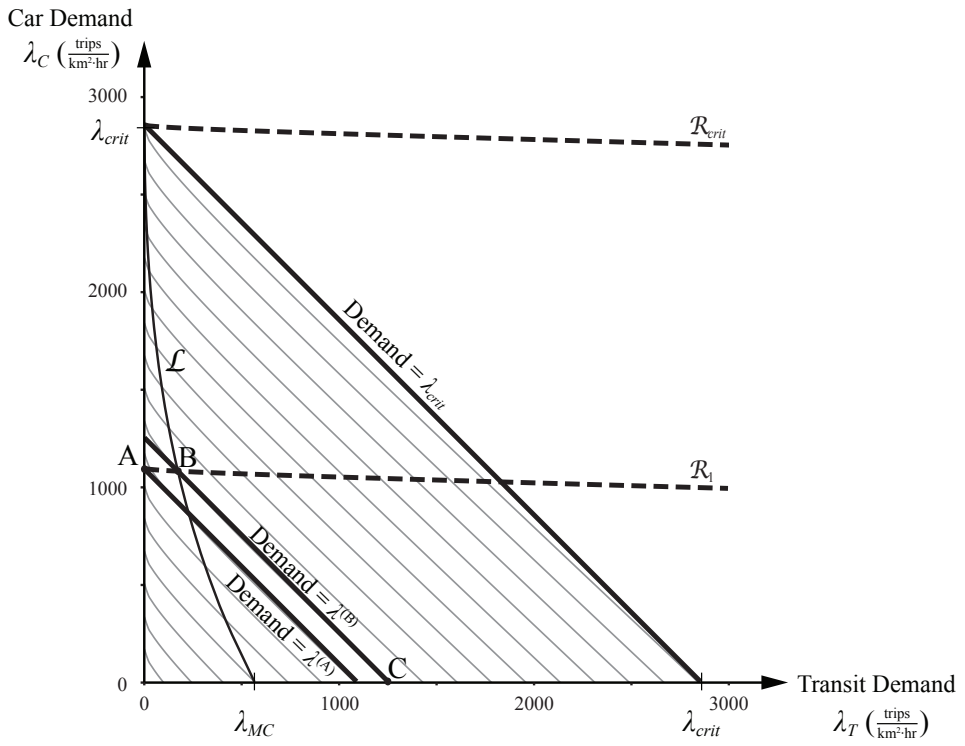


Figure 3.3. Cost contours and space constraints for a city with car and BRT

of all transit to the cost at the other extreme: all car or operating mixed modes with the available space fully utilized. We now show how this comparison can be made quickly by identifying the tipping point where the cost of mixing modes is the same as serving all trips by transit.

In Figure 3.3, the curve $\mathcal{L}$ is the locus of points (λ_T, λ_C) where $Z(\lambda_T, \lambda_C) = Z_T(\lambda)$; i.e., where the total cost is the same with mixed modes as if everyone was served by transit. To find $\mathcal{L}$, express each cost contour as:

$$Z_T(\lambda_T) + Z_C(\lambda - \lambda_T) = K. \tag{3.4}$$

For any value of λ, the mode split that yields the same cost as serving all trips by transit is attained by solving (3.4) for λ_T when $K = Z_T(\lambda)$. By substituting (2.17) into (3.4), the expression can be rewritten in the form:

$$Z_T(\lambda) - \alpha_C \lambda = Z_T(\lambda_T) - \alpha_C \lambda_T. \tag{3.5}$$

Note that because $Z_C(\lambda_C)$ is linear, the right side is the same function used in Proposition 1, which is shown for the car and bus case in Figure 3.4. If the axes cross at a cost of $Z_C(\lambda)$, the same figure hold for all λ.

Let us define λ_{MC} as the demand when (3.5) has a unique solution at $\lambda_T = \lambda_{MC}$. This is the density associated with the maximum cost in Figure 3.4. Then, for $\lambda \in (\lambda_{MC}, \lambda_{crit})$, there are two solutions which satisfy (3.5): $\lambda_T = \lambda$ and $\lambda_T = {\lambda_T}^{(\mathcal{L})} <$

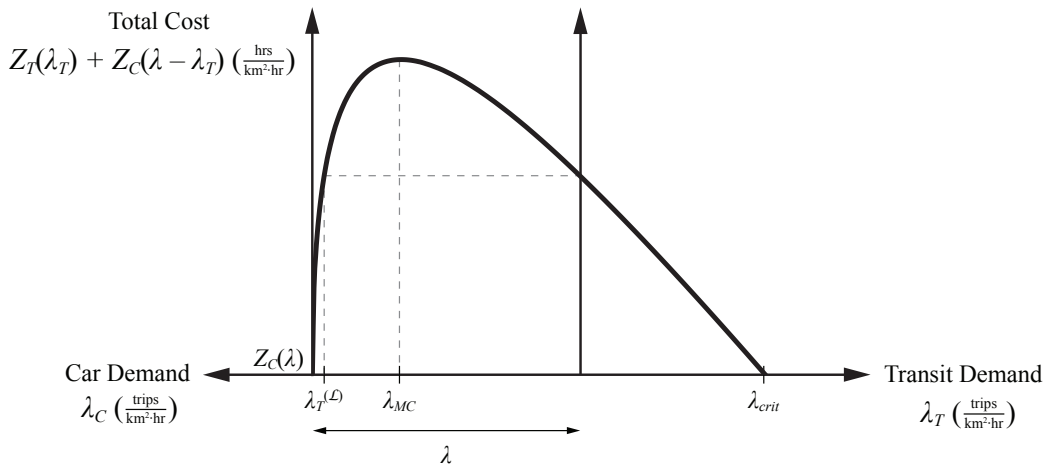


Figure 3.4. Total generalized cost of car and BRT system

λ_{MC}. At $\lambda = \lambda_{crit}$, ${\lambda_T}^{(\mathcal{L})} = 0$, and ${\lambda_T}^{(\mathcal{L})}$ increases as λ decreases until they are both equal at λ_{MC}. The values of ${\lambda_T}^{(\mathcal{L})}$ define $\mathcal{L}$ which must be a declining curve (see Figure 3.3). $\mathcal{L}$ extends from $(0, \lambda_{crit})$ to $(\lambda_{MC}, 0)$.

Note that for a BRT system with only cost coefficients for α_1 and α_2 in (2.28), the generalized cost function is:

$$Z_T(\lambda_T) = \alpha_1 \lambda_T + \sqrt{\alpha_2 \lambda_T} \tag{3.6}$$

and λ_{crit} and λ_{MC} are given by:

$$\lambda_{crit} = \frac{\alpha_2}{(\alpha_C - \alpha_1)^2} \tag{3.7}$$

$$\lambda_{MC} = \frac{\alpha_2}{4(\alpha_C - \alpha_1)^2} \tag{3.8}$$

where (3.7) is the result of setting car cost equal to transit cost as defined in Proposition 1, and (3.8) is obtained by setting their first derivatives (marginal costs) equal to one another. Therefore, $\lambda_{MC} = \lambda_{crit}/4$.

In Figure 3.3, the feasible demand segment for any $\lambda > \lambda^{(\mathrm{B})}$ intersects $\mathcal{R}_1$ to the right of $\mathcal{L}$. This means that the minimum feasible transit demand must lie in the interval $[{\lambda_T}^{(\mathcal{L})}, \lambda]$ which is associated with a greater cost contour than serving all trips by transit. This is shown in Figure 3.4 where the left side of (3.4) is greater than $Z_T(\lambda)$ in this interval. If a feasible demand segment lies entirely to the right of $\mathcal{L}$, then it is optimal to serve all trips with transit. Whenever the feasible demand segment extends to the left of $\mathcal{L}$ ($\lambda < \lambda^{(\mathrm{B})}$) it is optimal to operate at the left-most extreme point, because it is on a lower cost contour. This is achieved by serving all trips by car ($\lambda \leq \lambda^{(\mathrm{A})}$) or mixing modes such that the space is fully utilized ($\lambda^{(\mathrm{A})} < \lambda < \lambda^{(\mathrm{B})}$).

We define $R_{crit} \doteq R_C(\lambda_{crit})$, which is the road space required to serve all of λ_{crit} by car. Values of $R \geq R_{crit}$ never pose an active constraint since adequate space

is available to serve each λ with the least cost mode. Note that for cities where $R < R_{crit}$, the minimum demand at which all trips should be served by transit may be much less than the demand where transit otherwise becomes cost competitive; i.e., $\lambda^{(\mathrm{B})} < \lambda_{crit}$.

Graphical Representation of Solution

Figure 3.5 summarizes the results for the above analysis using Figure 3.3 for all possible values λ and R. The axes are scaled to normalize the demand, dividing by λ_{crit}, and the road space, dividing by R_{crit}. The axes represent dimensionless measures of demand and road space. For a BRT system with parameters α_1 and α_2, the scaled demand and road space are:

$$\text{Demand} = \lambda \frac{(\alpha_C - \alpha_1)^2}{\alpha_2} \tag{3.9}$$

$$\text{Road Space} = R \frac{(\alpha_C - \alpha_1)^2}{r_c \alpha_2}. \tag{3.10}$$

With the axes scaled in this way, a demand of 1 represents the critical demand, and a road space of 1 represents the critical road space. Now, the single figure represents the set of all city structures and incorporates all of the elements of the generalized cost functions. This single graphical solution thus represents all possible parameters when $\alpha_C > \alpha_1$.

Figure 3.5 shows whether a single mode or a mix of modes serves demand at the lowest total social cost. Consider a city with road space $R_1 < R_{crit}$, the space constraint is shown as $\mathcal{R}_1$ in Figure 3.3 and a horizontal line at $R = R_1(\alpha_C - \alpha_1)^2 / r_C \alpha_2$ in Figure 3.5. For low demand ($\lambda < \lambda^{(\mathrm{A})}$), the lowest cost is along the vertical axis of Figure 3.3 so these demands are in the *all car* regime in Figure 3.5. The range of demands that intersect $\mathcal{R}_1$ between A and B in Figure 3.3 are where the available road space should be fully utilized with a mix of modes provided, and in Figure 3.5 these demands fall in the *mixed car and transit* regime. The demand $\lambda^{(\mathrm{B})}$ passes through both points B and C in Figure 3.3, and this represents a tipping point where the cost of providing a minimal transit system is the same as serving all trips by transit. This transition is represented by the single point B, C in Figure 3.5 which lies on a line indicating the jump to an *all transit* system. For all greater demands ($\lambda > \lambda^{(\mathrm{B})}$), the total social cost of transportation in the city is minimized when all trips are carried by the transit system. There is a small set of cities with very constrained road space for which even a bus transit system cannot adequately meet demands. This case is most likely to emerge if there is a political reason why few streets can be dedicated to transit service.

The average generalized cost per trip in the system optimum is shown systematically for all cities in Figure 3.6. The generalized cost of each car trip, z_C, is always the same value as described by the car cost function (2.17). The magnitude of the cost depends of the parameters of the generalized cost functions. Intuitively, when road

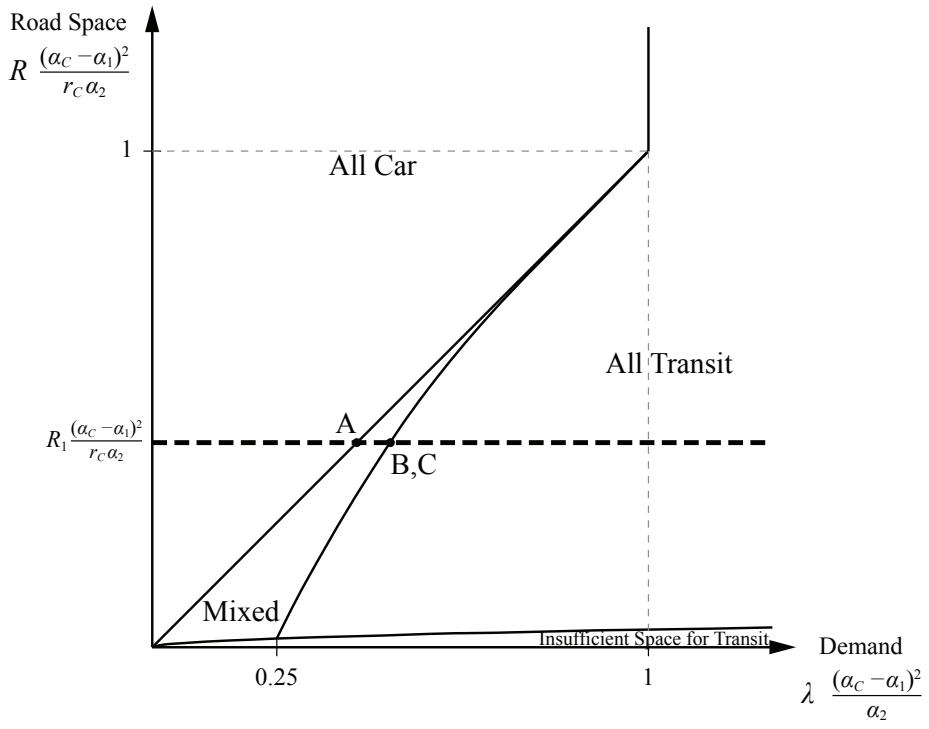


Figure 3.5. Summary of system optimum for all city structures with car and BRT

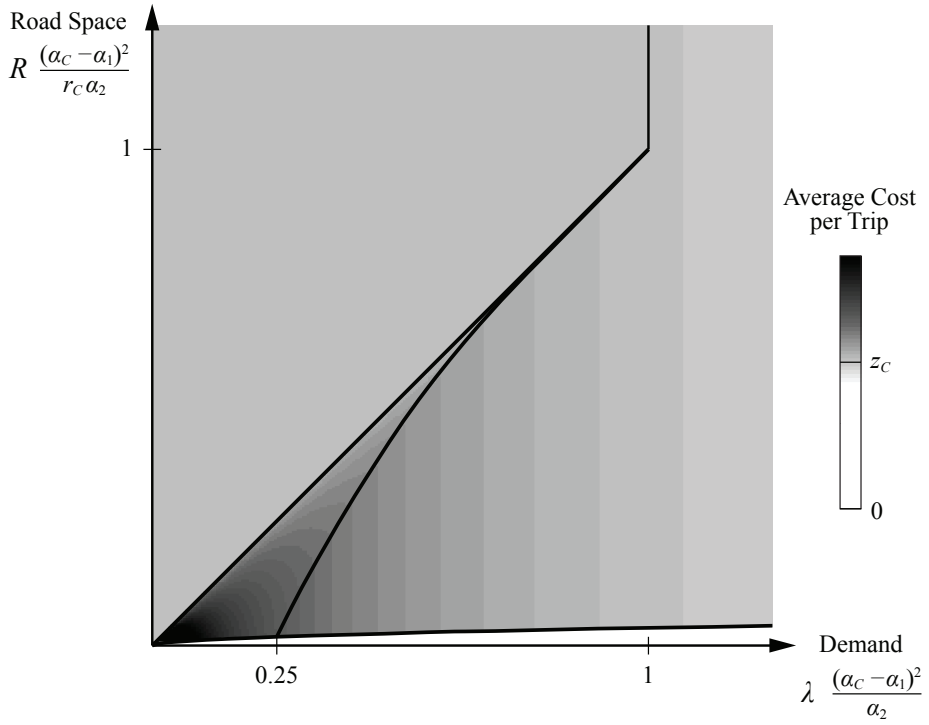


Figure 3.6. Average trip cost at system optimum for cars and BRT

space is a binding constraint, costs are greater than without. Note that once a city is served by *all transit*, increased demand lowers the cost per trip. This is the result of public transit's economies of scale. Costs are greatest for cities with contrained space and low demand. This is because transit service is required to meet demand, but the system is so small that it cannot be run cost-effectively.

3.2.2 Car and Standard Bus System

The results are similar for a standard bus service where the loss time per stop depends on the number of boarding and alighting passengers. For this type of transit system, the generalized cost function includes $\alpha_3 > 0$:

$$Z_T(\lambda_T) = \alpha_1 \lambda_T + \sqrt{\alpha_2 \lambda_T + \alpha_3 \lambda_T^2}. \tag{3.11}$$

The analysis with cost contours is the same, but the values of λ_{crit} and λ_{MC} are different. Applying Proposition 1, (3.7) becomes:

$$\lambda_{crit} = \frac{\alpha_2}{(\alpha_C - \alpha_1)^2 - \alpha_3}. \tag{3.12}$$

The value of λ_{MC} is sought using the same method as for (3.8) and setting the first derivatives of the car and transit cost functions to equal each other. So, λ_{MC} must satisfy:

$$\lambda_{MC} = \frac{(\alpha_2 + 2\alpha_3 \lambda_{MC})^2}{4(\alpha_C - \alpha_1)^2(\alpha_2 + \alpha_3 \lambda_{MC})} \tag{3.13}$$

We can now show that $\lambda_{MC} < \lambda_{crit}/4$. Since $\alpha_2 \alpha_3 \lambda_{MC} > 0$, then it is true that:

$$(\alpha_2 + 2\alpha_3 \lambda_{MC})^2 < (\alpha_2 + 2\alpha_3 \lambda_{MC})^2 + \alpha_2 \alpha_3 \lambda_{MC} = (\alpha_2 + \alpha_3 \lambda_{MC})(\alpha_2 + 4\alpha_3 \lambda_{MC}). \tag{3.14}$$

where the equality of the middle and last expressions can easily be verified by algebra. By manipulating (3.14), we see that $(\alpha_2 + 4\alpha_3 \lambda_{MC}) > (\alpha_2 + 2\alpha_3 \lambda_{MC})^2/(\alpha_2 + \alpha_3 \lambda_{MC})$. Substituting the left side of this inequality into (3.13) and isolating λ_{MC} establishes an upper bound for λ_{MC}:

$$\lambda_{MC} < \frac{\alpha_2}{4(\alpha_C - \alpha_1)^2 - 4\alpha_3} = \frac{\lambda_{crit}}{4}. \tag{3.15}$$

The implication of these changes compared to the BRT solution is that the system optimum can still be represented on the normalized axes as in Figure 3.5, but the curve delineating *mixed* and *all transit* operations shifts to the left. The lowest demand for which an *all transit* solution is optimal moves to a value $\lambda_{MC}((\alpha_C - \alpha_1)^2 - \alpha_3)/\alpha_2 < 0.25$.

This effect tends to get larger as α_3 increases, which means that larger loss times per passenger increase the relative range of cities which should be served by transit. This can also be interpreted as the effect of loss time per boarding and alighting passenger, which, all else held equal, makes low capacity transit systems less cost-efficient, and mixing modes less desirable.

3.2.3 Car and Metro System

Bus transit requires street space, so there can be high demands which cannot be served even with a surface transit system. This is shown in Figure 3.5 for cities where R is very low. In this case, demand can only be met with a transportation system which does not use streets (such as metro). The same methods described above still apply, although the cost function for metro has a discontinuity because there is a significant fixed cost component, α_0, for infrastructure.[1] The generalized cost function for metro is:

$$Z_T(\lambda_T) = \begin{cases} \alpha_0 + \alpha_1\lambda_T + \sqrt{\alpha_2\lambda_T} & \text{if } \lambda_T > 0 \\ 0 & \text{if } \lambda_T = 0. \end{cases} \tag{3.16}$$

The cost function jumps from 0 to α_0 when $\lambda_T > 0$, because tunnels, tracks, and stations must be built before the first passenger can be served.

In Figure 3.7, the demand line for λ_{crit} intersects both axes on the same cost contour satisfying $Z_C(\lambda_{crit}) = Z_T(\lambda_{crit})$. It follows from (2.17) and (2.28) that $\lambda_{crit} = Z_T(\lambda_{crit})/\alpha_C$, and substituting this into (3.1) for $K = Z_T(\lambda_{crit})$, the cost contour is:

$$\lambda_C(\lambda_T) = \lambda_{crit} - \frac{Z_T(\lambda_T)}{\alpha_C}. \tag{3.17}$$

For $\lambda_T \ll 1$ (approaching the vertical axis) this expression is approximately

$$\lambda_C(\lambda_T) \approx \lambda_{crit} - \frac{\alpha_0}{\alpha_C}. \tag{3.18}$$

So, the cost contour for $Z_T(\lambda_{crit})$ approaches the vertical axis at $\lambda_{crit} - \alpha_0/\alpha_C$ rather than λ_{crit} (see Figure 3.7). This difference is the amount of car demand that could be served for the fixed cost of metro.

The discontinuity of the cost function for metro also creates a truncation of the $\mathcal{L}$ curve. Let $\lambda^{(\mathrm{D})}$ be the demand at point D in Figure 3.7 where $\mathcal{L}$ meets the vertical axis. This is the demand satisfying $Z_C(\lambda^{(\mathrm{D})}) = Z_T(\lambda^{(\mathrm{D})}) - \alpha_0$. This point is also shown in Figure 3.8, where it is clear that for $\lambda > \lambda^{(\mathrm{D})}$, there is only one positive-valued solution to (3.5) at $\lambda_T = \lambda$. For $\lambda \in (\lambda_{MC}, \lambda^{(\mathrm{D})})$, there exists a ${\lambda_T}^{(\mathcal{L})} > 0$, and $\mathcal{L}$ slopes downward toward λ_{MC} just as described for the case of a bus transit system.

The same method to determine the optimal mode split for a city with cars and buses applies to a city with cars and metro. Figure 3.7 shows a space constraint, $\mathcal{R}_2$, for a city with road space $R_2 < R_{crit}$. The same three possible regimes of mode split are shown in Figure 3.9 as in 3.7: *all car* for $\lambda \leq \lambda^{(\mathrm{A})}$, *mixed car and metro* for $\lambda^{(\mathrm{A})} < \lambda < \lambda^{(\mathrm{B})}$, and *all metro* for $\lambda \geq \lambda^{(\mathrm{B})}$. The difference between metro and bus is that for $\lambda > \lambda^{(\mathrm{D})}$ (right of point D in Figure 3.9), it is never optimal to mix modes. In this case, demand is either not constrained by road space so all trips should be served by car, or the feasible demand segment (below $\mathcal{R}_2$) is entirely to the right of $\mathcal{L}$ and all trips should be served by metro.

[1] This discontinuity exists to some extent for all transit modes. Even a simple bus system will require some initial investment for signs at bus stops. However, this cost cannot be neglected for a mode with expensive fixed costs such as metro because metro tunnels and tracks must be built before the first train can run.

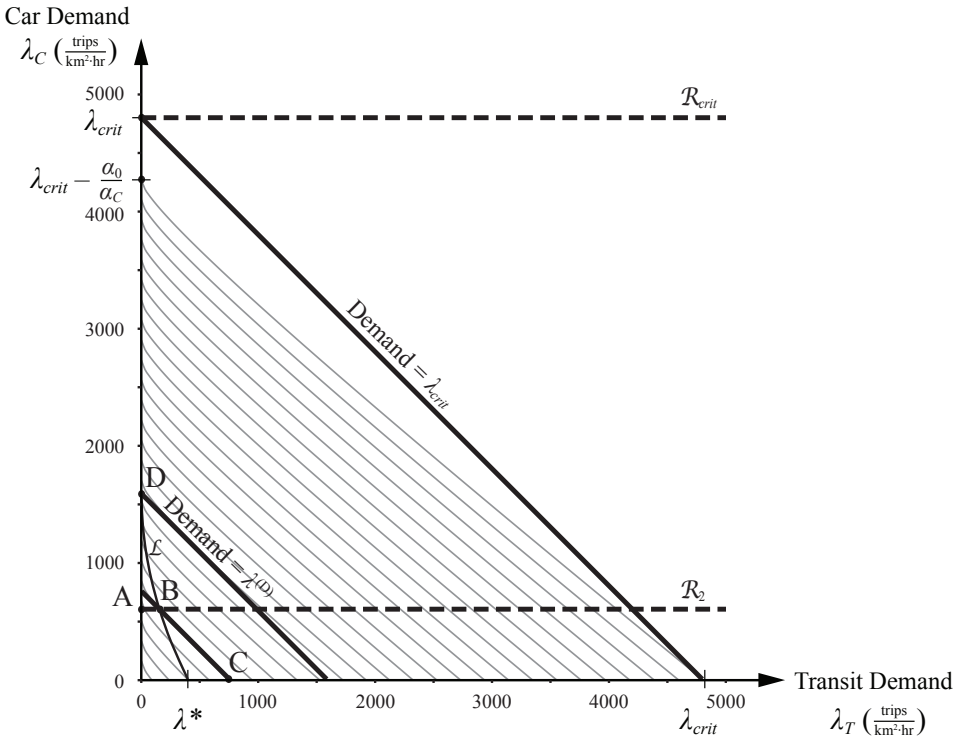


Figure 3.7. Cost contours and space constraints for a city with car and metro

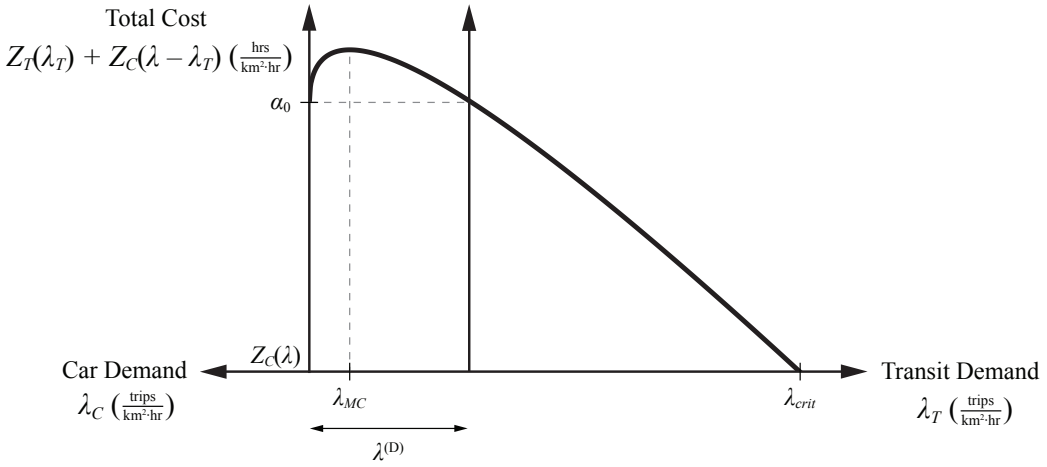


Figure 3.8. Total generalized cost of car and metro

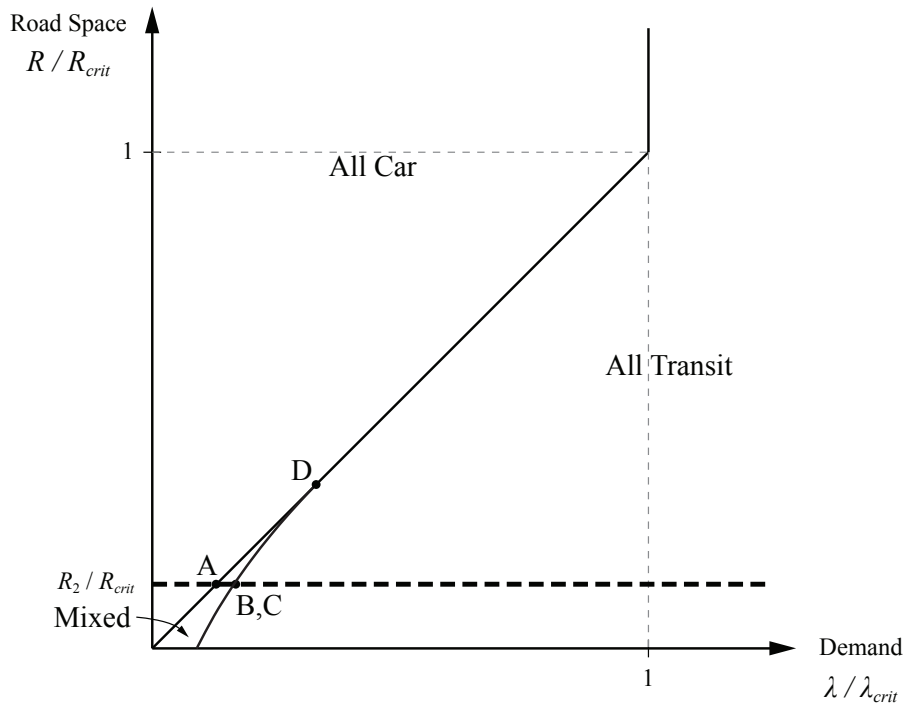


Figure 3.9. Summary of system optimum for all city structures with car and metro

An alternative interpretation of Figures 3.5 and 3.9 is to consider a city with fixed demand, λ. For a city with all cars, as road space is restricted (e.g., due to a policy of "road diets") there is always a range of R where all available space for roads should be full and mixed car and bus operations yield the least total social costs before switching entirely to transit. In a metro city, however, there are levels of demand ($\lambda > \lambda^{(D)}$) where everyone would be best off by switching suddenly to metro as soon as the road space is filled, so the available road space should never be utilized completely. This is expected, because in many cases, the enormous infrastructure cost for a metro system is only justified if it will be fully utilized and the infrastructure is shared by as many users as possible. Similar patterns as those presented here for bus and metro will arise for transit systems with more complex structures such as hierarchical routes or combined bus and metro operations, because $Z_T(\lambda_T)$ for these systems are also concave and increasing.

3.3 User Equilibrium

The system optimum solutions presented in the previous sections show how road space should be allocated and modes should be used to minimize the total generalized cost of the system. Individual users are not expected to choose this outcome on their own if they seek to minimize the generalized cost of their own trip. The generalized cost per trip is z_C for a free-flow car trip, and z_T for a transit trip. Note that in

Figures 3.5 and 3.9, transit is used for demands less than λ_{crit} when the road space is constrained. Therefore, following from the definition of λ_{crit} in Proposition 1:

$$z_T(\lambda) > z_C \quad \text{for } \lambda < \lambda_{crit}. \tag{3.19}$$

We will suppose that as a default, modes are priced to cover their own operating costs and externalities, so that each user pays the generalized cost of his or her own trip.

In equilibrium, users choose the mode with the lowest generalized cost. In the city with constant demand where everyone has the same trip characteristics and preferences, both modes are used simultaneously in equilibrium only if they are associated with the same generalized cost (Wardrop, 1952). If only one mode is used, then it must have the lowest generalized cost.

The system optimum solutions shown in Figures 3.5 and 3.9 are also user equilibrium solutions when the road space is not an active constraint (i.e., when $\lambda \geq \lambda_{crit}$ or above the diagonal when $\lambda < \lambda_{crit}$). When demand is a constraint, the system optimum is not in equilibrium because passengers would not choose the transit system if the streets are uncongested and a single car trip would be less costly.

Once the network capacity to serve cars at rate f_m is fully utilized, any greater demand requires the provision of a transit service. Otherwise congestion will result, lowering the network capacity and causing delays to grow without a bound. Therefore, even in user equilibrium, a transit service will operate when road space is constrained because the cost of transit is less than infinite delay. However, a small transit system is costly per passenger, so it will only be used when the generalized cost of driving including delays increases so that the generalized cost of both modes are equivalent. The equilibrium traffic state can be estimated using the Network Exit Function (NEF) introduced in Section 2.2.3, and reproduced here in Figure 3.10.

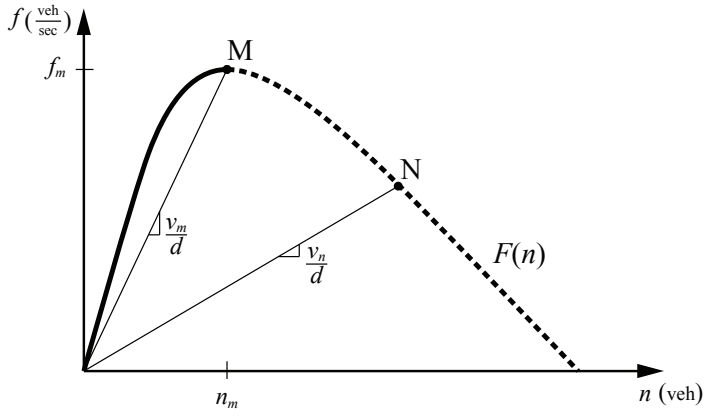


Figure 3.10. Network exit function showing congested user equilibrium state

The free flow trip time is represented on the NEF as the inverse of the slope from the origin to the traffic state. For example, when the network is operating at maximum capacity (point M), the trip time is $t_m = d/v_m$. Assuming that the transit system is operated with some dedicated space and priority to avoid congestion, then it

will have a consistent generalized cost per rider which depends on the transit demand, $z_T(\lambda_T)$. In equilibrium, when $z_T(N_T) > z_C$, then the traffic state will move to the right, down the congested branch of the NEF (see Figure 3.10) until the the delay satisfies the equilibrium condition: e.g., to a point N where $d/v_n - d/v_m = z_T(\lambda_T) - z_C$.

The resulting equilibrium solution can be summarized on the same scaled axes used to compare system optimum solutions, and this is shown in Figure 3.11. The difference between the user equilibrium and system optimum is that when space is constrained, modes are always mixed. The equilibrium mixing is suboptimal with wasteful, persistent congestion. Naturally, we would like to look for ways to push users to behave in the system optimum way. One way to do this is with pricing as described in the next section.

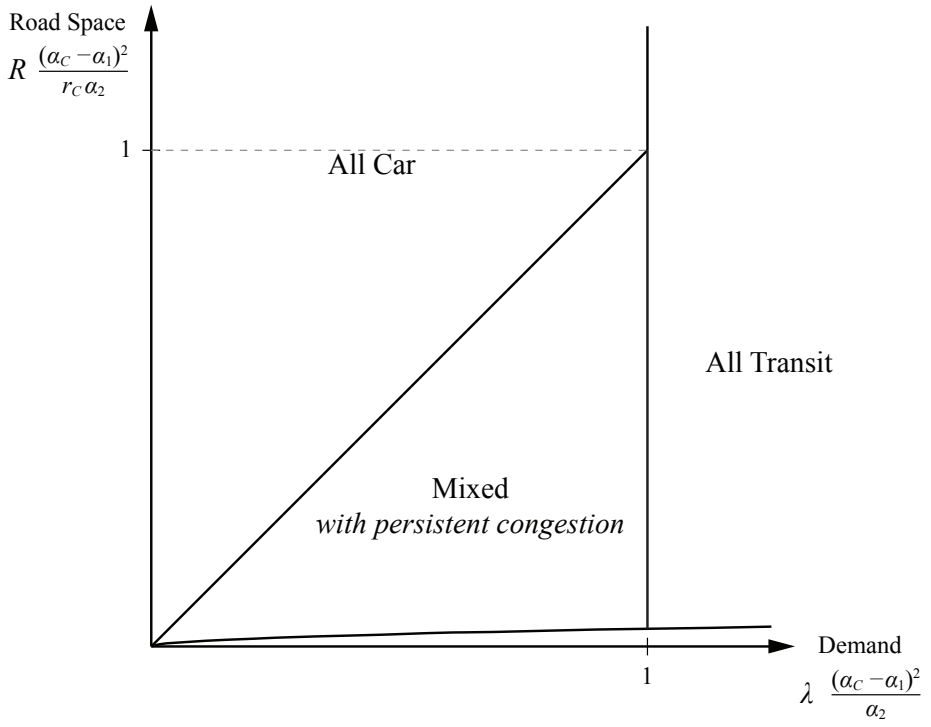


Figure 3.11. User equilibrium in city with constant demand

3.4 System Optimal Pricing

An optimal pricing strategy must move the user equilibrium mode choices to the system optimum presented in Section 3.2. Therefore, when the prices are set correctly, the system optimal use of modes will be a Wardrop (1952) equilibrium wherein users experience the same cost by either mode. The prices must make up the difference between the generalized cost of a free-flow car trip and the generalized cost of a transit

trip. The magnitude, \$, of the price must satisfy

$$\$ = z_T(\lambda_T^*) - z_C \tag{3.20}$$

where λ_T^* is the transit demand at system optimum, and z_C is the generalized cost of a car trip ($z_C = \alpha_C$ for an uncongested car trip with the cost functions presented). This price, \$, can be in the form of a toll for cars, a subsidy for transit fares, or some combination of the two. The optimal prices are not unique, because any pair which makes users indifferent between transit and car will allow both modes to be used simultaneously.

The magnitude of these prices is shown systematically for the full range of cities in Figure 3.12. The shading shows the prices scaled relative to $\$_{MC}$ which is the required subsidy when the marginal costs of both modes are equal. For cities where road space is not a binding constraint, no pricing intervention is needed because the system optimum is a user equilibrium. However, subsidies for transit are justified when the road space is constrained. The subsidy per trip is greatest when modes are mixed, because the transit system is serving a low transit demand, and the generalized cost per passenger is high. The *all transit* cases require much less subsidy per passenger, because the systems are larger and more efficient.

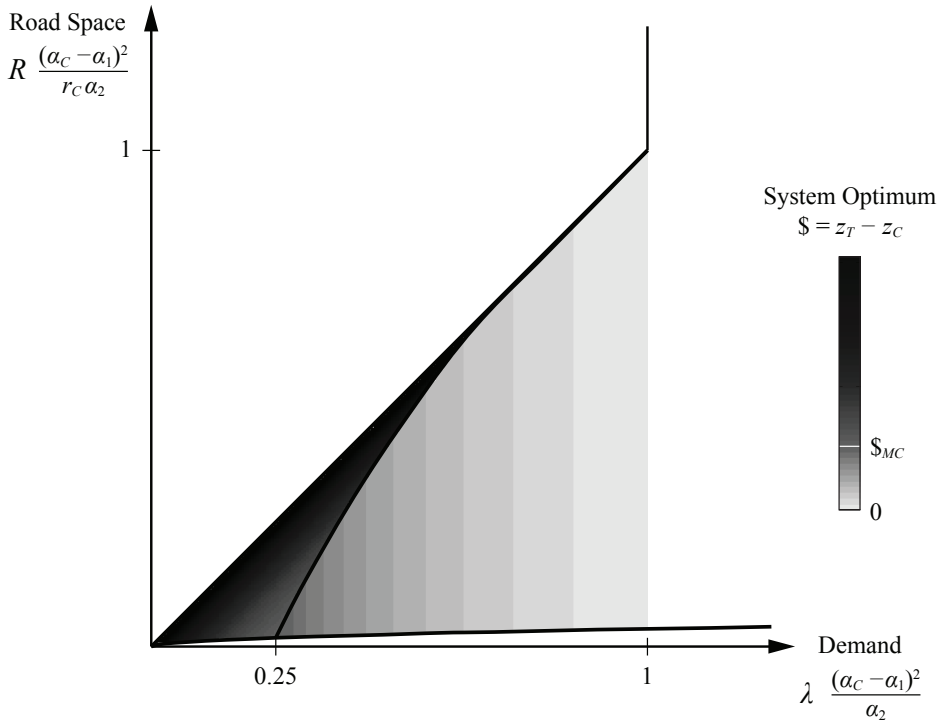


Figure 3.12. Required car toll or transit subsidy to achieve system optimum

The ability to choose how to make up the pricing difference with tolls and subsidies provides flexibility to achieve other policy objectives. For example, a city could seek to make the pricing system revenue neutral. These analyses could also be extended to consider elasticity of demand and the effect of the magnitude of user cost on total demand.

3.5 Influence of Parameters

The system optimum, user equilibrium, and pricing solutions in the preceding sections have all been presented on scaled axes which establish a plane representing the set of all city structures (i.e., all combinations of λ and R). The scale factors include the coefficients of the generalized cost functions and the car footprint so that one figure shows the solution for all possible parameter values. What this means is that such figures (i.e., Figures 3.5, 3.11, and 3.12) do not change if the properties of the trip or mode parameters change. What would change, however, is the location of a city on the solution plane.

Figure 3.13 shows the position of a hypothetical city on the scaled axes, and how it will move if various parameters are changed in isolation. Clearly, if the demand λ in this city increases, the city will move to the right, and if the road space available increases, the city will move up.

All parameters that affect the relative generalized cost of a car trip versus a transit trip result in a shift along the ray from the origin. As a car trip becomes more competitive on the basis of generalized costs, a city will move towards the origin (direction 1 in Figure 3.13). This effect can result from decreased costs associated with cars, increased costs associated with transit, or an increase in the value of time, β. As transit becomes more competitive, a city will tend to move away from the origin toward the critical demand at which transit is more efficient than an all car system (direction 2 in Figure 3.13).

Note that all of these shifts along the ray toward or away from the origin do not change whether a city's road space is an active constraint, because this is also represented by a diagonal with slope 1 passing through the origin. Essentially, this means that changing only the relative costs of modes does not change the severity of the road space constraint. What may change is the system optimum solution. As cities move toward the origin, cars serve trips more cost-effectively and if space is constrained, mixing modes is more likely to be desirable.

A different kind of shift happens as the result of changing trip length d, because this affects not only generalized costs, but also the footprint of a car trip. As trips grow longer, faster modes like car tend to become more competitive based on generalized costs (as shown in Figure 2.8). Longer trips also require a bigger footprint because road space must be occupied for a longer time. The result is that as trip lengths increase, a city will move down and to the left as illustrated in Figure 3.13. As expected, growing trip lengths can cause streets in cities to fill, activating the road space constraint.

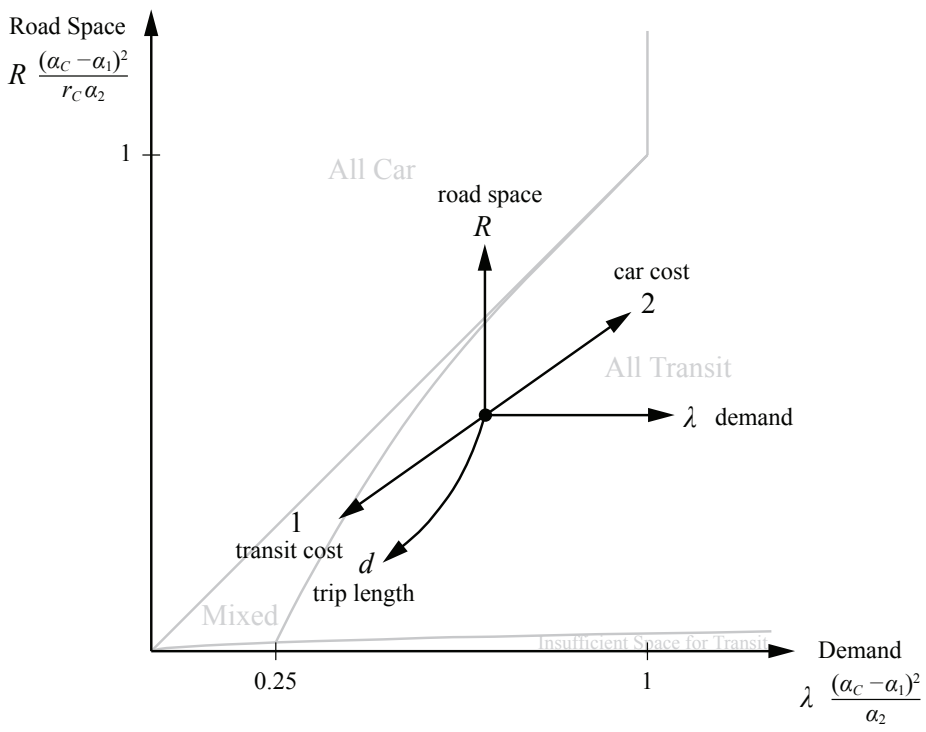


Figure 3.13. Effect of trip and mode parameters on a city's location on the scaled (λ, R) plane

By understanding the impact of various parameters, we can also compare real world cities. This analysis is idealized by flattening geography and assuming constant demand, so placing real cities on the figure is problematic, but we can discuss their relative positions qualitatively. Recall the 4 neighborhoods presented in Figure 2.1: Pleasant Hill, Berkeley, San Francisco, and New York. The population density and road area of each of these neighborhoods is increasing, so if all other parameter values are equal, they would fall on Figure 3.5 progressively to the right and above the one before it. Note that the road area increases less quickly than population density (which is a proxy for demand density). New York's Upper West Side almost certainly lies within the constrained region because the city already experiences traffic congestion even though a large number of trips are served by transit. Pleasant Hill, on the other hand, is likely situated well to left and above the diagonal representing full streets. Low density suburban environments are only likely to get congested if the network is poorly connected so that many streets are unable to serve through trips.

3.6 Summary of Findings

This chapter has shown how space should be allocated and modes should be priced in cities with constant demand that can be served by cars and transit. By explicitly

characterizing the shape of the cost functions for cars and transit, the system optimum solution is systematically identified for all possible city structures. The results for a city with cars and BRT are summarized on a single figure which incorporates all of the relevant parameter values. The effects of changing the cost function to model a standard bus or metro system reveal results which are qualitatively the same.

The system optimum for a city with two modes can be in one of three regimes. As demand increases, the regimes are passed in the following order:

1. *All Car* in which there is sufficient space for all trips to be served by cars when car trips have lower generalized cost than transit.

2. *Mixed Car and Transit* in which roads are fully utilized and only enough transit is provided to meet the road space constraint; the transition from cars to mixing is continuous.

3. *All Transit* in which the cost of serving all trips by transit is less than with mixed modes; the transition from mixing to all transit is a sudden abrupt change.

It has also been shown that when the road space constraint is active, the user equilibrium is suboptimal, but an optimal pricing strategy always exists to subsidize transit (or price cars) in order to achieve system optimum. In all of these cases, transit subsidies are justified in that they reduce the total generalized cost of the transportation system by eliminating congestion.

Chapter 4

Rush Hour City: Evening Peak

Up to this point we have been looking at the very idealized case of cities with constant demand. Real cities, of course, have rush hours, and the demand for transportation tends to be peaked in the morning and evening. The evening commute problem is more complicated than constant demand, because congestion in the network is dynamic so the travel decisions of commuters in the beginning of the rush affect the conditions faced by the commuters who follow.

This chapter investigates the evening commute in which the demand for trips varies over a rush period, and the only choice commuters have is their own transport mode. We will suppose for the evening commute that travelers do not adjust their departure time in response to traffic conditions. This is reasonable if we imagine that workers leave their place of employment at the end of the day and wish to get home as soon as possible.

Section 4.1 presents modifications to the cost functions to incorporate the time-dependent nature of the system. Section 4.2 shows how the Network Exit Function, introduced in Section 2.2.3, can be used to build queuing diagrams to model the evolution of traffic congestion on the network. The user equilibrium is presented in Section 4.3, and the system optimum which minimizes the generalized costs including delay is presented in Section 4.4. Section 4.5 shows a simple pricing strategy to achieve system optimum. Then, Chapter 5 will analyze a similar problem but with the added flexibility that commuters can also choose when they travel in addition to the mode choice.

4.1 Time-Dependent Mode Costs

The cost functions presented in Chapter 2 were based on the assumption that the demand is constant over time. When demand changes over time, the cost functions must be modified to reflect this. Suppose that within a period of the day of length t_{max} there is a peak in demand. We can think of t_{max} as the maximum length of the rush before the next peak period begins. The biggest change is that capital investments must be paid for the whole day even if infrastructure and vehicles are only used for

a short period in the peak. The infrastructure cost, for example, is decoupled from the mode cost function because it is determined by R, and it is not affected by the mode operations.[1]

Rather than expressing costs in terms of the demand rate, we are now considering a whole peak period, so costs are a function of the total number of trips during that period. The generalized cost for an uncongested car trip remains constant, and the total cost for cars (not including queuing delays) is:

$$Z_C(N_C) = \hat{\alpha}_C N_C \tag{4.1}$$

where $\hat{\alpha}_C$ is the generalized cost not including infrastructure costs, and N_C is the total number of car trips in the analysis period.

The cost function for transit changes a little more noticeably. Suppose that transit service is operated for a period of time $t_T < t_{max}$. Operating costs are only accrued for the time that the system is operational, t_T (e.g., fuel, labor). Capital costs (e.g., vehicles, maintenance facilities) are amortized over the whole day and must be paid whether transit service is running or not. The generalized cost function for an efficient bus system like BRT changes from (3.6) to become:

$$Z_T(N_T) = \hat{\alpha}_1 N_T + \sqrt{\hat{\alpha}_2 t_T N_T + \hat{\hat{\alpha}}_2 t_{max} N_T} \tag{4.2}$$

where $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\hat{\alpha}}_2$ are the generalized cost coefficients for this dynamic case. Note that as the t_T approaches t_{max} (which is the case for constant demand), the two terms in the square root can be combined into one α_2 coefficient which results in the same functional form as (3.6). Examples of these new cost coefficients for dynamic transportation systems are shown in Table 4.1. Further details about how the cost coefficients relate to the physical components of travel are shown in Appendix B.3.

Table 4.1. Cost Coefficients for Peaked Systems ($\beta = 20$ \$/hour, $d = 3000$ m)

Mode	$\beta\hat{\alpha}_C$, $\beta\hat{\alpha}_1$ [\$/trip]	$\beta^2\hat{\alpha}_2$ [\$^{2}/m^2·hour·trip]	$\beta^2\hat{\hat{\alpha}}_2$ [\$^{2}/m^2·hour·trip]
Car	8.64		
BRT	6.09	0.009	0.021

4.2 Dynamics of Network Congestion

The Network Exit Function (NEF), introduced in Section 2.2.3, describes the rate that vehicles are able to exit the network as a function of the number of vehicles in

[1]By separating infrastructure costs from mode costs, we are essentially supposing that the provision of road space in a city has been determined *a priori*, and the problem remains to optimize the use of the existing space. To consider the effect of changing road investment, we simply consider the effect of changing the road space constraint R.

the network. If trip lengths do not change over time, and we assume that a stable NEF exists, then we can model delay using queuing diagrams as if the network were a FIFO bottleneck with state-dependent capacity. This is consistent with observations in Geroliminis & Daganzo (2008) and analysis of adaptive driver behavior (Daganzo *et al.*, 2011). Using the NEF, we can estimate the total delay in the system for any arrival pattern of cars entering the network. We use the NEF rather than a strict constraint on footprints to evaluate the dynamic congestion effects of vehicles interacting on the network.

An example NEF is shown in Figure 4.1, and this function defines the exit flow of cars depending on the number of vehicles in the network, $f = F(n)$. For the evening commute, we will define delay as the travel time exceeding the travel time at the traffic state with maximum exit flow (point M).[2] To track only the excess travel time (delay), it is useful to define $n_e = n_n - n_m$ when traffic states are on the congested branch of the NEF. Since n_m is a fixed quantity, we can express exit flow as a function of n_e by:

$$f = F_e(n_e) \tag{4.3}$$

where $F_e(n_e) \doteq F(n_m + n_e)$. The state of traffic in the network is defined at any time by the number of vehicles in it, because the excess accumulation determines the rate at which people leave.

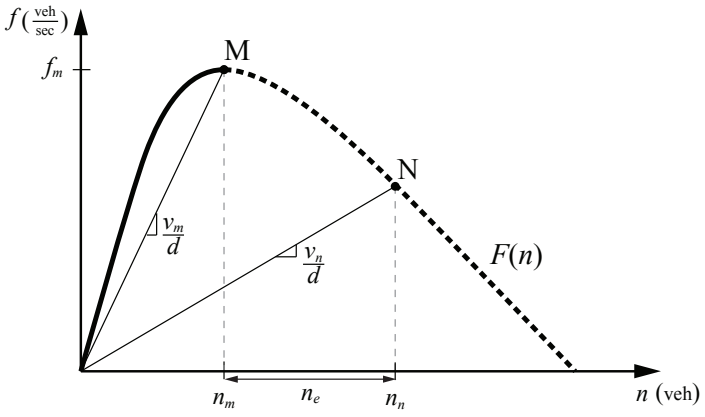


Figure 4.1. Network exit function showing congested traffic state N and excess vehicle accumulation

The excess accumulation of vehicles in the network can be shown graphically by plotting the cumulative number of vehicles that enter and exit the network over time. We will define two relevant cumulative curves: $A(t)$ is the cumulative number of vehicles that have entered or arrived in the network by time t, and $D(t)$ is the cumulative number of vehicles that have exited by t. The pattern of arriving vehicles

[2]In the morning commute, we will take a more general approach and allow a city to choose any uncongested state as a target operating traffic state. The restriction here to point M simplifies the analysis of the dynamics of congestion which, it will later be shown, should not occur in the morning commute system optimum.

is described by the demand. Since the demand is no longer a constant value, it is a function of time by $\lambda(t)$. Denoting the first derivative with respect to time with a dot, $\dot{A}(t) = \lambda(t)$ is the rate that vehicles enter the network. If these curves are plotted so that the vertical difference is the excess accumulation at time t, then $n_e(t) = A(t) - D(t)$.

If the vehicle accumulation is expressed as a function of time, $n(t)$, then the state of the network follows the mass conservation equation (Daganzo, 2007):

$$\frac{dn}{dt} = \dot{A}(t) - \dot{D}(t) \tag{4.4}$$

where $\dot{A}(t)$ is the rate that cars enter the network, and $\dot{D}(t)$ is the rate that they exit. Since the exit rate follows from the NEF, $\dot{D}(t) = F(n(t))$. We can use (4.3) to equivalently express this flow as a function of the excess number of vehicles, so n_e can be tracked as the state variable. The system delay is the total excess vehicle time spent in the system, and this is identified by integrating the excess accumulation over time.

For illustrative purposes, we will consider an evening peak characterized by a Z-shaped $A(t)$ which has a slope of 0 outside of the rush, and a slope of λ during a peak period which exceeds the network capacity. Therefore $A(t)$ satisfies:

$$\dot{A}(t) = \begin{cases} \lambda & t \in (0, t_p) \\ 0 & t \in (t_p, t_{max}) \end{cases} \tag{4.5}$$

where t_p is the end of the period of peak demand and t_{max} is the end of the total analysis period. The total demand over the time interval $(0, t_{max})$ is $N = A(t_{max})$.

This arrival curve is shown as a cumulative count in Figure 4.2. If the network fills quickly with cars at the start of the rush, then the excess accumulation will start to grow and the $D(t)$, whose slope is defined by the number of vehicles in the network, falls below $A(t)$. $D(t)$ can be estimated at all times by tracking the state of the network in short time intervals and using (4.3) to define the slope of the departure curve in each time step.

As n_e increases, the traffic state moves into the congested (right) side of the NEF, and the exit flow decreases. Therefore the slope, $\dot{D}(t)$ will diminish over the rush until the arrival rate $\dot{A}(t)$ drops at the end of the peak. If the vehicle accumulation ever reaches the jam value where the exit flow of the NEF is 0 at the far right, then accumulation can never diminish because $\dot{A}(t) \geq 0$ by definition as a cumulative count, and no vehicle can exit the network. This case results in permanent gridlock. If the queue does not clear by t_{max}, then the system will experience persistent congestion, and although demand is peaked, it will begin to behave more like a city with constant demand. In Figure 4.2, the excess number of vehicles eventually diminishes as $A(t)$ flattens, and $D(t)$ curves upward as vehicles exit the network and the congestion decreases. The total delay is simply the area between these curves, and we will call this total delay C.

The relationship between the road space available in a city and the NEF requires an assumed MFD. If the NEF is a scaling of the MFD based on the network size

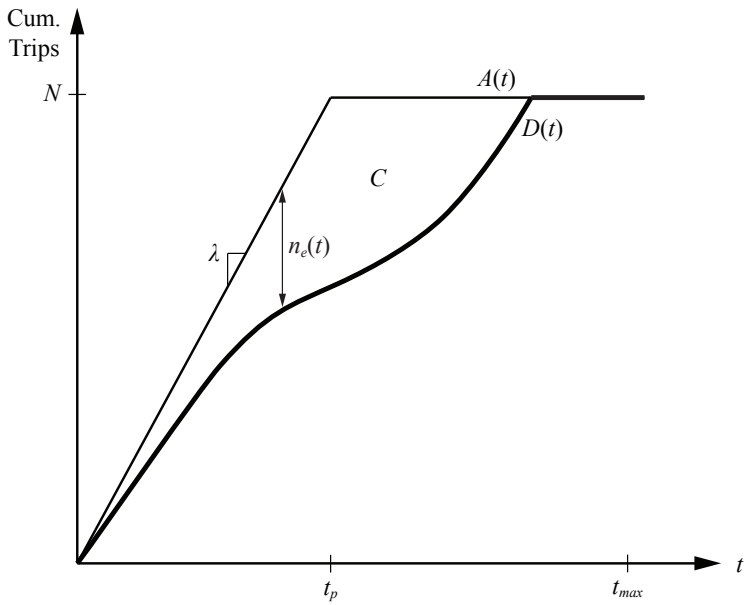


Figure 4.2. Queuing diagram for cars in a network in which exit flow declines with excess vehicle accumulation

and trip length (as described in Section 2.2.3), then its shape is related to R. By definition, $q_m = Q(k^*)$ is the maximum flow on the MFD. Then, using (2.7) and recognizing that the network length per area of city is $l = R/w$ (where w is the lane width), the maximum exit rate per area of city is:

$$f_m = \frac{Rq_m}{wd}. \tag{4.6}$$

When transit is operating, the NEF is assumed to scale as described in Section 2.2.3. A tilde (˜) indicates a value associated with the NEF while transit is operating; e.g., $\tilde{f}_m$ is the reduced capacity remaining for cars. In this way, the road space in a city is related to the dynamics of traffic states on the network. The demand density, of course, is described by $\lambda(t)$ which defines the slope of the arrival curve.

4.3 User Equilibrium

If the demand $\lambda(t)$ can be served by either cars or buses, we can separate the arrivals of travelers on each mode into two arrival curves, $A_C(t)$ for cars and $A_T(t)$ for transit, such that

$$A(t) = A_C(t) + A_T(t). \tag{4.7}$$

The problem is to determine how these modes will be split. In this section, the goal is to first understand the $A_C(t)$ curve which will arise at user equilibrium. Then, if we identify a system optimal $A_C(t)$, we can design a time-dependent pricing strategy

that is consistent with the Wardrop (1952) equilibrium conditions. We suppose for the evening commute that all queues are stored on streets in the network (rather than in garages), so all arriving cars force their way onto the streets and contribute to the total vehicle accumulation.

Recall that in a user equilibrium, each user will choose the mode with the lowest generalized cost (Wardrop, 1952). Therefore, if there is enough demand that serving all trips by transit makes the generalized cost of a transit trip, $z_T(N_T)$, less than the generalized cost of a free-flow car trip, z_C, everyone will choose transit. This tipping point defines the critical demand, N_{crit}, which is analogous to λ_{crit} for the city with constant demand. If $z_T(N) > z_C$, then people will choose to drive until the delays on the network make the cost of both choices equivalent. Then users will be indifferent between the two modes and both will be used simultaneously.

If multiple choices are used in equilibrium, they must have the same cost, so transit establishes an upper bound for the generalized cost of a car trip including delay. Therefore, the maximum delay for drivers that will be observed in user equilibrium, T, is:

$$T = z_T - z_C. \tag{4.8}$$

For this user equilibrium, we will assume that the transit service is given, so z_T is fixed. Then there is a unique user equilibrium travel pattern.

The user equilibrium is illustrated in Figure 4.3 for the Z-shaped $A(t)$ described by (4.5). At the beginning of the rush (point A in Figure 4.3), the delays are small, so the generalized cost of a car trip (including delay) is less than transit. The rush period begins with everyone choosing to drive; $A_C(t) = A(t)$. The traffic state will move down the congested branch of the NEF (see Figure 4.1) until the delay is equal to T. This will happen when the accumulation reaches n_n such that:

$$T = \frac{n_n}{F(n_n)} - \frac{d}{v_m}. \tag{4.9}$$

The time it takes to reach this point, t_B, depends on the shape of the NEF and the slope $\dot{A}_C(t)$ which determines how quickly the traffic state moves along the NEF. At point B transit becomes competitive, and users will choose to use it at a rate that maintains delay T. A transition must occur if transit takes up space and changes the NEF. Then, there will be additional congestion until n_n moves to the $\tilde{n}_n$ such that $T = \tilde{n}_n/\tilde{F}(\tilde{n}_n) - d/v_m$.

For the remainder of the rush, transit and cars will both be used. Cars will arrive at the same rate that they can be served, $\tilde{f}_n$, and the remaining demand, $\lambda - \tilde{f}_n$, will choose transit. The arrival curve for cars is therefore given by:

$$\dot{A}_C(t) = \begin{cases} \lambda & \text{for } t \in (0, t_B) \\ \tilde{f}_n & \text{for } t \in (t_B, t_p) \\ 0 & \text{otherwise} \end{cases}. \tag{4.10}$$

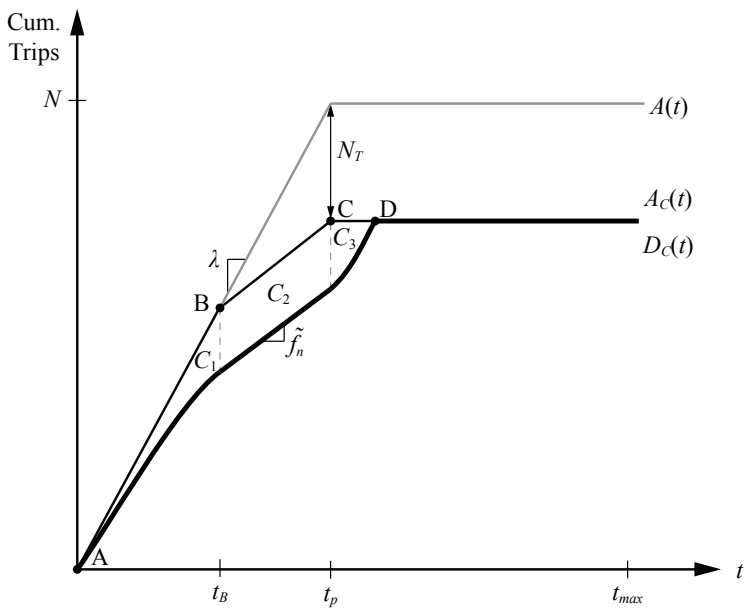


Figure 4.3. User equilibrium in the evening rush

The transit arrival curve is the complement following from (4.7), and it is given by:

$$\dot{A}_T(t) = \begin{cases} \lambda - \tilde{f}_n & \text{for } t \in (t_B, t_p) \\ 0 & \text{otherwise} \end{cases} . \tag{4.11}$$

This implies that the number of transit riders is $N_T = (t_p - t_B)(\lambda - \tilde{f}_n)$.[3]

In this user equilibrium, there are three phases listed below, each with a corresponding total delay as labeled in Figure 4.3.

Phase 1. Only cars are used in the interval $(0, t_B)$. The cost of each car trip is less than transit, and delays increase with each arriving commuter (segment $\overline{\text{AB}}$).

Phase 2. Cars and transit are used simultaneously in the interval (t_B, t_p) such that delay neither grows nor diminishes (segment $\overline{\text{BC}}$); the total delay associated with this phases is a parallelogram.

Phase 3. There are no more arrivals in the network after t_p (segment $\overline{\text{CD}}$); delay diminishes and the queues clear as vehicles exit the network.

As illustrated in Figure 4.3, the total delay is:

$$C = C_1 + C_2 + C_3 \tag{4.12}$$

where each of these delay components depends on $A_C(t)$ and the shape of the NEF. The user equilibrium is suboptimal, because the network becomes congested before transit is used, and then it remains congested throughout the rest of the rush.

[3]If the transit service were optimized for the number of users, then (4.8) and (4.9) can be used to define f_n and $\tilde{f}_n$ as a function of N_T. This requires knowing the function $F(n)$ which is the NEF.

4.4 System Optimum

In the system optimum, we face the same problem as presented in the preceding sections, except that we assume that the choice of mode can be controlled. The method for identifying the system optimum is to start by supposing that the number of transit users, N_T, is known, and then finding the optimal $A_C(t)$ and $A_T(t)$. Then, the system optimum is identified by choosing the value of $N_T^* \in [0, N]$ that minimizes the total system cost.

Minimum Generalized Cost for Given Transit Ridership

In the system optimum, the goal is to minimize the total system cost which includes the generalized costs of mode operations described by Z_C in (4.1) and Z_T in (4.2), as well as the total delay C. First an operating rule is proposed which minimizes the delay for cars, subject to a given number of transit riders. Suppose that the NEF has a maximum exit flow, $\tilde{f}_m$, when transit is operating, and that this traffic state is associated with the free-flow speed. This would be true of a triangular NEF, for example. With a more general concave NEF, points to the left of M may have shorter trip times. This possible trade-off between flow and travel time is addressed in detail in Chapter 5, Section 5.5.

Proposition 2. *Suppose that the following are given: a Z-shaped $A(t)$ with slope 0 outside of the rush and slope $\lambda > f_m$ for $t \in (0, t_p)$; network exit functions $F(n)$ and $\tilde{F}(n)$ where $\tilde{f}_m$ is associated with the free-flow speed; and N_T transit riders. The following strategy first minimizes the total system delay first, and then the generalized cost of the transit system (see Figure 4.4):*

If $N_T \geq t_p(\lambda - \tilde{f}_m)$:

$$\dot{A}_C(t) = \begin{cases} \lambda - \frac{N_T}{t_p} & \text{for } t \in (0, t_p) \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

$$\dot{A}_T(t) = \begin{cases} \frac{N_T}{t_p} & \text{for } t \in (0, t_p) \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

If $N_T < t_p(\lambda - \tilde{f}_m)$

$$\dot{A}_C(t) = \begin{cases} \tilde{f}_m & \text{for } t \in (0, t_T) \\ \lambda & \text{for } t \in (t_T, t_p) \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

$$\dot{A}_T(t) = \begin{cases} \lambda - \tilde{f}_m & \text{for } t \in (0, t_T) \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

where $t_T = N_T/(\lambda - \tilde{f}_m)$.

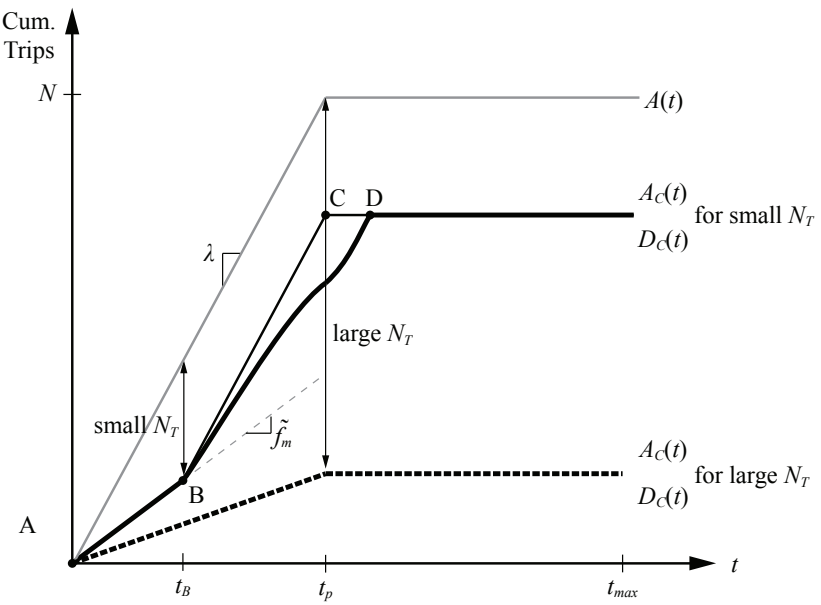


Figure 4.4. System optimum in the evening rush

Proof. The highest possible $A_C(t)$ during the rush interval $(0, t_p)$ that does not result in delays is $A_C(t) = \tilde{f}_m t$. By definition, $A_C(t_p) = \lambda t_p - N_T$. If $N_T \geq t_p(\lambda - \tilde{f}_m)$, then $A_C(t_p) \leq \tilde{f}_m t_p$. Therefore, it is possible to serve N_T without any delays for cars by running transit for the full interval $(0, t_p)$ so that $\dot{A}_C(t) \leq \tilde{f}_m$. The minimum capital investment for efficient transit service is associated with the steadiest arrival rate, because peaks in transit demand require additional vehicles for a short period in the rush. So, transit arrivals should be at a steady rate $\dot{A}_T(t) = N_T/t_p$ for $t \in (0, t_p)$. This establishes (4.14). Car arrivals must make up the difference which is $\dot{A}_C(t) = \lambda - N_T/t_p$ in the same interval, establishing (4.13).

If $N_T < t_p(\lambda - \tilde{f}_m)$, then delay cannot be avoided completely because $A_C(t_p) > \tilde{f}_m t_p$. There is no delay as long as $\dot{A}_C(t) \leq \tilde{f}_m$, and because $\tilde{f}_m$ is associated with the free-flow speed, there is no benefit for cars to arrive at a lower rate. Therefore, we should only consider arrival curves with rate satisfying:

$$A_C(t) \geq \tilde{f}_m t. \tag{4.17}$$

because delays will occur, and we seek to minimize them. The arrival curve for cars $A_C(t)$ is associated with the departure curve for cars $D_C(t)$ by the mass conservation equation, (4.4). In this case, the exit flow is given by $\tilde{F}_e(n_e)$ because transit is operating simultaneously. According to the NEF, $\dot{D}(t) \doteq \tilde{F}_e(n_e(t))$. $\tilde{F}_e(n_e)$ is a non-increasing function because the maximum exit flow is associated with the free-flow speed, so any slower speed (or greater n_e) cannot be associated with a greater exit flow.

Suppose that at some time t' in the rush, one and only one additional vehicle arrives and creates a new arrival curve denoted by hat, $\hat{A}_C(t') = A_C(t') + 1$.

Then, $\hat{A}_C(t) > A_C(t), \forall t > t'$, and the vehicle accumulation in the network at t' is $\hat{n}_e(t') = n_e(t') + 1$. The exit flow is non-increasing, so $F(\hat{n}_e(t')) \leq F(n)_e(t')$, which by definition implies that $\hat{D}_C(t') \leq D_C(t')$. Then at every subsequent time, the pattern perpetuates: $\hat{n}_e(t) > n_e(t)$ causing $F_e(\hat{n}_e(t)) < F_e(n_e(t))$ and $\hat{D}_C(t) < D_C(t)$ until the queue clears. Therefore, the additional vehicle arriving at time t' causes delays for every subsequent trip to be at least as great as before. This implies that the minimum delay impact is if additional trips are added as late as possible in the rush.

Let us start with the undelayed arrival curve $A_C(t) = \tilde{f}_m t$. Then to be in agreement with N_T total transit riders, all additional travelers must be served by car. The extra delay contributed by these car arrivals is minimized if they occur as late as possible in the rush. Since the arrival rate for cars cannot exceed the total rate of arrivals, $\dot{A}_C(t) = \lambda$ at the end of the rush, and the resulting patterns are given as (4.15) and (4.16). □

Once delay for cars is minimized, the total generalized cost of the transit system is minimized by serving transit riders at a constant rate. The transit cost model is a simplification of reality by assuming that transit service is an on/off system, and passengers ride transit at a constant rate while it operates. This approximation makes the cost function tractable, and it is near the optimum because a steady transit demand requires less capital investment than if there are additional peaks within the service period. The trade-off is that by not optimizing delay and transit simultaneously the operating interval may be slightly longer than the true optimum, but this effect is small because delays accumulate much faster than transit costs.

System Optimum Transit Ridership

Proposition 2 establishes a method to minimize the total cost of delay and the generalized cost of transit when N_T is given. These costs can be expressed as functions of N_T. The system optimum problem is to choose the value of N_T^* which minimizes the total generalized cost of the transportation system. $Z_T(N_T)$ is still a concave function, so it behaves similarly to one for the city with constant demand. Now, congestion could be present in system optimum, depending on its cost relative to the cost of the transit service required to avert it.

The delay is defined as a function $C(N_T)$, and it also depends on the demand rate λ and the shape of the NEF. Most realistic NEFs have a concave shape which makes an analytical expression for $C(N_T)$ difficult if not impossible to find. In practice, the delay can be estimated numerically by constructing arrival and departure curves governed by the exit function (4.3) and the mass conservation equation (4.4). The calculation procedure is the same as described for the user equilibrium in Section 4.3, except $A_C(t)$ is as defined by Proposition 2. The delay always increases as a convex function of the number of cars served in the peak, N_C.

Figure 4.5 shows the generalized car cost for different values of R.[4] This is the sum

[4]The curves in this figure were constructed using a simple triangular NEF, but the convex increasing shape will result from any concave NEF.

of the free-flow car cost and the total delay: $Z_C(N_C) + C(N_C)$, where the number of car trips is related to the number of transit trips by $N_C = \lambda t_p - N_T$. This is different from the constant demand case where a constraint on road space means that vehicles could either be served without congestion or not at all. The dynamic nature of peaked demand allows for congestion to emerge in the network when demand is high and decline when the demand subsides. Because networks can jam completely if the jam state at the bottom right of the NEF is reached, there are infeasible values of N_C for each R. Now, the amount of congestion in the network is a choice, and it depends on the road space available for the demand using it.

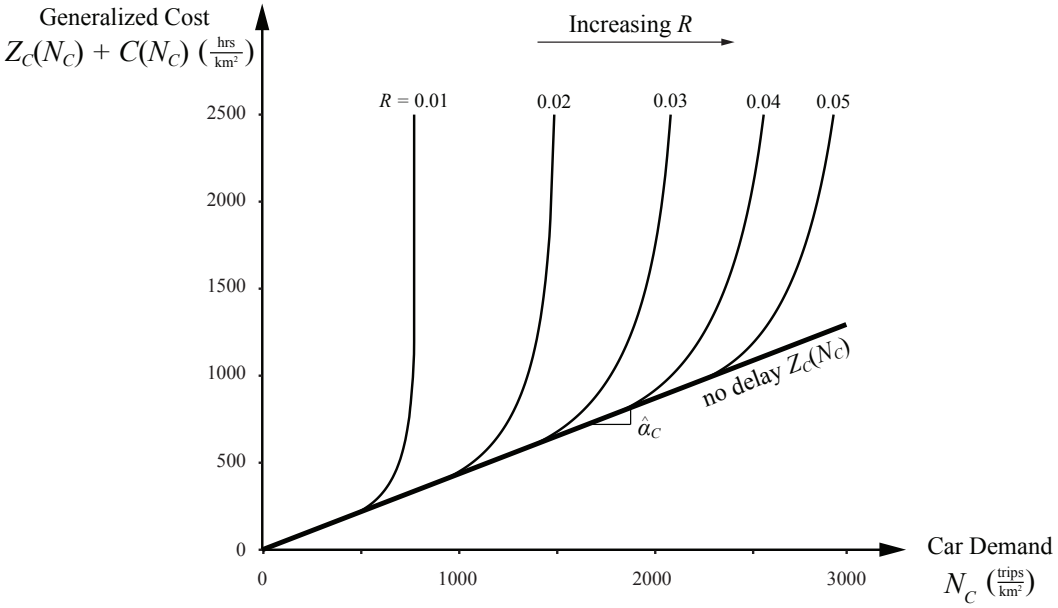


Figure 4.5. Total generalized car cost including delay ($t_p = 1$ hr)

In order to identify the system optimum N_T^*, it is useful to plot the total generalized cost as a function of the number of transit riders (as was done in Figures 3.4 and 3.8); an example result is shown in Figure 4.6. Demand is expressed as a total number of commuters N instead of the a rate λ. The effect of delay is clearly visible as shown for different values of R. This figure allows us to identify the solution for a range of R given the values of λ, t_p, and t_{max}. There are three possible system optimum solutions, and each is illustrated in Figure 4.6:

1. *All Car* – The minimum cost is achieved when all trips are served by car, so the lowest point is on the left axis. The entire function may be concave if R is large enough (Case A), or there may be a suboptimal local minimum with mixed modes if R is more restricted (Case B). In both cases illustrated, there is some limited congestion in the system optimum with all cars because the generalized cost functions do not intersect the left axis at 0.

2. *Mixed Modes* – The minimum costs is at a local minimum which is lower than the *all car* and *all transit* costs (Case C). Mixed modes is always associated with congestion because the local minimum must have its first derivative equal to 0, so it will not lie on the uncongested curve unless the total demand is N_{MC}. Also note that this is no longer a smooth transition as we observed for the city with constant demand.

3. *All Transit* – The minimum cost switches suddenly to the right axis. This can occur when there is another local minimum with mixed modes (Case D), or when R is so restrictive that costs are declining for all values of transit demand (Case E).

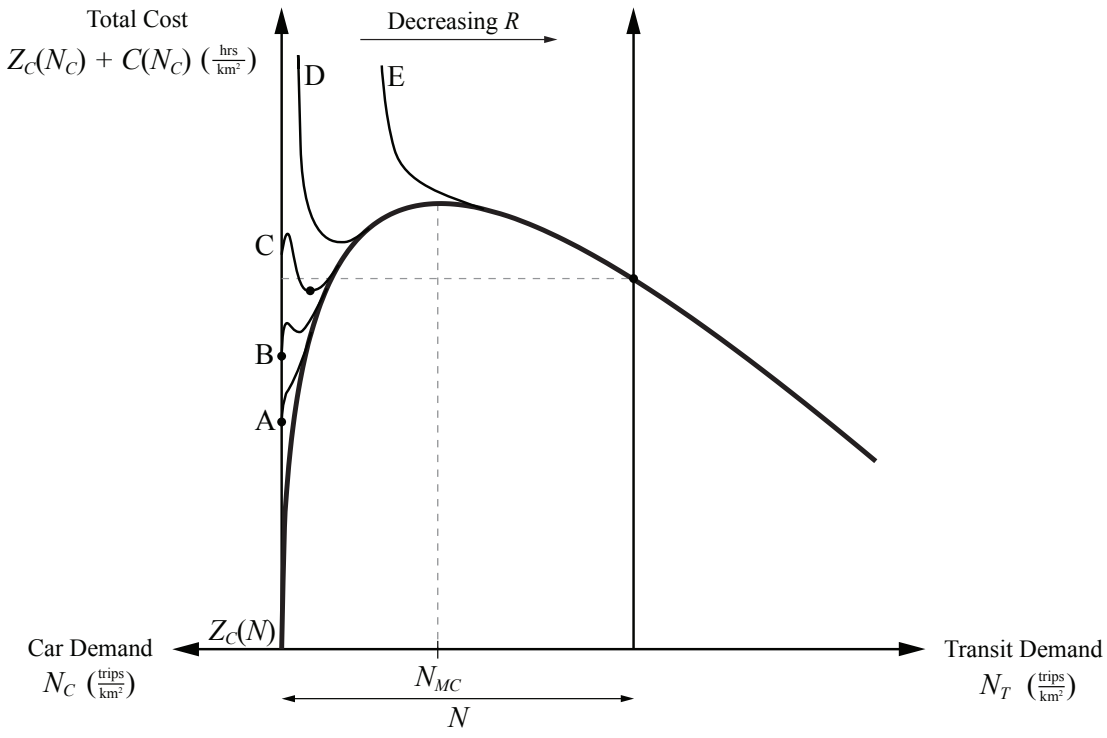
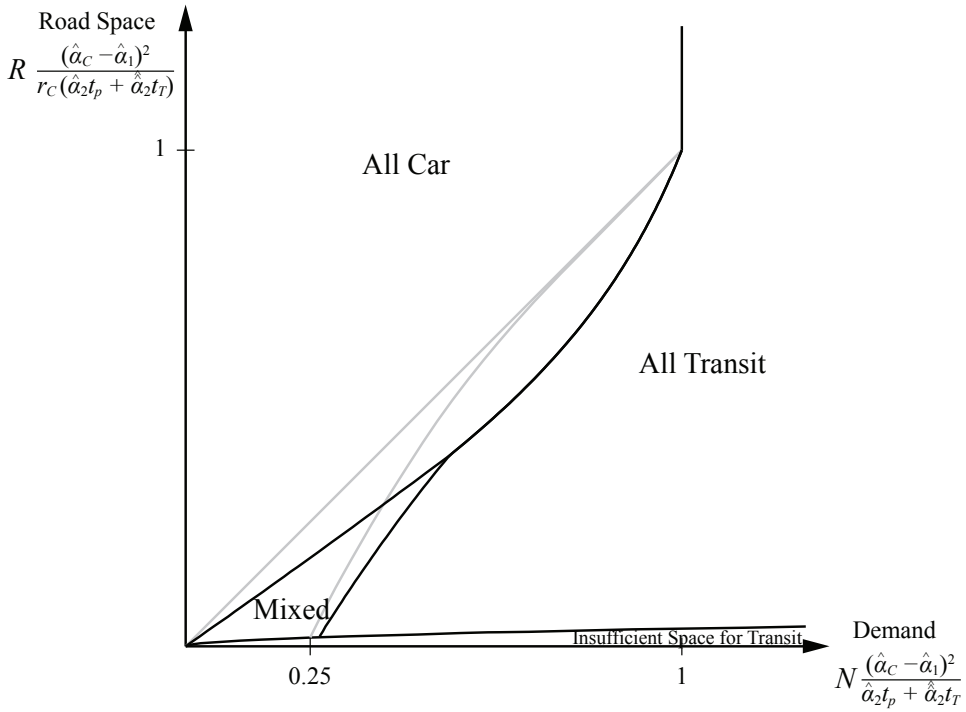
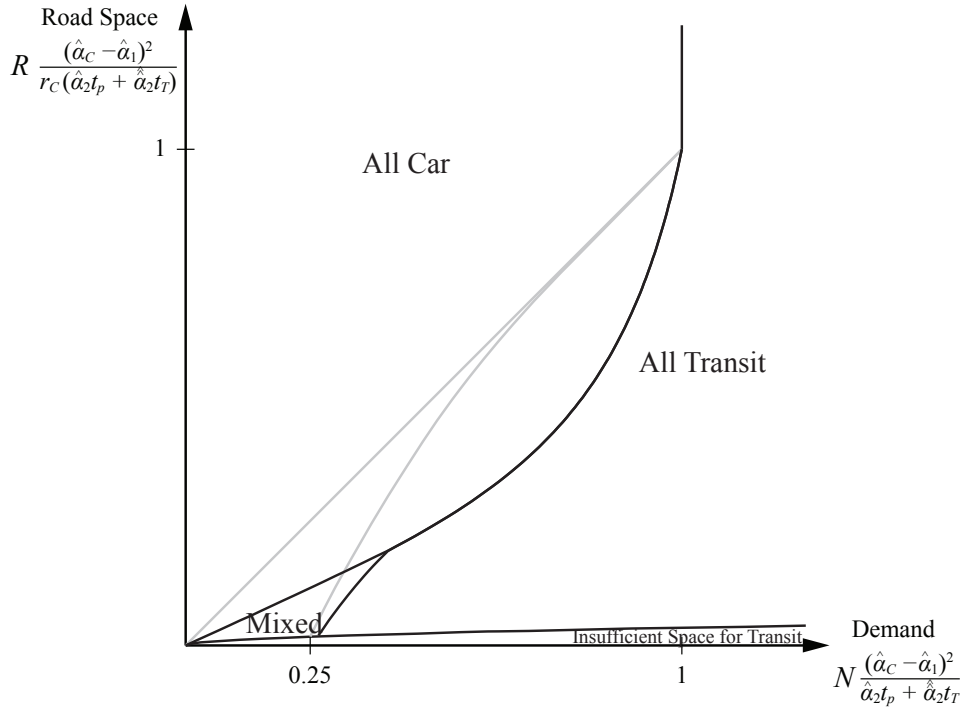


Figure 4.6. Total generalized cost of car and transit in the evening with congestion

The results are summarized on axes of scaled road space and demand, so that the system optimum solution for all possible city structures and input parameters is shown. In the evening rush, there is an additional variable of interest which is the peakedness of the rush (t_p compared to t_{max}). Two examples are shown in Figure 4.7 to show the effect of peaked demand. Notice that the shape is similar to the solution for the city with constant demand, resulting in the same three regimes. The difference is that peaked demand causes the edges delineating the regions to bow down and to the right, so there are more cities that should be served with all cars or a mix of modes. Below the diagonal, it is not possible to serve all cars without congestion. So, for some cities allowing congestion is optimal.



(a) Evening Peak ($t_p = 1$ hour, $t_{max} = 10$ hours)



(b) Evening Peak ($t_p = 0.5$ hour, $t_{max} = 10$ hours)

Figure 4.7. Summary of system optimum for cities with peaked evening demand served by cars and transit. Grey indicates no peak (constant demand) as a reference.

The effect of the peak becomes more pronounced as the peak gets shorter. This is intuitive, because an investment in transit capital and operating expenses is required in order to avoid congestion. If the peak is short, then the delays caused by limited road space are not very severe. When the congestion is very light, the investment in transit that will only be used for a short period of the day may not be worthwhile.

4.5 System Optimal Pricing

When the system optimum travel pattern involves using transit because of the road space constraint, users will not choose transit optimally in equilibrium. The system optimum can be achieved, however, by using pricing. Since the evening commute problem is posed so that users can choose which mode to use but not when they travel, the pricing strategy is very similar to the approach used in the city with constant demand (Section 3.4). At equilibrium, all modes that are used must have the same cost (Wardrop, 1952). Since the mode split changes over time, so should the pricing.

In the system optimum, the N_T^* transit riders should be served at the beginning of the rush as shown in Section 4.4 to prevent congestion from building up on the network. The user equilibrium differs, because congestion must develop and slow traffic before users are indifferent between the two modes. The optimal price at each time $\$(t)$ must make up the difference of the generalized cost per trip. So a toll for cars, a subsidy for transit, or some combination must make up the difference:

$$\$(t) = z_T(N_T^*, t_T^*) - z_C \quad \text{for } t \in (0, t_T^*) \tag{4.18}$$

where t_T^* is the duration of transit operations in the system optimum rush. There should be no congestion during this time period, so the price will be constant during the full time interval. This type of pricing can easily be implemented as static fares over the interval $(0, t_T^*)$ when transit service needs to be operated.

4.6 Summary of Findings

This chapter has shown how the cost and traffic models developed for an idealized city with constant demand can be extended to cities with peaked demand. Since traffic congestion on networks is a dynamic phenomenon, we make use of a dynamic model to track the evolution of traffic conditions over time.

This chapter has focused on a city with cars and a BRT network. For cases where the generalized cost of a transit trip exceeds the generalized free-flow cost of a car trip, the user equilibrium must become congested before transit becomes a competitive alternative. By letting the network become congested first, the total delays in the network increase substantially.

The system optimum solution is also identified for the evening, in which users can choose different modes but not arrival times in the network. The result is that

optimal transit use should occur at the beginning of the rush in order to prevent congestion from building up. The solution has been summarized graphically for the set of all cities and parameter inputs. As demand increases there are still 3 possible regimes: 1) *all car*, 2) *mixed modes*, and 3) *all transit*. The difference between the evening peak and constant demand is that the boundaries between the regimes are bowed downward and to the right which leaves more cities to be served by cars. The explanation for this is that with peaked demand, allowing a small amount of congestion is less costly than investing in a transit system to avert these delays. It is not surprising that as the peak gets concentrated (t_p gets much less than t_{max}), this bowing effect gets more pronounced and more cities should be served with car and suffer some congestion.

The evening commute problem suffers from depending heavily on the shape of the network exit function and operating in congested conditions which may be unstable. Some of these issues become less problematic in the morning commute problem. Chapter 5 addresses the more general morning commute problem in which user can choose when they travel in addition to their transport mode.

Chapter 5

Rush Hour City: Morning Peak

This chapter considers the morning commute for competing modes serving a population of travelers who are identical except for their wished travel time past a bottleneck. Whereas the evening peak problem works on the assumption that the arriving rate of users to the system is given, and they can only choose their mode, the morning commute recognizes that people also have a choice of when they travel. The results can be applied to urban networks modeled as bottlenecks whose capacity to serve vehicle trips declines as the network becomes crowded.

This chapter includes an analysis of the morning commute for a general S-shaped wish curve and a choice between passing a fixed-capacity bottleneck by car or using a general uncongestible alternative transit mode. Section 5.1 shows that the user equilibrium with a fixed capacity bottleneck is unique for a given transit service. Section 5.2 identifies the unique system optimal travel pattern which minimizes the social costs (or equivalently maximizes welfare). Section 5.3 presents a dynamic pricing strategy which moves the user equilibrium to system optimum. Qualitative insights are described. Finally, Section 5.5 shows that even though the user equilibrium for the network problem with state-dependent capacity is somewhat complex, it turns out that the system optimum version of the problem reduces to the fixed-capacity bottleneck model. Therefore, with suitably modified cost functions, the system optimal travel pattern, pricing strategies, and insights identified in Sections 5.2 and 5.3 apply to multimodal urban networks.

5.1 User Equilibrium

We first review the bottleneck model for a single mode from Hendrickson & Kocur (1981) and then add a transit mode. Consider the morning commute problem with a population of commuters who are identical (e.g., values of travel time and queuing delay) except for when they wish to get to their destination. If commuters drive, they must pass a bottleneck with capacity μ. The total number of commuters that wish to depart from the bottleneck by time t is described by a wish curve, $W(t)$, which is S-shaped. The slope of this curve is the time-derivative of $W(t)$ (denoted by a dot),

$\dot{W}(t)$, and it satisfies:

$$\begin{aligned} \dot{W}(t) &> \mu \quad \text{for } t \in (t_1, t_2) \\ \dot{W}(t) &\leq \mu \quad \text{otherwise} \end{aligned} \tag{5.1}$$

as shown in Figure 5.1. As a result of the first inequality, there will be a rush period starting at $t_e \leq t_1$ and ending at $t_L \geq t_2$ during which N commuters will experience queuing delay. Suppose that each commuter experiences a penalty for schedule deviation from their wished departure time which is described by a piecewise linear penalty function. Each minute of earliness is associated with a penalty of e equivalent minutes of travel time such that $0 < e < 1$, and each minute of lateness is equivalent to L minutes of travel time such that $L > 0$.

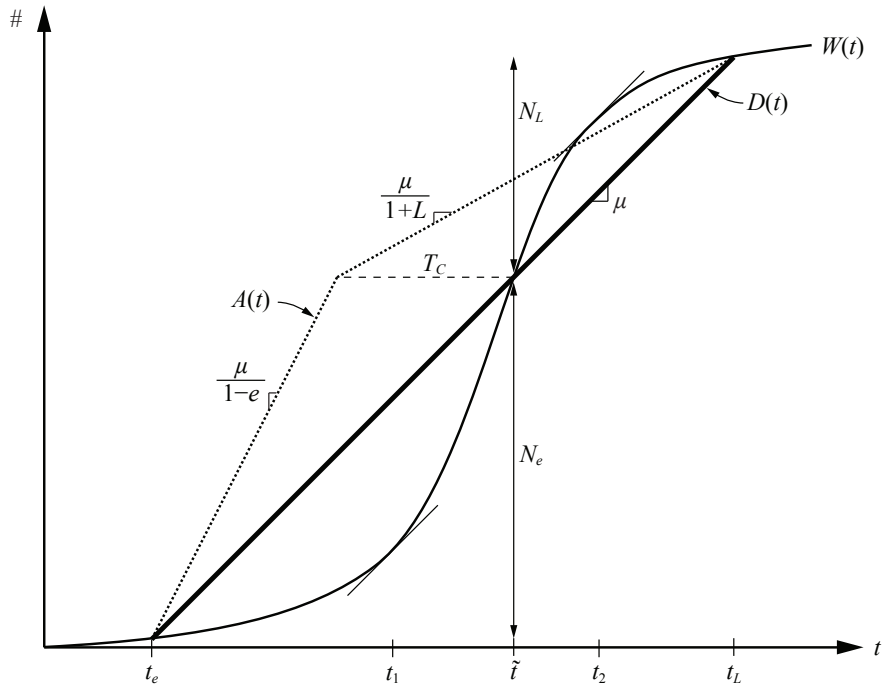


Figure 5.1. User equilibrium for a fixed capacity bottleneck using a single mode

In the absence of an alternative mode, and assuming that commuters arrive and pass the bottleneck in order of wished departure (first-wished, first-in, first-out or FWFIFO), we look for the beginning and end of the rush and for the equilibrium departure curve from the bottleneck which has slope $\dot{D}(t) = \mu$ for $t \in (t_e, t_L)$. This determines the time $\tilde{t}$ when a delayed commuter departs on time, as well as the number of commuters delayed by the bottleneck, N, N_e of which depart early and N_L depart late (see Figure 5.1). We also look for the user equilibrium arrival curve at the bottleneck, $A(t)$, which does not allow commuters to reduce their own travel costs by unilaterally changing their own arrival times. The slope of the arrival curve

in equilibrium, shown in the figure, must satisfy:

$$\dot{A}(t) = \begin{cases} \frac{\mu}{1-e} & \text{for commuters who depart early} \\ \frac{\mu}{1+L} & \text{for commuters who depart late.} \end{cases} \tag{5.2}$$

Otherwise, early and late commuters could reduce their travel costs by arriving earlier if the slope was greater or arriving later if the slope was less than those specified in (5.2). The result is that a critical commuter with wished time $\tilde{t}$ departs the bottleneck on time but experiences the maximum travel cost as queuing delay:

$$T_C = \frac{NeL}{\mu(e+L)}. \tag{5.3}$$

All travelers wishing to pass the bottleneck before $\tilde{t}$ are early in equilibrium, and all travelers wishing to pass after $\tilde{t}$ are late in equilibrium. Their excess costs (queuing and schedule penalty) are less than T_C.

If an alternative public transit mode becomes available, then commuters are able to choose when to travel and which mode to use. It is assumed in this section that the transit agency charges a fixed fare and operates on a fixed headway. Suppose that when transit is operating, it is fully segregated on its own lane so that transit service is not subject to traffic congestion. The transit system requires a fixed amount of dedicated space, so the bottleneck's remaining capacity to serve cars when both modes are operating is $\tilde{\mu} \leq \mu$. Transit users can always choose to pass the bottleneck at their wished times because use of the mode is not limited by congestion.[1] Therefore, given our assumptions, each transit rider has an identical generalized cost, z_T. This quantity and all costs appearing in this paper are expressed in units of equivalent queuing time (hours). A car trip without delay has a generalized cost of z_C (hours) which is independent of the number of car drivers. Thus, the total cost of driving through the bottleneck will be the sum of the cost of a free-flow trip and the excess costs of queuing delay and schedule penalty.

Following Wardrop (1952), it is assumed that at equilibrium each commuter chooses the mode and travel time which minimizes his or her own generalized cost. Transit will be competitive with the car for at least part of the rush hour if z_T is less than the generalized cost that the critical commuter would experience if transit is not provided: $z_C + T_C$. At equilibrium, the generalized cost of car and transit must be the same when both modes are used, and the generalized cost of a car trip cannot exceed that of a transit trip when only cars are used. Therefore, z_T is an upper bound for the cost of a trip by either mode. When competitive transit is provided, the maximum delay by car, T, satisfies:

$$T = z_T - z_C < T_C. \tag{5.4}$$

In order to distinguish between the travel patterns of cars and transit, we will consider the arrival and departure curves for each mode. Again, we assume FWFIFO

[1]This assumption is reasonable for a service using sufficiently large vehicles operated at regular headways but without a fixed schedule. The travelers cannot avoid the waiting time at a transit stop, but they can always board the next vehicle.

in both cases. $D_C(t)$ is the cumulative number of car departures at the bottleneck, and $A_C(t)$ is the cumulative number of car arrivals. $D_T(t)$ is the cumulative number of transit departures, and the arrival curve of transit is the same curve, $A_T(t) = D_T(t)$, because all transit trips can be completed on time.

An equilibrium is easy to find in two cases: if $z_T < z_C$, then a transit trip is less costly than even a free-flow car trip, and all trips will be made by transit; if $z_T > z_C + T_C$, then there is always a lower cost for traveling by car and the equilibrium will be the same as for the single mode problem. The following proposition addresses the remaining cases.

Proposition 3 (User Equilibrium, 2 Modes)**.** *If $W(t)$ is S-shaped, and each commuter can choose between traveling by car (with free-flow cost z_C per trip) through the bottleneck and an alternative transit mode with given cost per trip $z_T \in (z_C, z_C + T)$, there is a unique FWFIFO user equilibrium with the following properties (see Figure 5.2):*

1. *N_e, the number of early car commuters, is given by $N_e = \mu T/e$. They travel at the beginning of the rush, $t \in (t_e, \tilde{t}_e)$.*

2. *N_L, the number of late car commuters, is given by $N_L = \mu T/L$. They travel at the end of the rush, $t \in (\tilde{t}_L, t_L)$.*

3. *N_o, the number of on-time car commuters in the rush, is a strictly decreasing function of T, $N_o = N_o(T)$. They travel in the middle of the rush, $t \in (\tilde{t}_e, \tilde{t}_L)$.*

4. *N_T, the number of transit riders, is a strictly decreasing function of T, $N_T = N_T(T)$. They also travel in the middle of the rush, $t \in (\tilde{t}_e, \tilde{t}_L)$.*

Proof. Consider point A ($t = t_e$) where the first early commuter departs. Since the excess cost of driving (queuing delay and schedule penalty) is less than T shortly after this time, only cars are used and therefore $\dot{D}_C(t) = \mu$. For an equilibrium, the slope of the arrival curve for cars at the bottleneck should be as in Figure 5.2: $\dot{A}_C(t) = \mu/(1-e)$, so that queuing time increases at rate e/μ with each additional commuter. Clearly, the queuing time is T for the $N_e = \mu T/e$ early commuters in agreement with property 1. We choose the unique location of point A such that commuter N_e departs on time at $\tilde{t}_e$ (point B) as shown in the figure. This ensures that the excess cost of driving increases monotonically from 0 to T for commuters departing in $(t_e, \tilde{t}_e)$ in FWFIFO order. Therefore, no transit is used during the early interval. This establishes property 1.

A similar FWFIFO construction is used for the late part of the rush to identify the unique segment $\overline{\text{CD}}$ and the time interval $(\tilde{t}_L, t_L)$ where the excess cost of driving declines monotonically from T to 0. In this interval, queuing time declines at the rate L/μ with each departing commuter, so the number of late commuters is $N_L = \mu T/L$. This establishes property 2.

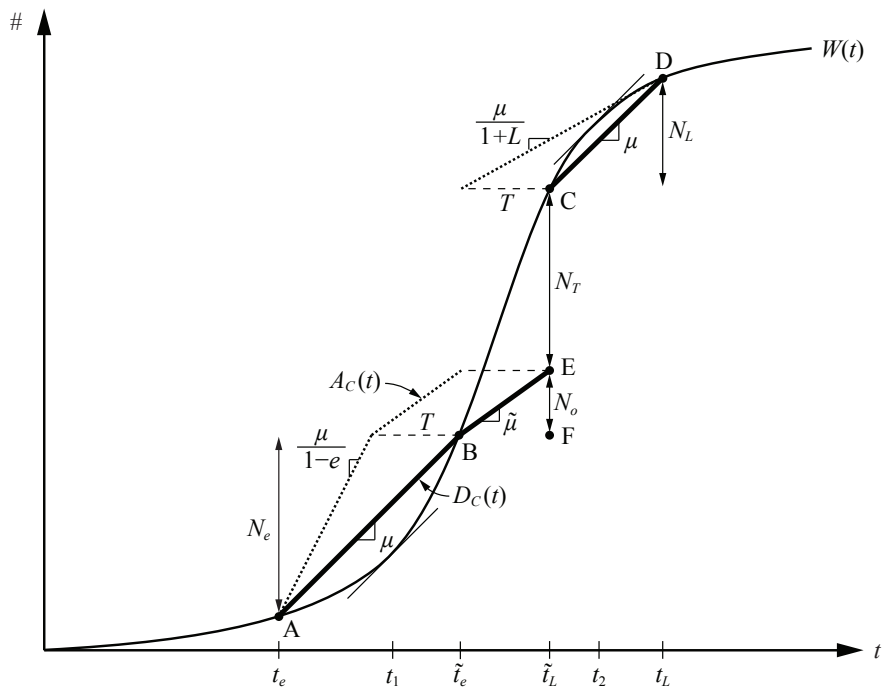


Figure 5.2. User Equilibrium for a bottleneck with a transit alternative

In the middle of the rush $(\tilde{t}_e, \tilde{t}_L)$, the number of commuters served is the length of segment $\overline{\text{FC}}$. Both cars and transit are used. Thus, cars depart in FWFIFO order at rate $\dot{D}(t) = \tilde{\mu}$. They experience a queuing delay T and no schedule penalty. Therefore, their number is $N_o = (\tilde{t}_L - \tilde{t}_e)\tilde{\mu}$ as shown by segment $\overline{\text{EF}}$. The number of transit users, N_T, is given by the length of segment $\overline{\text{CE}}$. They also pass the bottleneck in FWFIFO order, and experience no queuing or schedule delay. Note, N_T is always greater than 0 because $\dot{W}(t) > \mu \geq \tilde{\mu}$ in the middle of the rush.

Finally, note from the geometrical construction that if T increases, then $\tilde{t}_e$ increases and $\tilde{t}_L$ decreases; i.e., point B moves to the right along $W(t)$, and point C to the left. Clearly then, both N_o and N_T strictly decrease with T. This establishes properties 3 and 4. □

Note from Figure 5.2 that the departure curve for cars is piecewise linear in the rush with the slope always equal to the capacity of the bottleneck for cars. From Proposition 3, it follows that the number of commuters who depart early and late must satisfy:

$$\frac{N_e}{N_L} = \frac{L}{e}. \tag{5.5}$$

In this equilibrium, all commuters with wished times in $(\tilde{t}_e, \tilde{t}_L)$ travel on time and experience the same travel cost, $z_T = z_C + T$. The transit service is only used during this period. All early and late commuters travel only by car and experience a lower

cost. The total number of travelers in the rush is given by the sum:

$$N = N_e + N_L + N_o + N_T. \tag{5.6}$$

Each of these values, including N, is uniquely determined for any given $\{W(t), e, L, \mu, \tilde{\mu}\}$.

Note by comparing Figures 5.1 and 5.2 that the maximum cost of a trip in the two-mode equilibrium is less than that of a single-mode equilibrium. Since $N_T > 0$ implies $T < T_C$, it follows from properties 1 and 2 of Proposition 3 that there are fewer early and late commuters than for the single-mode equilibrium. These are represented by shorter segments $\overline{\text{AB}}$ and $\overline{\text{CD}}$ in Figure 5.2, which implies that the rush starts later and ends earlier with two modes than with all commuters traveling by car. Therefore, the rush period with multiple modes is shorter and involves fewer commuters. Provision of a competitive public transit alternative to congested driving is a Pareto improvement because every delayed commuter experiences a reduced travel cost, even those who travel by car at the beginning and end of the rush when no transit service is used.

5.2 System Optimum

The system optimal travel pattern will minimize the total system cost (or maximize welfare) associated with the bottleneck. Since queuing delay is an avoidable waste of time, $A_C(t)$ must equal $D_C(t)$ at system optimum. Thus, to find the system optimum, it suffices to identify the departure curves for cars and transit that minimize the monetary mode costs (e.g., vehicles, fuel, infrastructure, etc.), the free-flow travel time, and the schedule penalty.

In order to minimize the total system cost, we must consider the total generalized cost function of each mode. It is assumed that the transit spatial coverage is given, but its headway is chosen to minimize the sum of the agency and user costs (including the out-of-vehicle waiting time) for the given number of transit riders, N_T. Thus, the system optimum transit cost is a function of the number of transit users, $Z_T(N_T)$.[2] The system optimum problem is approached in two steps. First, we determine how car users and transit riders should behave if we are given that there is a total of N_T transit riders by the end of the peak period, t_{max}. The resulting costs are also determined. Then, the optimal number of transit riders, N_T^*, is identified by minimizing the system cost. All values associated with the system optimum are denoted with $*$.

To start, let us define the curve $W_L(t) \doteq W(t) - N_T$. This is a lower bound to $W_C(t)$, the number of car users that wish to depart the bottleneck by time t when there are N_T transit users. Logically, $W(t)$ is an upper bound for $W_C(t)$.

Proposition 4. *For a given wish curve, $W(t)$, and a given number of transit riders, N_T, there is a unique system optimal departure curve for cars, $D_C(t)$, and transit,*

[2] $Z_T(N_T)$ is a concave function that increases with $\sqrt{N_T}$ when the headway is determined endogenously to minimize the total generalized cost of the transit system as shown in Chapter 2.

$D_T(t)$. The $D_C(t)$ curve is piecewise linear with 3 segments going from $W(t)$ to $W_L(t)$ (see Figure 5.3):

Phase 1. $\dot{D}(t) = \mu$ while above $W(t)$, serving N_e^* trips (segment $\overline{AB}$); $W_C(t) = W(t)$; no transit is used.

Phase 2. $\dot{D}(t) = \tilde{\mu}$ from $W(t)$ to $W_L(t)$, serving N_o^* trips (segment $\overline{BC}$); $W_C(t) = D_C(t)$; and $D_T(t) = W(t) - W_C(t)$.

Phase 3. $\dot{D}(t) = \mu$ while below $W_L(t)$, serving N_L^* trips (segment $\overline{CD}$); $W_C(t) = W_L(t)$; no transit is used.

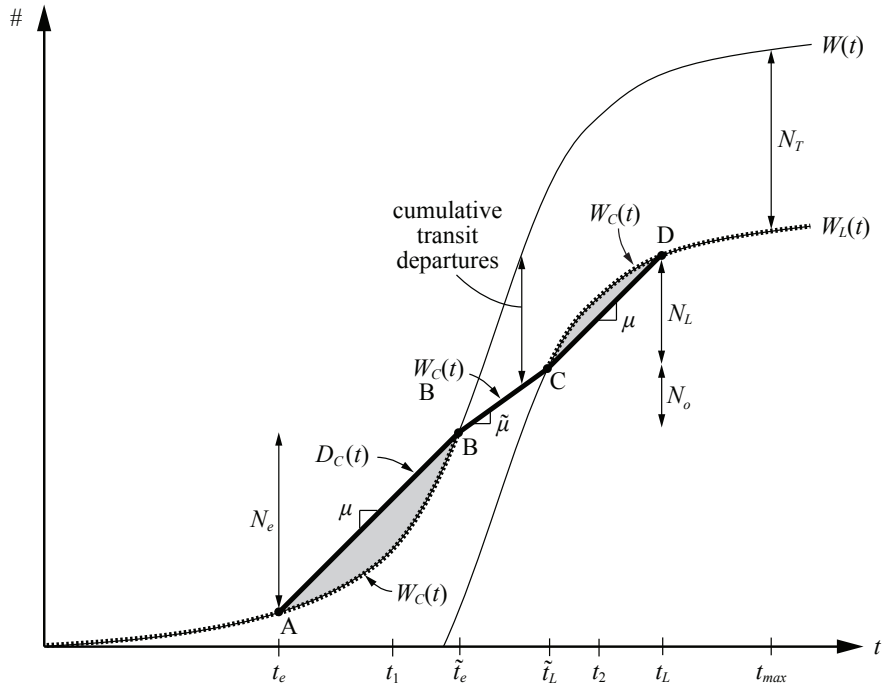


Figure 5.3. The departure curve for car that minimizes schedule delay given N_T

Proof. Since N_T is given, all monetary costs and free-flow travel times are fixed. Thus, the optimal departure curves must minimize only the remaining schedule delay for cars. In order to identify the optimal $D_C(t)$ and $D_T(t)$ we must also identify $W_C(t)$. This curve is bounded above by $W(t)$ and below by $W_L(t) = W(t) - N_T$ and must satisfy the following criteria (illustrated in Figure 5.3): $W_C(t)$ must start on $W(t)$ and end on $W_L(t)$; and for all t, $0 \leq \dot{W}_C(t) \leq \dot{W}(t)$.

We now show that there is a unique system optimal solution with the stated properties, as depicted in Figure 5.3. Consider the point B where the $W_C(t)$ diverges from $W(t)$. To the left of B, the schedule delay is minimized because $D_C(t)$ is as low as possible and $W_C(t)$ is as high as possible. Note, there is no transit use because

$W(t) - W_C(t) = 0$. Thus, for the given B, phase 1 is optimum. Likewise, to the right of point C where $W_C(t)$ joins $W_L(t)$, the schedule delay is minimized when $D_C(t)$ is as high as possible and $W_C(t)$ is as low as possible. There is also no transit because $\dot{W}_C(t) = \dot{W}(t)$. Thus for a given point C, phase 3 is optimum.

The schedule penalty can be made equal to 0 in phase 2 by choosing $W_C(t) = D_C(t)$. Therefore, for a given B, $D_C(t)$ should be chosen to minimize the schedule delay in phase 3. This is achieved by choosing the highest possible slope for $D_C(t)$. Note that $\dot{W}(t) > \mu$ in phase 2, so transit is used. Thus, $\dot{D}(t) = \tilde{\mu}$ in phase 2, as shown in Figure 5.3, and $D_T(t) = W_T(t) = W(t) - W_C(t)$ because all transit trips are served on time. Clearly, for a given B, there is a unique segment $\overline{\text{BC}}$ in phase 2, representing $W_C(t)$ and $D_C(t)$, that minimizes the schedule penalty.

Since point C is uniquely determined by point B, it only remains to pick the point B which corresponds to the minimum total schedule cost of earliness and lateness. An upward shift of B along $W(t)$ corresponds to a shift of the departure curve for early car commuters by dn_e and for late car commuters by dn_L (see Figure 5.4). Consideration of Figures 5.3 and 5.4 shows that this corresponds to an increase in earliness, $(eN_e/\mu)dn_e$, and a decrease in lateness, $(LN_L/\mu)dn_L$. Therefore, the schedule cost is minimized at the unique point when these two quantities (i.e., the shaded areas in Figure 5.4) are equal. This unique B defines the optimal solution. □

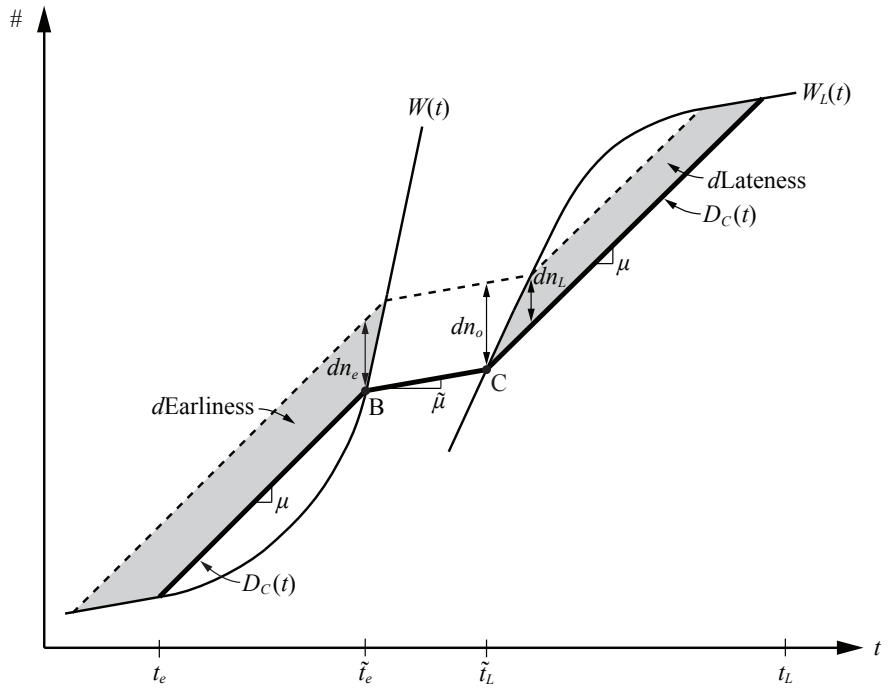


Figure 5.4. Decrease in earliness and increase in lateness resulting from a shift of B

Since Proposition 4 shows that there are unique pairs of wished curves and departure curves for cars and transit, given N_T, we can define three functions of N_T: the

total transit system cost, $Z_T(N_T)$; the total car cost, $Z_C(N_T)$; and the total schedule cost, $S(N_T)$. The value of N_T that minimizes the sum of these three functions is the system optimum transit ridership which we sought, N_T^*. We use N_e^* and N_L^* to denote the values obtained with the construction of Figures 5.3 and 5.4 for the system optimum N_T^*. We will also define $\lambda_e^* \doteq \dot{W}(\tilde{t}_e^*)$ and $\lambda_L^* \doteq \dot{W}(\tilde{t}_L^*)$, which are the slopes of $W(t)$ at the system optimum points B and C, respectively.

Proposition 5. *At system optimum, the wished curves and departure curves for cars and transit are such that:*

$$\frac{N_e^*}{N_L^*} = \frac{L(\lambda_e^* - \tilde{\mu})(\lambda_L^* - \mu)}{e(\lambda_e^* - \mu)(\lambda_L^* - \tilde{\mu})}. \tag{5.7}$$

Proof. Figure 5.4 shows that the effect of an incremental shift of B up and to the right along $W(t)$ is associated with shifting segment $\overline{\text{BC}}$ up by dn_o. This causes a upward shift of the departure curve for early car commuters by dn_e and for late car commuters by dn_L. Due to the geometry, these differentials are related by:

$$dn_o = dn_e \frac{\lambda_e - \tilde{\mu}}{\lambda_e - \mu} = dn_L \frac{\lambda_L - \tilde{\mu}}{\lambda_L - \mu}, \tag{5.8}$$

where $\lambda_e = \dot{W}(\tilde{t}_e)$ when the first on-time commuter departs the bottleneck and $\lambda_L = \dot{W}(\tilde{t}_L)$ when the last on-time commuter departs the bottleneck. At the system optimum, the schedule cost is minimized when the resulting change in total earliness balances the lateness:

$$\frac{eN_e}{\mu} dn_e = \frac{LN_L}{\mu} dn_L. \tag{5.9}$$

By manipulating (5.9) to express N_e/N_L in terms of dn_e and dn_L, then substituting expressions for these differentials from (5.8), it follows that the relative number of early and late commuters in the system optimum is:

$$\frac{N_e^*}{N_L^*} = \frac{L dn_L}{e dn_e} = \frac{L(\lambda_e^* - \tilde{\mu})(\lambda_L^* - \mu)}{e(\lambda_e^* - \mu)(\lambda_L^* - \tilde{\mu})}, \tag{5.10}$$

which establishes (5.7). □

If $W(t)$ is Z-shaped so that the demand is constant during the peak, then $\lambda_e^* = \lambda_L^*$, and $N_e^*/N_L^* = L/e$. Note that this not true in general, because λ_e^* may not equal λ_L^* (e.g., with the S-shaped $W(t)$ illustrated in Figure 5.4). Thus, the ratio of early to late commuters is different in the user equilibrium and system optimum. In order to identify the system optimum, it only remains to determine the optimal value of N_T^* which minimizes the total cost of the transportation system:

$$Z(N_T) = Z_T(N_T) + Z_C(N_T) + S(N_T). \tag{5.11}$$

Proposition 6 (System Optimum, 2 Modes)**.** *If $W(t)$ is S-shaped, and each commuter can be served either by car (with free-flow cost z_C per trip) through a bottleneck or using an alternative transit mode, then the system optimal travel pattern for both cars and transit is as in Figure 5.3, satisfies Proposition 5, and also satisfies:*

$$S'(N_T^*) = -\frac{eN_e^*(\lambda_e^* - \mu)}{2\mu(\lambda_e^* - \tilde{\mu})} - \frac{LN_L^*(\lambda_L^* - \mu)}{2\mu(\lambda_L^* - \tilde{\mu})}. \tag{5.12}$$

Proof. Refer to Figure 5.5, and note that an increase in transit ridership by dN_T results in a downward shift of $W_L(t)$ by dN_T. If instead we keep point C fixed, we obtain the lower departure curve corresponding to the segment $\overline{\mathrm{B}_e\mathrm{C}_e}$ (thinner curve). This maintains the same N_L before and after the shifts but N_e is decreased. If point B is not moved, we obtain the top departure curve corresponding to segment $\overline{\mathrm{B}_L\mathrm{C}_L}$ (thicker curve). This maintains the same N_e before and after the shift, but N_L is reduced. Obviously, neither of these solutions satisfy Proposition 5. The optimal departure curve (the dashed line in Figure 5.5) lies between these two extremes and corresponds to the segment $\overline{\mathrm{BC}}$ which balances N_e and N_L in the proper ratio as established by (5.10).

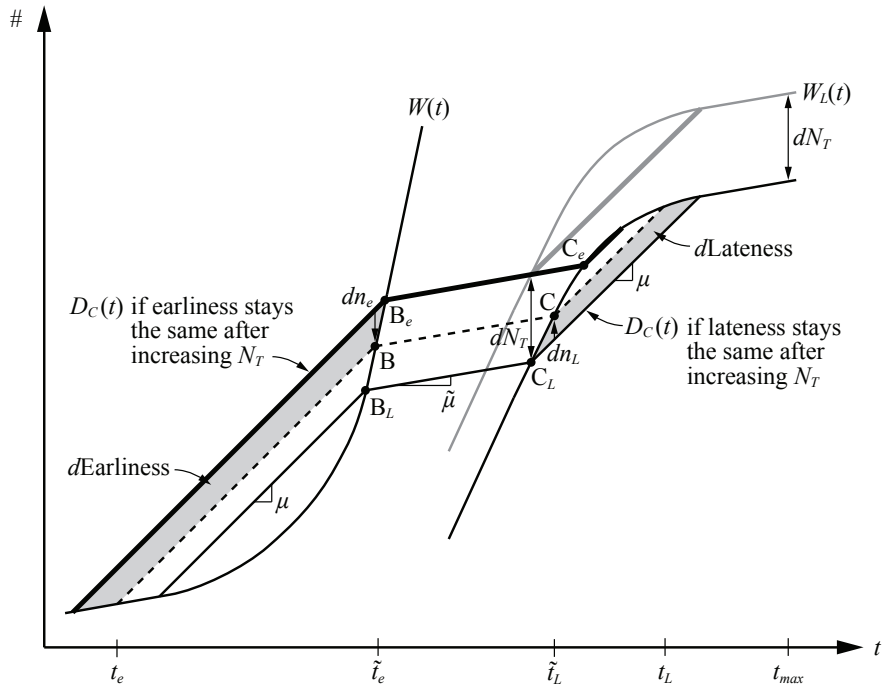


Figure 5.5. Change in earliness and lateness resulting from a change in N_T

The vertical difference between segments $\overline{\mathrm{B}_e\mathrm{C}_e}$ and $\overline{\mathrm{B}_L\mathrm{C}_L}$ is dN_T. Therefore, the sum of the vertical displacement between segments $\overline{\mathrm{B}_e\mathrm{C}_e}$ and $\overline{\mathrm{BC}}$ and between segments $\overline{\mathrm{BC}}$ and $\overline{\mathrm{B}_L\mathrm{C}_L}$ must be dN_T. Recall that (5.8) describes the relationship between a vertical shift of segment $\overline{\mathrm{BC}}$ by dn_o and the vertical shifts of the departure

curves for early commuters, dn_e, and late commuters, dn_L. Then using (5.8), the sum of vertical displacements can be expressed as:

$$dN_T = dn_e \frac{\lambda_e - \tilde{\mu}}{\lambda_e - \mu} + dn_L \frac{\lambda_L - \tilde{\mu}}{\lambda_L - \mu}. \tag{5.13}$$

where dn_e describes the shift from segment $\overline{\mathrm{B}_e\mathrm{C}_e}$ to segment $\overline{\mathrm{BC}}$ which contributes to the reduction in earliness, and dn_L describes the shift from segment $\overline{\mathrm{B}_L\mathrm{C}_L}$ to segment $\overline{\mathrm{BC}}$ which contributes to the reduction in lateness.

In order to maintain system optimum, the relative magnitudes of dn_e and dn_L must satisfy (5.9) so that the change in earliness is equal to the change in lateness. Using substitution from (5.10) into (5.13), we have:

$$\frac{dn_e}{dN_T} = \frac{\lambda_e - \mu}{2(\lambda_e - \tilde{\mu})} \tag{5.14}$$

$$\frac{dn_L}{dN_T} = \frac{\lambda_L - \mu}{2(\lambda_L - \tilde{\mu})}, \tag{5.15}$$

and the change in schedule delay with respect to N_T at the system optimum is:

$$\begin{aligned} S'(N_T^*) = \frac{dS(N_T^*)}{dN_T} &= \frac{d\text{Earliness}}{dn_e} \cdot \frac{dn_e}{dN_T} + \frac{d\text{Lateness}}{dn_L} \cdot \frac{dn_L}{dN_T} \\ &= -\frac{eN_e^*}{\mu} \cdot \frac{(\lambda_e - \mu)}{2(\lambda_e - \tilde{\mu})} - \frac{LN_L^*}{\mu} \cdot \frac{(\lambda_L - \mu)}{2(\lambda_L - \tilde{\mu})}. \end{aligned}$$

The first order necessary condition for optimization is given by setting the first derivative of (5.11) equal to zero and substituting $-z_C$ for $Z'_C(N_T)$. This gives the expression for $Z'(N_T^*)$ at system optimum.[3] □

Intuitively, the slope of the optimal departure curve for cars, $D_C(t)$, is always the bottleneck capacity during the rush period. Like the user equilibrium, when transit is being used, all car commuters should pass the bottleneck on time. The result is a piecewise linear departure curve for cars in the rush. Note, however, that the relative number of early and late commuters as expressed in (5.7) is not the same ratio as in (5.5) in the user equilibrium. Therefore, when cars and transit share road capacity, the system optimal departure curves are not the same as the user equilibrium departure curves. This is different from the result for a bottleneck serving a single mode in which case the system optimum is the same as the user equilibrium with delay removed by setting the arrival curve equal to the departure curve.

In order to identify the system optimum more concretely, we need to know the shape of the $W(t)$. Let us now consider the same Z-shaped demand that was analyzed in Chapter 4. For the morning commute, the demand defines $W(t)$ as shown in Figure 5.6.

[3]The second order necessary condition to minimize the cost requires that we calculate $d^2S(N_T^*)/{dN_T}^2$, so additional information is needed about the shape of $W(t)$.

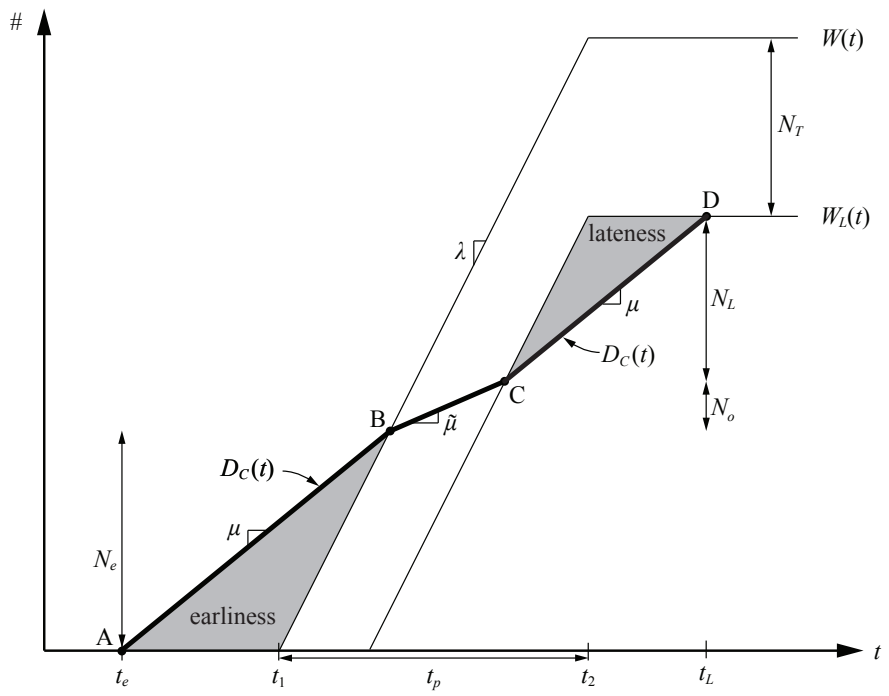


Figure 5.6. System optimal schedule delay for Z-shaped $W(t)$

Lemma 1. *If $W(t)$ is Z-shaped with slope λ during a peak of length t_p and 0 otherwise, then the schedule delay is given by:*

$$S(N_T) = \begin{cases} \left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right)^2 \frac{\lambda e L(\lambda - \mu)}{2\mu(e+L)} & \text{for } N_T < t_p(\lambda - \tilde{\mu}) \\ 0 & \text{otherwise.} \end{cases} \tag{5.16}$$

Proof. Consider Figure 5.6 illustrating the Z-shaped $W(t)$ and the optimal $D_C(t)$ for N_T as given by Proposition 4. In the middle of the rush, transit demand is $\lambda - \tilde{\mu}$, so to serve N_T commuters, transit is operated for a duration of time, $N_T/(\lambda - \tilde{\mu})$. All of the demand in the remaining time is served only by cars. This demand, $N_e + N_L$, is the difference between the total demand λt_p and the total demand in the middle of the rush $\lambda N_T/(\lambda - \tilde{\mu})$, i.e.:

$$N_e + N_L = \lambda \left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right). \tag{5.17}$$

Since the demand rate is always λ during the peak, then $\lambda_e^* = \lambda_L^* = \lambda$. According to Proposition 5, $N_e/N_L = L/e$ at system optimum. Substituting this ratio into (5.17), N_e and N_L are each defined by N_T as:

$$N_e = \frac{L}{e+L}\lambda \left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right) \tag{5.18}$$

$$N_L = \frac{e}{e+L}\lambda \left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right) \tag{5.19}$$

when $N_T < t_p(\lambda - \tilde{\mu})$. Otherwise, transit operates for the full duration of the rush and there are no early or late commuters.

The total earliness is the area between $D_C(t)$ and $W_C(t)$ when commuters depart early (the triangle below segment $\overline{\text{AB}}$): $N_e/2\,(N_e/\mu - N_e/\lambda)$. The cost of the earliness is the product of this area and e. Similarly, the total lateness is the area between $D_C(t)$ and $W_L(t)$ when commuters depart late (the triangle above $\overline{\text{CD}}$): $N_L/2\,(N_L/\mu - N_L/\lambda)$. The cost of the lateness is the product of this area and L. The sum of these two costs is the total schedule cost, S, and by simplifying we find:

$$S = \left(eN_e^2 + LN_L^2\right)\frac{\lambda - \mu}{2\lambda\mu}. \tag{5.20}$$

Now S is expressed in terms of N_e and N_L which are both functions of N_T. Substituting (5.18) and (5.19), the system optimal schedule cost is expressed as a function, $S(N_T)$. By simplifying the result, we obtain (5.16). □

Proposition 7. *If $W(t)$ is Z-shaped with slope λ during a peak of length t_p and 0 otherwise, then the total number of trips is $N = \lambda t_p$. The optimal transit ridership, N_T^*, is the value which minimizes:*

$$Z(N_T) = Z_T(N_T) + Z_C(N - N_T) + S(N_T). \tag{5.21}$$

N_T^* takes one of three possible values:

1. *$N_T^* = 0$; all trips served by car; then $Z = z_C N$.*
2. *$N_T^* = \lambda t_p = N$; all trips served by transit; then $Z = Z_T(N)$.*
3. *$N_T^* \in (0, t_p(\lambda - \tilde{\mu}))$; trips served by a mix of cars and transit; then $Z = Z(N_T^*)$. This happens only if $\exists N_T^*$ in the specified interval that satisfies:*

$$Z_T'(N_T^*) - z_C - \frac{eL}{\mu(e+L)}\left(t_p - \frac{N_T^*}{\lambda - \tilde{\mu}}\right) = 0 \tag{5.22}$$

$$Z_T''(N_T^*) + \frac{eL}{\mu(e+L)(\lambda - \tilde{\mu})} > 0. \tag{5.23}$$

Proof. The total cost is composed of three terms: $Z_T(N_T)$, $Z_C(N_T) = z_C(N - N_T)$, and $S(N_T)$. For real transit systems, $Z_T''(N_T)$ is negative but increasing because the total transit system costs increase with $\sqrt{N_T}$ when headways are optimized. Note from (5.16) that $S''(N_T)$ is constant when schedule delays exist (i.e., $N_T < t_p(\lambda - \tilde{\mu})$). Otherwise it is zero, and of course, $Z_C''(N_T) = 0$. Clearly then, the sum of these parts, $Z''(N_T)$, can cross 0 at most once while schedule delays exist, and it must be negative to left and positive to the right of the inflection point. For $N_T > \hat{N}_T$, $Z''(N_T) = Z_T''(N_T)$ which is negative. Thus, if there is range of N_T over which $Z(N_T)$ is convex, then it is contiguous and to the left of $t_p(\lambda - \tilde{\mu})$, and there can be at most one isolated local minimum. The first order optimality condition is expressed

explicitly for the Z-shaped wished curve by substituting $S'(N_T)$ into the relation, $Z'_T(N_T) + Z'_C(N_T) + S'(N_T) = 0$, and recognizing that $Z'_C(N_T) = -z_C$. The result is (5.22). A similar manipulation for the second order condition, $Z''_T(N_T) + Z''_C(N_T) + S''(N_T) > 0$, yields (5.23).

Note that N_T is bounded by the total demand λt_p. If $Z'(t_p(\lambda - \tilde{\mu})) < 0$ or $Z''(t_p(\lambda - \tilde{\mu})) < 0$, then there cannot be any local minimum with mixed modes, because $Z(N_T)$ is convex only for transit ridership exceeding that at which $Z''_T(N_T) = 0$ and less than $t_p(\lambda - \tilde{\mu})$. Thus, in these cases the optimal solution is a boundary point, either $N^*_T = 0$ (all car) or $N^*_T = \lambda t_p$ (all transit). Otherwise, there is a convex region in the feasible range of N_T which contains at most one unique local minimum, the optimal solution is either $N^*_T = 0$ (all car), $N^*_T = t_p\lambda$ (all transit), or the $N^*_T \in (0, t_p(\lambda - \tilde{\mu}))$ satisfying (5.22) and (5.23) (dual mode system optimum). □

For a Z-shaped $W(t)$, there is a unique system optimal solution for every demand and length of the peak. Following from Proposition 6, the optimal N^*_T is either with all trips served by car, all trips by transit, or the unique mix of modes satisfying the first and second order optimality conditions. Although $W(t)$ is not directly observable, if it is Z-shaped, $S(N_T)$ can be estimated from the observable values: N_T, μ, $\tilde{\mu}$, $\lambda - \tilde{\mu}$ (demand rate on transit), λt_p (total number of commuters). The values of e and L can be estimated from revealed preferences in equilibrium by measuring the rate at which delays increase and decrease over the rush. Thus, if the basic shape of $W(t)$ is assumed, all of the parameters required to identify system optimum can be estimated.

5.3 System Optimal Pricing

Now that the user equilibrium and system optimum have been identified for a bottleneck serving cars and transit, we will turn our attention to a pricing strategy that will achieve system optimal behavior in equilibrium. Commuters are assumed to choose when to travel and which mode to use based on the generalized cost of their own trip, the various components of which (travel time, vehicle costs, schedule delay, etc.) can be combined and expressed in units of equivalent queuing time. Since delay is time lost that cannot be otherwise productively used, an equivalent toll can be charged to road users so that they experience the same trip cost, but the delay is converted into money which can be redistributed back to society (Arnott *et al.*, 1990a). This section presents a time dependent pricing strategy for cars and transit which achieves system optimal (welfare maximizing) use of modes and infrastructure as identified in Section 5.2.

The system optimal pricing strategy adjusts commuter behavior so that $A_C(t) = D_C(t)$ which eliminates delays. Suppose that in the absence of pricing, the users of each mode must cover the mode's costs, so drivers pay z_C as a base rate and transit riders pay z_T. The pricing strategy will define the additional car toll $\$_C(t)$ and transit fare $\$_T(t)$. These prices are expressed in units of equivalent queuing time. Therefore,

the user cost of a free-flow car trip is $z_C + \$_C(t)$ (hours) and the user cost of a transit trip is $z_T + \$_T(t)$ (hours). A negative price represents a subsidy.

Proposition 8 (Optimal Prices)**.** *For any time-dependent car price satisfying*

$$\begin{aligned} \dot{\$}_C(t) &= e \quad \textit{for } t \in (t_e^*, \tilde{t}_e^*) \\ \dot{\$}_C(t) &= -L \quad \textit{for } t \in (\tilde{t}_L^*, t_L^*) \\ -L < \dot{\$}_C(t) &< e \quad \textit{otherwise,} \end{aligned} \tag{5.24}$$

the following time-dependent price for transit:

$$\$_T(t) = z_C - z_T + \$_C(t) \quad \textit{for } t \in (\tilde{t}_e^*, \tilde{t}_L^*), \tag{5.25}$$

achieves system optimum.

Proof. If an early driver departs dt later, then the schedule penalty for that trip is reduced by edt. We charge the driver an additional edt as a toll to cancel the benefit of departing later. Thus, the optimal toll must increase at rate e when commuters depart early in the system optimum, and by the same argument, the toll must decrease at rate $-L$ for commuters who depart late in the system optimum. Any commuter who departs on-time by car or transit will not have an incentive to change his or her own departure time if the change in cost is in $(-L, e)$.

The transit service is only used during the middle of the peak, $t \in (\tilde{t}_e^*, \tilde{t}_L^*)$, so its price must only be set for this interval. When car and transit are used simultaneously, the user cost of travel by both modes must be equal in order to maintain the Wardrop equilibrium; i.e., $\$_T(t) + z_T = \$_C(t) + z_C$. So, the price of transit which eliminates congestion and realizes the system optimal travel pattern is as in (5.25). □

Note that for the case without transit, the prices defined in Proposition 8 are the Vickrey (1969) prices which convert wasteful delay into toll revenue. Figure 5.7 illustrates optimal prices for a special case in which the car price is fixed at $\$_C(t) = \$_C^{\text{off-peak}}$ outside the rush. From the system optimum described in Section 5.2, N_e^* car commuters depart the bottleneck early at rate μ between t_e^* and $\tilde{t}_e^*$ (points A and B). Since the toll must increase at rate e during this interval, the car toll increases by $\Delta\$_e^* = eN_e^*/\mu$ from the first to the last early commuter. Likewise, the car toll decreases by $\Delta\$_L^* = LN_L^*/\mu$ for late commuters from $\tilde{t}_L^*$ to t_L^*. In the middle of the rush, $(\tilde{t}_e^*, \tilde{t}_L^*)$, all commuters are on time, so the optimal price can follow any curve from point B to C satisfying the third condition of (5.24) (e.g., the solid curve shown). Feasible prices are bounded by the dashed diamond. The system optimal price of transit is the same shape as $\$_C(t)$ translated down by $z_T - z_C$ as defined in (5.25).

Any vertical translation of the transit and car curves satisfies (5.24) and (5.25) and therefore will result in the same system optimal travel pattern. Thus, by shifting these prices up or down, it is possible to achieve additional policy objectives such as any particular car toll during the off-peak (including no toll) or revenue neutrality.

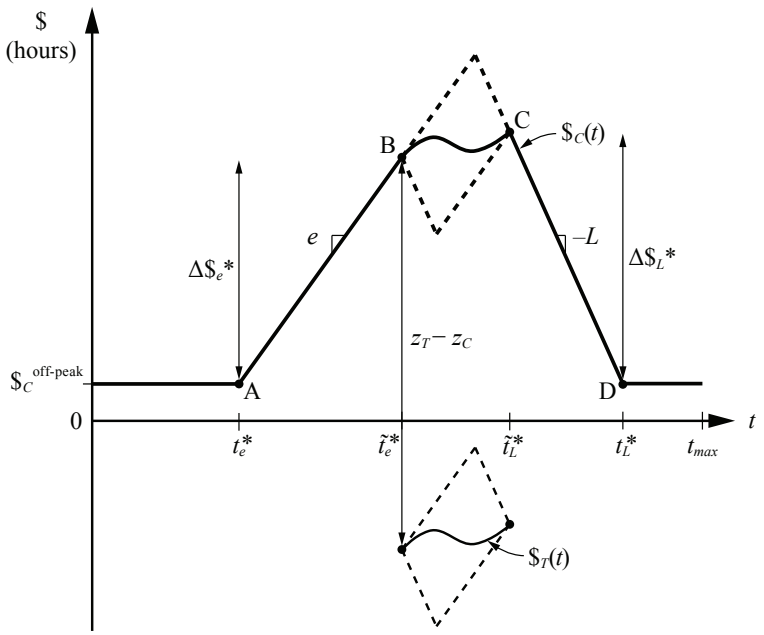


Figure 5.7. The system optimal time-dependent price for car and transit for the special case when the car toll is fixed at $\$_C(t) = \$_C^{\text{off-peak}}$ in the off-peak

From the system optimal cumulative departures of cars $D_C(t)$ and of transit $D_T(t)$, the net revenue $\$_{net}$ is given by:

$$\$_{net} = \int_0^{t_{max}} \$_C(t)\dot{D}_C(t) + \$_T(t)\dot{D}_T(t)dt, \tag{5.26}$$

where t_{max} is the amount of time until the next rush period. Since the car price can take any value in the off-peak, there is always a system optimal pricing strategy which is also revenue neutral.

5.4 Comparing Pricing Strategies for Bottlenecks

The optimal pricing strategy presented in Proposition 8 (Section 5.3) depends on the ability to charge a different price at every departure time from the bottleneck or the network. Although advancements in technologies are making fine time-dependent tolls increasingly feasible for bridges and congestion charge zones, this requires infrastructure investments and political support. In reality, tolling schemes are more limited. This is partly due to the difficulty of communicating and collecting tolls and fares which are changing in time and political resistance to dynamic congestion charges.

This section presents some alternative pricing strategies which are optimal for some real-world constraints. Without optimal time-dependent pricing, queues will

develop in the system, so the analysis for each of these strategies is only applicable to single roads with fixed-capacity bottlenecks. Although this is restrictive, the results are nevertheless applicable to facilities like bridges and tunnels. The insights are that there is a trade-off between efficiency and equity in pricing. More efficient strategies which eliminate more of the delay (and therefore reduce total system costs) also tend to result in bigger variation between the highest and lowest travel cost experienced.

The following sections present expressions for the total system cost Z and the difference between the lowest and highest user cost per trip which is either the maximum queuing delay T or the change in toll $\Delta\$$. The three pricing strategies compared are: $*$) optimal time-dependent prices, a) time-dependent transit prices with time-independent car prices, and b) time-independent prices for both cars and transit.

5.4.1 Optimal Time-Dependent Prices

The optimal prices which achieve the system optimum travel pattern for cars and transit are described in Proposition 8 and illustrated in Figure 5.7. The total system cost associated with this pricing scheme is the $Z(N_T^*)$ as given by (5.11) and the optimal departure curves described in Section 5.2.

The difference in car toll between the first early commuter who pays the minimum and the last early commuter who pays the maximum toll is $\Delta\$_e^*$ (as described in Section 5.3). Similarly, the difference in car toll between the first late commuter who pays the maximum and the last late commuter who pays the minimum is $\Delta\$_L^*$. Combine this result with Proposition 6, then express (5.12) in terms of N_e using (5.7). Substitute the result into the first derivative of (5.11), and then substitute $\Delta\$_e^*$ for eN_e/μ to get:

$$\Delta\$_e^* = (Z_T'(N_T^*) - z_C)\,\frac{\lambda_e - \tilde{\mu}}{\lambda_e - \mu}. \tag{5.27}$$

Use the same procedure, substituting N_L and then $\Delta\$_L^*$ to get:

$$\Delta\$_L^* = (Z_T'(N_T^*) - z_C)\,\frac{\lambda_L - \tilde{\mu}}{\lambda_L - \mu}. \tag{5.28}$$

Since (5.27) and (5.28) are the differences between car tolls for on-time travelers, the only other cost these travelers experience is the cost of a free-flow car trip. Therefore, $\Delta\$_e^*$ and $\Delta\$_L^*$ are the difference between the least and greatest user cost for early and late commuters respectively. Clearly, commuters experience different travel costs depending on their wished departure time. We can use these as a basis for comparing how equitable other pricing strategies are.

5.4.2 Time-Dependent Transit Prices, Time-Independent Car Price

An alternative pricing policy addresses the difficulty of dynamic pricing for individual car trips. Suppose that cars are priced at $\$_C$, but the transit fares can change over

time.[4] Recall that in equilibrium the user cost of all modes used must be the same, and from Proposition 8, the change in cost (as described by car price) must be bounded by $(-L, e)$ for on-time commuters. In Section 5.3, the argument was made that for any car price, the transit price must fluctuate according to (5.25).

Without time-dependent prices for cars, additional cost to drivers will be experienced as queuing delay which can be expressed as a function of time, $T(t)$. Therefore, we can re-write (5.25) to express the condition for equilibrium as:

$$\$_T(t) = z_C - z_T + \$_C + T(t). \tag{5.29}$$

For the period when both modes are used, and all commuters travel on-time, the equilibrium delay will depend on the price of transit.

By the same logic as presented for car tolls in Proposition 8, commuters will have no incentive to change their chosen departure time from the bottleneck as long as the transit price changes slowly. In Section 5.3 it was shown that the transit price should follow from the optimal car toll, but the optimal car toll can be identified following from the transit fare. Given a transit price satisfying $-L < \dot{\$}_T(t) < e$, the total cost per car trip must increase and decrease in the same way when both modes are used. This change in user cost is not a change in car toll but a change in the queuing delay at equilibrium. By changing the transit price slowly in the middle of the rush, commuters will choose the same departure time but with a varying amount of queuing delay. The consequence of this is that the transit price can be lowered during the peak to reduce delay.

Note that delay cannot be negative, so with fixed prices for cars, the transit price must hold constant when $T(t) = 0$. This is illustrated in Figure 5.8 for a case where the middle of the rush is long enough that delay can be completely eliminated for part of the rush as indicated by the section of flat cost between $\tilde{t}_e$ and $\tilde{t}_L$.

A superscript a denotes values associated with the optimal implementation of this pricing strategy. The price of transit at the beginning and end of the rush is given by:

$$\$_T(\tilde{t}_e) = z_C - z_T + \$_C + T_e^a \tag{5.30}$$

$$\$_T(\tilde{t}_L) = z_C - z_T + \$_C + T_L^a \tag{5.31}$$

where T_e^a and T_L^a are the queuing delays experienced by car commuters at $\tilde{t}_e$ and $\tilde{t}_L$, respectively. To minimize delay, the price should decrease at rate $\dot{\$}_T(t) = -L$ until $\$_T^{\text{peak}} = z_C - z_T + \$_C$, where it holds constant until increasing at the end of the rush at rate $\dot{\$}_T = e$ to satisfy (5.31). Thus, T_e^a is the maximum queuing delay experienced early in the rush, and T_L^a is the maximum queuing delay experienced late in the rush.

The total system cost when priced this way, Z^a, will now include a term for the queuing delay, so (5.11) becomes:

$$Z^a(N_T^a) = Z_T(N_T^a) + Z_C(N_T^a) + S(N_T^a) + C^a(N_T^a) \tag{5.32}$$

[4] Typically, transit services already charge users a fare. With the growing prevalence of automated fare collection and smart cards, it is not technically difficult to make the transit price time-dependent.

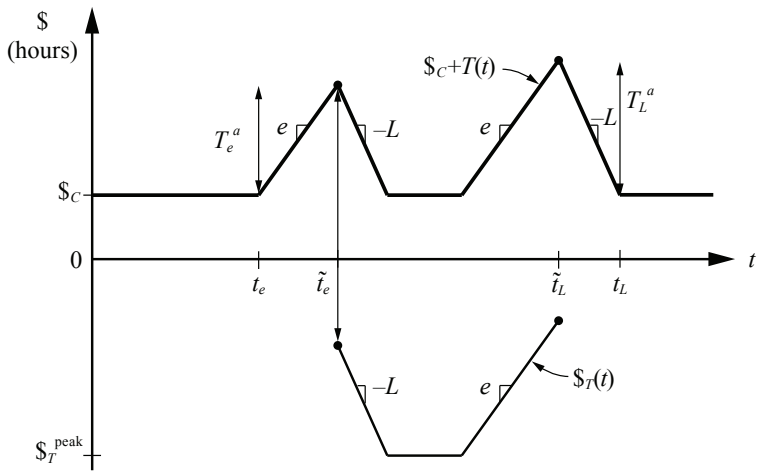


Figure 5.8. Time-dependent pricing for transit to minimize cost when the car toll is fixed at $\$_C$ at all times

where C^a is the queuing delay in equilibrium when transit is priced dynamically to minimize the delay as described above. Since the maximum delay at the beginning and end of the rush can be determined separately, the ratio of N_e and N_L can be fixed according to (5.7) in order to minimize schedule delay. Now, $T_e^a = eN_e^a/\mu$ and $T_L^a = LN_L^a/\mu$. Then, by taking the first derivative of (5.32) and substituting T_e^a and T_L^a into (5.12), he get:

$$T_e^a = (Z_T'(N_T^a) - z_C + C'^a(N_T^a))\,\frac{\lambda_e - \tilde{\mu}}{\lambda_e - \mu} \tag{5.33}$$

$$T_L^a = (Z_T'(N_T^a) - z_C + C'^a(N_T^a))\,\frac{\lambda_L - \tilde{\mu}}{\lambda_L - \mu}. \tag{5.34}$$

These values indicate the difference in user cost between car trips without delay and the maximum cost due to queuing delay experienced by the last early commuter and the first late commuter (see Figure 5.8). These magnitudes can be compared directly with (5.27) and (5.28) to assess the relative equity of this pricing strategy to the optimal time-dependent prices.

5.4.3 Time-Independent Pricing

Finally, we consider the most restrictive pricing policy in which only a fixed price can be applied to car and transit trips. Whether this is due to technical or physical constraints, time-independent pricing is the most common pricing strategy in cities. This includes the special case when car is unpriced ($\$_C = 0$). Since it will not be possible to price away congestion, prices must be picked to minimize the total cost including queuing delay that will result in equilibrium. We can add a fixed car toll

$\$_C$ to z_C and a transit fare $\$_T$ to z_T so that when both modes are used:

$$z_T + \$_T = z_C + \$_C + T \tag{5.35}$$

where T is the maximum queuing. Essentially, this is the user equilibrium problem presented in Section 5.1 where all drivers travel on time in the middle of the rush while both modes are used and experience the same queuing delay, T.

The cost function and optimal values for the time-independent pricing are denoted with a superscript b. The total cost function to be minimized is now:

$$Z^b(N_T^b) = Z_T(N_T^b) + Z_C(N_T^b) + S(N_T^b) + C(N_T^b) \tag{5.36}$$

where $C(N_T)$ is the total equilibrium queuing delay.

Since the prices for each mode are constant, the resulting equilibrium will be as illustrated in Figure 5.2 and the number of early and late commuters must satisfy (5.5). Therefore, N_e^b and N_L^b correspond to the same T^b as described by the correspondence between N_e, N_L, and T from Proposition 3. Then, the total cost is minimized by taking the first derivative of (5.36) and substituting T^b into (5.12):

$$T^b = \left(Z_T'(N_T^b) - z_C + C'(N_T^b)\right) \frac{2}{\frac{\lambda_e - \mu}{\lambda_e - \tilde{\mu}} + \frac{\lambda_L - \mu}{\lambda_L - \tilde{\mu}}}. \tag{5.37}$$

This maximum queuing delay is also the difference between the cost of an on-time free-flow car trip and the cost experienced by all on-time drivers in the middle of the rush. The relative equity of time-independent prices can be compared with the other pricing schemes by comparing T^b with the maximum delays that result from giving only transit time-dependent prices and the difference in optimal time-dependent car price.

5.4.4 Comparison of Equity and Efficiency

Three possible pricing strategies have been presented which minimize total system costs of the morning commute with varying degrees of restrictiveness. The optimal pricing strategy eliminates all queuing delays and achieves system optimum in cities where time-dependent prices can be charged to users of cars and transit. Two alternatives have also been presented which account for some real-world constraints but limit the amount of queuing delay which can be eliminated.

Now we can compare these strategies in terms of equity (the difference between the highest and lowest user cost of a trip) and efficiency (least total system cost). To evaluate equity, we only need to compare the difference in the cost of car trips, because the cost of a car trip equals the cost of a transit trip when both modes are used. Therefore, a comparison of equity is a comparison of the maximum change in car toll for early and late commuters ($\Delta\$_e^*$ and $\Delta\$_L^*$) with the maximum queuing delay for early and late commuters with time-dependent transit fares (T_e^a and T_L^a) and the maximum queuing delay with fixed prices (T^b). Comparing efficiency is even more

straightforward, because more efficient pricing strategies achieve lower total system costs, so this is simply a comparison of $Z(N_T^*)$, $Z^a(N_T^a)$, and $Z^b(N_T^b)$.

Before we can make these comparisons, we must first know something about the cost functions for each of the pricing strategies. Specifically, we need to know how the total queuing delay varies with the total number of transit riders, because this is needed to determine the change in queuing delay with optimal transit ridership. A proposition is presented which is useful for proving the relative efficiencies of the pricing strategies.

Proposition 9. *The total user equilibrium queuing delay, C, is a strictly decreasing function of the number of transit riders, $C = C(N_T)$, and $C'(N_T) < -T$. For the total queuing delay when car tolls are fixed and transit is charged an optimal time-dependent price, C^a, it is also true that $C^a(N_T) < C(N_T)$ and $C'^a(N_T) < 0$.*

Proof. From the geometry of the user equilibrium (see Figure 5.2), the total delay is the area between $A_C(t)$ and $D_C(t)$ which includes delay for early, late, and on-time drivers. At the beginning of the rush while drivers are early and queuing delay is growing, the total queuing delay is $\frac{1}{2}N_eT$, where $T = eN_e/\mu$ as described in Proposition 3. From this, a similar expression for queuing delay at the end of the rush can be expressed, and with N_o commuters experiencing $T = eL(N_e + N_L)/\mu(e + L)$, the total equilibrium queuing delay is:

$$C = \frac{eN_e^2}{2\mu} + \frac{LN_L^2}{2\mu} + \frac{eL(N_e + N_L)}{\mu(e + L)}N_o. \tag{5.38}$$

Following Proposition 3, $N_e(N_T)$, $N_L(N_T)$, and $N_o(N_T)$ are functions of N_T. By substituting T using the relations of Proposition 3, the change in total delay with respect to N_T can be written as:

$$C'(N_T) = T\left(N_e' + N_L' + N_o'\right) + \frac{eLN_o}{\mu(e + L)}\left(N_e' + N_L'\right). \tag{5.39}$$

From the S-shape of $W(t)$, a decrease in T makes the rush period indicated by the interval (t_e, t_L) start later and end sooner which corresponds to a reduced N. Therefore $N' \leq 0$, and from (5.6) it must be true that:

$$N_e' + N_L' + N_o' \leq -1. \tag{5.40}$$

Since Proposition 3 gives that N_o and N_T are both strictly decreasing functions of T, then $N_o' > 0$ and $N_e' + N_L' < -1$. Thus, $C'(N_T) < -T$.

Now, we turn our attention to the queuing delay associated with C^a. When the rush period is sufficiently long, then the delay occurs in two independent periods at the beginning and end of the rush (as illustrated in Figure 5.8). The pricing strategy actively reduces delay in the middle of the rush compared to the user equilibrium in which the queuing delay is constant while both modes are used. Therefore, it follows that $C^a(N_T) < C(N_T)$. The total queuing delay associated with the beginning of the

rush is $\frac{1}{2}N_eT_e(1+\tilde{\mu}/\mu)$ and similarly the total queuing delay at the end of the rush is $\frac{1}{2}N_eT_e(1+\tilde{\mu}/\mu)$. By substituting $T_e = eN_e/\mu$ and $T_L = LN_L/\mu$, the queuing delay is:

$$C^a = \frac{eN_e^2}{2\mu}\left(1+\frac{\tilde{\mu}}{\mu}\right) + \frac{LN_L^2}{2\mu}\left(1+\frac{\tilde{\mu}}{\mu}\right). \tag{5.41}$$

Since $N_e(N_T)$ and $N_L(N_T)$ are decreasing functions of N_T, and all of the other parameters are positive-valued, it follows that $C'^a(N_T) < 0$. □

We now consider how equity and efficiency compare in the most general case where $\tilde{\mu} = \mu$. Furthermore, we will assume that $Z'_T(N_T)$ does not increase as transit ridership grows. This is typically true for transit services which are designed to minimize the sum of user and agency costs; e.g., Holroyd (1967). The following results is true for typical cases where cars and transit systems share the same road space.

Proposition 10. *When $Z'_T(N_T)$ is non-increasing,*

$$T_e^a < \Delta\$_e^* \tag{5.42}$$
$$T_L^a < \Delta\$_L^* \tag{5.43}$$
$$T^b < \Delta\$_e^*, \Delta\$_L^* \tag{5.44}$$

and

$$Z(N_T^*) < Z^a(N_T^a), Z^b(N_T^b), \tag{5.45}$$

where $Z(N_T^*)$ is the total system cost function for system optimum as expressed in (5.11).

Proof. We begin by comparing the first two pricing strategies: optimal time-dependent prices and time dependent pricing only for transit. Let's start by supposing that $Z'_T(N_T^a) = Z'_T(N_T^*)$. Then,

$$Z'_T(N_T^a) - z_C + C'^a(N_T^a) < Z'_T(N_T^*) - z_C \tag{5.46}$$

which by comparing (5.33) to (5.27) and (5.34) to (5.28) implies that $T_e^a < \Delta\$_e^*$ and $T_L^a < \Delta\$_L^*$. Since N_T increases as T decreases, this also implies that $N_T^a > N_T^*$. $Z'_T(N_T)$ is non-increasing, so (5.46) holds, and (5.42) and (5.43) are true.

We now compare optimal time-dependent prices with fixed, time-independent prices. To simplify algebra, we define $\phi \doteq 2/\left(\frac{\lambda_e-\mu}{\lambda_e-\tilde{\mu}} + \frac{\lambda_L-\mu}{\lambda_L-\tilde{\mu}}\right)$. Note that $\phi > 0$. Proposition 9 shows that $C'(N_T^b) < -T^b$, so (5.37) gives us the relation:

$$T^b < \phi\left(Z'_T(N_T^b) - z_C - T^b\right)$$
$$T^b < \frac{\phi}{1+\phi}\left(Z'_T(N_T^b) - z_C\right) \tag{5.47}$$

where $\phi/(1+\phi) < 1$. Since $\tilde{\mu} \leq \mu$, it is clear that $(\lambda_e - \tilde{\mu})/(\lambda_e - \mu) > 1$ and $(\lambda_L - \tilde{\mu})/(\lambda_L - \mu) > 1$. Now, suppose that $Z'_T(N_T^b) = Z'_T(N_T^*)$. Then by comparing (5.47) to (5.27) and (5.28), it is true that

$$\frac{\phi}{1+\phi}\left(Z'_T(N_T^b) - z_C\right) < \frac{\lambda_e - \tilde{\mu}}{\lambda_e - \mu}\left(Z'_T(N_T^*) - z_C\right) \tag{5.48}$$

$$\frac{\phi}{1+\phi}\left(Z'_T(N_T^b) - z_C\right) < \frac{\lambda_L - \tilde{\mu}}{\lambda_L - \mu}\left(Z'_T(N_T^*) - z_C\right) \tag{5.49}$$

so it follows that $T^b < \Delta\$_e^*$ and $T^b < \Delta\$_L^*$. Since N_T increases as T decreases, then $N_T^b > N_T^*$. With a non-increasing $Z'_T(N_T)$, (5.48) and (5.49) still hold, and (5.44) is true.

Finally, it is clear that $Z(N_T^*) < Z^a(N_T^a)$ and $Z(N_T^*) < Z^b(N_T^b)$ because $Z(N_T^*)$ is the system optimal cost which is by definition the minimum system cost. □

The alternative pricing strategies result in lower maximum cost than optimal time-dependent pricing. The trade-off for more equitable prices is greater total system cost. For the same off-peak price of car, the alternative pricing strategies are associated with lower transit fares as indicated by (5.25), (5.30), (5.31), and (5.35). This means that greater subsidies for transit are justified with alternative pricing strategies than with optimal pricing. With a lower relative cost of transit compared to cars brought on by higher subsidies, correctly implemented alternative pricing should result in greater transit use. This is expected, because less of the queuing delay can be eliminated for car drivers, so total system costs are minimized by serving more trips on transit.

We can now gain some additional insights by considering the special case where the bottleneck capacity to serve cars is not affected by transit operations ($\tilde{\mu} = \mu$). This will be the case if transit uses a separate guideway, such as a rail system, or if bus lanes are dedicated all of the time. Now, we can make comparisons with the unpriced user equilibrium for which relevant values are denoted by a superscript *UE*. The total system cost of the user equilibrium is described by the same function as presented for fixed prices: $Z^b(N_T^{UE})$.

Proposition 11. *When $Z'_T(N_T)$ is non-increasing and $\tilde{\mu} = \mu$,*

$$\Delta\$_e^* = \Delta\$_L^* < T^{UE} \tag{5.50}$$

where T^{UE} is the maximum queuing delay in the unpriced user equilibrium, and

$$Z^a(N_T^a), Z^b(N_T^b) < Z^b(N_T^{UE}), \tag{5.51}$$

where $Z^b(N_T^{UE})$ is the total system cost in the unpriced user equilibrium.

Proof. When $\tilde{\mu} = \mu$, then (5.27) and (5.28) are equivalent, so:

$$\Delta\$_e^* = \Delta\$_L^* = Z'_T(N_T^*) - z_C. \tag{5.52}$$

Suppose that the user equilibrium had the same number of transit riders such that $N_T^* = N_T^{UE}$. The transit system exhibits economies of scale in that $z_T(N_T) > Z'_T(N_T)$ because $Z'_T(N_T)$ is non-increasing. Then from (5.4) we see that:

$$Z'_T(N_T^*) - z_C < z_T(N_T^{UE}) - z_C \tag{5.53}$$

which implies that $\Delta\$_e^*, \Delta\$_L^* < T^{UE}$. Since T decreases as N_T increases (Proposition 3), $N_T^* > N_T^{UE}$. So, $Z'_T(N_T^*) < Z'_T(N_T^{UE})$ which still satisfies (5.53). Therefore (5.50) is true.

Now, let's turn our attention to the total system cost with sub-optimal pricing and the unpriced equilibrium. The comparison between $Z^b(N_T^b)$ and $Z^b(N_T^{UE})$ is straightforward because the time-independent pricing problem is simply to minimize the user equilibrium cost function. We know from (5.44) and (5.50) that $T^b < T^{UE}$, so it must also be true that $N_T^b > N_T^{UE}$. Since N_T^b minimizes $Z^b(N_T)$, any other value (such as N_T^{UE}) must be associated with a greater total system cost, so $Z^b(N_T^b) < Z^b(N_T^{UE})$.

Finally, we compare $Z^a(N_T^a)$ and $Z^b(N_T^{UE})$. We know from (5.42) and (5.43) and (5.50) that $T_e^a, T_L^a < T^{UE}$, and $N_T^a > N_T^{UE}$ because T decreases with increased N_T. Since N_T^a minimizes $Z^a(N_T)$, any other value must be associated with a greater total system cost. Therefore, $Z^a(N_T^a) < Z^a(N_T^{UE})$. We can also see that the only difference between (5.32) and (5.36) is the queuing delay term. Proposition 9 states that $C^a(N_T) < C(N_T)$, so $Z^a(N_T^{UE}) < Z^b(N_T^{UE})$, and it must be true that $Z^a(N_T^a) < Z^b(N_T^{UE})$. □

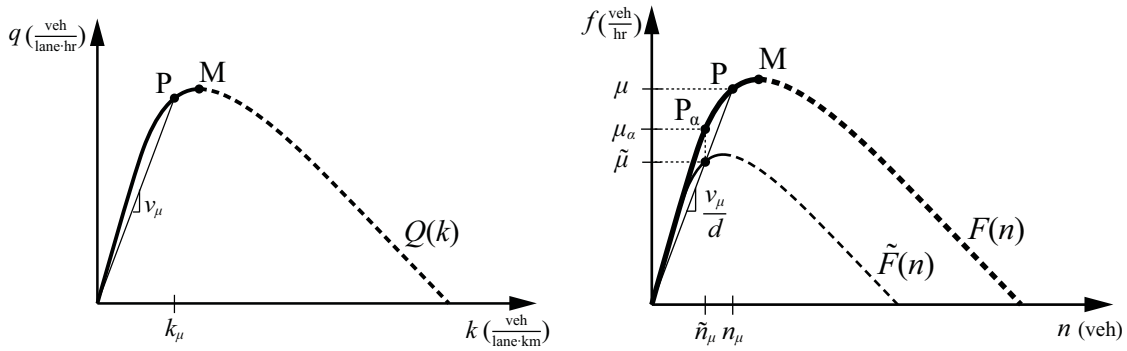
In the special case that $\tilde{\mu} = \mu$, all the pricing strategies reduce the maximum user cost compared to the unpriced user equilibrium. The total system cost for all of the pricing strategies presented is also lower than without pricing, so it is clear that it is more efficient and more equitable to use prices to change commuter behavior. Even if optimal time-dependent prices cannot be charged, there are benefits for equity and efficiency when using alternative pricing strategies including charging fixed tolls and transit fares.

5.5 Urban Networks with Competing Modes

The previous sections have shown the user equilibrium and system optimum for two competing modes using a single bottleneck with fixed capacity. Here, we extend the results to urban networks. A bottleneck on a road will discharge vehicles at fixed capacity as long as there are vehicles in a queue feeding it.[5] However, the capacity of an urban network to discharge vehicles to their destinations depends on the number of vehicles circulating in the network. Unlike a bottleneck on a single road, queues of vehicles in a network tend to block other streets and impede network flow.

[5]This is approximately true, although evidence suggests that the queue discharge rate is reduced when queues grow very long (Koshi *et al.*, 1992).

Recent work identifying a consistent Macroscopic Fundamental Diagram (MFD) that relates flow to vehicle density on a network is discussed in detail in Section 2.2.1. The connection between the MFD and a similarly consistent relationship between the rate that cars exit the network and number of vehicles in it called the Network Exit Function (NEF) is described in Section 2.2.3. Using these tools, we will study this system assuming that the instantaneous exit flow depends only on the number of vehicles in the network at that time. In this way, a vehicle exiting a network is analogous to a vehicle departing a bottleneck, so we can think of the network as a bottleneck with the state-dependent capacity given by the exit function. Figure 5.9 shows a general concave MFD and two NEFs associated with it: $F(n)$ when there is no transit, and $\tilde{F}(n)$ when transit is operating.



(a) Macroscopic Fundamental Diagram (MFD) (b) Network Exit Function (NEF)

Figure 5.9. The MFD and NEF are shown for a network with and without transit. Dashed lines indicate congested traffic states, and solid lines are uncongested states where a target operating state P may be reasonably chosen.

The maximum feasible exiting flow is associated with point M in Figure 5.9. For a given traffic state on the MFD (such as point P), the slope from the origin represents the average vehicle speed across the network, v_μ, which includes time spent at signals and in queues. The total time required to complete a trip of length d is the reciprocal of the analogous slope on the NEF, v_μ/d. Traffic states to the right of M (dashed lines in Figure 5.9) are congested and should always be avoided because the same network and exit flow can be achieved with greater traffic speeds and fewer vehicles on the road to the left of M (solid lines in Figure 5.9).

Geroliminis & Levinson (2009) provides a numerical method to construct the user equilibrium for a single mode on a network with a stable, single-valued NEF. The user equilibrium problem in networks is complicated by the reduced exit flow when the network becomes congested. Fortunately, the system optimal network problem is not affected by this complication because only uncongested traffic states (to the left of M) should occur. Geroliminis & Levinson (2009) also presents the system optimum and optimal pricing for a network with a single mode taking advantage of this result. Now, by keeping the traffic states only on the uncongested side of the NEF, we look

at how the morning commute problem with two modes applies to the network system optimum.

Although point M corresponds to the maximum feasible exit flow, a city could choose to put a limit on exit flow by capping it at a target exit flow μ associated with point P to the left of M. A lower target exit flow lengthens the rush but serves each vehicle with less travel time. Figure 5.9(b) shows that at point P, μ is associated with a critical accumulation of vehicles on the network, n_μ, such that $\mu = F(n_\mu)$. We will define delay as the excess travel time over d/v_μ for a trip of length d. So, in system optimum where delays are avoided, d/v_μ can be interpreted as the maximum travel time guarantee.

If the transit service uses a separate right of way and has no impact on the street network (e.g., metro or permanent dedicated bus lanes), then $F(n)$ does not change, and $\tilde{\mu} = \mu$. By applying system optimal pricing as described in Section 5.3, car commuters will choose to travel at rate μ, so the network will maintain a steady accumulation of n_μ vehicles. No delay will be experienced.

In reality, transit services often use the same street space as other vehicles, so deploying buses will reduce the remaining capacity available for cars. This effect on the NEF is represented by (2.8) in Section 2.2.3. Examples of NEFs with and without transit operations are shown in Figure 5.9(b). Note that the point P associated with the target exit flow moves along the ray with slope v_μ/d towards the origin so the travel time per trip does not change. This peak is associated with the same density k_μ as before, so the optimal car accumulation when both modes are operating, $\tilde{n}_\mu$, and the exit flow (capacity) for cars, $\tilde{\mu}$, are given by:

$$\tilde{n}_\mu = \alpha n_\mu \tag{5.54}$$

$$\tilde{\mu} = \alpha \mu. \tag{5.55}$$

Note that $k_{\tilde{\mu}} = k_\mu$, because the network is managed to operate at the same point P on the MFD (Figure 5.9(a)) with and without transit operations. Expressions (5.54) and (5.55) describe the traffic state for cars when transit and cars are operating together on the network in the middle of the rush. This is shown in Figure 5.10 by the slope of the departure curve for cars exiting the network in the middle of the rush (segment $\overline{\text{BC}}$).

The procedure for identifying the system optimum is the same described in Section 5.2. Conditional on the segment $\overline{\text{BC}}$, the total earliness and lateness are minimized by serving car trips at the maximum possible rate before $\tilde{t}_e$ and after $\tilde{t}_L$. Then, segment $\overline{\text{BC}}$ can be slid up or down until the sum of the schedule penalties for all early and late commuters is minimized.

For most of the rush, early and late commuters can be served at rate μ associated with point P, and vehicle accumulation n_μ. In the middle of the rush, when both transit and cars operate together on the street network without delay, the average vehicle density in the network cannot exceed k_μ and the total car accumulation cannot exceed $\tilde{n}_\mu$. Therefore, just before transit service begins at $\tilde{t}_e$, the vehicle accumulation

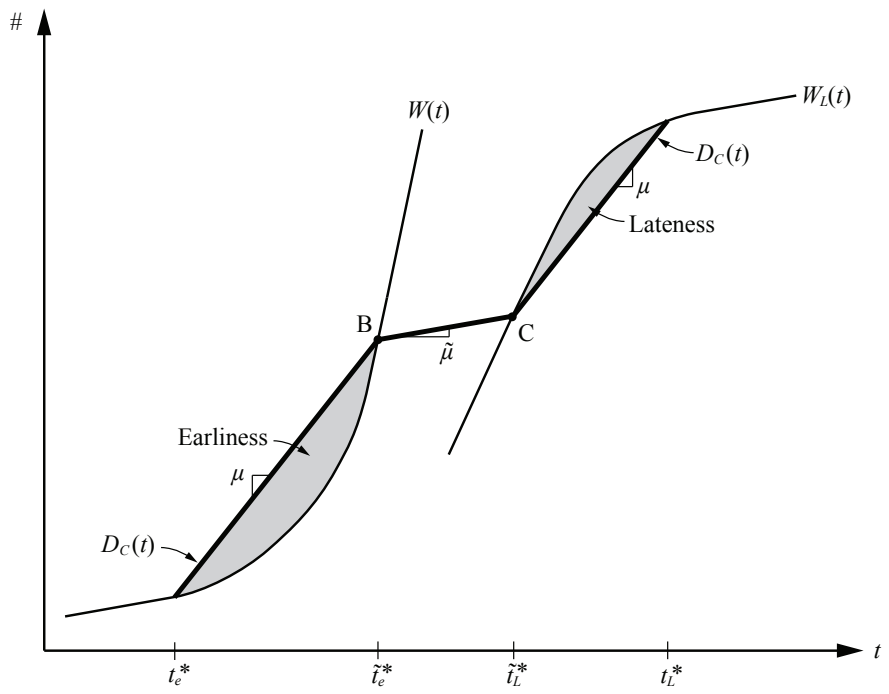


Figure 5.10. System optimal departure curve for cars exiting a network when transit operates

must be reduced to $\tilde{n}_\mu$ and the network exit rate μ_α is given by:

$$\mu_\alpha = F(\tilde{n}_\mu). \tag{5.56}$$

This results in a shift of the traffic state to the left from P to P_α along $F(n)$ in Figure 5.9(b). If the duration of this transition is very short compared to the length of the rush, then the effect is small and the departure curve for cars in the network system optimum (see Figure 5.10) is approximately the same piece-wise linear pattern identified in Proposition 4.

In reality, there are two competing secondary effects of the transition from n_μ to $\tilde{n}_\mu$: 1) increased total earliness, and 2) travel time savings for faster trips. More details about the transition and these effects are presented in Section 5.5.1. This transition effect does not occur when transit service ends because cars are able to freely enter and rapidly raise the vehicle accumulation to n_μ at $\tilde{t}_L$.

In order to eliminate queuing delay, the optimal prices presented in Proposition 8 are required so that users choose to travel at the system optimal departure time. In equilibrium, the price for car users must increase at rate e for early drivers and decrease at rate L for late drivers; otherwise queuing delays will emerge as commuters seek to adjust their travel times in order to reduce their own experienced cost. In order to achieve the transition from P to P_α, some additional network control (e.g., adjusting signal timings) is required, but the optimal prices ensure that congestion does not develop. The time intervals $(t_e, \tilde{t}_e)$ and $(\tilde{t}_L, t_L)$ are determined by the target

exit flow μ as shown in Figure 5.10. These determine the optimal car toll and transit fare as described in Section 5.3.

5.5.1 Traffic State Transition

The Network Exit Function (NEF) describes the relationship between the number of cars in a network and the rate that vehicles exit the network as described in Section 5.5. In order to prevent delays for traffic when transit service begins, the vehicle accumulation in the network must be reduced to $\tilde{n}_\mu$ immediately before the start of transit service at $\tilde{t}_e$. This corresponds to a transition in the rate that cars discharge from the network from μ to μ_α as shown by points P and P_α (see Figure 5.9(b)). Since the NEF is concave, $\tilde{\mu} \leq \mu_\alpha$, and the slope from the origin to μ_α is no less than to $\tilde{\mu}$. Therefore, vehicle trips before transit starts operating are at least as fast as when both modes are operating together and no delays are incurred. The effect on vehicle departures is illustrated by $D_C(t)$ to the left of point B in Figure 5.11.

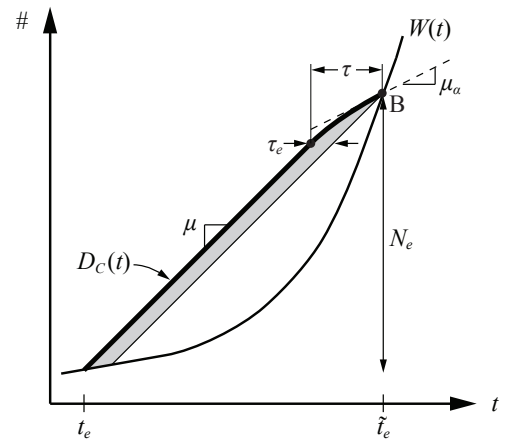


Figure 5.11. System optimal departure curve for early cars in a network transitioning from departure rate μ to μ_α.

If the vehicle accumulation is expressed as a function of time, $n(t)$, then recall that the state of the network follows the mass conservation equation (Daganzo, 2007) as expressed in (4.4):

$$\frac{dn}{dt} = \dot{A}(t) - F(n(t))$$

where $\dot{A}(t)$ is the rate that vehicles enter the network. We define τ as the transition time for the vehicle accumulation to drop from n_μ to $\tilde{n}_\mu$. Then, τ is minimized if no vehicles enter the network ($\dot{A} = 0$), and trips exit according to the NEF. The conservation equation (4.4) is an ordinary differential equation which can be solved as a boundary value problem to obtain τ:

$$\tau = -\int_{n_\mu}^{\tilde{n}_\mu} \frac{1}{F(n)} dn. \tag{5.57}$$

Recall from (5.54) that $\tilde{n}_\mu = \alpha n_\mu$.

The transition from n_μ to $\tilde{n}_\mu$ results in two competing effects. First, the total earliness cost is increased because the maximum departure rate for early commuters cannot be sustained at μ for the entire interval $(t_e, \tilde{t}_e)$. The reduced exit flow immediately preceding transit service adds τ_e additional earliness to nearly every early commuter (see Figure 5.11). This is the difference between the transition time, and the time it would have taken for the same $(1-\alpha)n_\mu$ trips to exit at rate μ:

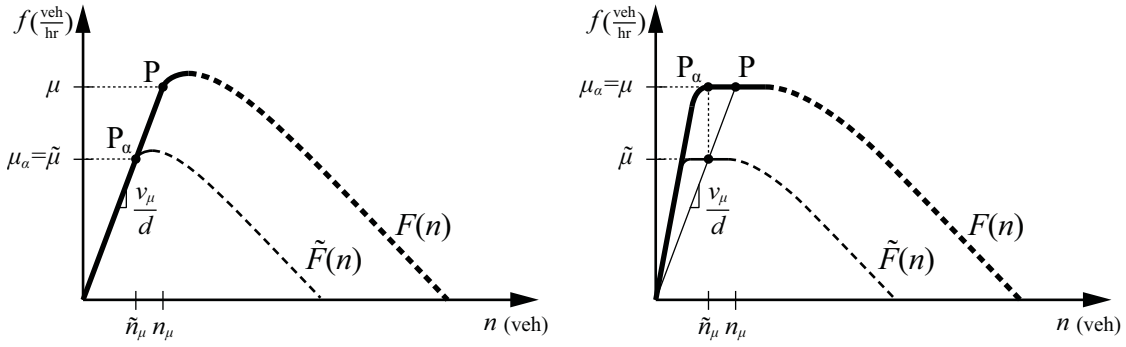
$$\tau_e = \tau - \frac{(1-\alpha)n_\mu}{\mu}. \tag{5.58}$$

Since nearly every early driver experiences τ_e additional earliness, the total system cost increases by approximately $eN_e\tau_e$.

Second, some travel time savings are experienced by early commuters in the transition period of length τ which reduces the total system cost. This occurs because the transition from point P to P_α decreases the exit flow to μ_α. Since $\tilde{\mu} \leq \mu_\alpha$ and both flows are associated with $\tilde{n}_\mu$, the travel time for P_α will be at least as short as for P, if not shorter. The aggregated travel time savings, TT_s, is the difference between the total travel time for $(1-\alpha)n_\mu$ trips to exit while the network accumulation is n_μ and the total travel time during the transition period for the same number of trips to actually exit:

$$TT_s = \frac{(1-\alpha)n_\mu^2}{\mu} - \int_0^\tau n(t)dt \tag{5.59}$$

The first term of (5.59) is the product of the time it takes $(1-\alpha)n_\mu$ to exit the network at rate μ and the n_μ vehicles which are present in the network at all times. Upper bounds for the magnitude of these effects can be determined by considering two NEFs: $\mu_\alpha = \tilde{\mu}$, and $\mu_\alpha = \mu$ (see Figure 5.12).



(a) Case 1: NEF when $\mu_\alpha = \tilde{\mu}$

(b) Case 2: NEF when $\mu_\alpha = \mu$

Figure 5.12. Example NEFs for cases with maximum earliness cost (Case 1) and maximum travel time savings (Case 2).

Case 1: Maximum Earliness The largest possible change in exit flow from P to P_α is a transition from μ to $\mu_\alpha = \tilde{\mu}$. In this case, NEF must be linear to the left of P as shown in Figure 5.12(a). This will result in the greatest possible transition time τ and additional earliness τ_e, because the exit rate can be no lower for each vehicle accumulation if $F(n)$ is concave. For this case, the exit flow is given by:

$$F_1(n) = \frac{\mu}{n_\mu} n \quad \text{for } n \in (0, n_\mu). \tag{5.60}$$

We can solve for τ by substituting (5.60) into (5.57), and solving the integral with $\tilde{n}_\mu = \alpha n_\mu$:

$$\tau = -\frac{n_\mu}{\mu} \ln \alpha. \tag{5.61}$$

Then, by substituting (5.61) into (5.58), the added earliness for each early commuter is:

$$\tau_e = \frac{n_\mu}{\mu} \left(-\ln \alpha - 1 + \alpha\right). \tag{5.62}$$

This is an upper bound for the τ_e associated with any concave NEF. Note that n_μ/μ is the average travel time for a trip of length d, and τ_e will be small for many reasonable values of α (e.g., τ_e is less than 3% of the uncongested travel time for values of $\alpha > 0.8$).

The NEF in this case is linear to the left of P, so all traffic states are associated with the same slope to the origin for $n \leq n_\mu$ (see Figure 5.12(a)). The average travel time per trip in the network does not change over the course of the transition, and therefore there are no travel time savings experienced. This can be verified by solving the mass conservation equation, (4.4), with (5.60) and substituting the result into (5.59).

Case 2: Maximum Travel Time Savings The largest possible reduction in travel time from P to P_α is when the exit flow transitions from μ to $\mu_\alpha = \mu$. In this case, the NEF has a constant value between P_α and P as shown in Figure 5.12(b). Although we would expect the point P always to be chosen as the left most point with exit flow μ, this case provides an upper bound for the total travel time savings as μ_α approaches μ.

Since the exit flow is always μ, the number of vehicles in the network at any time during the transition is given by:

$$n(t) = n_\mu - \mu t. \tag{5.63}$$

We also know that the duration of the transition for $(1-\alpha)n_\mu$ vehicles to depart will be:

$$\tau = (1-\alpha) n_\mu / \mu. \tag{5.64}$$

Substituting (5.63) and (5.64) into (5.59), and solving the integral, results in a total travel time savings of:

$$TT_s = \frac{(1-\alpha)^2 n_\mu^2}{2\mu}. \tag{5.65}$$

This is an upper bound for the TT_s associated with any choice of P on the uncongested side of a concave NEF. Note that this value is independent of the number of early commuters as long as the length of the period when commuters travel early is longer than the transition period.

The NEF in this case does not contribute any additional earliness to the other early commuters ($\tau_e = 0$). This result can be easily verified by substituting (5.64) into (5.58) and occurs because the exit flow is always maintained at μ until transit service starts operating. Therefore, the departure curve for cars in this case is still represented by Figure 5.10, and the system optimal solution will be exactly the same as the travel pattern identified in Section 5.2.

5.5.2 System Optimum and City Structure

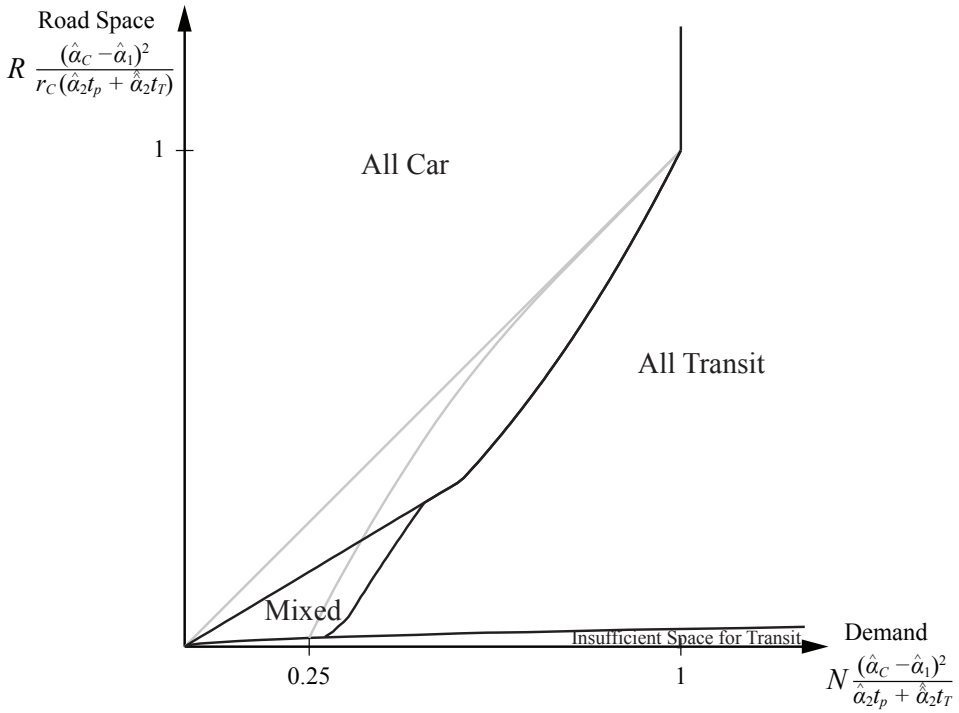
To relate this work back to the characteristics of city structure (road space and demand), we can summarize the system optimum result on scaled axes just as was done for the city with constant demand and the evening peak. For these illustrations, we will focus on a special case where the wished curve is Z-shaped, and then we can use the result of Proposition 7. Since the rate that riders will use transit is always $\lambda - \tilde{\mu}$, the duration of transit service, t_T, is uniquely determined by N_T. Therefore the transit cost function, (4.2) is a function only of N_T, so it is straightforward to identify if and where a possible multimodal optimum exists.

The results are shown in Figure 5.13 and they are qualitatively very similar to those of the evening commute. There are still three regimes: 1) *all car*, 2) *mixed modes*, and 3) *all transit*. The difference is in how the modes are operated in the various regimes. In the morning commute system optimum there should be no congestion, but there may be extra costs associated with schedule penalty which occur in the system optimum travel patterns identified in Section 5.2. The delineations between the regimes bow increasingly downward and to the right as the demand gets more peaked. The interpretation for this is similar to that of the evening commute, except in this case the trade-off is that for some cities the total schedule penalty associated with everyone driving is too small to justify investing in a transit system to avoid these costs.

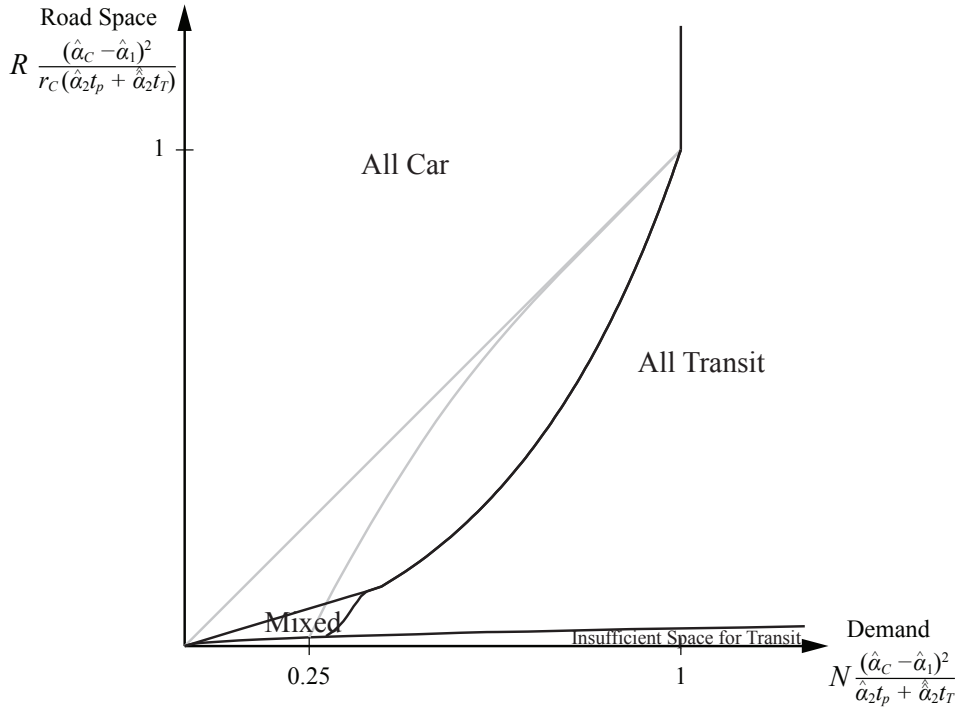
5.6 Summary of Findings

This chapter includes the most general analysis in this dissertation. The morning commute problem on a network with multiple modes consists of individuals whose desired departure times from the network are distributed in time. Using the framework introduced in Vickrey (1969) and developed in Hendrickson & Kocur (1981), a single bottleneck used by multiple modes is studied, and the results are applied to urban networks.

As a starting point, a single bottleneck with fixed capacity that serves both cars and transit has been studied. Commuters choose which mode to use and when to



(a) Morning Peak ($t_p = 6$ hours, $t_{max} = 10$ hours)



(b) Morning Peak ($t_p = 3$ hours, $t_{max} = 10$ hours)

Figure 5.13. Summary of system optimum for cities with peaked morning demand served by cars and BRT.

travel in order to minimize the generalized cost of their own trip. The transit agency chooses the headway and when to operate. The following results are shown for this type of bottleneck:

1. If the transit agency charges a fixed fare and operates at a given headway, and only when there is demand, then there is a unique user equilibrium.

2. If the transit agency chooses its headway and time of operation for the common good, then there is a unique system optimum.

3. Time-dependent prices exist to achieve system optimum.

4. Optimal prices are not always the difference between the system optimum and user equilibrium costs.

Finally, it is also shown that results 2 and 3 apply to urban networks.

For both the single bottleneck and network, the provision of public transit is a pareto improvement because everyone experiences a lower cost than in the single mode case. Public transit has also been shown to reduce the duration of the rush period in user equilibrium. When cars and transit share the same road capacity, the system optimum travel pattern differs from the user equilibrium. Although optimal time-dependent prices always exist, they do not simply price the delay out of the equilibrium result. The optimal prices are also not unique, so there is flexibility to pursue other policy objectives.

Other results included in this chapter are a comparison of various pricing strategies for the bottleneck with cars and transit. These alternative prices address realistic constraints that make optimal time-dependent pricing difficult to implement. The relative efficiency and equity of these strategies are ranked based on analysis of a general S-shaped wished curve. Although none of the alternatives are as efficient as the optimum, some are more equitable in that the difference between the maximum trip cost and minimum trip cost can be diminished.

Chapter 6

Conclusion

Transportation is of critical importance to the life and operation of cities worldwide because it provides people with access to opportunities. With limited resources to continue expanding urban transportation infrastructure, cities need to be able to manage their streets, and the transportation systems which use them, cost effectively. This dissertation has ought to address this problem by answering the question: *What are the costs of providing accessibility for cities of different structures?* The focus has been on looking at how urban space should be allocated to different transport modes, and how these modes should be operated and priced efficiently.

Viewing cities at the macroscopic level, this work focuses on revealing the physical relationships between city structures, transportation systems, and the costs of serving travel demand in cities. Unlike existing work which is either site-specific or empirically shows a relationship between city structure and the costs of mobility, this research builds a theory of urban physics that can be applied to any city or any mode. Using macroscopic models of traffic and transit systems, a systematic analysis of the allocation of space and use of modes for any city has been presented. These models recognize that vehicles require space, and demonstrate the effect of spatial limitations of cities on how transportation systems can and should be operated and priced.

The reality is that no city will operate with perfect efficiency, and no two cities will behave exactly alike. The models developed in this dissertation establish bounds for what is physically possible. By understanding what can be achieved with a transportation system in a city, and the mechanisms that affect performance outcomes, the results of this work can be used as a basis for developing plans to address real challenges related to the growing demand for limited urban space.

The remaining sections of this chapter include a summary of the key contributions of the dissertation research, some alternative applications that could be further developed, and directions for future research.

6.1 Contributions

The contributions of this dissertation are connecting recent advances in understanding the behavior of urban transportation networks with economic theory. Starting with the modeling methodology, this dissertation develops functional relationships between characteristics of city structure, travel demand, and transport modes, and the various types of costs associated with transportation systems. This is done recognizing that vehicles require space, and that the physical extent of queuing on networks affects the networks' ability to move vehicles.

Using these models, several insights are gained from studying the idealized case of cities with constant demand, and more realistic cities with evening and morning peak demand. The analysis has focused on a comparison of two modes (individual cars and collective public transit) in a city where all trip characteristics are homogeneous among the population. For each of the cases, analysis has been done for the user equilibrium and system optimum, and pricing strategies have been developed to achieve the system optimum.

In summary, the main contributions of this dissertation are:

1. There are three possible operating regimes for the system optimum: *all car*, *mixed modes*, and *all transit*. As demand increases, cities that start with plentiful street space eventually fill their roads and make a gradual transition to mixed modes where just enough transit is provided to fit the demand on the road space available. Then, the transition to all transit is a sudden jump when the cost of mixing modes exceeds serving everyone on transit.

2. The availability of competitive public transit reduces the trip costs for all users. In the morning commute, this is caused by making the rush period shorter, involving fewer people. Furthermore, every singe commuter experiences a reduced trip cost even if he or she travels when no transit is available.

3. Transit subsidies (or car tolls) can make everyone better off in cases where road space is constrained.

4. The effect of peaked demand is to increase the range of cities for which it is optimal to serve all trips by car. The result is that there are additional costs associated with car trips in equilibrium (schedule delay in the morning, congestion in the evening). These costs may not be large enough to justify the expensive investments in transit systems needed to avoid those costs.

User equilibrium, system optimum, and optimal pricing solutions are identified for cities with constant demand, evening peaks, and morning peaks. The morning peak results are important because they provide insights on how to efficiently manage a network without congestion. The following are identified for the morning commute:

1. Unique user equilibrium when the transit service level is given;

2. Unique system optimum when the transit agency can choose the headway and operating time for the common good;

3. Existence of time-dependent prices to achieve system optimum;

4. Results for the fixed capacity bottleneck apply to the system optimum network problem.

The broader contribution of this work is that normative models have been constructed with realistic physics in order to identify the efficient allocation of space and operation of modes.

6.2 Other Applications

All of the modeling in this dissertation was built on the same cost function and constraint: minimize some combination of time, money, and externalities, subject to a road space constraint. This is a reasonable framework for a city that is already well-developed with street infrastructure essentially built. In this city, the goal may be to find ways to make the existing transportation infrastructure operate more efficiently.

An alternative formulation is alluded to in the dissertation, but not discussed in depth: e.g., to view some type of environmental impact such as greenhouse gas emissions as a constraint on the transportation system. Since the functional forms are the same for many different types of costs, the qualitative solutions will be the same. The changing magnitudes of cost parameters merely moves cities to different positions on the solution space.

6.3 Future Work

There are many ways to improve the applicability of this work by relaxing assumptions which restrict how well the results describe real cities. Although the analytical models can only be pushed so far, the theory developed in this dissertation can be used as a basis for numerical or simulated approaches. Some areas for future related research are:

1. Investigating the effects of heterogeneity among users, particularly with respect to trip lengths and value of time. These are both important determinants of mode choice, and yet the values vary from person to person and from trip to trip. An expected outcome of this heterogeneity is that cities should see a richer variety of modes than the results of this dissertation suggest.

2. Considering the effects of using different modes in combination. The models are structured to account for the characteristics of the access mode, but more sophisticated ways to operate cars and transit together, or transit systems with different types of services, for example, could provide insights for designing hierarchical transportation systems.

3. Studying the effects of vehicle interactions on network performance. This dissertation research is built on assumptions that transit lanes are well-managed and space is dedicated so that there are no conflicts between different vehicle types which reduce capacity. A better understanding is needed of how networks are able to serve multiple modes. This is an important step to identifying the passenger-carrying rather vehicle-carrying capacity of transportation networks.

4. Understanding the environmental performance of transportation as a system rather than merely the sum of its components. Most of the existing work on the life-cycle impacts of transportation focuses on studying individual vehicles and scaling them up, but the macroscopic approach used in this research is well suited for studying large complex systems like transportation systems in cities.

Bibliography

AAA. 2007. *Your Driving: How much are you really paying to drive?* Tech. rept. American Automobile Association.

Ahn, K. 2009. Road pricing and bus service policies. *Journal of Transport Economics and Policy*, **43**(1), 25–53.

Alonso, W. 1964. *Location and Land Use: Toward a General Theory of Land Rent.* Cambridge, Mass.: Harvard University Press.

Anas, A., & Xu, R. 1999. Congestion, land use, and job dispersion: A general equilibrium model. *Journal of Urban Economics*, **45**, 451–473.

Anas, A., Arnott, R., & Small, K.A. 1998. Urban spatial structure. *Journal of Economic Literature*, **36**(3), 1426–1464.

Arnott, A., De Palma, A., & Lindsey, R. 1990a. Departure time and route choice for the morning commute. *Transportation Research Part B*, **24**(3), 209–228.

Arnott, R., & Yan, A. 2000. The two-mode problem: Second-best pricing and capacity. *Review of Urban and Regional Development Studies*, **12**(3), 170–199.

Arnott, R., De Palma, A., & Lindsey, R. 1990b. Economics of a bottleneck. *Journal of Urban Economics*, **27**(1), 111–130.

Arnott, R., De Palma, A., & Lindsey, R. 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, **83**(1), 161–179.

Axhausen, K.W. 1998. *Theoretical Foundations in Travel Choice Modeling.* Oxford: Elsevier Science Ltd. Chap. Can we ever obtain the data we would like to have?, pages 305–323.

Beesley, M.E. 1965. Tha value of time spent travelling: Some new evidence. *Economica*, **32**(126), 174–185.

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

Boyd, J.H., Asher, N.J., & Wetzler, E.S. 1978. Nontechnological innovation in urban transit. *Journal of Urban Economics*, **5**, 1–20.

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

Burns, L.D. 1979. *Transportation, Temporal, and Spatial Components of Accessibility.* Lexington Books, Lexington, Mass.

Cameron, I., Kenworthy, J.R., & Lyons, T.J. 2003. Understanding and predicting private motorised urban mobility. *Transportation Research Part D*, **8**, 267–283.

Chester, M.V., & Horvath, A. 2009. Environmental assessment of passenger transportation should include infrastructure and supply chains. *Environmental Research Letters*, **4**, 024008.

Currie, G., Sarvi, M., & Young, W. 2004. *Urban Transport X: Urban Transport and the Environment in the 21st Century.* WIT Press. Chap. A new methodology for allocating road space for public transport priority, pages 375–388.

Currie, G., Sarvi, M., & Young, W. 2007. A new approach to evaluating on-road public transport priority projects: Balancing the demand for limited road-space. *Transportation*, **34**, 413–428.

Daganzo, Carlos F., Gayah, Vikash V., & Gonzales, Eric J. 2011. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness, and instability. *Transportation Research Part B*, **45**(1), 278–288.

Daganzo, C.F. 1985. The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science*, **19**(1), 29–37.

Daganzo, C.F. 1997. *Fundamentals of Transportation and Traffic Operations.* Pergamon.

Daganzo, C.F. 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B*, **41**, 49–62.

Daganzo, C.F. 2010. Structure of competitive transit networks. *Transportation Research Part B*, **44**(4), 434–446.

Daganzo, C.F., & Geroliminis, N. 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transportation Research Part B*, **42**(9), 771–781.

Danielis, R., & Marcucci, E. 2002. Bottleneck road congestion pricing with a competing railroad service. *Transportation Research Part E*, **38**(5), 379–388.

Delucchi, M.A. 2007. Do motor-vehicle users in the U.S. pay their way? *Transportation Research Part A*, **41**(10), 982–1003.

Department of Planning, Building, and Code Enforcement. 2004. *Off-street parking design standards*. Tech. rept. City of San Jose.

DeSerpa, A. C. 1971. A theory of the economics of time. *The Economic Journal*, **81**(324), 828–846.

Doherty, S.T. 2003. Should we abandon activity type analysis? *In: Moving through nets: The physical and social dimensions of travel, 10th International Conference on Travel Behaviour Research*.

Eichler, M., & Daganzo, C.F. 2006. Bus lanes with intermittent priority: Strategy formulae and an evaluation. *Transportation Research Part B*, **40**, 731–744.

Ferrari, P. 1999. A model of urban transport management. *Transportation Research Part B: Methodological*, **33**(1), 43–61.

Ferrari, P. 2005. *Variational Analysis and Applications*. Springer. Chap. An Optimization Problem with an Equilibrium Constraint in Urban Transport, pages 393–408.

Geroliminis, N., & Daganzo, C.F. 2007. Macroscopic modeling of traffic in cities. *In: TRB Annual Meeting*.

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. Chap. Cordon pricing consistent with the physics of overcrowding, pages 219–240.

Gonzales, E.J., Chavis, C., Li, Y., & Daganzo, C.F. 2011 (23–27 January). Multimodal transport in Nairobi, Kenya: Insights and recommendations with a macroscopic evidence-based model. *In: Transportation Research Board 90th Annual Meeting*.

Goodwin, P.B. 1981. The usefulness of travel budgets. *Transportation Research Part A*, **15**, 97–106.

Gordon, P., & Richardson, H.W. 1997. Are compact cities a desirable planning goal. *Journal of the American Planning Association*, **63**.

Gordon, P., Kumar, A., & Richardson, H.W. 1989. The influence of metropolitan spatial structure on commuting time. *Journal of Urban Economics*, **26**, 138–151.

Hägerstrand, T. 1970. What about people in regional science? *Pages 7–21 of: Ninth Congress of the Regional Science Association*, vol. 24.

Handy, S.L., & Niemeier, D.A. 1997. Measuring accessibility: An exploration of issues and alternatives. *Environment and Planning A*, **29**, 1175–1194.

Hansen, W.G. 1959. How accessibility shapes land use. *Journal of the American Institute of Planners*, **25**(2), 73–76.

Hanson, S. 2004. *The Geogrpahy of Urban Transportation*. Guilford Press. Chap. The context of urban travel: Concepts and recent trends, pages 3–29.

Harris, B. 1967. The city of the future: The problem of optimal design. *Papers of the Regional Science Association*, 185–195.

Harris, B. 2001. Accessibility: Concepts and applications. *Journal of Transportation Research*, **4**(2/3), 15–30.

Hendrickson, C., & Kocur, G. 1981. Schedule delay and departure time decisions in a deterministic model. *Transportation Science*, **15**(1), 62–77.

Hensher, D.A. 2001. Measurement of the valuation of travel time savings. *Journal of Transport Economics and Policy*, **35**(1), 71–98.

Holroyd, E.M. 1967. The optimum bus service: A theoretical model for a large uniform urban area. *Pages 308–328 of: Proceedings of the Third International Symposium on the Theory of Traffic Flow*.

Huang, H.J. 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), 267–284.

Illich, I. 1974. *Energy and Equity*. London: Marion Boyars.

IPCC. 2007. *Climate Change 2007: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. New York: Cambridge University Press. Chap. Summary for policymakers.

Kain, J.F. 1968. Housing segregation, negor employment, and metropolitan decentralization. *The Quarterly Journal of Economics*, **82**(2), 175–197.

Keeler, T.E., & Small, K.A. 1975. *The Full Costs of Urban Transport, Part III: Automobile Costs and Final Intermodal Comparisons*. Berkeley, California: Institute of Urban and Regional Development, University of California.

Kenworthy, J.R., & Laube, F.B. 1999. Patterns of automobile dependence in cities: an international overview of key physical and economic dimensions with some implications for urban policy. *Transportation Research Part A*, **33**, 691–723.

Kenworthy, J.R., & Laube, F.B. 2001. *Millenium Cities Database for Sustainable Transport*. [CD-Database] International Union of Public Transport (UITP), Brussels.

Kitamura, R., Nakayama, S., & Yamamoto, T. 1999. Self-reinforcing motorization: can travel demand management take us out of the social trap? *Transport Policy*, **6**(3), 135–145.

Koenig, J.G. 1980. Indicators of urban accessibility: Theory and application. *Transportation*, **9**, 145–172.

Koshi, M., Kuwahara, M., & Akahane, H. 1992. Capacity of sags and tunnels on Japanese motorways. *ITE Journal*, **62**(5), 17–22.

Kwan, M., Murray, A.T., O'Kelly, M. E., & Tiefelsdorf, M. 2003. Recent advances in accessibility research: Representation, methodology, and applications. *Journal of Geographical Systems*, **5**, 129–138.

Lago, A. 2003. *Spatial Models of Morning Commute Consistent with Realistic Traffic Behavior*. Ph.D. thesis, University of California, Berkeley.

Laube, F.B., Kenworthy, J.R., & Zeibots, M.E. 1999. *Siedlungsstrukturen, räumliche Mobilität und Verkehr: Auf dem Weg zur nachhaltigkeit in Stadtregionen?* IRS Institut für Regionalentwicklung und Strukturplanung. Chap. Towards a science of cities: City observation and formulation of a city theory, pages 99–118.

Levinson, D.M. 1998. Accessibility and the journey to work. *Journal of Transport Geography*, **6**(1), 11–21.

Li, Z.C., Huang, H.J., Lam, W.H.K., & Wong, S.C. 2007. A model for evaluation of transport policies in multimodal networks with road and parking capacity constraints. *Journal of Mathematical Modelling and Algorithms*, **6**(2), 239–257.

Lighthill, M.J., & Whitham, G.B. 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. *In: Proceedings of the Royal Society of London.* A, Mathematical and Physical Sciences.

Litman, T.A. 2003. Economic value of walkability. *Transportation Research Record*, **1828**, 3–11.

McKean, J.R., Johnson, D.M., & Walsh, R.G. 1995. Valuing time in travel cost demand analysis: An empirical investigation. *Land Economics*, **71**(1), 96–105.

Meyer, J.R., Kain, J.F., & Wohl, M. 1965. *The Urban Transportation Problem*. Harvard University Press.

Miller, H.J. 1991. Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Science*, **5**(3), 287–301.

Mishan, E.J. 1967. Interpretation of the benefits of private transport. *Journal of Transport Economics and Policy*, **1**(May), 184–189.

Mogridge, M.J.H. 1997. The self-defeating nature of urban road capacity policy: A review of theories, disputes and available evidence. *Transportation Policy*, **4**(1), 5–23.

Mohring, H. 1972. Optimization and scale economies in urban bus transportation. *The American Economic Review*, **62**(4), 591–604.

Mohring, H. 1979. The benefits of reserved bus lanes, mass transit subsidies, and marginal cost pricing in alleviating traffic congestion. *Current issues in urban economics*, 165–95.

Mokhtarian, P.L., & Chen, C. 2004. TTB or not TTB, that is the question: A review and analysis of the emprical literature on travel time (and money) budgets. *Transportation Research Part A*, **38**, 643–675.

Moritz, W.E. 1997. Survey of North American bicycle commuters: Design and aggregate results. *Transportation Research Record*, **1578**, 91–101.

Murphy, J.J., & Delucchi, M.A. 1998. A review of the literature on the social cost of motor vehicle use in the United States. *Journal of Transportation and Statistics*, **1**(1), 15–42.

National Transit Database. 2009. *National Transit Profile*. Tech. rept. Federal Transit Administration.

Navarro, R.A., Heierli, U., & Beck, V. 1985. *La Bicicleta y los Triciclos*. St. Gallen, Switzerland: SKAT, Centro Suizo de Technologia Apropiada.

Newell, G.F. 1971. *Applications of queueing theory*. London: Chapman and Hall.

Newell, G.F. 1979. Some issues relating to the optimal design of bus routes. *Transportation Science*, **13**(1), 20–35.

Newman, P.W.G., & Kenworthy, J.R. 1989. Gasoline consumption and cities: A comparison of U.S. cities with a global survey. *Journal of the American Planning Association*, **55**(1), 24–37.

O'Regan, K.M., & Quigley, J.M. 1998. *Accessibility and economic opportunity*. Tech. rept. University of California at Berkeley.

Pendyala, R.M., Yamamoto, T., & Kitamura, R. 2002. On the formulation of time-space prisms to model constraints on personal activity travel engagement. *Transportation*, **29**(29), 73–94.

Pickrell, D.H. 1985. The cost of constructing new rail transit systems. *Transportation Research Record*, **1006**, 48–55.

Pirie, G.H. 1979. Measuring accessibility: A review and proposal. *Environment and Planning A*, **11**, 299–312.

Pushkarev, B.S., & Zupan, J.M. 1977. *Public Transportation and Land Use Policy*. Indiana University Press.

Richards, P.I. 1956. Shockwaves on the highway. *Operations Research*, **4**(1), 42–51.

Rossi-Hansberg, E. 2004. Optimal urban land use and zoning. *Review of Economic Dynamics*, **7**, 69–106.

Sharp, C.H. 1967. The choices between cars and buses on urban roads. *Journal of Transportation Economics and Policy*, **1**, 104–111.

Small, K.A., & Chu, X. 2003. Hypercongestion. *Journal of Transport Economics and Policy*, **37**(1), 319–352.

Smith, M.J. 1984. The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science*, **18**(4), 385–394.

Solow, R.M. 1972. Congestion, density and the use of land in transportation. *The Swedish Journal of Economics*, **74**(1), 161–173.

Solow, R.M. 1973. Congestion cost and the use of land for streets. *The Bell Journal of Economics and Management Science*, **4**(2), 602–618.

Solow, R.M., & Vickrey, W.S. 1971. Land use in a long narrow city. *Journal of Economic Theory*, **3**, 430–447.

Sparks, G.A., & May, A.D. 1971. A mathematical model for evaluating priority lane operations on freeways. *Highway Research Record*, **363**, 27–42.

Tabuchi, T. 1993. Bottleneck congestion and modal split. *Journal of Urban Economics*, **34**(3), 414–431.

Tanner, J.C. 1981. Expenditure of time and money on travel. *Transportation Research Part A*, **15**(1), 25–38.

Thoreau, H.D. 1854. *Walden; or, Life in the Woods*. Boston: Ticknor and Fields.

Toole-Holt, L., Polzin, S.E., & Pendyala, R.M. 2005. Two minutes per person per day each year: Exploration of growth in travel time expenditure. *Transportation Research Record*, **1917**, 45–53.

Vickrey, W.S. 1969. Congestion theory and transport investment. *The American Economic Review*, **59**(2), 251–260.

Wachs, M., & Kumagai, T.G. 1973. Physical accessibility as a social indicator. *Socio-economic Planning Science*, **7**, 437–456.

Wadhwa, L.C., & Wirasinghe, S.C. 2003. *Urban Transport IX: Urban Transport and the Environment in the 21st Century*. Southampton: WIT Press. Chap. True cost of road travel, pages 525–534.

Wardrop, J.G. 1952. Some theoretical aspects of road traffic research. *ICE Proceedings: Engineering Divisions*, **1**(3), 325–378.

Wheaton, W.C. 1998. Land use and density in cities with congestion. *Journal of Urban Economics*, **43**, 258–272.

Whitelegg, J. 1993. Time pollution. *The Ecologist*, **23**(4).

Wirasinghe, S., Hurdle, V.F., & Newell, G.F. 1977. Optimal parameters for a coordinated rail and bus transit system. *Transportation Science*, **11**(4), 359–374.

Zahavi, Y., & Ryan, J.M. 1980. The stability of travel components over time. *Transportation Research Record*, **750**, 19–26.

Zahavi, Y., & Talvitie, A. 1980. Regularities in travel time and money expenditures. *Transportation Research Record*, **750**, 13–19.

Appendix A

Glossary of Symbols

Introduced in Chapter 2

a	=	reachable area associated with trip length d [dist^2]
A	=	accessibility per trip [reachable opportunities]
α	=	proportion of network for cars when transit operates [–]
α_C	=	cost per uncongested car trip [time/trip]
α_0	=	cost coefficient for transit [time/dist^2·time]
α_1	=	cost coefficient for transit [time/trip]
α_2	=	cost coefficient for transit [time/dist^2·trip]
α_3	=	cost coefficient for transit [time^2/trip^2]
β	=	value of time [$/time]
c	=	vehicle occupancy [trips/veh]
c_a	=	vehicle occupancy of access mode [trips/veh]
c_d	=	vehicle operating cost per distance [$/veh·dist]
c_i	=	infrastructure cost per unit area (e.g., paving streets) [$/dist^2·time]
c_t	=	vehicle cost per time [$/veh·time]
d	=	length of a trip [dist]
D	=	population density, number of people per area [ppl/dist^2]
δ	=	trip-making rate per person [trips/ppl·time]
e_d	=	environmental impact of vehicle operation per distance [#/veh·dist]
e_i	=	environmental impact per unit area of infrastructure [#/dist^2·time]
e_t	=	environmental impact of vehicle per time [#/veh·time]
E_Z	=	total external impact of interest [#/dist^2·time]
f	=	flow of vehicles exiting the network [veh/time]
γ	=	monetized value of external impact (policy variable) [$/#]
h	=	clear headway required for transit vehicle operation [time/veh]
H	=	headway of transit service [time/veh]
k	=	average network density [veh/lane·dist]
k^*	=	critical network density associated with q_m [veh/lane·dist]
κ	=	constant relating accessibility to D and d^2 [opportunities/ppl]

l = total network length [lane·dist]
λ = demand rate, trip generation per area per time [trips/dist^2·time]
λ_C = demand rate for cars [trips/dist^2·time]
λ_T = demand rate for transit [trips/dist^2·time]
M_Z = total money cost [\$/dist^2·time]
n = number of vehicles accumulated in the network [veh]
ψ = vehicle time per trip [veh·time/trip]
q = average network flow [veh/time·lane]
q_a = maximum flow for access mode [veh/time·lane]
q_m = maximum flow [veh/time·lane]
r = vehicle footprint per trip [dist^2·time/trip]
r_C = vehicle footprint per car trip [dist^2·time/trip]
r_p = physical area required per parking space [dist^2]
r_1 = road space coefficient for transit [dist^2·time/trip]
r_2 = road space coefficient for transit [dist^2·time/trip]
r_3 = road space coefficient for transit [dist^4·time^2/trip^2]
R = road space available for moving transportation per area of city [–]
$\bar{R}$ = road space for moving required per trip [dist^2·time/trip]
R_C = total road space required by the car transportation system [–]
R_p = parking space available per area of city [–]
R_T = total road space required by the transit system [–]
$\bar{R}_p$ = parking space required per trip [dist^2·time/trip]
s = route and stop spacing for transit system [dist]
t_a = access time [time/trip]
t_m = travel time in-vehicle for trip of length d [time/trip]
$\bar{T}$ = total travel time per trip [time/trip]
t_w = waiting time out of vehicle [time/trip]
T_Z = total time cost per time [time/dist^2·time]
v = average transit speed including stops for passengers [dist/time]
v_a = average speed of access mode [dist/time]
v_m = average network speed associated with q_m [dist/time]
V_d = total vehicle distance operated per time [veh/dist·time]
$\bar{V}_d$ = vehicle distance per trip [veh·dist/trip]
V_t = total vehicle time (fleet size) [veh/dist^2]
$\bar{V}_t$ = vehicle time per trip [veh·time/trip]
w = lane width for mode [dist]
w_a = lane width for access mode [dist/lane]
x = loss time per passenger for boarding and alighting [veh·time/trip]
y = loss time per stop [time]
Z = total generalized cost of the transportation system [time/dist^2·time]
Z_C = total generalized cost of the car system [time/dist^2·time]
Z_T = total generalized cost of the transit system [time/dist^2·time]

Introduced in Chapter 3

f_m	=	maximum exit flow when network serves only car [veh/time]
K	=	total generalized cost for iso-cost contour [time/dist^2·time]
$\mathcal{L}$	=	locus of points (λ_T, λ_C) where $Z(\lambda_T, \lambda_C) = Z_T(\lambda)$
λ_{crit}	=	critical demand at which $Z_C(\lambda_{crit}) = Z_T(\lambda_{crit})$ [trips/dist^2·time]
λ_{MC}	=	demand rate at which marginal costs are equal [trips/dist^2·time]
n_m	=	accumulation of vehicles associated with f_m [veh]
\$	=	system optimal price per trip in units of time [time/trip]
$\$_{MC}$	=	optimal price when transit is λ^* [time/trip]
v_n	=	average network speed associated with state N [dist/time]
$\mathcal{R}$	=	boundary of road space constraint
z_C	=	average generalized cost per car trip [time/trip]
z_T	=	average generalized cost per transit trip [time/trip]

Introduced in Chapter 4

$A(t)$	=	cumulative arrival curve to the network
$A_C(t)$	=	cumulative arrival curve of car trips
$A_T(t)$	=	cumulative arrival curve of transit trips
$\hat{\alpha}_C$	=	cost per peaked uncongested car trip [time/trip]
$\hat{\alpha}_1$	=	cost coefficient for peaked transit [time/trip]
$\hat{\alpha}_2$	=	cost coefficient for peaked transit [time^2/dist^2·time·trip]
$\hat{\hat{\alpha}}_2$	=	cost coefficient for peaked transit [time^2/dist^2·time·trip]
C	=	total queuing delay [time]
$D(t)$	=	cumulative departure curve from the network
f_m	=	maximum exit flow for network with only car [veh/time]
$\tilde{f}_n$	=	exit flow associated with congested state $\tilde{N}$ [veh/time]
$\tilde{f}_m$	=	maximum exit flow for network with cars and transit [veh/time]
n_e	=	excess accumulation of vehicles [veh]
n_m	=	accumulation of vehicles associated with f_m [veh]
n_n	=	accumulation of vehicles associated with state N [veh]
$\tilde{n}_n$	=	accumulation of vehicles associated with state $\tilde{N}$ [veh]
N	=	total number of trips in rush period [trips]
N_C	=	total number of car trips [trips]
N_T	=	total number of transit riders [trips]
N_T^*	=	total number of transit riders in system optimum [trips]
$\$(t)$	=	system optimal price per trip at t in units of time [time/trip]
t_{max}	=	maximum length of the rush [time]
t_p	=	duration of peak demand [time]
t_T	=	duration of transit operations [time]
T	=	maximum delay for drivers [time]

v_n	=	average network speed associated with state N [dist/time]
w	=	lane width for mode [dist]
z_C	=	average generalized cost per uncongested car trip [time/trip]
z_T	=	average generalized cost per transit trip [time/trip]

Chapter 5

e	=	schedule penalty for earliness [equivalent queuing delay]
k_μ	=	average network density associated with μ [veh/dist]
L	=	schedule penalty of lateness [equivalent queuing delay]
λ_e^*	=	demand rate at $\tilde{t}_e^*$ [veh/time]
λ_L^*	=	demand rate at $\tilde{t}_L^*$ [veh/time]
μ	=	maximum departure flow with only cars [veh/time]
$\tilde{\mu}$	=	max. departure flow of cars when transit operates [veh/time]
μ_α	=	departure flow associated before transit operates [veh/time]
dn_e	=	incremental vertical shift of early departures associated with dn_o
dn_L	=	incremental vertical shift of late departures associated with dn_o
n_μ	=	accumulation of vehicles associated with μ [veh]
$\tilde{n}_\mu$	=	accumulation of vehicles associated with $\tilde{\mu}$ [veh]
dn_o	=	incremental vertical shift of on-time departures of cars
N_e	=	total number of early commuters [trips]
N_L	=	total number of late commuters [trips]
N_o	=	number of on-time car commuters in middle of rush [trips]
S	=	total cost of schedule penalties [time]
$\$_C(t)$	=	price (toll) for a car trip at time t [time/trip]
$\$_C^{\text{off-peak}}$	=	price (toll) for car trips in the off-peak [time/trip]
$\Delta\$_e^*$	=	difference in optimal car price between first and last early trip
$\Delta\$_L^*$	=	difference in optimal car price between first and last late trip
$\$_{net}$	=	net revenue of pricing policy
$\$_T(t)$	=	price (fare) for a transit trip at time t [time/trip]
$\tilde{t}$	=	departure time of critical commuter in 1-mode equilibrium
t_e	=	beginning of the rush when first early commuter travels
$\tilde{t}_e$	=	departure time of first on-time commuter in rush
t_L	=	end of the rush when last late commuter travels
$\tilde{t}_L$	=	departure time of last on-time commuter in rush
t_1	=	first time when $\dot{W}(t)$ exceeds capacity
t_2	=	last time when $\dot{W}(t)$ exceeds capacity
T_C	=	maximum delay for drivers in 1-mode equilibrium [time]
T_e	=	delay experienced by the last early driver [time]
T_L	=	delay experienced by the first late driver [time]
TT_s	=	aggregated travel time savings in transition period [time]
τ	=	transition time to shift from n_μ to $\tilde{n}_\mu$ [time]
τ_e	=	additional earliness from transition time [time]

v_μ	=	average network speed associated with μ [dist/time]
$W(t)$	=	cumulative wished departure curve from the network
$W_C(t)$	=	cumulative wished departure curve of car trips
$W_T(t)$	=	cumulative wished departure curve of transit trips

Appendix B

Cost Model Coefficients

The general structure of the cost model for different types of modes is presented in Chapter 2. In Sections 2.4.1 and 2.4.3, detailed expressions for the individual components of travel time, footprints, and vehicles are presented. These contribute to the generalized cost functions and road space requirements of the modes, which are presented in a general form in Sections 2.4.2 and 2.4.4. In this appendix, the expressions relating the generalized costs to the various components are presented.

B.1 Individual Modes

Each component associated with a trip using an individual mode is independent of the demand. So, the total system generalized cost and road space required are attained by multiplying the values for a single trip by the demand λ_C.

Generalized Cost Coefficient

Each of the costs contributing to the generalized cost function are combinations of physical components with cost coefficients shown in Section 2.3.1. These are then combined into a generalized cost function as described in Section 2.3.2. First we substitute cost components from (2.15), (2.16), (2.13), and (2.14) into (2.9) and (2.10). Then, we substitute these quantities along with (2.12) into (2.11) to get the coefficient for individual modes such that $Z_C(\lambda_C) = \alpha_C \lambda_C$:

$$\begin{aligned} \alpha_C &= t_a + \frac{d}{v_m} \\ &\quad + \frac{1}{\beta}\left(c_t\psi + c_d d + \frac{c_i w_a v_a t_a}{c_a q_a} + \frac{c_i w d}{q_m c} + c_i r_p\left(\psi - \frac{d}{v_m}\right)\right) \\ &\quad + \frac{\gamma}{\beta}\left(e_t\psi + e_d d + \frac{e_i w_a v_a t_a}{c_a q_a} + \frac{e_i w d}{q_m c} + e_i r_p\left(\psi - \frac{d}{v_m}\right)\right) \end{aligned} \tag{B.1}$$

Required Road Space Coefficient

The road space coefficient for moving cars is simply the footprint identified by (2.15), so $r_C = \bar{R}$. The road space for parking is given by $\bar{R}_p$ from (2.16).

B.2 Public Transit Modes

The public transit modes involve more complex relationships between the demand and the generalized costs of the system, because the cost of each trip depends on the total demand for the system, λ_T.

B.2.1 Surface Transit on Shared Streets (e.g., Buses)

Generalized Cost Coefficients

The costs of the public transit system depend on its designed route/stop spacing, s, and headway, H. By substituting components from (2.23), (2.26), (2.24), and (2.25) as done above produces a generalized cost function of three variables $Z_T(\lambda_T, s, H)$. This is a convex function of H, and a closed form solution for the optimal headway, H^*, which minimizes the generalized cost can be found by setting the first derivative with respect to H equal to zero:

$$H^* = \sqrt{\frac{(c_t + \gamma e_t)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d + (c_i + \gamma e_i)\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)}{\frac{\lambda_T s \beta}{2} + \frac{dx\beta\lambda_T^2 s^2}{8}}}. \tag{B.2}$$

This optimal headway can then be substituted back into the total generalized cost function. The route and stop spacing could also be optimized, but this is not straightforward to do analytically, and furthermore, it is unrealistic for transit agencies to be changing the value of s frequently. Since the optimal s is insensitive to demand, it is sufficient to treat the physical structure of the network as fixed.

The generalized cost is of the form expressed in (2.28). By collecting terms, each of the coefficients is defined as follows:

$$\alpha_1 = \frac{s}{v_a} + \frac{d}{v_m} + \frac{yd}{s} + \frac{1}{\beta}\left(2(c_t + \gamma e_t)x + (c_i + \gamma e_i)\left(\frac{w_a s}{q_a c_a} + 2r_p x\right)\right) \tag{B.3}$$

$$\alpha_2 = \frac{32}{s\beta}\left((c_t + \gamma e_t)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d + (c_i + \gamma e_i)\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)\right) \tag{B.4}$$

$$\alpha_3 = \frac{8xd}{\beta}\left((c_t + \gamma e_t)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d + (c_i + \gamma e_i)\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)\right) \tag{B.5}$$

Required Road Space Coefficients

The required road space has the same functional form as the generalized cost. Assuming that the transit agency always operates the optimal headway to serve demand,

we simply substitute (B.2) into (2.26). The resulting road space coefficients are:

$$r_1 = \frac{w_a s}{q_a c_a} + 2 r_p x \tag{B.6}$$

$$r_2 = \frac{\frac{8\beta}{s}\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)^2}{(c_t + \gamma e_t)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d + (c_i + \gamma e_i)\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)} \tag{B.7}$$

$$r_3 = \frac{2dx\beta\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)^2}{(c_t + \gamma e_t)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d + (c_i + \gamma e_i)\left(wh + \frac{r_p}{v_m} + \frac{r_p y}{s}\right)} \tag{B.8}$$

As described in Section 2.4.4, this model represents a transit system which operates on a grid network and shares street space with other modes. If the loss time per stop becomes independent of the number of passengers boarding and alighting (e.g., a BRT with prepaid fares) then x approaches zero, so r_3 and α_3 are eliminated.

B.2.2 Transit on Dedicated Guideway (e.g., Metro)

Generalized Cost Coefficients

To consider systems with a significant fixed infrastructure investment per route length (e.g., dedicated guideway systems), the cost function requires further modification. In this case, the effective headway of the transit vehicle in traffic equals the operating headway, $h = H$, so the headway drops out of the expression for required road space for moving vehicles. Thus, all of the infrastructure costs for routes are included in α_0 and are eliminated from α_1, α_2, and α_3.

$$\alpha_0 = \frac{4w(c_i + \gamma e_i)}{\beta s} \tag{B.9}$$

$$\alpha_1 = \frac{s}{v_a} + \frac{d}{v_m} + \frac{yd}{s} + \frac{(c_i + \gamma e_i) w_a s}{\beta q_a c_a} \tag{B.10}$$

$$\alpha_2 = \frac{32}{\beta s}\left((c_t + \gamma e_t)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d + (c_i + \gamma e_i)\left(\frac{r_p}{v_m} + \frac{r_p y}{s}\right)\right) \tag{B.11}$$

Required Road Space Coefficients

If transit is operated on streets, such as a BRT system with fully dedicated lanes, the required road space is simply the route length per area multiplied by the lane width:

$$r_0 = \frac{4w}{s}. \tag{B.12}$$

For modes that operate on entirely separate guideways that do not require street space, this term is 0. For example, metro systems which are built in tunnels require huge infrastructure investments which are accounted for in the generalized cost function, but these systems require almost no surface space.

Table B.1. Model parameters (typical values used for analysis)

Parameter	Symbol	Units	Walk	Bike	Car	Bus	BRT	Metro
Vehicle occupancy	c	trips/veh	1	1	1			
Operating cost of vehicle per distance	c_d	\$/veh·m	0	0	2.58×10^{-4}	0.001	0.001	0.006
Infrastructure cost per time	c_i	\$/m^2·hr	4.57×10^{-4}	4.57×10^{-4}	4.57×10^{-4}	4.57×10^{-4}	0.0257	0.9
Operating cost of vehicle per time	c_t	\$/veh·hr	0	1.41×10^{-12}	0.644	78.0	78.0	120
Vehicle ownership (time per trip)	ψ	veh·hrs/trip		6	6			
Capacity flow in traffic	$q = 1/h$	veh/hr	1800	450	450	150	150	
Area of parking space	r_p	m^2		1	36	150	150	
Access time	t_a	hr		0.1	0.1			
Vehicle un-delayed average speed	v_m	m/hr	3600	12,000	30,000	30,000	30,000	60,000
Width of right of way	w	m	1	1	3.5	3.5	3.5	
Time per boarding/alighting passenger	x	hr/trip				0.0014		
Loss time per stop	y	hr				0.0111	0.0125	0.0125

Values were estimated from the following sources: AAA (2007); Department of Planning, Building, and Code Enforcement (2004); Geroliminis & Daganzo (2008); Gonzales *et al.* (2011); Kenworthy & Laube (2001); Moritz (1997); National Transit Database (2009).

B.3 Costs for Time-Dependent Demand

In Chapters 4 and 5, the demand for cars and transit peaked over the course of the day. As explained in Section 4.1, there are some costs, such as capital investments in vehicles and infrastructure which must be paid for whether they are used all day long or sit empty. The following sections show how the parameters of the generalized cost functions change when this is taken into consideration.

B.3.1 Individual Modes

The costs of road infrastructure are incurred whether the road is being used or not, so these components are decoupled from the cost parameter. The remaining costs of an uncongested car trip are still the same for each trip. We revise (B.1) to be:

$$\hat{\alpha_C} = t_a + \frac{d}{v_m} + \frac{1}{\beta}\left(c_t\psi + c_d d\right) + \frac{\gamma}{\beta}\left(e_t\psi + e_d d\right). \tag{B.13}$$

The capital costs associated with unused cars is incorporated by the hours of vehicle ownership attributed to each trip, ψ, which is also used in the generalized cost function for the constant demand case.

B.3.2 Public Transit Modes

The generalized cost function for transit requires a little more modification for the time-dependent case, because in addition to the street infrastructure, each vehicle is a capital investment that must be paid for whether in service or empty. The expressions are similar to those in Section B.2.1, omitting the infrastructure costs. The difference is that the operating costs per time are slightly modified to $\hat{c}_t$ and $\hat{e}_t$ to reflect that the fixed capital investment in vehicles is considered separately as $\hat{c}_v$ and $\hat{e}_v$ which are the costs incurred over the entire analysis period of length t_{max} and are independent of t_p. The time-dependent cost coefficients for a BRT system (where $x = 0$) are:

$$\hat{\alpha}_1 = \frac{s}{v_a} + \frac{d}{v_m} + \frac{yd}{s} \tag{B.14}$$

$$\hat{\alpha}_2 = \frac{32}{s\beta}\left(\left(\hat{c}_t + \gamma\hat{e}_t\right)\left(\frac{1}{v_m} + \frac{y}{s}\right) + c_d + \gamma e_d\right) \tag{B.15}$$

$$\hat{\hat{\alpha}}_2 = \frac{32}{s\beta}\left(\left(\hat{c}_v + \gamma\hat{e}_v\right)\left(\frac{1}{v_m} + \frac{y}{s}\right)\right) \tag{B.16}$$

In the case of constant demand these two components combine together at an hourly rate to make up about \$78 per hour as reported in Table B.2.2. Based on estimates from the National Transit Database (2009), approximately \$70 of this is hourly operating expenditures for labor, fuel, and maintenance. The remaining \$8 per hour is attributed to the capital cost of the vehicle.